UNIVERSITE DE NICE - SOPHIA ANTIPOLIS

ECOLE DOCTORALE STIC

SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THÉSE

pour l'obtention du grade de

Docteur en Sciences

de l'Université de Nice-Sophia Antipolis

Mention : Traitement Automatique du Signal des Images

présentée et soutenue par

Miriam Redi

Novel Methods for Semantic and Aesthetic Multimedia Retrieval

How Can Computers Appreciate and Interpret Pictures?

Nouvelles Méthodes pour la Recherche Sémantique et Esthétique d'Informations Multimédia

Comment Peuvent les Ordinateurs Apprécier et Interpréter des Images?

Thèse dirigée par : Bernard Merialdo

preparée à EURECOM, Departement Multimedia

soutenue le 29 Mai 2013

Jury :

Jenny BENOIS-PINEAU	-	LABRI, (AIV)	*Rapporteur*
Philippe-Henri GOSSELIN	-	INRIA (Texmex)	*Rapporteur*
Georges QUENOT	-	LIG (MRIM)	*Examinateur*
François BREMOND	-	INRIA (Stars)	*Examinateur*
Yoann POULAIN	-	Amadeus	*Invité*
Bernard MERIALDO	-	EURECOM (MultiMedia)	*Encadrant*

Abstract

What is that object depicted in that image? Very easy question, if asked human beings. But if we ask the same question to a computer, it will answer us with the values of the pixels composing the image: it would not be able to *recognize* the content of an image.

However, in the internet era, computerized categorization and classification of images and their properties (objects, scene, emotions generated, aesthetics and aritstic traits) became of crucial importance for the automatic organization, selection and retrieval of the huge amount of visual content surrounding us. But how can computer *see* the meaning of an image?

Multimedia Information Retrieval (MMIR) is a research field that helps building "intelligent" systems that automatically recognize the image content and its characteristics. MMIR systems takes as input an image and give as output a set of automatically assigned *labels* describing its properties.

In general, this is achieved by following a chain process: first low-level *features*, namely numerical vectors summarizing the image statistics, are extracted; features are then *pooled* into compact image *signatures*; based on such signatures, *machine learning* techniques are then used to build models able to distinguish between different image categories . Such model is finally used to recognize the properties of a new image, and output the corresponding labels.

Despite the advances in the field, human vision systems still substantially outperform their computer-based counterparts. In this thesis we therefore design a set of novel contributions for each step of the MMIR chain, aiming at improving the global recognition performances.

In our work, we explore techniques from a variety of fields that are not traditionally related with Multimedia Retrieval, and embed them into effective MMIR frameworks. For example, we borrow the concept of image saliency from visual perception, and use it to build low-level features. We employ the Copula theory of economic statistics for feature aggregation. We re-use the notion of graded relevance, popular in web page ranking, for visual retrieval frameworks.

In the following, we will explain in detail our novel solutions and prove their effectiveness for image categorization, video retrieval and image aesthetics assessment.

Acknowledgments

Those who do not know the torment of the unknown cannot have the joy of discovery. - Claude Bernard

I believe this is the right quotation to start my acknowledgements section: curiosity is the reason why I made this PhD and curiosity will drive my future as a researcher.

Some years back, my advisor Professor Bernard Merialdo and the on-line travel company Amadeus decided to trust me and give me the possibility to become a PhD student in Eurecom. They trusted me, supported me and helped me, and now I am graduating, presenting a thesis that I could not believe possible before, and I want to thank them very much for leading me to this point.
How did I get into research? All this started with research project for my Master Thesis at UC Santa Barbara, and I will always be very thankful to my Master advisors Gabriella Olmo and BS Manjunath for teaching me first the joy of discovery.

I also need to thank another professor that works in TU Delft. She is not *my* professor, she is not my supervisor, she is my "life advisor" that during these years supported me emotionally and technically, somehow, for some weird reason, believing in me. She's my sister and she looks incredibly like me, but she's wiser. Then, of course, Mamma, she always gave me the strength to go and explore what is going on outside our little world, and Dad, to teach me his non-rhetorical way of life, and to find explanations to unknown things. As a result of Mum and Dad, at age 5 I wanted to become an astronaut.
I would also like to express my love and gratitude to all my old friends: Elena, Vale, Lella, Nico,.... Fortunately, they are too many, since my earliest childhood they accepted the nerd that was in me and made me laugh when I needed it, helped me reasoning when I was loosing my rationality, and share with me the craziest moments ever.

Before coming to France, I was researching on Image Forensics under the California sun. Not bad at all, especially if you are 24. Deciding about your future at this age is not easy, and choosing to make a PhD and come to France would have been much harder, if I didn't have people like Paola, Lorenzo, Simona and Aristide that listened to me all the time when I was looking for the right path to take.

And it turned out to be the right path. Since the first day in Eurecom, I felt like home. The atmosphere in this institute was friendly and welcoming, and I have to thank all the colleagues, the secretaries and the IT people for all the help I received during this years in Eurecom.
The region of Sophia Antipolis is full of bright smart people that became very close to me and I can say right now to have a new family here.
I need to first thank my five stars: five girls that, on one way or the other, became

really close and trustable friends, they encouraged me for all these years in Sophia, giving me strength to achieve this PhD, and many other things. 5 different girls, from 4 different continents, they are all special in their own way and I am very grateful to be their friend: Leyla, Claudia, Carolina, Marianna and Christelle.
Then, there are the lighthouses of my PhD, the Italian Postdocs of Eurecom, they are not only very good friends, but also very good motivators and supporters when dealing with research. I always got wise advices from both of them, they were very important for me and my research and I am happy to have them: Matteo and Andrea.
There are then a huge amount of friendly, amazing people who encouraged me whenever I had bad moments and celebrated with me the good ones, understanding me when I was disappearing during deadlines or when I was stressed for this or that reason: all the people from Eurecom, especially Carmelo, Tony, Paul, Pavlos, Davide B., Davide C., Martina, Jonas, ..., and all the other smart friends I am honored to have, most of all the people of International Dinner!
Lastly, maybe the most important acknowledgement. Someone that truly believes in me no matter how much I don't make sense. He is my boyfrieand and the reason why I made it through this thesis with a big smile on my face.

This thesis is the product of many factors: the passion for multimedia information processing, the California sun, the Cote d'Azur sun, the love of my friends and family, the sound of rock music, the sound of reggae music, the need for traveling, the culture blending, the engineering education. I hope it is a good work.

Contents

1 Introduction **1**
- 1.1 The Importance of Multimedia Information Retrieval Today 2
- 1.2 How Does it Work? The Image Analysis Pyramid 5
- 1.3 Our Contributions at each Level of the Pyramid 9

2 State of The Art and Baselines for Multimedia Information Retrieval **13**
- 2.1 State of The Art 13
  - 2.1.1 Low-level Features: Local, Global, Aesthetic 13
  - 2.1.2 Feature Aggregation: Feature Encoding and Pooling 17
  - 2.1.3 Model Learning and Kenrels 19
  - 2.1.4 Which Evaluations? Which Applications? 23
- 2.2 Baselines and Datasets 26
  - 2.2.1 Baselines: Low-Level Features 27
  - 2.2.2 Baselines: Local Feature Aggregators 28
  - 2.2.3 Baselines: Learning Frameworks and Kernels 28
  - 2.2.4 Benchmarking Datasets and Their Experimental Setup 29
  - 2.2.5 Evaluation Measures 32

3 Level 0: Saliency-based Hybrid Features for Image Categorization **35**
- 3.1 Visual Attention and Saliency 38
  - 3.1.1 Computational Models for Saliency Detection 39
  - 3.1.2 The Spectral Residual Saliency Detector 39
- 3.2 Initial Analysis: Saliency-Aware Color Moments 40
  - 3.2.1 Why Adding Saliency Information to Color Description? . . . 41
  - 3.2.2 The Color Moments Feature 43
  - 3.2.3 Saliency-Aware Color Moments 43
  - 3.2.4 Experimental Validation 45
- 3.3 Saliency Moments for Scene Categorization 48
  - 3.3.1 A Biologically-Inspired Hybrid Descriptor 49
  - 3.3.2 Saliency in a Holistic Signature: Motivation and Key Elements 50
  - 3.3.3 Saliency Moments for Image Categorization 52
  - 3.3.4 Experimental Validation 56
- 3.4 Summary and Future Work 58

4 Level 1: Aggregating Local Features through Marginal analysis and Copulae **61**
- 4.1 An Introduction to Feature Pooling: Statistics and Existing Approaches 63
  - 4.1.1 LIDs as Random Vectors 64

4.1.2 The Major Existing Feature Aggregators and Their Statistic Analysis . . . 65
4.2 Our Approach: Marginal Modeling for Feature Aggregation . . . 67
4.3 Marginal Modeling: Visual Alphabets for Descriptor Aggregation (MEDA) Model . . . 68
4.3.1 The Signature: Marginals Estimation for Descriptors Aggregation . . . 70
4.3.2 Alphabet Construction . . . 71
4.3.3 Experimental Validation . . . 74
4.4 Multidimensional Modeling of Marginals: MultiMEDA Kenrel . . . 78
4.4.1 Peculiarities of The MultiMEDA Kernel . . . 80
4.4.2 MEDA from a Kernel Perspective . . . 81
4.4.3 Kernelized Multi-MEDA: Multidimensional Probability Estimation From Marginals . . . 83
4.4.4 Experimental Validation . . . 84
4.4.5 Video Retrieval . . . 87
4.5 Multivariate Modeling of Marginals: Copula Signatures . . . 88
4.5.1 COpulae and Marginal Signatures: an Overview . . . 89
4.5.2 Copulae: Linking Marginals with Joint Distributions . . . 90
4.5.3 COMS: Multivariate LID Analysis from Marginal Values . . . 92
4.5.4 Experimental Validation . . . 95
4.6 Summary and Future Work . . . 99

5 Level 2: A Multimedia Retrieval Framework Based on Automatic Relevance Judgments 101
5.1 Graded Relevance for Visual analysis: Motivations and Contributions 104
5.1.1 Why Graded Relevance? . . . 104
5.1.2 Automatic Graded Relevance Assignments for Multimedia Retrieval . . . 105
5.2 The Baseline: Binary-Relevance Learning Frameworks . . . 107
5.3 Our Approach: A Graded-Relevance Learning Framework . . . 108
5.3.1 Decision Values as Relevance Indicators . . . 109
5.3.2 A Multi-Level Training Set with Different Relevance Levels . 110
5.3.3 Multi-Level Prediction and Fusion . . . 110
5.4 Experimental Validation . . . 111
5.5 Summary and Future Work . . . 115

6 Level 3: Beyond Pure semantics: the Synergy with aesthetic analysis 117
6.1 Related Work . . . 121
6.2 A New Set of Compositional Features Modeling the Photographer's "Intent" . . . 121
6.3 Retrieving Appealing Images and Videos by Learning Flickr-based Graded Judgments . . . 125

6.3.1 Training and Test Datasets: from Flickr Interestingness to Video Appeal . . . 127
6.3.2 Our Proposed System . . . 128
6.3.3 Evaluation . . . 131
6.4 Enhancing Semantic Features with Compositional Analysis for Scene Recognition . . . 132
6.4.1 Analyzing Compositional Attributes for Scene Recognition . . . 133
6.4.2 Experimental Results . . . 134
6.5 Summary and Future Work . . . 136

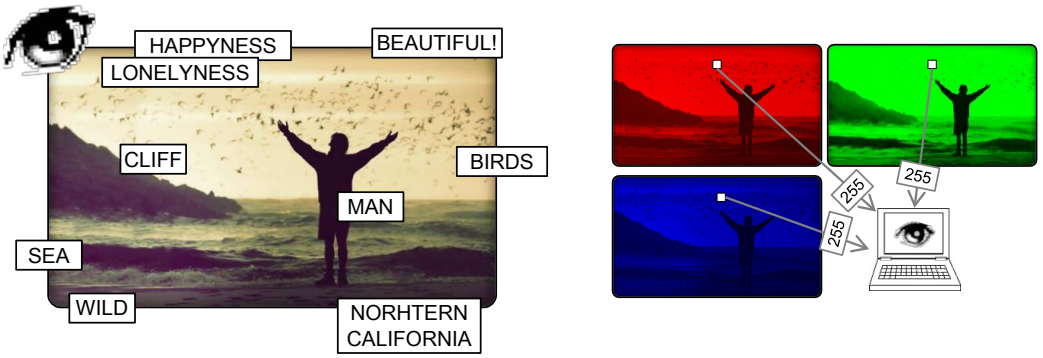
7 Conclusions and Future Perspectives 139
7.1 Our contributions From a Multidisciplinary Point of View: the Lessons Learnt . . . 139
7.2 Future Perspectives: Does Content matter? . . . 141

Bibliography 143

A Nouvelles Méthodes pour la Recherche Sémantique et Esthétique d'Informations Multimédia 157
A.1 Résumé . . . 157
A.2 Introduction . . . 163
A.2.1 L'importance de la Recherche d'Information Multimédia Aujourd'hui . . . 164
A.2.2 Comment ça Marche? La Pyramide des Analyse des Images . . . 167
A.2.3 Nos Contributions à Chaque Niveau de la Pyramide . . . 171
A.3 Conclusions et perspectives d'avenir . . . 176
A.3.1 Les Contributions d'un Point de vue Multidisciplinaire: les Leçons Apprises . . . 176
A.3.2 Perspectives d'Avenir: une Question de Contenu? . . . 178

Chapter 1

Introduction



"A picture is worth a thousand words". Unquestionable consideration, when dealing with human beings and their *human* vision system. When looking at a picture, we not only identify objects, actions, scenes, landmarks, but we also link the image content with our memories, with a set of emotions, movies, books, sensations.... However, when dealing with *computer* vision systems, the scenario changes. For an artificial vision system, the same picture is simply worth a thousand or more *pixels*, namely triplets of discrete numbers representing the amount of green, red and blue in an image point.

How can we allow computers to see the world like we do, and infer thousands of real words from a digital image? How can we make machines that transform *pixels* to *semantics* and other relevant information about the image content? One set of solutions is provided by **Multimedia Information Retrieval** (MMIR), a research discipline that helps bridging the gap between pixel-level values and semantic-level understanding.

MMIR techniques aim to automatically extract information about objects, scenes, emotions depicted in the image, based on the analysis of its *visual content*. MMIR researchers design frameworks that *learn* how to link the pixel values, summarized into non-redundant *features*, to a set of *intelligible concepts*.

A MMIR system is able to automatically **classify** or **categorize** an image based on its visual appearance, by automatically **recognize** the image content and properties. Given an image, a classification framework outputs a set of short descriptions of its content, that we call *labels* or *annotations*. Labels can be seen as positive or negative judgments regarding the **relevance** of an image with respect to a given concept (e.g. "there is a cat", "there is not a mouse"). Given the automatically

assigned annotations, MMIR systems can also go beyond classification, allowing for **content-based image retrieval**.

The *goal* of a Multimedia Retrieval framework, i.e. its *application*, depends on the nature of the annotations assigned to the image. A MMIR system can automatically classify or retrieve images not only based on their *semantics*, but also on the *emotions* it arouses, the degree of *beauty* of the given picture or how much an image is *interesting*.

MMIR systems are therefore complex frameworks whose performances depend on many factors: the type of features used, the learning framework, the redundancy reduction techniques, the quality of the annotations, the application etc.. Inspired by very diverse disciplines, in this thesis, we will discuss many different contributions to the improvement of the quality of MMIR systems, from global features based on saliency to graded-relevance learning frameworks. For each of the factors playing an important role in MMIR, we present novel techniques that improve the global *performances* of a Multimedia retrieval system, namely the *accuracy* of the labels predicted, or the *precision* of the retrieved results. We will mainly focus on two types of applications for our MMIR studies: **semantic analysis**, namely the automatic extraction of object, scene and general concepts labels, and **aesthetic analysis**, namely the automatic prediction of the image beauty degree.

In the remainder of this introductory chapter, we will give a broad overview of the motivation and the structure of this thesis. We will first explain in detail the importance (see Sec. 1.1) of MMIR. We will then give an overview of the key processes and steps that an MMIR system needs to automatically assign image annotations, and finally highlight our contributions to the field (Secc. 1.2 and 1.3).

1.1 The Importance of Multimedia Information Retrieval Today

In 2012, the Flickr[1] users have uploaded on the photo management website an impressive amount of pictures, 517.863.947 uploads. And Flickr is just one of the many on-line services that allow sharing digital visual content.

We live in a **digital visual world**: news, movies, pictures, user-generated images and videos... With the widespread diffusion of portable device and broadband internet connections, we produce, edit, and share huge amounts of image and videos almost *instantaneously*, allowing other users to access fresh, original visual content from every point of the interconnected world at the same time. Accessing this content means having an eye on the world: we develop ideas and concepts by receiving and transmitting visual information, we share emotions and memories through images because "an image is worth a thousand words".

Millions of users every day explore such multimedia space, by searching in the multimedia collections the media items that better suit their needs. And Multimedia Information Retrieval is about helping them to explore such space.

[1] www.flickr.com

As a matter of fact, in order to better organize, search, and select portions of this huge amount of visual information, we need efficient and effective tools to index and *retrieve* the elements in such data collections. One of the most intuitive ways for exploring visual information is to search and select image based on their *content*, by looking for visual data containing given *semantics*, e.g. "images with cats". One could also like to look at images by *beauty degree* or according to the *emotions* that the images generate, e.g. "scary images".

However, semantic, aesthetic and affective exploration is in practice not so simple. An image is worth a thousand words, but those words are in our human eyes, while the thousands of pixels forming an image are meaningless to a machine, if the machine does not know how to understand them. The digital visual world is very difficult to order and organize: we need therefore automatic procedures to support such exploration.

While textual documents can be indexed and retrieved, for example, by counting the word frequencies, when dealing with digital images the plot changes, because there are no actual "words" to index. One solution would be to order the visual data manually, by assigning a set of words to each image describing their characteristics, and then categorize and retrieve them based on such descriptions. However, given the volume of multimedia data we are dealing with, it is practically infeasible to ask humans to label the whole digital visual world by hand.

The general solution of on-line services for image retrieval is to work with textual information related to visual data. Given a textual query, such systems look for relevant images given their *contextual* information, namely all the collateral text related to the image but coming from external sources, and then rank the images based on their relevance to the concepts expressed in the query. For example, in popular web image search engines, such as Yahoo! or Google, the semantics of the image are inferred given the intelligible concepts inferable from the text surrounding the picture in the webpages related to the image. Another popular approach, used by the most common multimedia repositories such as Flickr, YouTube, or Facebook, is to retrieve and organize images given user-generated "tags", namely short textual labels that are assigned by the image owners and that somehow reflect some image properties such as content, location, emotions, etc..

Both approaches have several practical drawbacks. First, the concepts inferred from the image surrounding text are often unreliable when dealing with complex semantics, and especially with emotions. While the user-generated tags might be less noisy, manual semantic labeling of visual data is time-consuming and can be often incomplete, and most of the times totally absent. Moreover, discriminative contextual information appears in web images only, while for offline image collections, the only source of information that can help the image categorization is in its pixels. In order to infer the thousands of words expressed by an image, we would require therefore *intelligent* tools that can automatically deduce the image content given its visual appearance: we need frameworks that model the human vision and recognition system, and that can *recognize* the perceivable objects, scenes, aesthetics and emotions depicted in the image.

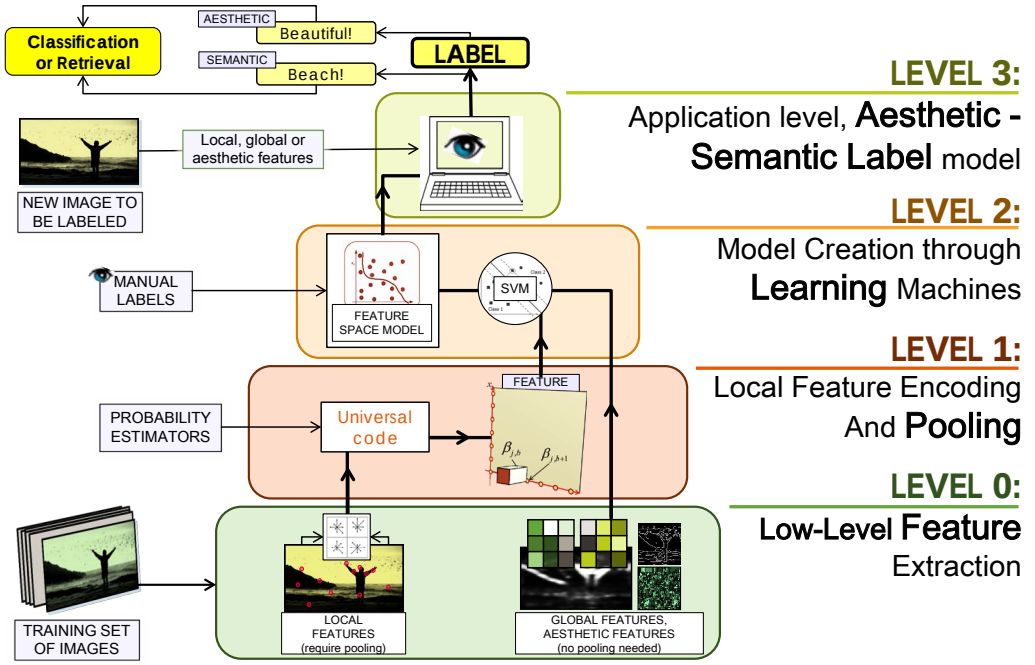


Figure 1.1: We represent a Multimedia Retrieval framework using a pyramidal structure: from feature extraction from label prediction and image ranking.

Multimedia Information Retrieval is a research discipline investigating these issues. MMIR indeed studies how to link the pixel intensity values of an image with its general meaning. Multimedia Retrieval systems automatically label the images with information about their content, by using learning machines that "see" the semantics, aesthetics and emotions generated by the picture, and translate them into intelligible words.

How does it work? In the following, see Sec. 1.2, we will illustrate a MMIR system with a simple pyramidal structure having different *level of abstractions*, where the base is the pixel-level processing for feature extraction, and the top corresponds to the application level, namely the label assignment (see Fig. 1.1). Everything start with a **groundtruth**, a training set of images annotated with their corresponding *known* labels, namely manually assigned annotations reflecting the presence/absence of given concepts. Given such groundtruth, MMIR systems first extract a set of features (level 0), namely a small set of very informative values regarding its visual appearance (e.g. what are the most dominant color in the image?). They then use such features, or their reduced version (level 1), together with the image labels as input for machine learning techniques, that learn a model (level 2) able to associate the feature values to the presence/absence of a given concept. Such intelligent system will be (level 3) then able to automatically annotate new, *unknown*, test images given their features and the computed model.

For each level of the MMIR pyramid, in this thesis we give a broad overview of

the existing techniques and design one or more novel solutions that aim to enrich the global visual analysis. We will learn how to increase the performances of the MMIR systems for semantic analysis by building new low-level features, novel feature aggregators, and a particular learning framework. We will then re-use the lessons learnt in our semantic analysis studies to build a Multimedia Retrieval System for aesthetic analysis. We will see an overview of our contributions in Sec. 1.3.

1.2 How Does it Work? The Image Analysis Pyramid

In this thesis, we look at a general Multimedia retrieval system as a layered pyramidal framework. In the **MMIR pyramid**, each level corresponds to a different stage of the visual information transformation process. Each level re-processes the outputs of the lower levels, starting from the raw image pixels and aiming at the automatic understanding of the image meaning. The higher the layer, the higher the level of abstraction, from the discrete, meaningless, pixel integer values, to the semantic/aesthetic image label intelligible to humans. The higher the level, the smaller the amount of information processed, from the complete pixel map to the simple image label. In the following, we will take a closer look at the characteristics of each layer.

Level 0: Low-Level Feature Extraction

Low-level features are the roots of any intelligent system for image analysis. Features, or *signatures, descriptors* are descriptive sets of numbers summarizing important visual properties of the image. This means that images with similar semantics or aesthetics should have similar low-level features. Features therefore help the machine to perceive the similarity or dissimilarity between images as humans do.

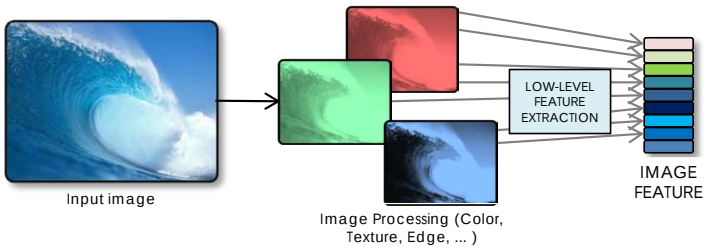


Figure 1.2: Feature Extraction: a small set of numbers describing the image properties is extracted from the image pixels.

At this stage of the MMIR process every pixel in the image is examined, and relevant information is summarized into a smaller set of numbers, stored in the final image signature. Feature extraction techniques can range from simple pixel average or counts, to the detection of edges, corners, and texture properties. Generally,

the types of features extracted changes with the goal, the application of the MMIR system. The pixel information can be "relevant" to a given task with different degrees: for example, the value of the *contrast* of a given image tells us more about the image aesthetic characteristics rather than highlighting the objects depicted.

But which features are important for semantics and aesthetics? *semantic features* are in general designed to describe the image *content*: the objects and their placement in the scene. In the literature of MMIR for semantic analysis we can find two opposite approaches for feature extraction: **Local Features**, such as SIFT [105] or SURF [6] namely statistical descriptions of the edges distribution around local interest points, and **Global Features**, that summarize general properties of the image into a single descriptor, such as color [175] or texture [187]. **Aesthetic features** are instead designed to describe the image *composition* and *style*, such as its contrast [116], its level of details [110] or its low depth of fields indicators [32].

Low-level features are the basement of every MMIR system and their *informativeness*, or *discriminative ability* namely the quantity of reliable information about the image content/aesthetics they carry, is crucial for the development of effective MMIR frameworks for image analysis.

Level 1: Feature Encoding and Pooling

In some cases, low-level features cannot be used directly as input for the third level of the pyramid, namely the learning framework, but they need to pass through an intermediate step that aggregates them into a compact image signature. This is the case of local semantic features such as HoG [31], SIFT [105], and SURF [6].

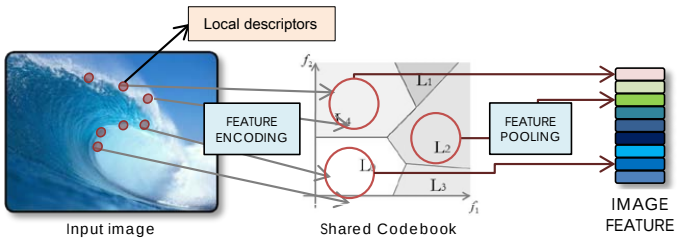


Figure 1.3: Feature Pooling: local features are aggregated into a compact image descriptor.

The reason for this issues is that learning framework require low dimensional, *fixed-length* signature, while, in local feature extraction, a *variable amount* of local descriptors are extracted from each image, making the direct extraction of signatures with equal dimensionality for all images impossible.

The general solution is therefore to *aggregate* all the local image descriptors into a compact fixed-length signature representing the behavior of the image keypoints. This is achieved by first **encoding** a training set of local descriptors into a smaller set of values, namely a shared codebook with a given number of "visual words". For a new image, the variable amount of descriptors is then **pooled** into a new image

signature that aggregates their properties by looking at their distribution given the shared codebook, having therefore a fixed dimensionality equal to the number of visual words.

Pooling is another important level of processing in MMIR: the aggregation process, generally achieved through vector quantization and high-dimensional clustering may cause losses of information regarding the global distribution of the local image descriptors, thus reducing the global informativeness of the features. The final image signature has to retain the original information as much as possible, while keeping the dimensionality low and equal for all the images.

Level 2: Model Learning

Once obtained a fixed length signature, by directly processing pixels with global semantic or aesthetic features, or by pooling the local descriptors, the next step towards the automatic understanding of the image characteristics is the **learning** step. At this level, we use supervised learning frameworks that *learn* how to distinguish between images containing different characteristics (i.e. different content or aesthetic degree), and then *predict* the labels corresponding to new images.

Similar to human brains, that recognize the world based on the association with their memories, supervised machine learning requires a **groundtruth**, namely a set of **training** images (features) for which the corresponding labels to be predicted are known. In order to express this knowledge, such annotations are generally previously assigned by hand, by asking humans to indicate the presence or absence of a given semantic concept, the emotion generated, or an aesthetic degree.

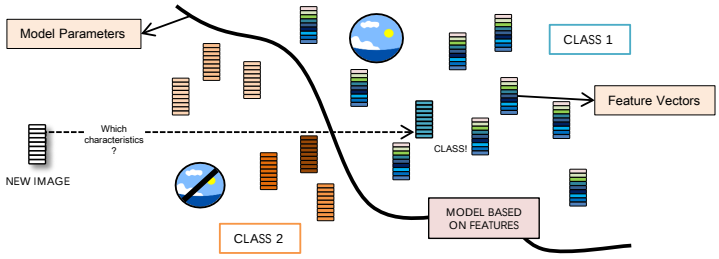


Figure 1.4: Model Learning: based on the feature values, the learning frameworks learn the parameters to distinguish between different image categories.

In general, each property to be predicted, i.e. each label, represents a separate learning problem. By processing the groundtruth, the learning framework determines the links between the feature values and the label (for example, a high value of blue color is likely to be correlated with the presence of water, and unlikely to represent the presence of grass). In case of aesthetic degree prediction, the learning framework learns to distinguish between appealing and non-appealing pictures given the aesthetic feature values.

The output of this step is a set of *models*, one for each label to be predicted, that

contains all the equations and parameters to partition the feature space into groups having similar labels (e.g. similar aesthetic degree, presence of the same object...).

The accuracy of the partitions in the learning step often depends on the quality of the training dataset: both the groundtruth annotations and the feature informativeness are crucial for the construction of a meaningful model. Moreover, the precision of the model can vary depending on the type of learning algorithm used, and how well it finds common patterns between the features. In general, for MMIR systems, we use Support Vector Machines [17], namely simple binary learning frameworks based on kernel similarity measures.

Level 3: Tasks and Applications

What is left in the Multimedia Information Retrieval chain? We are missing the prediction step, namely the automatic assignment of semantic/aesthetic labels to new, unseen images. At this level of the chain, we **test** the performances of our model by computing the accuracy of the predictions on the new images.

At the top of the pyramid, we do not deal with any training set nor large-scale feature analysis: for a new given image, the MMIR system *predicts* one or more labels, given the concept-specific models built in the lower level and the feature values of the new image. Such labels are predicted with a given degree of *confidence*, namely a value representing the reliability of the labels automatically assigned.

Depending on the **task** the system is designed for, MMIR systems can then present the results into two different ways. When the aim is to *classify* a set of images into pre-defined mutually exclusive categories (for example, scenes or objects), the MMIR system, given an image, outputs its corresponding category, and the performances are evaluated through the overall **accuracy** measure, namely the number of correctly classified images over the total number of images. When dealing with image *search*, the MMIR system presents the results of a given textual *query* from a pre-defined set of non-exclusive image labels. Given the confidence score assigned by the predictor, the MMIR system *retrieves* a list of images ranked according to their pertinence with respect to the query. A common way to evaluate the quality of such ranked lists is the Mean Average Precision, namely a measure that takes into account the amount of relevant image that are retrieved given the query, and the order in which they appear in the ranked lists of returned results.

At this level of the chain, an important aspect is the type of **application** the system is built for. Traditionally, MMIR frameworks are designed by researchers for **general semantic analysis**, i.e. object and scene recognition. Given the established importance of automatic semantic analysis over the years, the crucial element to determine the quality of a general semantic MMIR system is the **evaluation** step. Evaluation is generally performed using well known publicly available benchmarking datasets (mostly developed for scene [143, 133, 205] or object [41] recognition), built to compare the performances of various techniques using a common groundtruth, or by participating to international challenges and evaluation campaigns such as TrecVID [168] or Pascal VOC [39].

But retrieval and classification techniques can go also beyond the general semantic prediction, by applying, for example, semantic analysis techniques to narrower, **specific domains** that benefit from the automation brought by MMIR (for example, medical images, cooking videos, satellite images...). Moreover, MMIR techniques can be used to predict information that is not strictly related to the image object and scenes, abandoning semantics and focusing on **emotions**, **artistic** traits, **aesthetic** assessment.

1.3 Our Contributions at each Level of the Pyramid

Given the properties and crucial elements of the MMIR systems, in this thesis we propose a set of novel contributions at all levels of the MMIR chain, with the aim of improving the global visual analysis and the accuracy of MMIR systems built for semantic and aesthetic analysis.

The peculiarity of our work is its intrinsic multidisciplinary: we borrow, study, extend and re-use techniques from fields that are traditionally external or not directly related to Multimedia Retrieval. By introducing these new cues into MMIR systems, we build solutions that are not only very effective in terms of recognition, but generally also complementary to the existing approaches.

For level 0, 1 and 2 (feature extraction, pooling and learning) we design a set of new techniques that we test on general semantic classification/retrieval benchmarking datasets. In our first studies, the main application our techniques is therefore mainly semantic analysis. At level 3, we go beyond pure semantic applications and we build a system for the prediction of the image appeal and beauty, by embedding-using many of the techniques we proposed in the lower levels into an MMIR system we build for aesthetic analysis. In the following we will see an overview of our contributions.

Level 0: Saliency-based Hybrid Features for Image Categorization (Chapter 3)

In Chapter 3 we operate directly at a pixel-level, building a set of new, very discriminative low-level features for semantic analysis inspired by visual perception theory.

As we have seen, semantic features can be classified into two groups: *local* and *global* features. While the first ones are very informative regarding the image details and contours, and invariant to transformations, global features gather the general image behavior, somehow losing some accuracy when the illumination/rotation conditions change. Despite their effectiveness, the major drawback of local features is their computational cost and the pooling requirement; on the other hand, global features are extremely low-dimensional and fast to compute.

Our features stand in an intermediate point between the two mentioned approaches: we design a set of **hybrid features**, namely *global*, low-dimensional efficient features that embed some *locally*-parsed information. The local information

we integrate into a global descriptor arises from the **image saliency maps**, namely grayscale matrices with higher intensity values corresponding to the regions that more probably attract the human fixations in the image. We present two features following the hybrid approach: the **Saliency-Aware Color Moments** descriptor, namely an improvement of a color feature based on saliency techniques, and the **Saliency Moments** descriptor, namely a gist-based [133] descriptor embedding saliency information.

We test the effectiveness of our features for object categorization, scene recognition and video retrieval, and we show that they not only outperform existing features for MMIR, but that saliency bring also *complementary* information to the pool of existing descriptors generally used in MMIR.

Level 1: Aggregating Local Features through Marginal analysis and Copulae (Chapter 4)

In Chapter 4 we dedicate to the feature pooling and encoding, by proposing a set of new techniques for fast and effective local image descriptors aggregation, inspired by economic statistics modeling.

Our observation is that traditional techniques for feature aggregation need expensive procedures for the encoding techniques: they need to estimate the *joint* distribution of the components of the local image descriptors in a training set in order to build a shared codebook that will allow to produce a compact image representation. The compact image representation is then calculated by computing the *joint* distribution of the image local descriptors *given* the global codebook. This approach has been proved to be very effective [30, 79] for MMIR applications. However, one of the major drawbacks is its intrinsic storage and computational cost, together with the loss of information due to the pooling-encoding step.

The solutions we propose differ significantly from the traditional approach, and improve both efficiency and accuracy of traditional feature pooling methods. First, we design the **MEDA** descriptor, that describes the behavior of the image local descriptors based on the approximation of their **marginal** distribution, leading to an image signature that is extremely light to compute but that keeps high accuracy for classification and retrieval. We then improve this method by building **MultiMEDA**, a kernel for Support Vector Machines that is able to extract a *multidimensional* probability of the local image descriptors given the product of the marginal approximations stored in MEDA. Finally, we use Copula theory [166] to compute the real joint probability of the local image descriptors, based on the pure marginal information stored in MEDA. We model the *multivariate* probability of the image keypoints without involving any encoding process in the multidimensional space. The resulting **COMS** vector is proved to be much more effective than state-of-the art technologies for local feature aggregation applied to scene recognition and video retrieval.

Level 2: A Multimedia Retrieval Framework Based on Automatic

Relevance Judgments (Chapter 5)

In Chapter 5, we focus on re-designing the learning framework used for classification and retrieval, by introducing in MMIR some concepts from web information retrieval.

The main observation is that, in the learning step of MMIR systems, training labels are assigned using a *binary scale* (Relevant/Non Relevant). This means that the user-generated annotations identify the mere presence or absence of a given concept in the visual data, without allowing for intermediate options. However, a picture can be relevant to a semantic category with different degrees, depending on the way such concept is represented in the image.

Different from the most common frameworks, in Chapter 5 we build a learning framework that supports **graded relevance judgments**, namely multiple degrees of annotations reflecting the different levels of image relevance with respect to a given concept. Since manual annotation is an expensive and imprecise process, in order to quickly build graded ground truths, we propose a measure to reassess binary-labeled databases without involving manual effort: we automatically assign a reliable relevance degree (Non, Weakly, Average, Very Relevant) to each sample, based on its position with respect to the hyperplane drawn by Support Vector Machines in the feature space.

We test the effectiveness of our system on two large-scale databases, and we show that our approach outperforms the traditional binary relevance-based frameworks in both scene recognition and video retrieval.

Level 3: Beyond Pure semantics: the Synergy with aesthetic analysis (Chapter 6)

In the last technical chapter, we investigate a new emerging application of Multimedia Retrieval: aesthetic analysis. We apply new and existing MMIR techniques, traditionally used for semantic tasks, to the problem of automatic image appeal assessment. We re-use many of the lessons learnt in the previous Chapters and apply it to the prediction of the aesthetic degree of visual content. We add semantic cues to aesthetic analysis frameworks, and we see the improvement brought by the content-based features to the global image appeal prediction.

Moreover, we also explore the other way around: are aesthetic analysis tools useful for semantic analysis? We investigate the importance of aesthetic analysis for semantic applications, by testing the effectiveness of aesthetic features for a scene recognition MMIR framework.

In our contribution at the application level, we therefore enrich semantic and aesthetic visual analysis by exploring the synergy of those two applications for MMIR. The main idea is that semantic analysis and aesthetics are two closely related applications in Multimedia Retrieval. We show the benefits and the limits of this synergy, and propose some improvements in this direction.

This manuscript is structured by following the pyramidal structure we illustrated. First, in Chapter 2, we give a detailed overview of the state of the art techniques for MMIR systems applied to semantic and aesthetic analysis. Chapter 3 to 6 explain our contributions at each level of the MMIR pyramid, from feature extraction to the application level. Finally, in Chapter 7 we draw conclusions about the work carried out for this thesis, and picture some possible future tracks to follow in order to improve the existing and new technologies, aiming at the global enrichment of the automatic visual analysis.

Chapter 2

State of The Art and Baselines for Multimedia Information Retrieval

Multimedia Information Retrieval is a complex research discipline that involves different, multidisciplinary fields: from low-level signal processing to machine learning, passing through statistical modeling, many different research domains play an important role in the development of MMIR systems.

In this Chapter, we will review the most important research work in the field, looking at the key methods that have been proved to have a substantial impact for MMIR throughout the years. Moreover, we will take a close look at the most important datasets and baseline techniques for semantic analysis in MMIR. We will highlight in particular the methods and image collections that we will use throughout this manuscript to evaluate the effectiveness of our contributions for MMIR.

2.1 State of The Art

In this Section, we will outline the most important works for Multimedia Information Retrieval. We will structure this Section following the layers of the MMIR pyramid: for each of the level, we will highlight significant related works, helping the reader to have a clear idea of our reference background, on top of which we will build our new techniques outlined in this dissertation.

The main application of the techniques outlined in this Section is semantic analysis, due to its key role in the MMIR literature. We will discuss other possible applications in Sec. 2.1.4. We will start by focusing on low-level features (global, local, aesthetic, see Sec. 2.1.1), then look, in Sec. 2.1.2 at the relevant methods for feature encoding and pooling. We will then see key techniques for learning discriminative models in Sec. 2.1.3 and finally look at the evaluation benchmarks for semantic analysis, and at other possible applications for MMIR systems that have been proposed until now (see Sec. 2.1.4).

2.1.1 Low-level Features: Local, Global, Aesthetic

Low level features lie at the base (level 0) of the MMIR pyramid and represent the key elements for the development of effective MMIR system, since they store those image properties that should allow machines to discriminate between different image categories. The construction of the feature vector depends on the final application

of the MMIR system: specific features are deployed given the goal of the framework, aiming at modeling peculiar useful image properties. We will review here the most important features for semantic and aesthetic analysis, namely the two MMIR applications that we consider in this thesis.

2.1.1.1 Semantic Features

Semantic features aim at modeling the image *content*. They summarize into a few set of numbers the objects, shapes, contours, scene attributes that are depicted in an image. In MMIR literature, we can distinguish between two main approaches for low-level feature extraction for semantic analysis, namely local and global features.

Global Features

Global image features model the general properties of the image, namely holistic attributes that describe images "as a whole", without involving therefore detailed local analysis. Such low-level global features generally use basic signal processing techniques to directly process the pixel values and extract salient information about the image color distribution, texture patterns, or, for example, the edge distribution. Global features are generally used for scene recognition [142, 181], since they represent holistic information about the image, or as complementary features for complex content-based video retrieval systems [153, 128].

Information about the image **color distribution** is quite straight-forward to compute, since the pixel values represent the amount of red, green and blue colors in a given image point. The first, intuitive approach that represents the chromatic information, namely the color histogram, has been proved in the early years to be an effective way to describe images [180, 58]. The simple histogram structure has been improved by considering spatial information by Rao et al. [145], and by using kernel density estimation techniques for histogram modeling in [192]. Color histogram has been successfully improved also by Smith and Chang [169] by considering a more biologically-inspired color space, namely the HSV space, and then using the resulting histogram for fast image search.

Following the histogram idea, a faster and more robust descriptor has been proposed in [175], where the first three moments of the color distribution are stored in the Color Moments (CM) feature. Color correlograms [71] represent a step further towards the accurate modeling of the image color properties, by considering the spatial correlation of the various colors. In [138], Pass et al. further improve the statistical modeling of color information, by proposing color coherence vectors, resulting from classifying each pixel as coherent/non coherent based on whether the pixel and its neighbors have similar colors. Recently, very efficient color distribution entropy measures have been shown [178] to be very effective for image retrieval.

Texture properties are also very important for the description of image content. They represent the description of the patterns of the surfaces depicted in the image, bringing therefore precious information for image discrimination. Tamura textural features are among the most popular features, they were first presented in [182]

as a way to model texture using a computational approach following the human visual perception principles. Popular texture models are also the Gray-level co-occurrence matrices, proposed by Haralick [60], that extract feature information using second order statistics. Wavelet packets have been more recently used in [187], to characterize textures at multiple scales, and efficiently re-used for Video Retrieval in [153]. Similarly, Manjunath et al. [115] use Gabor wavelets for browsing and retrieval of image data.

The **edge distribution** is another relevant cue for content-based image retrieval, since it describes the amount of edges and their orientations in the image. The most widely used edge descriptor is the MPEG7 Edge Histogram [200], where the amount of edges at each orientation is stored in a window-based histogram. Recently, edge distribution has been improved by computing edge orientation autocorrelograms [111] and by computing geometric distributions of edge pixels along the dimensions of angle and radius, generating the Angular Radial Edge Histogram [144].

Holistic scene features are a particular type of features that aim at summarizing the global shape of the scene by computing a general description of the image contours. The most popular holistic feature for scene recognition is the Gist descriptor [133], that samples the values of the Fourier transform of a given image to obtain a "spatial envelope" of the image. Similarly, Schyns and Oliva[161] represent the global scene gist by using oriented blobs in a particular spatial organization to obtain a coarse description of the input scene. Another holistic approach is presented in [10], where "Geons" are used to represent the scene shape as arrangement of basic geometrical forms.

Local Features

Local Image Descriptors are extracted from localized areas in the image, and generally they are built to describe the surroundings of local interest [44] or densely sampled [42] points. Since the amount of Local Image Descriptors varies depending on the image structure, such descriptors cannot be directly used as input for learning machines, but they instead need to pass through an aggregation process (see next Section).

One first issue about local image analysis is the detection of the **points of interest**, namely salient points in the images containing important information regarding the image content. Besides dense sampling [42], that assumes that interest points are equally spaced throughout the image, two main interest point detector have been widely used in literature, namely the Harris detector [64], invariant to image rotation, and the Difference-of-Gaussian (DoG) detector, presented in [106]. Later on, Mikolajczyk and Schmid [117] apply scale selection after using a multi-scale framework to detect interest points, invariant to scale changes. Image intensities are used in the work Tuytelaars and Van Gool [186] to detect affine invariant regions. More recently, affine invariant interest points have been extracted using an iterative algorithm then modifies location, scale and neighborhood of each point in [118] and proved to be very effective for image categorization. This work was further improved

by [119], where A Harris-Laplace detector is proved to be invariant to scale, rotation and affine transformations.

Once extracted the interest point, the second issue for local feature extraction is how to **describe** their surroundings. Various techniques have been proposed in the recent years, the most widely used being the SIFT [106] descriptor, designed by Lowe, that first detects interest points using the DoG method, then rotates the corresponding surrounding patches in the sense of their dominant orientation, and finally describes each patch by capturing the local magnitudes and orientations for each subwindow resulting from a 4x4 subdivision of the image gradient of the patch area. The resulting SIFT descriptor is 128-dimensional, having 16 sub-regions described by 8 orientation bins. Such high dimensionality of the SIFT descriptor is reduced in the PCA-SIFT [87], where principal component analysis [84] is applied to the normalized gradient patch.

Several other local descriptors have been proposed to overcome some disadvantages of the SIFT vector. For example, SURF [6] (Speeded-Up Robust Features) have been presented to improve the efficiency of local analysis by using gradient images and Hessian detectors. GLOH (Gradient and Location Oriented Histograms) have been proposed by Mikolajczyk and Shmid in [120] to improve the SIFT descriptor accuracy, by re-arranging the grid of the patch and allowing for more orientation bins, and by finally reduce the dimensionality of the resulting descriptor using PCA [84]. Shape Contexts have been also used to describe local shapes, by collecting in a histogram the distribution of the orientations of the vector connecting a set of points sampled from a shape contour. HOG (Histograms of Oriented Gradients) features [6] represent nowadays one of the most accurate way to represent images (for a comparison of various descriptors, see for example [205]). The HOG method compute local edge orientations over image cells, namely radial or rectangular small image regions, and then collects them in a weighted histogram.

2.1.1.2 Aesthetic Features

Aesthetic Features have been used in literature for MMIR system applied to aesthetic analysis, namely the prediction of the image aesthetic and appeal degree.

In general, aesthetic features aim at modeling the photographic rules using a computational approach, namely highlighting the properties of the image *composition*, such as symmetry, contrast, etc... Pioneer work in aesthetics was the paper from Datta et al. [32], where various features inspired by photography theory were invented. For example, Light Exposure, computed using the average pixel intensity, Image Colorfulness, namely the image relative color distribution, the Rule of Thirds, calculated by averaging the Hue, Saturation and Brightness of the inner rectangle resulting from a 3x3 division of the image, the Shape Convexity, obtained through the covex hull of local patches, the Low depth of Field Indicators, inferred from the wavelet coefficients of local rectangular image blocks, and the Region Composition, namely a measure evaluating how many distinct color blobs and how many disconnected signi?cantly large regions are present in the image.

The pool of aesthetic features is enriched by [100], where, in order to describe the composition of paintings, images are segmented into smaller areas and shape features such as mass center, variance, and skewness are calculated. Moreover, global and segment-based contrast measures are extracted. Later on, Obrador et al. compute golden mean and triangle rules [131] as an extension of the Rule of Thirds feature in [32], and create new features such as Simplicity, based on the number regions resulting after segmentation, Layout's Pleasantness, given by the average distance between centroids of the relevant regions, and Visual Balance. Recently, Bhattacharya et al. [8] used high-level features for the enhancement and assessment of photo quality: for example, Relative Foreground Position, i.e. the distance between the foreground's center of mass, to each of four corners of the inner rectangle resulting by a 3x3 division of the image, and Visual Weight Ratio, computed by counting the number of pixel of the sky region divided by the number of pixel in the foreground region.

2.1.2 Feature Aggregation: Feature Encoding and Pooling

Local Image Descriptors, as mentioned in Sec. 2.1.1.1 need generally to be re-processed before passing to the next step, the learning stage. This is achieved by aggregating them into compact image signatures that represent their global distribution over the image using a few discriminative values. In order to obtain such representation, the general approach employed by the most common feature aggregators is to follow a two-step process: first, a set of training local descriptors is *encoded* into a few values representing the global distribution of the keypoints, i.e. a *universal model*, and then, for each image, local descriptors are *pooled* into an image signature representing their distribution given the universal model. Several approaches have been proposed to accurately design the two steps of the local descriptor aggregation process.

2.1.2.1 Encoding Methods

The creation of a universal model that reflects the global density of the local image descriptors is of crucial importance. Such universal model should capture the general behavior of the image keypoints so that the final image signature obtained with pooling can be discriminative enough to distinguish between different image categories. The general approach is to cluster or analyze the keypoints of a training set of images, and generate a global model through a shared codebook or a parametric model of the multivariate probability distribution.

One of the most popular solutions for feature encoding is the **Vector Quantization** performed by the **Bag of Words (BoW) model** [30]. In this scenario, the keypoints are first clustered into a **visual dictionary** partitioning the descriptor space into areas containing equal numbers of keypoints. The visual dictionary, or codebook, contains a set of shared **visual words**, namely feature vectors representing the centroids of the clusters in the feature space. The pooling process will

construct the final image signature by counting in each image the occurrences of such visual words. The BoW was first introduced by Csurka et al. in [30], applying k-means clustering on a training set of local image descriptors and then using the centroids of the resulting clusters as visual words. Various techniques have been proposed later on to vector quantize the keypoint space and improve the construction of the visual codebooks. For example, in [98] mean-shift clustering is used, [129] hierarchically quantizes LIDs in a vocabulary tree and [122] uses Extremely Randomized Clustering Forests to build efficient visual codebooks. Another way to define visual codebooks is proposed in [185], where the codebook is composed of the hypercubes resulting from the quantization of each dimension of the LID into a fixed lattice. In a different approach, vector quantization efficiency is improved in [206] by generalizing it to sparse coding.

While all the mentioned methods use *unsupervised* clustering, encoding approaches based on **supervised encoding** have been recently proposed to improve the performances of MMIR systems. More discriminative codebooks have been proposed by Winn et al. [199] that merge codewords that are proved to be less discriminative for the MMIR task, and later by Lazebnik and Raginsky [95] that maximize the mutual information between features and labels in the encoding step. Supervised visual dictionaries for sparse coding have been proposed by Mairal et al. [112] and optimized by Boureau et al. in [20].

A universal model can be also constructed by **generative parametric** models, namely by computing the global probability density function of the descriptors in a training set by fitting a known parametric distribution. This approach have attracted a lot of attention in the recent years [79, 72, 40], especially when coupled with a pooling step based on discriminative approaches, since treating multivariate distributions becomes non feasible when using learning machines. In these approaches, a Gaussian Mixture Model [135] is used as a way to describe the global distribution of the image keypoints. In this approach, each Gaussian in the mixture could be seen as the equivalent of a visual world in the Bag of Word model.

2.1.2.2 Pooling Methods

Once the encoding step has created a universal shared model reflecting the global behavior of the keypoints in a training set of images, the pooling step needs to be performed in order to aggregate the local descriptors in an image into a visual signature that reflects their joint distribution. This step is performed in different ways, depending on the type of universal model built in the encoding step (visual dictionary or parametric model).

When the universal model is a visual **codebook** of n visual words, the most simple approach is the *hard assignment* [30], that approximates each keypoint in an image to the closest visual word and then collects the occurrences of the visual words in a n dimensional histogram. This intuitive approach has been improved in various ways. For example, Van Gemert et al. in [190] apply soft assignments by taking into account the distance of each visual word to the closes cluster centroids.

Similarly, Jegou et al. in [77] use Hamming Embedding to refine the visual words assignment by including a binary signature representing the position in the Voronoi cell. Jegou et al. in [78] also improve the hard assignment approach by computing, for each point, the element-by-element distance with the closest visual word, and store in the VLAD vector the resulting values.

One major issue regarding classical visual words assignment is the lack of spatial information. Lazebnik and Raginsky in [96] address this problem by introducing the concept of "Spatial Pyramid", where a different histogram of visual words is assigned to different image regions. This work is improved in [194] where each descriptor is projected into its local-coordinate system, and then max pooling is performed to integrate the projected coordinates and generate the final image representation. The Spatial pyramid approach is further improved by Boureau et al. in [19], that restricts the pooling process not only to localized areas in the image space, but also to local hypercubes in the descriptor space.

When the universal model is a **generative parametric model**, the output of the encoding step is a continuous probability distribution function describing the global keypoints distribution. Given that learning machines require a finite, fixed length image signature, how to transform this continuous model into a discrete set of numbers, i.e. an image features? Two, very popular approaches solve the problem by coupling generative encoding with **discriminative pooling process**. For example, Perronnin et al. in [79] first estimate the global density using Gaussian Mixtures, and then use Fisher Kernels [79] over image keypoints to generate the Fisher Vector signatures, that reflects the way in which the parameters of the distribution of the image keypoints should be changed to fit the global Gaussian Mixture. Fisher Vectors are proved to be one of the most effective solutions for LID-based image analysis. Another approach, inspired by the supervectors of speaker recognition [22] is proposed by Inoue et al. in [72], where, after fitting a GMM with the training descriptor, each image is represented by a supervector containing the adapted mean values.

2.1.3 Model Learning and Kenrels

Learning is a fundamental step for MMIR, since it represents the "intelligence" of the system, the way in which the machines can memorize how to associate given feature values to given image properties. Single similarity measures are too weak to determine such complex links: the general approach is therefore to embed similarity measures in more complex learning frameworks for classification and retrieval. Depending on the task of the MMIR systems, two types of learning frameworks can be used: unsupervised and supervised models. The first one works on absence of labeled data, and it is generally used for the ranking or re-ranking for query-by-image retrieval, namely retrieval systems where the query is directly a digital image, and the aim is to retrieve similar examples. Supervised learning is used instead for classification, categorization, and for ranking of the retrieval results.

2.1.3.1 Unsupervised Learning Methods

Unsupervised learning methods have been widely used to visually separate the images in multimedia collections, and new, ad-hoc clustering methods have been created for this purpose. For example, Gordon et al. in [56], cluster images by exploiting the information bottleneck (IB) principle, namely by maximizing the mutual information between clusters and image content. Another example can be found in [26], where a dynamic clustering approach is proposed, by applying a graph-theoretic clustering algorithm to a collection of images in the vicinity of the query.

Moreover, various unsupervised methods have been used in MMIR for **web images exploration**. For example, in [50], consistent bipartite graph co-partitioning is used to cluster Web images based on the consistent combination of visual features and image surrounding texts. Similar approach, but more text-oriented is presented by Jing et al. with the IGROUP interface for web image clustering [82]. A re-ranking approach is proposed in [191], where image search results are diversified using lightweight clustering techniques in combination with a dynamic weighting function of the visual features.

Clustering techniques have been also recently used to boost image annotation and classification systems, such as frameworks for automatic annotation of personal albums [81], object recognition [45], and as we have seen, for local feature encoding into visual words [30, 129, 122].

At the edge between unsupervised and supervised Learning, several approaches have tackled the problem of **semi-supervised learning**. Semi-supervised training is a way for reducing the effort needed to annotate the groundtruth, by training the model with a small number of fully labeled examples and an additional set of unlabeled or weakly labeled examples. For example, Rosenberg et al. [156] use self-training for object detection, and Zhou et al. in [212] improve content-based image retrieval performances using an approach inspired by co-training that incorporates unlabeled data. Other works use semi-supervised boosting [114, 99] to improve the learning with unlabeled examples for image retrieval.

2.1.3.2 Supervised Learning Methods

When groundtruth annotations are fully available, one of the most used learning approaches for both retrieval and classification of digital images is the supervised learning. In this scenario, a training set of images is previously annotated with labels indicating to the presence/absence of given image properties. The problem often reduces to a classification problem where the task of the learning framework is to find common patterns between the feature values of the images belonging to the same category.

Classification methods can be grouped into two major approaches: generative and discriminative models. In generative modeling, each class is represented with its *probability distribution*; then, for a new image, the Bayes formula is used to compute the posterior probabilities, that represent how likely is that the image belongs to each of the classes. Discriminative models estimate the class *boundaries*

by looking at similarities and differences between features belonging to different categories.

Generative Models

Generative Models have been used for scene and object categorization in MMIR for their ability to deal with many classes. A particular approach is the one of Bosch et al. [15], that model scenes classes based on the object distribution discovered using Probabilistic Latent semantic analysis over bag of visual words, and then classify the test images using a k-nearest neighbor classifier. Another popular approach is to use Conditional Random Fields, as proposed by [66] to classify natural images.

One of the more practical ways to model the class density is the Gaussian Mixture Model. For example, in [140] GMMS are computed based on color and texture features, and then used in various classification tasks. Generative Models using Dirichlet Mixtures have attracted a lot of attention in the recent years. For example, Kivinen et al. in [90] use Dirichlet Processes for marginal modeling and Markov trees for feature dependency modeling in order to represent the density of each class, and then use marginal likelihood to assign the scene category. Latent Dirichlet Allocation is also used in [43] to discover mid-level "themes" based on which each class is modeled and then to classify new images based on Bayesian rule.

Discriminative Models

Discriminative learning frameworks directly model the separation between the classes in the feature space, and store the parameters useful to characterize such separation.

The easiest discriminative learning framework is the **Nearest-Neighbor** technique, where no learning on the training set is required, and new sample is classified by calculating the distance to the nearest training case, and assigning the label accordingly. In MMIR literature, one of the works successfully employing Nearest Neighbor, that have been received substantial attention from the community, is the one in [13] by Boiman et al. for object categorization.

Decision trees [158] are very efficient and discriminative tools for classification: the aim is to minimize the global entropy of the training set by splitting the data at the optimum threshold. In semantic analysis, decision trees in their simple form have been used for region-based classification in [104], and as an ensemble in a random forest in [16].

Neural networks are more complex learning schemes aiming at modeling the biological functioning of our brain. They have been widely used in MMIR due to their ability of solving multiclass classification problems. For example, Hopfield Neural Networks have been used in [125] for object recognition and an extended version of neural network, namely a dynamic link architecture that group neurons dynamically into higher-order entities was proposed in [93] for person recognition. More recently, convolutional neural networks have been used for ImageNet classifi-

cation by Krizhevsky et al.[92].

Support Vector Machines are probably the most widely used learning algorithms for image classification and retrieval. The basic idea of SVMs is to find a hyperplane that separates positive from negative examples. This is done by evaluating the similarity between training examples using specific **kernel** measures. It is a powerful tool for visual-feature based learning because of its optimized, fast algorithm and its high generalization ability.

Support Vector Machines have been durably used for several MMIR applications [32, 110, 205] and in several evaluation campaigns such as TrecVID [168] due to their effectiveness in binary classification. When dealing with multiple classes, the common approach [69] is to reduce the problem to a set of binary classification problems, by 1-vs-all or 1-vs-1 groups of classifiers. In the first case, a model is learn from each class that separates it from all the other classes, while in 1-vs-1 support vector machines, a model is learnt to distinguish between each pair of classes.

SVMs can also be used to go beyond the simple global image labeling: for example, in Multiple Instance Learning [25], used for region-based image categorization, images are viewed as bags, each of which contains a number of regions resulting from image segmentation. MIL defines a bag as "positive" if at least one of the regions in the bag is positive, otherwise, the bag is labeled as negative. Then, a Diverse Density function is used define a set of instance prototypes, namely pattern of instances that are more likely to appear in given classes, and then such prototypes are learnt using SVMS.

SVMs are also used for non-classification tasks. For example, active learning frameworks, [53, 183, 195] extensively use SVMs to choose the examples to interactively query the users that are manually labeling a new dataset. Similarly, Relevance Feedback algorithms [210, 67, 27], take the results that are initially returned from a given query and analyze through SVMs the relevance of those results with respect to that query given the user response.

One of the core elements for the effectiveness of the SVMs learning is the **kernel** used to evaluate similarities between feature vectors and define an optimal decision boundary, namely a hyperplane in the feature space. When the input samples are linearly separable, the similarity between two features v and w is computed with a simple dot product $v \cdot w$. However, in many cases, e.g. in multimedia data representation, decision boundary is not linear: one common solution is to define a transform ϕ that maps the input space in the feature space $v \rightarrow \phi(v)$ and then use a kernel function $k(v, w) = \phi(v) \cdot \phi(w)$ to represent the dot product in the high-dimensional feature space.

One of the most common choices for kernels is the Radial Basis Function kernel, that has been proved [210, 148] to be very effective for image retrieval. However, due to the diversity of the features used for image recognition and retrieval, several work focused on building ad-hoc kernels for specific descriptors. For example, Histogram Intersection kernels, originally built in [180] to match color histograms, have been proved in [113] to be efficient tools for image classification. Another example is also

brought by the work of Lazebnik et al. in [97], that use Spatial Pyramid Kernels to add the spatial information in the BoW model learning. A particular approach for kernel modeling is represented by the work in [12] where local image descriptors are mapped into a low dimensional feature space, then averaged to obtain a set-level random features, and finally using a linear classifier to model the resulting vectors. Probabilistic approaches, such as Bhattacharaya kernels and Kullback-Lieber divergence kernels, have been used in [40] to compute similarities between images described by fitting GMMs with the image keypoints.

2.1.4 Which Evaluations? Which Applications?

At the end of the MMIR chain, at the top of the pyramid, the system takes as input an image, and, given its learnt model and the image feature, assigns a label to the new image. Generally, labels are output together with a confidence score representing the likelihood that the image can take such label, based on the position of the image feature in the feature space. Such confidence score is generally used to rank results in retrieval frameworks, where images are ranked according to their relevance to a given concept or query.

The nature of the labels assigned, and therefore the type of the user query, determines the *application* of the MMIR system. While the *structure* of the MMIR system is fixed (feature extraction, pooling, learning, prediction), the *type of information* processed by the system is adapted according to the *goal* of the Multimedia Retrieval Framework. For example, if the *application* of the system is to predict the emotions that the image arouses, then the underlying features, annotations, and learning frameworks will be adapted to reflect the affective content of the images.

MMIR systems can be built for different applications, the most popular one being semantic analysis for general concept detection: almost every technique listed until now have been built for this purpose. In such research works, at the application level the performances are evaluated by comparing new and existing techniques on benchmarking datasets.

Semantic MMIR can be also applied to a particular domain (e.g. medical imaging, space imaging, ...), showing the usefulness of semantic analysis outside the pure research context. MMIR techniques can also go beyond semantic image classification: they can be employed for more diverse applications such as artistic, aesthetics affective image analysis.

In the following, we will analyze key aspects of the application level: we will first look at the general semantic analysis techniques from an evaluation point of view, and then show various works that apply semantic analysis to narrow semantic domains. We will then look at how MMIR systems can be used for a variety of applications different from semantic analysis.

2.1.4.1 Semantic Analysis

The main, traditional application of MMIR is **semantic analysis**, namely the automatic recognition of image objects and scenes. The majority of the techniques we have seen before are designed to fulfill this tasks. At an application level, two main interesting elements can characterize the novelty of semantic analysis techniques: their effectiveness compared to state-of-the art techniques, and their application to novel, relevant sub-domains.

Determining the Effectiveness of Semantic Analysis Techniques: Datasets and Evaluation Campaigns

In order to evaluate the effectiveness of semantic analysis approaches, one of the most common procedures is to compare the performances with existing, state of the art approaches. Several databases for object [41, 57] and scene [143, 133, 205] categorization have been built as **benchmarking image collections** that can support the development and the comparison of new and existing techniques for MMIR. In this thesis, we extensively use such image collections and in the next Section, we will look at some of this datasets in details.

Moreover, there exists several **evaluation campaigns and competitions** that aim at gathering the works in the field from different research groups around the world, and evaluating their efforts on a common, large scale database. In general, such campaigns provide the participants with training sets of annotated data, and the task is to build system addressing semantic MMIR problems. Such systems are then employed to label a test set of unlabeled images, and results are evaluated by matching them with manual judgments. Examples of such competitions are the Pascal Visual Object Classes Challenge [39], for object recognition, the ImageCLEF [123] for cross language annotation and retrieval of images, the TrecVID [168] evaluation campaign for semantic indexing, search and retrieval of video collections, and the recently appeared MediaEval [94] benchmark for multimedia retrieval focusing on multimodal approaches involving text, speech, social information, etc.. All the mentioned evaluation campaigns are useful to share new ideas and establish permanent knowledge for semantic MMIR.

Applying Semantic Indexing to Real Problems

While traditional semantic analysis aim at recognizing objects and scenes with general, large-scale semantics, many curious and useful applications for semantic analysis have been explored throughout the years to automatically classify images coming from narrow semantic domains. In the following, we will outline a non-exhaustive list some of these applications, giving an idea of the potentialities of MMIR tools for semantic analysis.

One of the most popular applications for semantic analysis is **medical image classification**, namely the automatic identification of sicknesses of bodies part given medical visual data such as echography, RMIs, x-rays, etc. As an example of its

importance, the ImageCLEF challenge has an entire task dedicated to the annotation and retrieval of medical images [124], and participants achieve impressive results for this task, in particular when combining visual and textual information together. Another very useful application for semantic analysis techniques is about **satellite images**, see [198] for a review, namely the identification of objects and regions given remote sensing images. Multimedia Information Retrieval can be also useful for **computer security**, see for example the works for Captcha generation [33] and breaking [21] using visual features.

One application that recently attracted the attention of the MMIR researchers, brought in the ImageCLEF benchmark in 2011, is **plant identification** [52], namely the recognition of plant and flower species, useful for the automatic monitoring of the environment. Similarly, a substantial amount of research work have been carried out [172] to develop image classification techniques for **ecologic activity monitoring**, such as underwater activity identification, the categorization of animal species, and so on. Another interesting way to apply MMIR techniques for semantic analysis is to build classification systems supporting **cooking activities** [37], for example cooking gesture recognition, or ingredient identification.

2.1.4.2 MMIR for Other Applications

Multimedia Information Retrieval is not only about semantics: many different types of labels can be predicted given a visual recognition system, the structure is similar, but the information extracted from the images varies.

One of the most widely explored branches of MMIR using content-based techniques is **aesthetic image analysis**, aiming at building systems that automatically classify the image beauty and appeal. Pioneer work in this field is the one from Datta et al. [32], that learn features that model photography rules, and use a groundtruth of web images from Photo.net annotated with aesthetic judgments averaged over a large number of users to predict the image beauty. Wong et Al improve it in [201] by adding saliency information in the prediction framework to distinguish between amateur and professional pictures. Obrador et al. in [131] further improve the work in [32] by adding more compositional features. A step towards the incorporation of image semantics into an aesthetic framework is represented by the work of Obrador et Al. [130], that build different aesthetic models for different image categories, using pre-defined manually labeled image categories. The use of semantic features for aesthetic prediction has been explored also in [35], where semantic concepts such as animals, scenes, people, are detected and the probability of their presence is used as an attribute to predict image aesthetics and interestingness.

Multimedia Retrieval frameworks have been recently extensively used for **affective image classification**, namely the categorization of images based on the emotions they arouse. A first, very simple approach for emotion recognition based on color was presented by Colombo et al. in [29], and expanded by [9] by including textural and shape features for a complete affective-based image retrieval system. Textual and visual information are later combined in [204] for affective image re-

trieval by defining an affective space, and similarly in [197] an affective space is determined through psychological studies, and then features such as luminance-warm-cool fuzzy histogram, saturation-warm-cool fuzzy histogram integrated with color contrast and luminance contrast integrated with edge sharpness are used to learn affective models. In [59], Hanjalic et al. apply aesthetic techniques for video classification and presentation. In the recent years, a research work that became very popular for affective analysis is the one from [110], that infer a set of affective features from psychology and art theory for a complete affective classification system. Similarly, Lu et al. in [107] recently propose to model image emotions based on shape features.

Another interesting application of MMIR, is **artistic image analysis** that has been widely used for painting analysis and cultural heritage preservation studies. Several types of knowledge can be automatically inferred from paintings, and this branch of MMIR aims at building systems that model such knowledge. For example, given the painting contained in specific collections, [83] build systems that automatically infer who is the artist that painted them. Another interesting work was presented in [75], where MPEG7 descriptors were used to create the profiles of art painting images (i.e. artist, current, ...). Similarly, in [74] Ivanova et al. group paintings by color harmonies and color using learning techniques. Panting cracks are classified into typical patterns in [2] using local and global features. artistic Image analysis combines with aesthetics in [100], where the beauty of paintings is evaluated using aesthetic features. Similarly, we find the presence of both artistic and affective analysis in the work of Zhang et al. [209], that assess the affective content of painting using low-level features.

2.2 Baselines and Datasets

In order to clearly understand our contributions throughout the manuscript, and understand the novelty of our approaches, we present here a detailed analysis of the benchmarking datasets and baseline techniques that we use to compare and evaluate the performances of our descriptors, aggregators, learning frameworks and applications.

For each of the level of the pyramid, corresponding to each Chapter in each thesis, we outline here the implementation details of a set of techniques that have been widely used in MMIR literature for Semantic analysis. These techniques represent our **baselines**, namely the reference methods whose performances we compare with our new proposed solutions. We will therefore focus on the most popular descriptors, feature aggregators, learning techniques for semantic MMIR. We then discuss here the peculiarities and experimental set-up of commonly used benchmarking databases that we use to evaluate our techniques, together with the commonly used evaluation measures we employ for this purpose.

2.2.1 Baselines: Low-Level Features

We detail here the properties of very popular low-level features, whose performances for MMIR we will compare with our methods. Following to the feature definition we proposed, we chose both global and local features that have been proved to be very effective for image and video classification and retrieval.

The **Global Features** we extract for evaluation are as follows:

- **Color Moments [175]** is a color descriptor incorporating higher order statistics. This global descriptor computes, for each color channel in the LAB space, the mean, variance and skewness of the pixel values lying each subregion resulting from the division of an image with a 5×5 grid.

- **Wavelet Feature [187]**. This texture-based descriptor calculates the variance in the Haar wavelet sub-bands for each window resulting from a 3×3 division of a given keyframe.

- **Edge Histogram [200]**. The MPEG-7 edge histogram describes the edges' spatial distribution through the amount of vertical, horizontal, diagonal, minor diagonal, and non-directional edges in each of the 16 blocks resulting from a 4×4 subdivision of the image. Each image block is then further divided into smaller regions, the direction of the most prominent edge in each region is taken and a 5-dimensional (1 for each direction) histogram is updated accordingly.

- **Gist Descriptor**. This holistic descriptor has been first introduced in [133] as a powerful descriptor for scene recognition. Its values represent the average over the windows in a 4×4 of the Fourier spectrum sampled with a set of Gabor filters.

Moreover, we also consider a set local descriptors that we will then use as input for existing and new feature encoding/pooling methods. In particular, we choose to extract the **SIFT** [105] descriptors using the VIREO system [1]. We compute three different types of sift descriptors using different salient points detectors and descriptors:

- **Sift DoG**. We detect the interest points in an image using Difference of Gaussians, as originally proposed in [105], and then describe them with 128-dimensional sift resulting from the 8 orientation bins of the histogram of each subwindow of the patch around the interest point.

- **Sift HLD**. We detect points using Hessian-Laplacian Detector, proposed in [119] as a scale and affine invariant interest point detection , and then describe them with 128-dimensional sift resulting from the 8 orientation bins of the histogram of each subwindow of the patch around the interest point.

- **PCA-SIFT**. We detect points u using Difference of Gaussians, and then describe them using the PCA-SIFT technique outlined in [87], that applies PCA to the values in the patch surrounding the interest point, generating a 36-dimensional descriptor.

2.2.2 Baselines: Local Feature Aggregators

The local descriptors presented in Sec. 2.2.1 cannot be used directly as input for the learning machines we will use to compare the effectiveness of our approaches. This data need to be aggregated into compact image representations. We will use here two of the most important approaches for feature aggregation, namely the Bag of Words (BoW) model and the Fisher Vectors.

In order to build the **Bag of Words** signature, we follow the approach outlined in [30]. We use a K-means algorithm to cluster a subset of the training set descriptors in a vocabulary of n visual words (the amount of visual word will change based on the dataset). Then, for each SIFT point in an image, the nearest neighbor in the vocabulary is calculated; based on this statistics a n-dimension feature vector is built collecting the number of points in the image that can be approximated by the nth visual word. Typical sizes of BoW are around $n = 500$.

Moreover, we will also use the PCA-SIFT descriptors as input to a very powerful generative-discriminative aggregation technique, namely the **Fisher Vectors** methods [79]. In order to extract the compact Fisher Vector Signature, we will use the fast implementation proposed in [79]. First, we estimate the global distribution of the keypoints given a subset of the training set descriptors using a Gaussian Mixture Model with m Gaussians (typically 32). We then obtain the Fisher vectors by computing the gradient of the log-likelihood of the image keypoints with respect to the with respect to mean of the global GMM. The final signature has therefore dimensionality $m \times 36$ (number of Gaussians multiplied by descriptor dimensionality).

2.2.3 Baselines: Learning Frameworks and Kernels

Here, we look at the choices we made in our work regarding the learning machines and their similarity measures. As mentioned, in our experiments, we mainly use **Support Vector Machines** [17] to learn both the global descriptors (see Sec. 2.2.1) and the aggregated descriptors (see Sec.2.2.2). We will use the same frameworks to test the effectiveness of our descriptors and aggregators.

We chose Support Vector Machines for their ease of practical use when dealing with large features and large-scale data, and for their proven effectiveness for MMIR [153, 210]. The most important element of a Support Vector Machine is the kernel, namely the similarity measure used for comparing 2 features v and w belonging to different or similar classes. Among the various approaches available, we chose a set of kernels that better fit our needs.

We will use, mainly for global features, a **polynomial kernel** with degree d (in

general, we choose d=2)

$$k(\mathbf{v}, \mathbf{w}) = (\mathbf{v} \cdot \mathbf{w})^d$$

We will also employ, mainly for kernel re-designing and aesthetic feature learning, a **Radial Basis Function Kernel**:

$$k(\mathbf{v}, \mathbf{w}) = \exp(-\lambda ||\mathbf{v} - \mathbf{w}||^2).$$

When learning $n-$dimensional signatures aggregating local descriptors, we will use the **exponential chi-squared kernel**:

$$k(\mathbf{v}, \mathbf{w}) = \exp(-\lambda \sum_{i=1}^{n} \frac{(v_i - w_i)^2}{\frac{1}{2}(v_i + w_i)}).$$

2.2.4 Benchmarking Datasets and Their Experimental Setup

Many of the solutions we propose for MMIR are tested on benchmarking datasets for semantic analysis. In particular, we chose a set of very challenging datasets, very popular in our field for scene recognition, object categorization and concept detection for video retrieval (see Figg. 2.1 and 2.2 for visual examples). In the following, we will detail the content of such image collections, and show how we process them to train and test our techniques.

2.2.4.1 Scene Recognition

MMIR techniques for Automatic Scene Recognition aim at automatically predict the image scene category (where was the image taken?) based on a pre-defined set of scene classes (generally, mutually exclusive). In order to evaluate the performances of our techniques for scene recognition, we selected three databases that have been widely used in literature to study the impact of global and local low-level features or scene analysis.

The **Outdoor Scenes Database**has been used in [133] to evaluate the performances of the Gist descriptor and to describe the properties of the spatial envelope. It is composed of 8 categories of natural scenes and a total of 2600 color images, with a resolution of 256x256 pixels. For each feature, we trained the classifier with our baselines techniques on 100 images per class and used the rest for testing.

The **Indoor Scenes Database** was proposed in [143] as a new, unique database for indoor scene recognition, collecting around 15000 images from various sources, and considering 67 different image categories related to indoor environments. For the indoor scenes experiments, we follow the approach outlined in [143]: we use 20 images for testing and the remaining for training over the baseline descriptors.

The **Scene Understanding Database (SUN)** is a large-scale scene recognition databases. It was proposed in [205] as a complete dataset for scene understanding, with a variety of indoor and outdoor scene environments, spanning 899 categories for more than 130,000 images. As in [205], for benchmarking purposes,

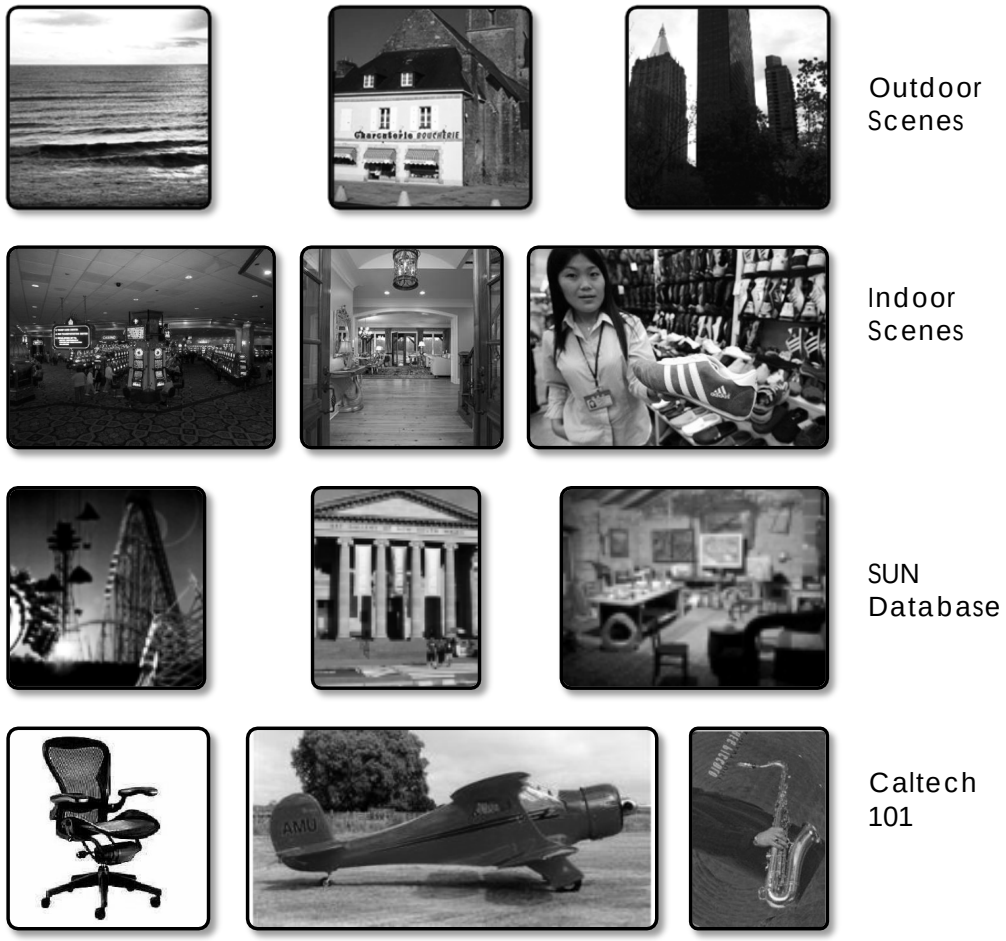


Figure 2.1: Typical pictures from our selected datasets for scene and object recognition

we select a pool of 397 scenes out of the categories proposed, and we use a subset of the SUN dataset consisting 10 folds that contains, for each category, 50 images for test and 50 for training. Results are obtained by averaging the performances of the descriptors over the 10 partitions considered.

2.2.4.2 Object Recognition

Object Classification algorithms aim at labeling an image with an object category selected out of a defined set of mutually-exclusive classes. One of the most popular databases for this task is probably the **Caltech 101** [41] database , a widely-used dataset that contains images of various resolutions labeled with 101 different semantic object categories. Despite its limited amount of highly cluttered images and its lack in pose variation, we chose this database because it is one of the most di-

Dataset ID	Task	Images/Shots	Classes/Concepts	Evaluation Measure
`Indoor Scenes`	Scene Recognition	2600 images	8 classes	Average Accuracy
`Outdoor Scenes`	Scene Recognition	15000 images	67 classes	Average Accuracy
`SUN database`	Scene Recognition	130,000 images	899 classes	Average Accuracy
`Caltech 101`	Object Recognition	9146 images	101 classes	Average Accuracy
`TrecVID`	Video Retrieval	around 100,000 shots	10 concepts	Average Precision

Table 2.1: Overview of the benchmarking datasets used in this thesis

verse multi-object set of labeled images publicly available. For object categorization tests, we follow the experimental approach explained for the indoor scene images (20 images per class for test, the rest for training).

2.2.4.3 Video Retrieval

We also test the performances of our techniques in a large-scale video retrieval framework. We use as a database the **TrecVID 2010 IACC.1.tv10.dev** set, which is composed of 3200 Internet Archive videos (a total of around 100,000 shots).

In particular, we focus on the **Light Semantic Indexing Task (SIN)**, where we are required to build a retrieval system that can produce a ranked list of relevant shots for a set of 10 semantic concepts proposed (*Airplane_Flying, Boat_Ship, Cityscape, Classroom, Demonstration_Or_Protest, Hand, Nighttime, Singing, Telephones*).

Our baseline run is composed as follows. First, we identify, for each video, the keyframe of each shot, representing the central frame of the sequence. For each keyframe/shot, we then extract a pool of low-level features (Color Moments [175], a Wavelet Feature [187], and the MPEG7 edge histogram [200]) together with pooled features (typically, SIFT DoG+BoW and SIFT HLD+BoW). We then use them as input for a set of concept-specific classifiers, namely SVMs with polynomial kernel of degree 2 for global features, and SVMs with chi-squared kernel for aggregated features. The output of this step is a set of feature-specific models separating the feature space in relevant/non relevant examples for each concept. For each concept c, we have therefore defined a model based on each feature extracted from the training data.

Such model is then used to detect the presence of c in a new sample s based on each feature. The classifiers parameters are selected via exhaustive grid search: their value is chosen based on the Mean Average Precision maximization on the development set. For each concept and each feature f, we obtain concept scores (the label confidence) representing the probability of the label given the shot $p_f(c|s)$.

All the concept scores coming from the different features are linearly combined to obtain the final concept score for each shot, that we will use to build the ranked list of shots (see Figg. 2.2 for a visual explanation).

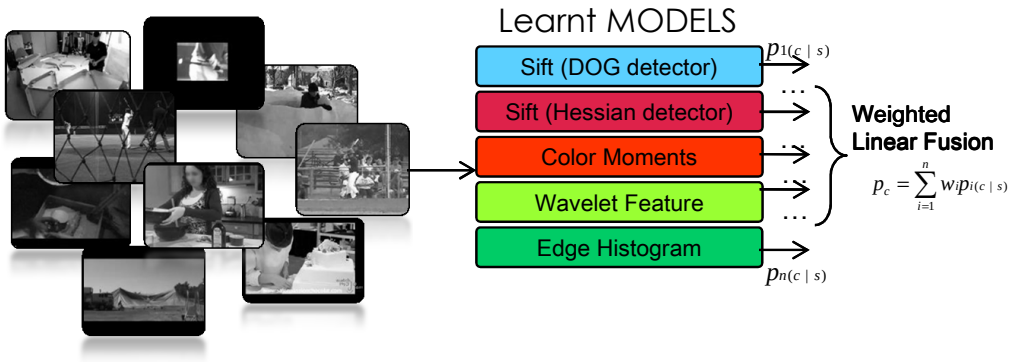


Figure 2.2: Typical Baseline Run and Snapshots from the Light semantic Indexing Task of TrecVID 2010

2.2.5 Evaluation Measures

We have seen the baseline descriptors, aggregators, learners, and datasets, but, how to evaluate the performances of our systems? In this dissertation, we mainly use two different evaluation measures: Average Accuracy and Average Precision. The use of one measure or the other depends on the specific task that we are evaluating.

The **Average Accuracy** measure is used in case of classification and categorization (for scene and objects data). For this problem, we have a pre-defined set of possible c classes. Average Accuracy represents the percentage of correct predictions that a given system makes, compared with the actual labels of the test data. Since we are dealing with multiclass classification problems, the accuracy is calculated *per class*. That is, the accuracy is the averaged ratio between the true positives for a given class (correct matches), divided by all the examples belonging to that class.

$$\text{average accuracy} = \frac{1}{c} \cdot \sum_{i=1}^{c} \frac{\text{number of true positives for class i}}{\text{number of text examples in class i}}$$

The **Mean Average Precision** measure represents the precision of the ranked results of a retrieval set. We will therefore use it to evaluate the performances of our methods tested on the Video Retrieval Task. In order to understand the concept of MAP, we first have to understand the concept of **precision** and **recall**. Given a *query*, a retrieval system returns a set of *documents* that can be either *relevant* or *non-relevant* to the user query. Precision represent the amount of relevant

documents retrieved, compared to all the documents retrieved

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

while recall represents the portion of relevant documents retrieved, compared to all the relevant documents in the collection

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

Precision and recall are useful measures for information retrieval. However, they lack of considering *the order* in which the documents have been retrieved: more relevant documents should be retrieved first, while at the lower level of the list less relevant documents should appear. The **Average Precision** (AP) measure solves this problem by considering the performances of the retrieval framework at each document k of the ranked list of K results, by averaging the precision for a single query:

$$\text{AP} = \frac{\sum_{k=1}^{K} \{\text{precision at document k}\} \cdot \{\text{relevance of document k } (0|1)\}}{|\{\text{relevant documents}\}|}$$

Mean Average Precision (MAP) is the mean of Average Precision over all queries (in our case, it will be the mean of the AP of each concept in the TrecVID SIN Task).

Chapter 3

Level 0: Saliency-based Hybrid Features for Image Categorization

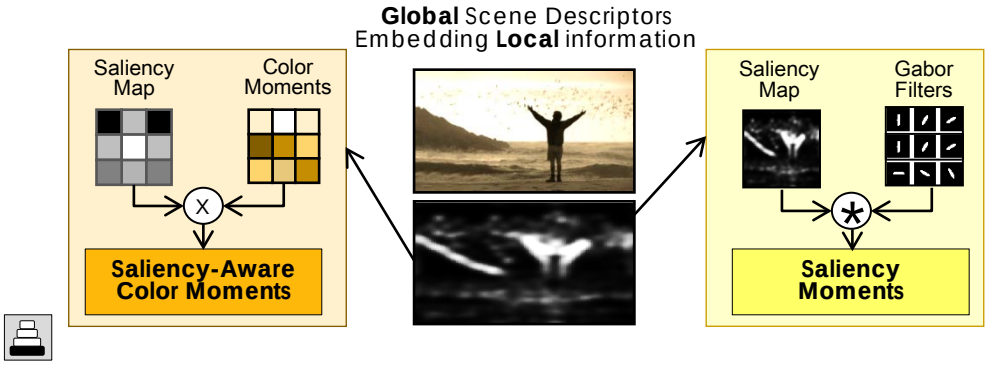


Figure 3.1: Our two hybrid features for image recognition.

In this Chapter we design two features for Scene Recognition, Object Recognition and Video Retrieval. The peculiarity of our features is that the information they carry is both ***global*** *and* ***local****. As a matter of fact, our* ***hybrid*** *features embed local, discriminative, visual information into global, efficient image features. This is achieved by exploiting the discriminative information contained in the* ***saliency maps****. We show that our features are not only very discriminative for the proposed tasks while keeping a light-weight structure, but also that saliency represents a new, complementary source of information regarding the image content, that can be effectively combined with the existing low-level features.*

Low-level features are the basement of every MMIR framework: they represent one of the crucial elements for the development of effective, automatic categorization and retrieval systems. The performances of MMIR frameworks are therefore substantially constrained by the type of features used for representing image and videos. Given the importance of such features, in this chapter we present a set of contributions to improve the low-level image representation using a new set of visual features. We focus here on the development of semantic features for MMIR systems with semantic analysis applications. We will see in Chapter 6 their application to aesthetic analysis, together with a set of new and existing specific aesthetic features.

Low-Level Features for semantic Multimedia Information Processing

Generally, low level semantic features for image recognition model, using a computational approach, the information that our brain (the retina, lateral geniculate nucleus (LGN) and primary visual cortex (V1)) processes in the earliest vision stages, such as color information and oriented lines and edges. Visual features represent low-level information that generates the recognition of image properties. Features are sets of numbers extracted by directly processing the pixel values, similar to the early stage of the human vision, where no spatial integration or pooling is performed. Such pixel-level features reflect therefore the natural statistics of the image and provide information about simple image attributes at a local and a global level (e.g. red on color, round on shape, and dirty on texture). In the MMIR literature for object, concept and scene recognition, namely semantic analysis, we can find two opposite approaches for low-level feature extraction:

1. **Features arising from Local analysis**: as mentioned in Chapter 2, descriptors such as SIFT [105], HOG [31], and SURF [6] describe the surrounding of local interest [44] or densely sampled [42] points in an image, by performing local edge and orientation analysis. Generally, such locally extracted descriptors cannot be used straight-forward for classification, because of the high-dimensionality of the global image representation (see next Chapter for further details). "Raw" local descriptors are generally grouped (see Chapter 4), using keypoint aggregators [30, 146], into a single feature vector, that is then used for modeling and categorization (see Chapters 5 and 6). The resulting signature carries very discriminative information caused by a detailed invariant analysis of local edges performed in relevant image regions, achieving a very precise model of the visual input. However, one major disadvantage of this class of features is their high computational cost arising from both the detailed local invariant analysis and the pooling process.

2. **Global analysis**: image descriptors such as Color Moments and Correlogram [175, 70], Wavelet feature [187], Edge Histogram [200] extract holistic properties of the image (respectively, color distribution, texture, and edge patterns) into a single descriptor without requiring segmentation, interest point detection or grouping operations. One particular type of global feature is the GIST [133] descriptor, whose aim is to represent the "shape of the scene", a fingerprint containing the general layout of the image structures. This class of features is very computationally efficient, but the lack of detailed analysis and robustness to transformations might deal to a loss of discriminative ability for MMIR tasks.

These two classes of features have been widely used for scene categorization (global features) , object recognition (local features) or general semantic concept detection (both types of features together).

Our Contribution: Hybrid Features Based on Saliency

Opposite to this subdivision, we propose here a set of new, biologically plausible, **hybrid features** for content representation that stand in an intermediate point between the mentioned approaches. Our hybrid techniques try to eliminate the weaknesses of both approaches by embedding some *local* analysis into a *holistic* representation of the image. By doing so, we ensure that our features are, like global features, very efficient, while keeping the detailed analysis and accuracy typical of local analysis.

How can we perform local analysis at a pixel level without introducing computational complexity and while keeping the dimensionality low, by embedding *local* information into *global* features? We rely here on a particular type of local information: the **image salient regions**, namely the selected subset of very informative areas that attract the human eyes when glancing a scene (see Sec. 3.1 for a wide explanation of visual attention and saliency). Automatic saliency detectors are available in literature to automatically summarize the distribution of the salient regions into a **saliency map**. Using a specific, light-weight spectral saliency detector [68], we exploit the coarsely localized information arising from the saliency maps and integrate it with global features to build our discriminative, efficient hybrid descriptors.

Our features represent one of the first attempts to **enhance *global* features with saliency information**. As a matter of fact, the role of saliency for semantic MMIR has been mainly explored to improve *local* features-based analysis. For example, in [193] Walther et al. show that object recognition performances are improved by extracting keypoints in subregions corresponding to salient proto-objects: a similar approach is used by Lowe et al in [46] for a mobile robot vision system. Saliency information is also used by Moosman et al in [121] to sample image subwindows and classify image patches for object recognition. On the other hand, besides the mentioned studies integrating saliency with MMIR, visual attention studies has been rarely re-used for *global* image description and recognition. We can find attempts of fusing holistic data with visual attention outside the MMIR context: Torralba et al in [184] combine the *gist* information with the local saliency map to perform object search and detection. Visual attention features have been used for mobile robotics scene recognition in [164], where a low dimensional feature vector is used to represent each feature map extracted from orientation, color and intensity channel.

These work suggests us that saliency is a promising cue for semantic image analysis using global features. In this Chapter, we design two different visual features that arise from the integration of local and global features, and we test their effectiveness for semantic analysis, namely scene, object and concept recognition (see Fig. 3.1):

1. We first make an initial analysis about the role of saliency as a weighting factor to enhance a color descriptor, resulting in an enriched color descriptor named **Saliency-Aware Color Moments** [147].

2. We then explore the discriminative ability of the saliency distribution as a hole, using the image saliency map as a fingerprint representing the global shape of the scene, and summarizing it into a single holistic descriptor that we name **Saliency Moments** [148].

We test the performances of a traditional MMIR system embedding our features, in a variety of tasks, namely indoor and outdoor scene recognition, object categorization and concept detection for video retrieval. Results show that hybrid features outperform the existing local and global approaches for both tasks.

In the following, we will first give a brief overview of the notion of saliency, with some highlights on the existing saliency detectors, including a detailed explanation of the saliency detector we use for our experiments in Sec. 3.1. We will then detail the proposed approaches that embed saliency into low-level global features, namely the Saliency-Aware Color Moments and the Saliency Moments descriptors, respectively in Secc. 3.2 and 3.3.

3.1 Visual Attention and Saliency



Figure 3.2: (a) Saliency distribution can be seen as a coarse-resolution representation of the image layout; (b) Multi-resolution saliency represents different level of details in visual attentional selection

Some regions in the image are more informative than others for the human eye: our visual cortex, when looking at a picture, focuses on few areas in the images, that pop-out from the background, clustering around high-contrast regions and image singularities [207, 137]. The human brain analyses a scene by gathering a reduced but sufficient amount of information from such **salient regions**, i.e. very informative areas that support the long-term recognition process. Various attention-based computational models have been proposed emulating the human way of parsing the visual space with attentional selection, namely performing local parsing of image regions, looking for image singularities. As pointed out in [14], the extraction of salient regions differs from a segmentation problem, because saliency maps highlight foreground objects with respect to their background, while segmentation algorithms

generate a partition of the image into regions of consistent properties. The general output of a saliency detector is a **visual saliency map** (see Fig 3.2 for visual examples) highlighting the regions that pop out when observing the image (e.g. areas where the image shows high contrast or statistical singularities), corresponding to the areas of *human fixations*. We will see in the next Section an overview of the most common automatic saliency detectors.

3.1.1 Computational Models for Saliency Detection

Automatic saliency detectors aim at translating into a computational model the neurobiological process on early visual perception theory. Various ways to implement visual attention processes have been proposed. The first pioneer work for saliency detection using a computational approach was the one from Itti et al. [73] that used *center-surround differences* statistics to define the image saliency map, based on color, orientation, and intensity features. Such work was improved in [109] with a local contrast with fuzzy growth model. Harel et al. in [63] further improve Itti's approach by adding a graph-based analysis for map normalization achieving a very efficient model for saliency detection.

Later, *learning methods* were used by Liu et al. [103] to model the differences between foreground and background objects through multi-scale contrast, center-surround histogram, and color spatial distribution. Learning techniques were also used in [86], where object and faces detectors are used to predict human fixations, and by Torralba et al. in [184], that build a "contextual guidance model" for predicting salient regions combining global features, bottom-up saliency, and top-down mechanisms at an early stage of visual processing.

Frequency-based saliency detectors have recently attracted a lot of attention in the field. Such spectral-based methods are much lighter and more efficient compared to the previously mentioned saliency detectors. For example, Achanta et al. [3] proposed an effective and efficient method for saliency detection based on the difference of the pixel color from the average image color. Spectral components in an image have also been used in [68] by computing the difference of the image spectrum with the average image spectrum, further improved by [54], that uses the phase spectrum instead of its magnitude.

For the development of our efficient global image analysis features, we chose to employ a method from this last class of detectors.

3.1.2 The Spectral Residual Saliency Detector

Saliency maps based on frequency analysis represent therefore a fast and effective tool for extracting coarsely-localized information about the image objects and their locations, without recurring to pure local analysis or interest point detection. Among those, we chose for our experiments the Spectral Residual approach [68] for its efficiency and effectiveness, and for its light implementation that well fits the type of analysis we perform with our hybrid features. The Spectral Residual technique

aims to detect coarse salient regions using a fast, straight-forward approach, that does not require parameter selection or multi-channel features weighting, and it is therefore suitable for being the basic component of a global feature.

This method exploits the properties of the amplitude $A(f_x, f_y)$ of the Fourier Spectrum, observing that statistical singularities in the frequency domain correspond to salient proto-objects in the pixel domain.

In order to obtain a saliency map, the Spectral Residual method computes the following steps on the input image:

1. the luminance channel of the input image I is downsized to a $i \times i$ coarser resolution.

2. The log-spectrum $L(f_x, f_y) = log(A(f_x, f_y))$ and its smoothed version $F(f_x, f_y) = L(f_x, f_y) \star h_n(f_x, f_y)$, where h_n is an average filter of size n, are computed on the grayscale matrix. As a matter of fact, it is showed that the log-spectra of different images are described by frequency-amplitude curves with very similar shapes. $F(f_x, f_y)$ represents therefore an approximation of such general behavior of the log spectra. If all the natural images share a general log-spectrum behavior, the spectral elements that produce discrimination between different images, and that therefore imply visual attention, can be found in the local peaks in the curve that deviate from such general trend.

3. the log spectral residual
   $L_R(f_x, f_y) = L(f_x, f_y) - F(f_x, f_y)$
   is obtained therefore by subtracting the two signals computed in the previous step.

4. the linear version of the spectral residual

$$R(f_x, f_y) = exp(L_R(f_x, f_y) + P(f_x, f_y)) \tag{3.1}$$

   is obtained by joining $L_R(f_x, f_y)$ with its original phase $P(f_x, f_y)$

5. The saliency map is then obtained by applying the Inverse Fourier Transform (IFT) on $R(f_x, f_y)$, giving, for image I, the saliency map

$$S(I) = IFT(R(f_x, f_y)) \tag{3.2}$$

3.2 Initial Analysis: Saliency-Aware Color Moments

In this Section, we show how we exploit the spectral residual signal to enhance the discriminative power of existing window-based color indexing techniques.

Our observation is that traditional color descriptors treat all the image regions with equal importance. However, we know from visual perception theory that some

image areas carry more information about the image content (e.g. the scene foreground). Therefore, higher importance should be given to the chromatic characteristics of more informative windows when building low-level color features. We present here an *informativeness-aware color descriptor* based on the Color Moments (CM) feature [175]. We first define a saliency-based measure to quantify the amount of information carried by each image window; we then change the window-based CM feature according to the computed local informativeness. Finally, we show that this new hybrid feature outperforms the traditional Color Moments in a variety of challenging dataset for scene categorization and video retrieval.

In the following, we will first give a brief introduction on color indexing techniques, motivating our choices and giving a high-level description of our approach (See Sec. 3.2.1), we will then recall the principles of the CM descriptor in Sec. 3.2.2, and propose our Saliency-Aware Color Moments enhancement in Sec. 3.2.3. Finally, we will validate our theory with some experimental results in Sec. 3.2.4.

3.2.1 Why Adding Saliency Information to Color Description?

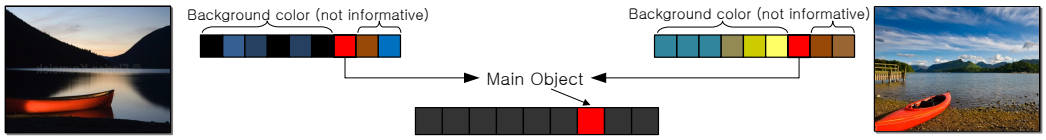


Figure 3.3: Color indexing issue: even if the two images depict the same thing and the main object (the canoe) has the same color, the two backgrounds vary and the feature vectors are completely different.

Color Descriptors have been widely used an automatic image analysis. Color is a necessary, powerful feature for the recognition of scenes and objects for both biological and computational visual systems. Color-based features play an important role in Multimedia Retrieval for semantic analysis.

The most intuitive representation of the chromatic information, namely the color histogram, has been proved to be an effective way to describe images [180, 58]. Following this idea, a faster and more robust descriptor has been proposed in [175], where the first three moments of the color distribution are stored in the Color Moments (CM) feature. Generally, in MMIR, the CM is used in its localized version, where the index is built by dividing the image into an $n \times n$ grid and collecting the moments of the resulting image sub-windows.

Despite the proved effectiveness of chromatic features for object and concept recognition, two main elements can cause the decrease of their discriminative ability. First, images *semantically dissimilar* (i.e. depicting completely different concepts) might have *similar color composition*, thus introducing noise in visual class separation. This first issue can be partially solved by combining the color index with other sources of visual description (texture, edge, ...) in a complete MMIR system.

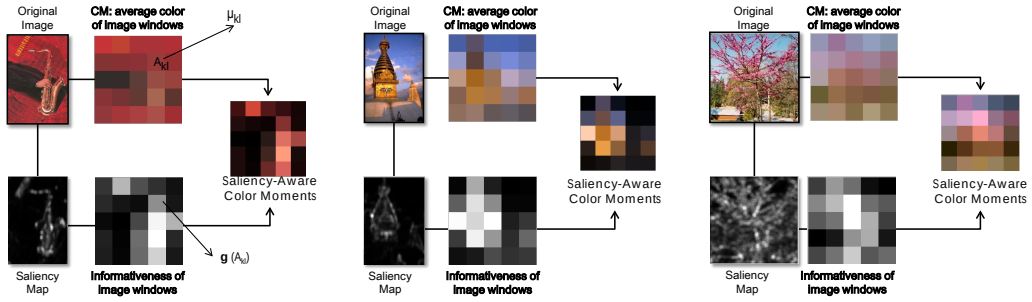


Figure 3.4: The effect of adding saliency and informativeness measures in the color moments computation: the chromatic information is extracted from the main region in the image and more importance is given to the salient regions color components.

Second, traditional window-based color analysis does not take into account the fact that *some regions* (e.g. the foreground) could *contain more information than others*. Treating all image windows with equal importance might cause inconsistencies in color description, especially when the amount of informative regions is small compared to the less important regions, i.e. when the main object is small compared to the background (an example is shown in Fig 3.3).

Here, we propose a solution for this second issue; the main observation is that we can improve the discriminative power (partly removing the mentioned inconsistencies) of the color features by *collecting the chromatic components of the informative subregions only*. An attempt of weighting image areas for color indexing was proposed in [174], where users were required to indicate a value for each subregion representing its importance for image matching. Another solution to this problem was brought by Sebe et al. in [162], where CM vectors were extracted from image patches surrounding interest points. These works show that the informativeness of image regions can be a meaningful way to improve color-based image retrieval.

Different from these approaches, we design an informativeness-aware color feature that *automatically* weights the image regions according to their importance, thus differing from the *manual* measures used in [174]. How can we automatically measure the image sub-windows importance? Given the relationship between the amount of information and the probability of a region to attract our attention, we propose here a means of measuring image areas informativeness based on the local saliency distribution,

We then use it to improve the Color Moments feature for image recognition and retrieval, building a new descriptor that we call **Saliency-Aware Color Moments (SACM).** This results in a low-dimensional representation of the image that allows meaningful/salient regions to be taken more into account when performing color-based matching and retrieval (see Fig. 3.4 for a visual explanation). Since we use coarsely-localized information, we ensure computational efficiency, different from the refined interest-point analysis proposed in [162].

With our approach we build therefore a hybrid feature that adds some localized

information (i.e. the saliency distribution) in a typically global feature, without involving any parameter tuning, learning or image segmentation. With a fast pre-processing step, we change the localized CM values according to the amount of information carried by each window, that we calculate with easy operations.

3.2.2 The Color Moments Feature

The traditional window-based Color Moments [175] is one of the most widely used chromatic descriptors in image analysis and retrieval. It is based on the statistical analysis of the distribution of pixel values at given locations.

First, an image $I \in R^{X \times Y}$ is divided into a set of rectangular image subregions

$$A_{kl} \in R^{M \times N}$$

where $k = 1 \dots \frac{X}{M}$ and $l = 1 \dots \frac{Y}{N}$ are the region indexes and $M \times N$ is the window resolution [1].

For each window A_{kl} the color feature in [175] extracts color information and builds the window index

$$cm_{kl}^{(I)} = \{\mu_{kl}, \sigma_{kl}, \eta_{kl}\} \tag{3.3}$$

where μ_{kl} represents the average pixel value over the subregion A_{kl}, and σ_{kl}, η_{kl} correspond to the second and third moment of the distribution drawn from the pixel values, namely standard deviation and skewness. Finally, as shown in Fig.3.5 , the feature describes the color components of an image by gathering the chromatic information of each image subregion in a global image signature $cm^{(I)} = \{cm_{kl}^{(I)}\}$.

3.2.3 Saliency-Aware Color Moments

In its original framework, the CM feature is homogeneously calculated over the whole set of image regions, without considering that not all the sub-windows are equally important. As shown, various computational models [3, 73, 68] have been built that highlight such regions in a saliency map, a matrix that represents the distribution of the saliency over the image surface, or, equivalently, the probability that a specific location attracts the visual attention of an observer, with higher values where the image shows high contrasts or statistical singularities.

The main idea (see Fig.3.5) is that we can quantify the informativeness of an image sub-window by calculating the amount of saliency in it. The more the saliency concentrated in its rectangular area, the more the information carried by such sub-window. Having calculated each sub-window importance, a scalar value that goes from 0 (not informative) to 1 (very informative), we can then use it to weigh its corresponding CM index. In this way, less informative regions do not give an important contribution in the final feature vector, and the description is mainly based on the chromatic components of the salient objects.

In the remainder of this Section we explain in details our proposed approach for color indexing. A window-based informativeness measure is proposed in Sec.

[1] window sizes are chosen so that $mod(X, M) = 0, mod(Y, N) = 0$

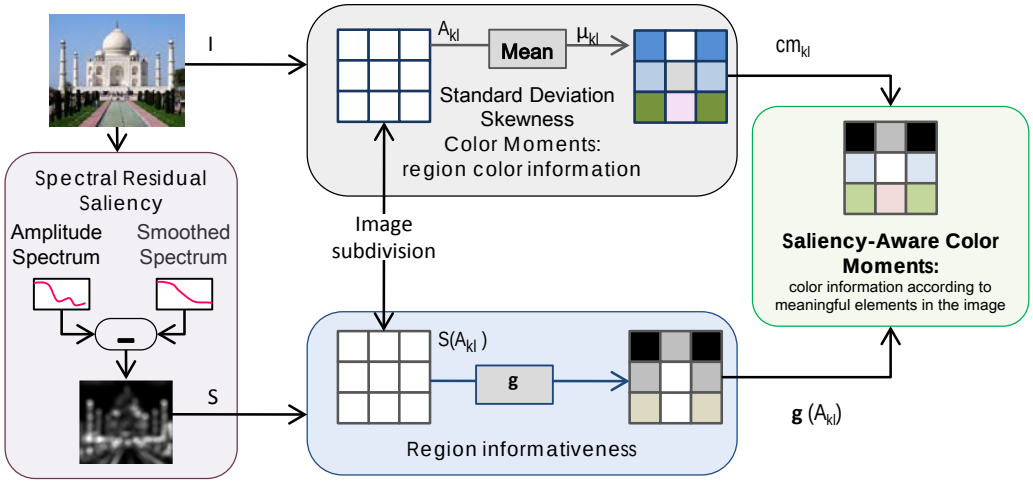


Figure 3.5: SACM Algorithm: CM are extracted from each of the $M \times N$ windows and weighted by the informativeness value that is obtained, for each window, by averaging the saliency map.

3.2.3.1, so that, for each rectangular area described by CM, we also have a value that quantifies the information carried. Finally, in Sec. 3.2.3.2 the two analysis are combined to build a Saliency-Aware Color Moments feature.

3.2.3.1 Image Regions Informativeness

How can we extract the importance of an image region using a quick computational approach? As said, such value should represent the amount of salient regions in each image window, in order to represent the amount of information carried.

From the previous subdivision, we have a set of $M \times N$ rectangular region A_{kl}, and we need to find a function

$$g : R^{M \times N} \rightarrow R$$

that maps the image window in a scalar value representing its informativeness, by exploiting the local saliency information.

We know that the saliency distribution can be obtained by using visual attention algorithms. No matter the approach used, the output of such models is a saliency map, namely a matrix with higher pixel values corresponding to higher probability of the pixel to fall into the visual attention space. In our case, we compute the saliency distribution using the Spectral Residual approach [68] as in Sec. 3.1.2, that gives, for an image I, its corresponding frequency-based saliency map $S(I)$ in Eq. (3.2).

Our proposed procedure is as follows (see Fig. 3.5 for a visual explanation):

1. From the image I, we obtain a $X \times Y$ saliency map $S(I)$ (to simplify, we assume same dimension for input image and output map).

2. We can then find the window-based saliency distribution by dividing $S(I)$ into subregions $S(A_{kl}) = \{s_{ij}\}$, being $i = Mk, \ldots, Mk + M - 1$, and $j = Nl, \ldots, Nl + N - 1$ the pixel indexes inside the saliency sub-window, whose dimension is again $M \times N$.

3. Given the windowed saliency map $S(A_{kl})$, the informativeness γ_{kl} of the rectangular area A_{kl} can be obtained by averaging its value over the sub-window surface:

$$\gamma_{kl} = g(A_{kl}) = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} s_{ij}}{M \times N}$$

The function g will have higher values when the image window considered contains more salient regions (higher values in the map), and lower values when the window considered carries little information.

3.2.3.2 Adding Informativeness to the Color Feature

We now have a window-based color analysis cm_{kl} and a window-based informativeness measure γ_{kl}. How do we integrate these two sources of information in a meaningful feature for image recognition and retrieval?

Our aim is to extract from the image the color information generated mostly from its salient regions (see Fig. 3.4). A straightforward way to obtain this effect is to weigh the window-based color statistics with the scalar value representing the amount of information carried by that window (the value of function g, as explained in the previous Section). We therefore change Eq. (3.3) in order to "switch off" the less important windows, obtaining a new set of components for each A_{kl}:

$$sacm_{kl}^{(I)} = \{\mu_{kl} \cdot \gamma_{kl}, \sigma_{kl}, \eta_{kl}\} \tag{3.4}$$

By weighting the first moment of each window, we modulate its average color brightness based on the local informativeness value, allowing salient regions to pop-out from the image background and mitigating the effect of less important regions.

Finally, we gather in a single descriptor the region-based indexes by concatenating them in a feature vector $sacm^{(I)} = \{sacm_{kl}^{(I)}\}$ that we use as input for the recognition and retrieval systems.

3.2.4 Experimental Validation

We evaluate here the improvement brought by adding our informativeness measure into a classical color indexing technique, experimenting its effectiveness for scene recognition, object recognition and video retrieval. We use as baseline for comparison the Color Moments descriptor.

3.2.4.1 Experimental Setup

For our experiments, we divide each image (or keyframe) into 25 rectangular subregions ($k = 1, \ldots, 5$ and $l = 1, \ldots, 5$) and extract the CM feature from each of them, as shown in the baselines description (Chapter 2).

In parallel, we extract the map containing the salient locations in the image, as shown in Sec. 3.1.2. In order to ensure computational efficiency, we chose to compute the map with the spectral residual method [68], which produces fast saliency measures, perceptually comparable to the state of the art methods. In Sec. 3.2.3 we assumed for simplicity that the map $S(I)$ has the same resolution $X \times Y$ as the input image. In practice, for most of the saliency detection algorithm[2], $S \in R^{X' \times Y'}$, with $X' < X$ and $Y' < Y$, therefore, having the same number of subregions (i.e. the ratio between image and window resolution), the windowed saliency distribution will have dimension $M' \times N'$, where $M' < M$ and $N' < N$.

We test our new descriptor and compare it with the Color Moments feature on a variety of dataset and tasks for MMIR semantic analysis. For both CM and SACM we learn a model using SVMs with polynomial kernel of degree 2 and test it on the following datasets:

- for the **scene recognition** task, we considered the outdoor scene categories database, introduced by Torralba et al.in [133].
- for the **object recognition** task, we chose the widely used Caltech 101 database [41].
- Moreover, we compare the effectiveness of CM and SACM for **video concept detection**, comparing the performances of the two features for the TrecVID 2010 semantic Indexing Task.

For all the datasets considered, for both features, we use for experiments the same training/test experimental setup as our baselines.

3.2.4.2 Experimental Results

Our descriptor represents an initial study about the possibility of introducing saliency in MMIR for semantic Image analysis. Despite its simplicity, experimental results show that our feature brings substantial improvement to the Color Moments feature, for all the tasks considered

Outdoor Scene Categories

In Fig. 3.6 we show the results on the test set for each class of the outdoor scenes database. When looking at the average multiclass accuracy, we can see that our approach, that boosts the color feature with saliency measures, actually improves

[2]For example, the Spectral Residual method in [68] gives saliency maps at resolution 128×128 pixels.

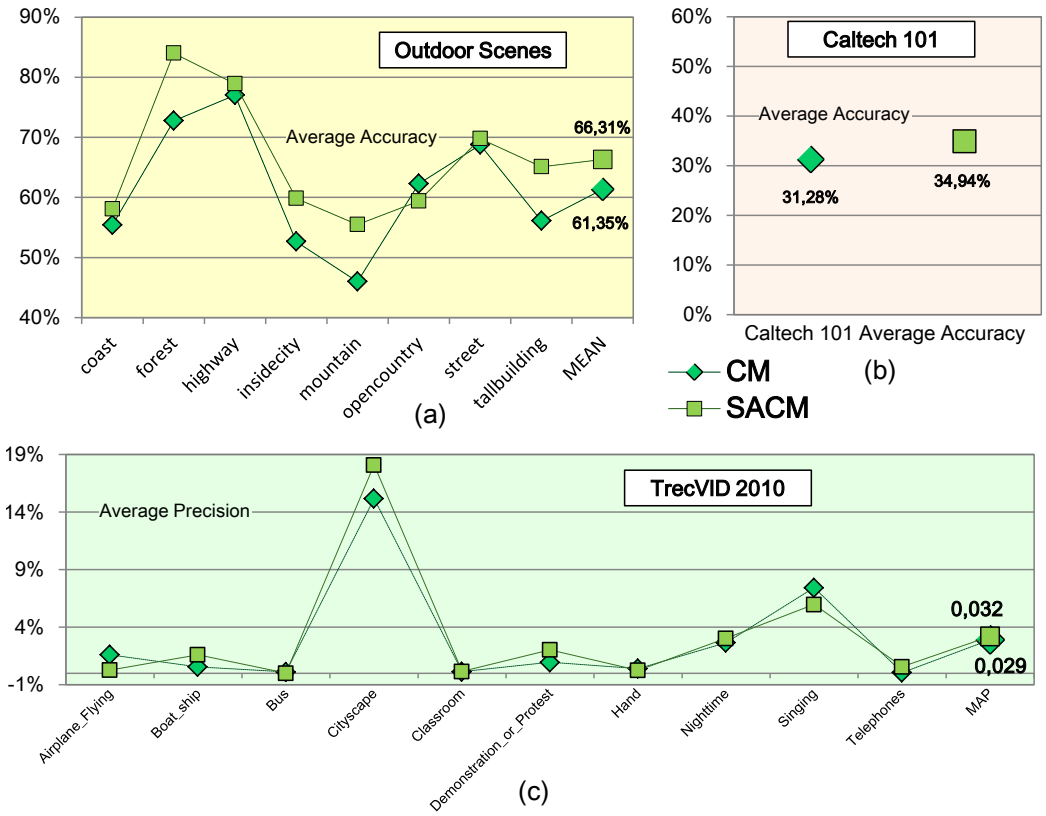


Figure 3.6: Comparison of Saliency-Aware Color Moments and Color Moments for all the three semantic MMIR tasks considered. (a) Outdoor Scenes, results shown in term of average accuracy on the test set. (b) Caltech 101, results shown in terms of average accuracy on the test set. (c) TrecVID 2010 semantic Indexing Task. Results shown in terms of average precision for each concept and Mean Average Precision for all the concepts considered.

the average accuracy for outdoor scene recognition, with SACM that brings an improvement of about 10% over the standard CM feature.

We can explain this improvement because outdoor scene categories such as *city*, *mountain*, and *tallbuilding*, are clearly identifiable by one single object (buildings, mountains). Since with our approach we consider only the most informative regions, the resulting color signature is describing the image areas containing the most important objects in the image (i.e. a mountain), namely the regions that are crucial to distinguish between the different image categories.

Caltech-101

In Table 3.1 we show the per-class accuracy on the test set of the Caltech 101 dataset, while the comparison of the two features in terms of average accuracy is shown in Fig. 3.6 (b)Results show that by considering the color of the main object

	CM	SACM		CM	SACM		CM	SACM		CM	SACM
Motorbikes	100,0%	**100,0%**	brain	45,5%	**45,5%**	wrench	27,3%	**54,5%**	bass	10,0%	**10,0%**
minaret	100,0%	**100,0%**	windsor_chair	45,0%	**50,0%**	saxophone	26,3%	**31,6%**	garfield	5,6%	**5,6%**
Leopards	100,0%	**100,0%**	grand_piano	45,0%	**50,0%**	joshua_tree	26,3%	**31,6%**	water_lilly	5,0%	**5,0%**
Faces_easy	100,0%	**100,0%**	dolphin	45,0%	35,0%	gramophone	26,3%	21,1%	stapler	5,0%	**10,0%**
Faces	95,0%	**100,0%**	umbrella	42,1%	**57,9%**	butterfly	26,3%	21,1%	hedgehog	5,0%	**40,0%**
airplanes	90,0%	**95,0%**	buddha	42,1%	36,8%	pyramid	25,0%	**35,0%**	crocodile	5,0%	**10,0%**
pagoda	85,0%	60,0%	starfish	40,0%	**55,0%**	flamingo	25,0%	**30,0%**	chair	5,0%	**10,0%**
bonsai	75,0%	65,0%	scorpion	40,0%	**40,0%**	ewer	25,0%	20,0%	barrel	5,0%	**5,0%**
trilobite	72,2%	**83,3%**	revolver	40,0%	**46,7%**	dalmatian	25,0%	15,0%	wild_cat	0,0%	0,0%
pizza	70,0%	65,0%	lotus	40,0%	30,0%	stegosaurus	21,1%	**36,8%**	snoopy	0,0%	0,0%
sunflower	65,0%	**70,0%**	kangaroo	40,0%	**55,0%**	schooner	20,0%	**26,7%**	sea_horse	0,0%	**5,0%**
cellphone	63,2%	**63,2%**	inline_skate	36,8%	**47,4%**	okapi	20,0%	**25,0%**	scissors	0,0%	**11,1%**
accordion	63,2%	47,4%	electric_guitar	36,8%	**52,6%**	mandolin	20,0%	**40,0%**	platypus	0,0%	**5,0%**
watch	60,0%	**60,0%**	dollar_bill	36,8%	**47,4%**	llama	20,0%	**45,0%**	octopus	0,0%	**5,9%**
stop_sign	57,9%	**63,2%**	soccer_ball	35,7%	**42,9%**	lamp	20,0%	15,0%	mayfly	0,0%	0,0%
hawksbill	55,0%	50,0%	wheelchair	31,6%	**42,1%**	emu	20,0%	15,0%	lobster	0,0%	0,0%
menorah	52,9%	**64,7%**	euphonium	31,6%	**52,6%**	crab	15,8%	15,8%	gerenuk	0,0%	**5,0%**
ketch	52,6%	**63,2%**	camera	31,6%	**36,8%**	headphone	15,0%	**30,0%**	cup	0,0%	0,0%
helicopter	50,0%	40,0%	binocular	31,6%	21,1%	nautilus	13,3%	**20,0%**	crocodile_head	0,0%	**10,0%**
chandelier	50,0%	**60,0%**	strawberry	30,0%	**45,0%**	panda	10,5%	**15,8%**	ceiling_fan	0,0%	**10,5%**
rooster	47,4%	**57,9%**	ibis	30,0%	20,0%	ferry	10,5%	**21,1%**	cannon	0,0%	0,0%
metronome	47,4%	**63,2%**	elephant	30,0%	**35,0%**	rhino	10,0%	**15,0%**	brontosaurus	0,0%	0,0%
laptop	47,4%	42,1%	tick	27,8%	22,2%	pigeon	10,0%	**20,0%**	beaver	0,0%	**5,0%**
dragonfly	47,1%	41,2%	crayfish	27,8%	22,2%	flamingo_head	10,0%	**10,0%**	ant	0,0%	0,0%
yin_yang	46,7%	**53,3%**	cougar_face	27,8%	**27,8%**	cougar_body	10,0%	5,0%	anchor	0,0%	0,0%

Table 3.1: Per-class results of the Saliency-Aware Color Moments compared to CM on the Caltech 101 dataset.

only, SACM improves the color indexing performances for the object recognition task: the average classification accuracy improves of about 10%, when compared to the CM descriptor. This is due to the fact that, in the Caltech 101 database, the images depict objects that clearly detach from a uniform background, making therefore the salient object detection an easy task for the Spectral Residual Algorithm, and allowing SACM to describe the color of the main object only.

TrecVID 2010

We show in Figure 3.6(c) that the retrieval performance of SACM is in average 10% better than CM, with some peaks for concepts like *Cityscape* (+20 %), *Boat_Ship (+190%)*, *Demonstration (+113 %)*. A reasoning similar to the previous tasks can be done to explain such performances. Despite the substantial noise in the images and in the annotations of the TrecVID data, the saliency detector can identify the most important objects in the images, decisive to identify the keyframe labels, such as buildings, boats, and the typical flags and writing of the demonstrations.

3.3 Saliency Moments for Scene Categorization

With SACM, we showed the importance and discriminative power of saliency information for global image analysis. In this Section we will make a step further.

We present **Saliency Moments**, a holistic descriptor for image recognition based on saliency and inspired by another biological vision principle: the *gist* perception. The *gist* of a scene is the coarse-level representation of the visual input that the brain performs in the very first glance of a scene. Our idea is to generate a *gist* of the image based on saliency maps.

We extract the saliency information with the Spectral Residual approach, and

create a hybrid, low-dimensional image signature by globally processing the saliency map directly in the frequency domain. Results show that this new type of image description outperforms and complements the traditional global features on scene and object categorization, for a variety of challenging datasets.

In the following, we will first introduce in Sec. 3.3.1 the Saliency Moments descriptors and the notion of *gist*. We will then motivate with visual perception theory principles our choice to use saliency information as the *gist* of the scene, see Sec. 3.3.2, and then give in Sec. 3.3.3 the implementation details of the Saliency Moments descriptor. We will finally show some experimental results on applying the Saliency Moments descriptor for image categorization (Sec. 3.3.4).

3.3.1 A Biologically-Inspired Hybrid Descriptor

Biological visual systems can be a useful source of inspiration for the development of effective computational vision systems. By analyzing how humans process the real word scenes and objects in their early vision stage, we can build more discriminative image features. Given this intuition, we propose here a new, biologically plausible, hybrid feature for content representation that is inspired by the visual perception theory. In particular, we explore two processes of the visual cortex, the (local) already mentioned **selective visual attention** and the (global) **gist** perception for scene recognition.

As mentioned, visual attention refers to the fact that the human eye, when recognizing the content of a scene, focuses on a subset of selected **salient regions** that attract its attention (local process). Such process is modeled in computer vision by existing *local* analysis algorithms [73, 3, 63] that output saliency maps based on predicted human fixations. On the other hand, various studies [126, 134] proved that the brain is able to recognize images under very brief exposures (less than 100 ms), gathering a **coarse representation of the image** contours and structures: the *gist* of the scene (global process). Various *global* image descriptors have been proposed in literature modeling such low-resolution, holistic summarization of the image spatial layouts and components (e.g. the spectrum-based Gist [133], texture-based[154]).

Visual attention and gist perception both refer to early stages of human vision. However, while the first one is based on a *local parsing* of the image regions, the *gist* is a *global* fingerprint of the visual information reaching the brain when looking at a scene. Our aim is to explore the interaction between these two principles using a computational approach and apply it to MMIR. Even if both these two aspects of visual perception have inspired computational models for image understanding and categorization, the interaction between the two has been rarely explored for MMIR. The Gist descriptor has been successfully used in [134] to enhance saliency detection, showing that the synergy between these two notions leads to an enriched visual analysis.

Given these observations, we evaluate here the contribution of adding locally-extracted saliency information in a global feature for image categorization and retrieval. Following the idea that the *gist* of the scene is not a pre-attentive task (see

Sec. 3.3.2 for further explanations), we build a robust hybrid feature based on a low dimensional representation of the shape of the salient region. Our image signature, called Saliency Moments, embeds some locally-parsed information, i.e. the salient regions and objects in the scene, in a holistic representation of the scene. This is achieved by abstracting the salient region shape as a *whole* for a global, *gist*-based[3], discriminative description of the image. In order to ensure computational efficiency, we choose a frequency-based light-weight algorithm, namely the Spectral Residual, [68] for the extraction of the saliency distribution and perform the image signature construction via spectral sampling (directly in the Fourier domain) and higher order statistics.

The final hybrid descriptor takes advantage of the discriminative power of local analysis details (i.e. the saliency map) while keeping a low dimensionality and fast computation. Moreover, the key aspect of our descriptor is that saliency is a new source of discriminative information compared to traditional features for image categorization (e.g. color and edge distribution). Therefore, when we combine Saliency Moments with existing local and global descriptors MMIR, we add complementary, meaningful information that improves the overall performances of the system.

Saliency Moments represents one of the first attempts in literature to use saliency maps for image categorization. The closest method related to our descriptor is the one presented in [164], where center-surround difference saliency maps are averaged over local windows, and then used in scene recognition for robot navigation. Our descriptor differs from this approach because first of all we extract a light saliency map, thus ensuring the computational efficiency typical of global features. Moreover, instead of performing averaging operations, we extract the saliency principal components by analyzing the map directly in the frequency domain, thus keeping its discriminative power and using the saliency information as a whole signal representing the shape of the scene. Moreover, while in [164] the method is tested on a few, ad-hoc created scene categories for robot navigation, we present here an extensive set of experiments conducted in challenging benchmarking datasets for scene and object recognition.

3.3.2 Saliency in a Holistic Signature: Motivation and Key Elements

Here we motivate the use of saliency as a *gist*-based image fingerprint with studies from the visual perception theory.

How do we process the information coming from the visual space?

A plausible answer can be found in [126]: the human brain synthesizes the image *globally* before understanding the *local* details (i.e. it sees the "forest before seeing the trees"). According to this model, Oliva and Schyns in [161] showed that the visual information is organized in a set of *spatial frequencies* that correspond to different resolutions and levels of detail of the visual space. When first looking at a

[3]In the following, we will use "*gist*" to identify a coarse representation of the image and "Gist" to refer to Torralba's descriptor in [133]

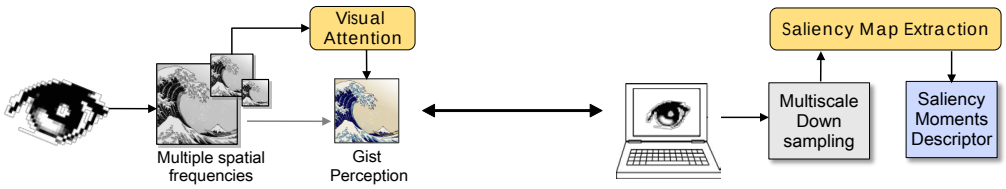


Figure 3.7: Interaction of attention and *gist* in visual perception theories: a multi-resolution input is parsed to obtain salient frequencies when gathering the spatial envelope of the scene. Our proposed implementation: a multi-resolution saliency map is extracted and summarized into a global signature.

scene, we perceive the holistic, most coarse-grained representation of the image, i.e. its *gist*, which is enough for the human brain to categorize the visual space after a very brief exposure (100ms or below). In this phase, we do not rely on segmentation or local analysis operations but we gather the meaningful information into a low-resolution *gist* of the scene. According to the definition of *gist*, such "holistic envelope" should represent an "impoverished version of the *principal contours and textures*" [134].

On the other hand, a well-studied aspect of the human visual perception is the selective visual attention, i.e. the process by which the human brain analyses a scene by gathering a reduced but sufficient amount of information from the multi-dimensional visual space.

Traditionally (see, for example [62]) visual attention is considered to be independent and posterior to the *gist* perception. As pointed out in [164], apparently *gist* and saliency rely on opposite procedures, as the first one is a global, fast summary of the image structures, while visual attention requires slow local analysis to highlight image singularities. Nevertheless, the human cortex bases the visual input understanding on both these components, and some perception-based experiments proved the interaction between these two elements for rapid scene analysis. These studies (see [134] [177] [28]) report that, similar to the traditional attentional perception, scene understanding under brief exposures involves an attentional stage that selects different frequencies from different spatial scales (see figure 3.7 for a visual explanation). Following these theories, there would be an early attentional selection before the *gist* perception that directs the fixations to particular salient region, supporting that contribute to the recognition process.

Another issue regarding early stages of vision is: does a chromatic component come into the picture under brief exposures? different studies showed that color can play an important role in the rapid recognition of object and scenes. According to these studies, conducted by Oliva et al in [132] and by Castelhano et al in [23], the human brain, when gathering the *gist* of an image, synthesizes and uses the color information for the classification task.

Given all these observations, we want to test the importance of the visual atten-

tion component in the *gist* perception using a computational approach that:

1. Represents the input as a **multi-resolution visual signal**, according to the spatial frequencies organization of the visual information mentioned in [161].
2. Extracts the **saliency distribution** for every spatial scale considered, simulating the pre-*gist* attentional stage.
3. Analyzes the visual **saliency as a whole**, summarizing the previous analysis in a *gist*-based image signature.
4. Explores the role of the **chromatic component** by adding a coarse representation of the locally dominant color information.

3.3.3 Saliency Moments for Image Categorization

We therefore build our hybrid descriptor by implementing the four requirements outlined in the previous Section (see Fig. 3.8 for a visual explanation of our algorithm).

The idea is to use the saliency shape (the ensemble of contours of the salient objects and regions in a digital image) as an image fingerprint, in order to represent the visual attention information in a *gist*-based image signature. Despite from its local nature, using the saliency maps as a signature of the scene does not contrast with the definition of spatial envelope seen in [134]. In fact, the saliency map is a grayscale matrix, with higher pixel values that cluster around strong edges or object of interest, outlining, as a whole, a coarse representation of the spatial composition of the scene. Moreover, Fig. 3.2 shows that different objects and scenes generate different saliency maps: the saliency shape can be seen as a discriminative source of information for image categorization.

According to point (1) and (2), we downsample the image at different scales and compute a multi-resolution map of the perceptually relevant areas (implementation details can be found Sec. 3.3.3.1). We use for this purpose a Fourier-domain saliency detector proposed in [68] that highlights different salient shapes for different resolutions (see Fig.3.2(b)).

We then propose an approach for the global image signature construction (requirement (3)): we decompose the signal in what we call the "saliency components", obtained by sampling the spectral maps directly in the frequency domain. We then extract various statistics from these samples, building an image index that we call "Saliency Moments" (SM) (see Se. 3.3.3.2 for details).

Finally, following requirement (4), we describe (Sec. 3.3.3.3) a color-opponents based chromatic feature that is merged with the previous index to build the Color Saliency Moments (CSM) feature.

3.3.3.1 Multi-Resolution Visual Attention

In this Section we show how we extract the saliency information signal, based on which we will build the Saliency Moments descriptor. Of the many computational

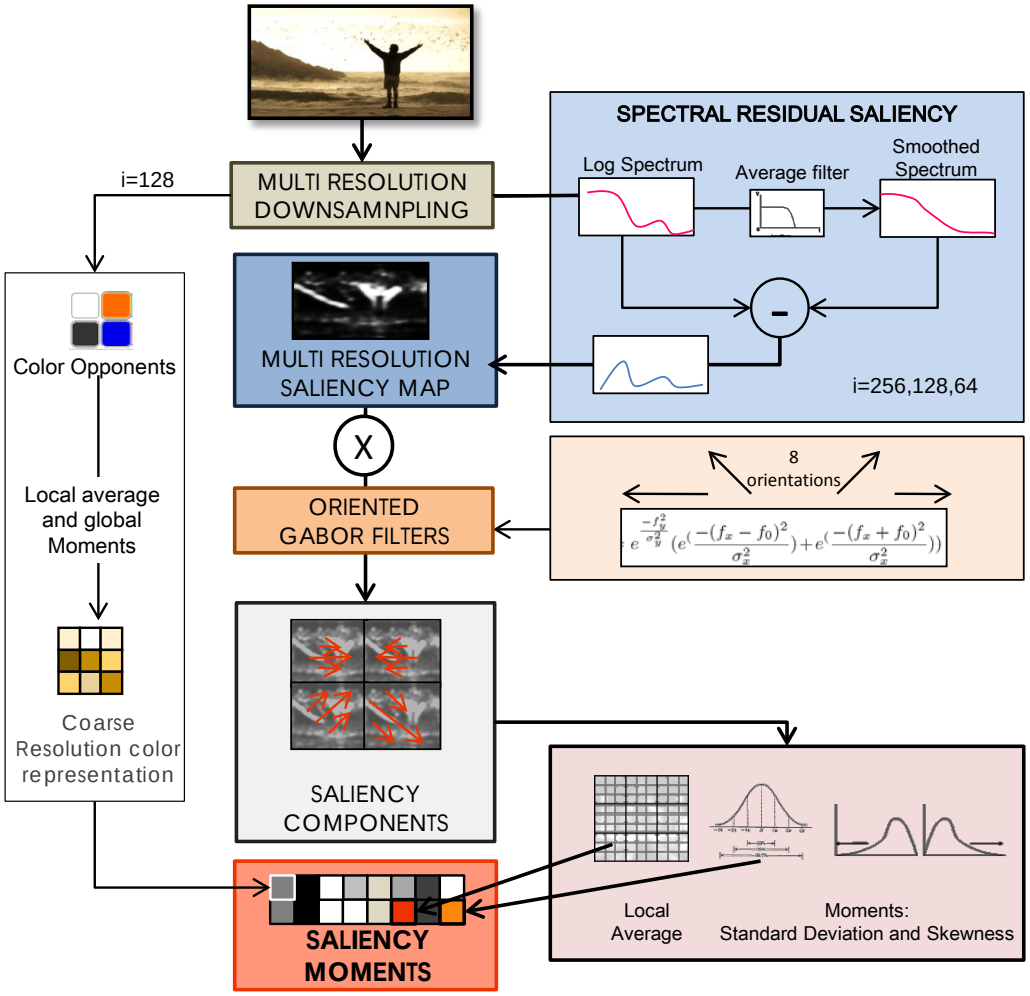


Figure 3.8: SM Algorithm: The spectral saliency at multiple resolutions is convolved with Gabor filters. The resulting "saliency components"' are then downsampled using local mean, and global standard deviation and skewness

models available in literature, we chose to compute the visual attention map with a spectrum-based approach presented in [68] by Hou et al., that we extensively explained in Sec. 3.1.2. Besides its efficiency and accuracy, we chose the Spectral residual detector for another peculiarity, namely its ability to *capture saliency at different scales.*

As pointed out in [68], Spectral Residual can detect salient regions under various scales of the image, depending on the size selected in the resizing preprocessing step. Different spatial scales lead to different saliency maps, detecting proto-objects with a level of details that increase with the resolution chosen, as shown in Fig.3.2 (b).

In our global feature, we compute the spectral residual $R(f_x, f_y)_i$ on three $i \times i$

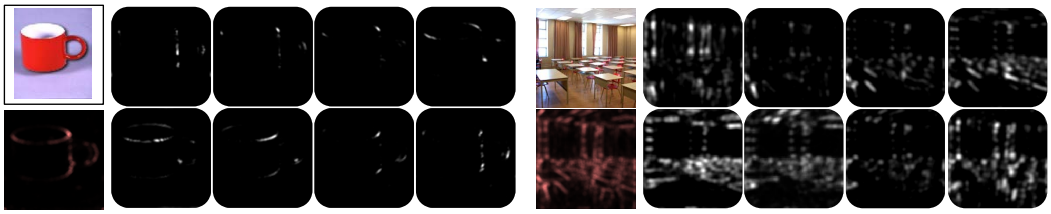


Figure 3.9: Saliency components: sampling the frequency-domain spectral residual with oriented Gabor filters we obtain different views of the saliency map in the pixel domain

rescaled versions of the input images ($i = 64, 128, 256$), simulating the variety of possible salient spatial frequencies (and salient shapes), from coarse to fine, available to the observer when recognizing a scene.

3.3.3.2 The Image Signature: Saliency Components and Saliency Moments

We now construct a coarse representation of the image based on the salient spectrum. We use as input of this step the Fourier-transformed Saliency Map, in Eq. (3.1), we process it with a Gabor wavelet in the frequency domain; finally, we compute average and higher order statistics in the pixel domain.

In fact, $R(f_x, f_y)_i$ is a very high dimensional signal (86016 variables: each component of the 3d-matrix $R(f_x, f_y)_i$, for all values of i) that we want to use as a whole to discriminate different image categories for the scene and object recognition task. We want to reduce the dimensionality of such information, finding a smaller set of variables that allow to preserve the variation between different image categories.

However, as shown by Torralba et al in [133], traditional techniques for dimensionality reduction, like Principal Component analysis [84], do not estimate the most informative components reliably, when applied on such spectral, high-dimensional signals. We therefore use a Gabor filter-based approach, proposed in [133] for the power spectrum dimensionality reduction, that approximates, as shown in [85] , and [163] , the behavior of the primary visual cortex cells receptive fields. Our proposed procedure is as follows:

1. **Gabor Filters Computation:** Here, we analyze the saliency distribution directly in the frequency domain with a set of oriented Gabor filters[49] that are described by the function:

$$G(f_x, f_y)_{i,\theta} = e^{\frac{-f_y^2}{\sigma_y^2}} (e^{\frac{-(f_x - f_0)^2}{\sigma_x^2}} + e^{\frac{-(f_x + f_0)^2}{\sigma_x^2}}) \tag{3.5}$$

where f_0 is the central frequency chosen to be 0.3 cycles/pixel, and σ_x, σ_y are filter parameters. As R_i is already a multi-resolution signal, and saliency is already created at different scales, the same central frequency f_0 is selected

for each scale i. For each band i, we considered 8 orientations by changing the value of θ to rotate the components in equation (3.5).

2. **Spectral Sampling:** We have therefore a set of 24 (8 orientations x 3 resolutions) filters that sample the spectral residual in the following way:

$$F(f_x, f_y)_{i,\theta} = R(f_x, f_y)_i \cdot G(f_x, f_y)_{i,\theta} \tag{3.6}$$

3. **Saliency Components:** Why Gabor filters? The convolution between the saliency map at a given scale and the filters (a product in the frequency domain) at different orientations gives us a set of "saliency components", that represent different points of view about the saliency shape matrix, "shots" of its discriminative regions, according to the orientation of the filter considered. We can obtain the pixel-domain equivalent $M(x, y)_{i,\theta}$ of the frequency-domain samples in Eq. (3.6), by applying the IFT to the samples $F(f_x, f_y)_{i,\theta}$. As shown in Fig.3.9, they represent fundamental, highly informative components of the saliency shape.

4. **Averaging Operations:** We now want to summarize, in a shorter index, meaningful information about the spatial distribution of such saliency components. We use a simple approach suggested in [164] for the downsampling of the saliency map: the local averaging. We divide each of the 24 saliency components into 16 non-overlapping sub-regions. We then consider each block as a sample and take the average value of every image block, and we store it in the image feature vector, as in equation (3.7):

$$V_{i,\theta}^{k,l} = \frac{1}{16} i^2 \sum_{x=\frac{1}{4}ik}^{\frac{(k+1)i}{4}-1} \sum_{y=\frac{1}{4}il}^{\frac{(l+1)i}{4}-1} M(x, y)_{i,\theta} \tag{3.7}$$

where k, l represent respectively the horizontal and vertical block indexes, and $i \times i$ is the saliency component resolution. We therefore obtain a 384-dimensional (16 blocks x 24 components) image index.

5. **Saliency Moments:** In order to make the feature more robust, and similar to the Color Moments feature [175], we interpret each saliency component as a probability distribution and calculate 2nd and 3rd moment, namely standard deviation and skewness, on the whole matrix $M(x, y)_{i,\theta}$, for all the i and θ considered. The result is a 48-dimensional vector storing the higher order statistics, that we concatenate with the previously computed index $V_{i,\theta}^{k,l}$ obtaining a descriptor composed of 432 elements: the SM descriptor.

3.3.3.3 The Color Contribution

The proposed approach, until now, receives as input a single-channel, grayscale image and builds a descriptor based on the luminance values only.

We add the chromatic information in our descriptor by concatenating a summarized representation of the dominant colors in the image, following an approach similar to the one in [175]:

1. We transform the RGB input image (at resolution i=128) into an opponents-based color space, namely the L*A*B*. The choice of this color space is again due to its biological plausibility: the LAB system is built to map the perceptual distances between colors, as explained extensively in [132]. Moreover, the channels A and B represent colors along the green-red and yellow-blue opponents, similarly to how the visual cortex gathers the chromatic information.

2. We perform averaging operations over subwindows obtained from the A and B channels.

3. Similar to the SM approach, we then calculate 2nd and 3rd order statistics on the global image matrix.

3.3.4 Experimental Validation

In this Section we present a set of experiments that we carry out to test the discriminative power of the Saliency Moments descriptor for semantic analysis.

3.3.4.1 Experimental Setup

We compare our Saliency Moment Descriptor with the most widely used global features for MMIR, by building MMIR systems for two different categorization tasks:

- **Scene Categorization**, for which we use two datasets, namely outdoor scene categories [133] and indoor scenes [143].

- **Object Recognition**, using the Caltech-101 dataset [41].

In particular, we consider for comparison the Gist descriptor [133], the wavelet feature [187], the Color Moments feature [175] and the Edge-Histogram based descriptor [200]. We also experiment with the two different versions of our image signature to test the influence of color opponents for scene and object recognition, by computing both SM and CSM and comparing them to the other descriptors (respectively, *Saliency Moments* and *CSM* in Fig. 3.10). For all datasets, for all features, we learn a model of the feature space using SVM with polynomial kernel of degree 2. For all the datasets considered, we use the same training/test setup as our baselines.

Moreover, we also show the effectiveness of SM for **Concept Detection for Video Retrieval**, by embedding it in a high-level feature extraction system tested on the TrecVID 2010 [168] dataset. We use as baseline the standard TrecVID baseline [153], as described in Chapter 2, and add the contribution of SM by linear fusion, computing the improvement in terms of mean average precision.

Moreover, we also present the results for our participation to the TrecVID 2012 Evaluation Campaign [128], where we presented a set of runs for the semantic Indexing Task, including one based on the addition of Saliency Moments to the baselines of features (similar to the one we use for TrecVID 2010).

3.3.4.2 Experimental Results

Outdoor Scene Categories

The first dataset considered is the 8-categories outdoor scenes dataset [133]. Results in Fig. 3.10(a) show that, despite its lower dimensionality, our visual attention-based feature outperforms the Gist descriptor, and that adding a coarse representation of the dominant colors further improves the prediction accuracy. By adding some light-weight local information, our hybrid descriptor outperforms the pure global descriptor while keeping similar efficiency.

Indoor Scene Categories

The second group of experiments is based on a dataset that has been first proposed in [143] as a new, unique database for indoor scene recognition. Despite the challenging task, results shown in Fig.3.10(b) confirm the discriminative power of saliency for image description: the CSM feature brings an improvement of 33% over the Gist descriptor, which is already substantially outperforming the other existing global descriptors.

Caltech-101

We evaluate also the effectiveness of our approach for object recognition on the Caltech 101 database. Despite from its limited amount of highly cluttered images and its lack in pose variation, we chose this database because it is one of the most diverse multi-object set of labeled images publicly available. Same trend can be spotted in the results for this task, with the SM obtaining again very good results on the average accuracy with the SM (+35% compared to the Gist descriptor and +21% compared to the edge histogram feature). Fig.3.10(c) shows the classification results for the proposed set of descriptors.

TrecVID 2010

We show here the results for the TrecVID 2010 semantic indexing task. Results in Fig. 3.10(d-e) show the per concept average precision and the MAP. By adding a new, discriminative source of information, namely the Saliency Color Moments, we introduced complementary knowledge on the image representation. Therefore, by combining the concept score of the five features with the saliency-based classifiers output we improve significantly the performances of the final retrieval framework. In particular, global scene concepts such as *Classroom*, *Cityscape* or *Nighttime* benefit from the introduction of SM in the pool of traditionally used features for this task.

TrecVID 2012

In Fig. 3.10 (f) the performances (MAP) of the various systems submitted by EURECOM for the light-SIN task of TrecVID 2012 edition are presented. In 2012, 15 out of 50 concepts were evaluated for this light task. Our run *Eurecom_ Videosense_ SM* is built on top of the baseline *Eurecom_ Baseline*, by adding two new descriptors including the Saliency Moments descriptor. With this run, our system is able to retrieve the videos of this challenging task with a MAP improvement of more than 28% compare to the baseline. This is again due to the complementarity brought by our saliency-based descriptor compared to existing global and local descriptors for MMIR.

3.4 Summary and Future Work

We have proposed a novel approach to model low-level features for scene recognition. We developed *hybrid* descriptors, by building holistic, coarse-grained representations of the image using saliency information. We ensured both efficiency and discriminative power by embedding some *locally*-parsed information in a fast, low dimensional, *global* description of the image. We created two different descriptors: **Saliency-Aware Color Moments** (SACM) [147], that uses the local saliency distribution as a measure to weigh the color moments index, resulting in a color feature that gives more importance to more informative image subregions; **Saliency Moments** (SM) [148], where we first extract a saliency map, namely a grayscale matrix highlighting the perceptually salient regions, we analyze its shape with Gabor filters, and finally extract from the obtained samples mean and higher order statistics. In both cases, we showed that the resulting hybrid descriptors outperform the state-of-the-art global descriptors for scene recognition. As mentioned in Sec. 3.1, the analysis in our work relies on a spectral saliency detector [68]. It was not in the aim of this work to compare different visual attention computational models for image indexing and retrieval. However, both SACM and SM performances could be further improved by using more complex saliency measures, e.g. the model proposed by Itti et al. in [73].

An idea from the future extensions of SACM comes from the observation that, similar to Color Moments, many of the global descriptors included in a CBIR systems are computed on a window basis (e.g. Wavelet Features [187] or the Edge Histogram [200]), in order to add some spatial constraint in the holistic representation of the image. Therefore, a possible extension of the Saliency-Aware Color Moments may involve the use of our informativeness measure to boost other window-based global features (e.g. the MPEG Edge histogram [200]).

On the other hand, despite the spectral sampling and the moments extraction, Saliency Moments is still quite high-dimensional compared to traditional low-level features (e.g. Color Moments and Wavelet Feature). Therefore, part of the future work to improve the SM descriptor will focus on more effective dimensionality reduction techniques. Another related topic to be explored is the chromatic com-

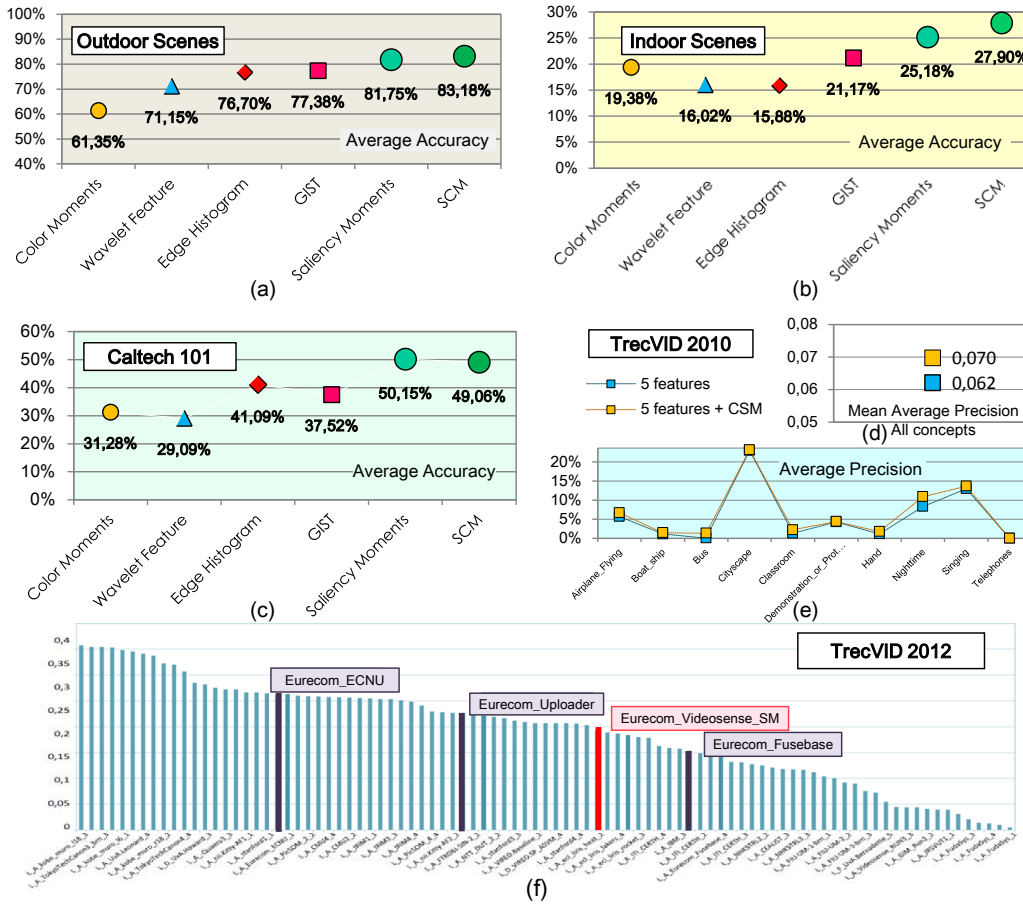


Figure 3.10: Performances on the test set for the different descriptors. Accuracy in scene recognition on the (a) outdoor scene dataset, (b) indoor scene dataset and in object recognition on the (c) Caltech-101. (d) mean average precision (MAP) and (e) per-concept average precision (AP) for TrecVID 2010: we show the improvement brought by adding Saliency Moments to the pool of visual descriptors (f) ranking of the participants to the TrecVID 2012 semantic Indexing task

ponent. By adding a simple, low dimensional representation of the dominant color we achieved very good performances for scene recognition, while the CSM in the Caltech 101 dataset performs slightly worse than SM. The proposed color contribution that we merge with our saliency-based descriptor is just one of the many biologically-plausible possibilities, and our future research will study how to relate the dominant color extraction with the visual attention information.

Chapter 4

Level 1: Aggregating Local Features through Marginal analysis and Copulae

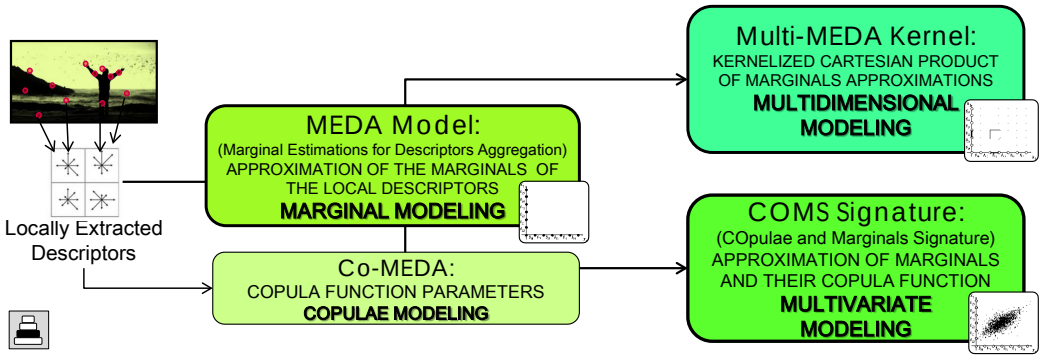


Figure 4.1: Our marginal-based aggregators for image recognition.

In this Chapter we propose three different methods for the aggregation of local image descriptors, and we prove their effectiveness for Scene Recognition and Video Retrieval. Unlike traditional techniques that aggregate the local descriptors using multivariate modeling approaches, the peculiarity of our pooling algorithms is that they stem from marginal analysis. We first aggregate the local descriptors through their ***marginal approximations****, and then build a* ***kernel*** *that can process the marginal approximations and infer a* ***multidimensional probability****. Finally, we use* ***Copula structures*** *to derive a* ***real multivariate probability*** *of the local descriptors arising from pure marginal information, without involving therefore expensive multivariate modeling techniques. We show that the resulting aggregated signatures are much more efficient and very discriminative compared to the existing methods for feature encoding and pooling.*

We have seen how low-level *global* features can be used to describe the visual content and automatically detect semantics by using supervised learning techniques.

In some cases however, pixel-level features cannot be used *directly* as input for traditional learning frameworks. In general, this occurs when performing *local* analysis with descriptors such as such as SIFT, HOG, or SURF[105, 31, 6].

The reason for this issue is that learning machines typically need as input image features that:

- has *low* dimensionality

- has *fixed* dimensionality

However, in local analysis, multiple Local Image Descriptors [105, 31, 6] (LIDs), are computed to describe the surrounding of either (I) *interest points* [44] or (II) *densely sampled points* [42]. When dealing with interest point description (I), a *variable amount* of local descriptors, or *keypoints*, are extracted from each image, making therefore the direct extraction of a *fixed-length*, global image signature unfeasible. Likewise, in case of densely extracted local features (II), the mere concatenation of the descriptors from a dense fixed grid would result in a redundant, high-dimensional, untreatable feature, while learning machines need low-level, informative features. In both cases, if we want to build a discriminative signature representing the image content based on the LIDS analysis, we would need a way to reduce the high dimensional data from local features into a smaller set of values of fixed length.

The General Solutions for Feature Aggregation

In order to describe the behavior of the image LIDs while preserving information and properly describe the image content, we cannot use simple operations for dimensionality reduction such as averaging neighboring LIDs values, because neighboring LIDs can carry very different values, since they describe different regions of interest in the image, and averaging them would mean canceling their discriminative power. In order to build a fixed-length signature for each image, the *global behavior* of the image LIDs needs to be captured: we want a feature describing which are the values the LIDs in every image, and with which magnitude,. We need to model the LIDs **probability distribution**. The general solution is therefore to **aggregate** the local descriptors into a new, statistically relevant, low dimensional image signature that gathers their properties and reflects their joint probability distribution function (PDF).

In general, this stage of the image analysis chain is performed in two steps: first, the local descriptors of a training set are **encoded** through multivariate modeling techniques into a universal model representing the global, natural keypoints distribution. For a new image, keypoints are then **pooled** into a fixed length signature based on their behavior with respect to the universal model (see Sec. 4.1.2 for further details).

Feature aggregation is an important level of processing in MMIR, since the final image signature must retain as much as possible all the amount of rich information stored in the locally extracted descriptors, and avoid as much as possible information loss during the aggregation process. Several works have tackled this problem under different point of views. For example, the Bag of Words model in its different versions [30, 129, 98] uses Vector Quantization (encoding step) and multidimensional histogram counting for keypoints multivariate PDF estimation (pooling step). Similarly Fisher Vectors [79] compute the distance (pooling) between the image PDF and the global probability distribution (previously encoded) of the keypoints through Fisher Kernels (see Sec. 4.1.2 for further details).

Feature aggregation is generally used in pure semantic analysis, and in particular for local features. On the other hand, the features used for aesthetic analysis are more similar to global features, and they can therefore be directly used as input for the Level 2 of the MMIR pyramid, i.e. the learning step. We will therefore focus our attention on the creation of new local features aggregators for semantic MMIR.

Major Issues and Our Contribution

Despite their good performances in image recognition and retrieval, one of the major drawbacks of all these approaches is their complexity and computational cost, for both clustering high dimensional feature space (determined by the descriptors distribution) and visual words assignment. Moreover these types of aggregators, since they base the image LIDs aggregation on a universal model (such as a codebook or a mixture) , do not directly reflect the real PDF of the image keypoints, dealing to a decrease of the discriminative power of local image descriptors.

In this chapter we present a substantial set of contributions to the improvement of the feature aggregation process, both from efficiency and an effectiveness point of view. While generally the pooling algorithms [79, 30, 78] aim to model the *multivariate* probability of the local image descriptors, we present a set of aggregators that stem from marginal, *monovariate* descriptor analysis. With our marginal-based approaches [146, 149, 152], we overcome most of the problems related to traditional feature pooling, such as low computational efficiency and loss of informativeness.

We test the effectiveness of our pooled descriptors for scene recognition and video retrieval, and we show that our monovariate analysis is a new, complementary, discriminative and efficient cue for automatic visual analysis.

In the following, we will first in Sec. 4.1 recall some principles of probability theory, useful for understanding how LIDs are treated, and then give a detailed introduction of the statistical analysis performed by the existing pooling methods. We will then provide a high-level overview of our contributions, and stress the differences and the complementarities with the existing approaches, see Sec. 4.2. We then detail the theory behind MEDA (Sec. 4.3), MultiMEDA (Sec. 4.4) and COMS (Sec. 4.5), namely our three proposed algorithms.

4.1 An Introduction to Feature Pooling: Statistics and Existing Approaches

In this Section, we will first introduce some notations and definitions, looking at the statistical properties of the image keypoints, by analyzing them as *random vectors*. We will then see how the most common feature encoding and pooling methods apply these principles and process the keypoints in order to obtain a fixed length image representation.

4.1.1 LIDs as Random Vectors

Generally, in local image analysis, for an image I we extract a set of m k-dimensional *local image descriptors*, or *local features*, or *keypoints* using a variety of existing methods from local image analysis [105, 6, 31]:

$$x^{(I)} = \{x_j^i\}_{j=1,\dots,k}^{i=1,\dots,m}$$

where k is typically 36 or 128 dimensions.

In order to understand the statistic properties of the image LIDs, we have to look at the descriptors $\{x_j^i\}$ as realizations of a general *random vector* x. A random vector is a set of correlated or uncorrelated random variables that share the same probability space. In our case the random variables are represented by the LID *components* x_j. A random vector of length k generates a probability in the $\mathbb{R}^k$ space,

$$p(x) = p(x_1, \dots, x_k) = P(X_1 = x_1, \dots, X_k = x_k)$$

We will call this measure *joint distribution*, or *multivariate distribution*.

Depending on the relations between the components, the multivariate PDF can be estimated in two ways. When components are **independent**, meaning that it does not exist a variable in the vector influencing the probability of the others, the multivariate probability $p(x)$ can be computed with the product of their probabilities, namely:

$$\prod_{j=1}^{k} p_j(x_j). \tag{4.1}$$

When components are **correlated**, the joint PDF needs to be estimated by looking at the behavior of the random vector in the k-dimensional space. There are several options for multivariate density estimation. *Parametric methods* assumes the vector distribution follows a specific density model such as Multivariate Gaussians or Gaussian mixtures [135]. *Non-parametric methods* such as Vector Quantization [4] try to model the global PDF without assuming any knowledge on the data distribution.

The **multivariate distribution** can give a lot of information regarding the behavior of the random vector. In our local image analysis case, the real shape of the LIDs PDF $p(x^{(I)})$ can give substantial discriminative information regarding the image content, representing the likelihood that the LID components take at the same time a given combination of values. However, the estimation of the shape of the joint PDF of a random vector is a non-trivial problem for high values of k.

On the other hand, in a random vector, the **monovariate** distributions of each component of the vector are the *marginal distributions*, and they represent the likelihood of each component to take a value in the $\mathbb{R}^1$ space, namely

$$p_j(x_j) = P(X_j = x_j)$$

Estimating the marginal probabilities of a random vector is pretty straight-forward,

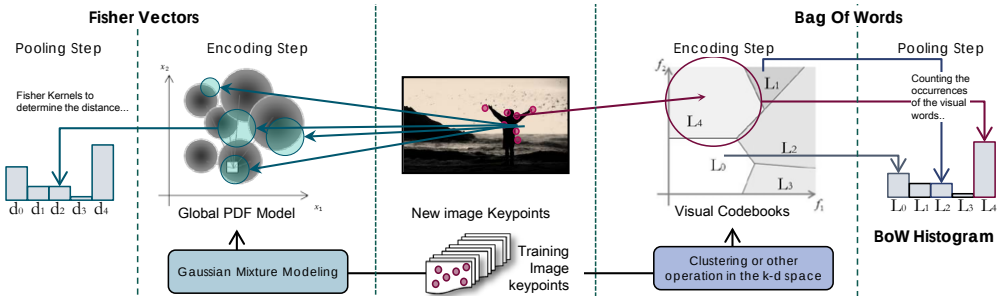


Figure 4.2: Bag of Words and Fisher Vectors: the state of the art in feature pooling.

since there is no need for operation in the $\mathbb{R}^k$ space, for example using histogram counts [176] or kernel density estimation [165]. The *marginal distributions* $p_j(x_j^{(I)})$ of local keypoints of an image I represent the behavior of each component of the LIDs: which values in its range are the most probable? They carry discriminative information about the single components, but they lack of describing the multidimensional properties of the LIDs.

4.1.2 The Major Existing Feature Aggregators and Their Statistic Analysis

While the marginal modeling of LIDs distribution has rarely been explored for semantic feature extraction, the estimation of the multivariate distribution of image LIDs has been extensively studied in the Multimedia Indexing field. The idea is to represent the content of an image through the shape of the distribution (or an approximation) of its LIDs, building discriminative signatures that can be used for learning and classification. But how is the LID information gathered so that their distribution in the $\mathbb{R}^k$ can be properly approximated?

In general, *LIDs aggregators do not directly represent the distribution of the image keypoints*, because this would lead again to computationally expensive, variable length, high dimensional image signatures, that could not be used as input for the learning machines. Traditional methods for LIDs aggregation *approximate the joint PDF* of the image keypoints by a 2-step process (see Fig. 4.2.

1. First, a **universal shared model** $u \approx p(x)$ of the LIDs space is learnt using parametric or non-parametric techniques such as vector quantization or Gaussian mixture models. In this step, the LIDs of a training set of images are **encoded** into a smaller set of values (e.g. codewords or Gaussian multivariates) representing a universal model of the general behavior of the keypoints of natural images.

2. For a new image I, the global PDF of their LIDs is then approximated based on the distribution of the image keypoints given the universal model $p(x^{(I)}) = p(x^{(I)}|u)$. This is obtained by **pooling** them (generally, by histogram count

[30] or, for example, soft assignment [190]) in fixed-length image signatures that can be easily compared and matched by traditional kernel machines in the subsequent learning phase in a practical, efficient way.

One of the most popular models for feature pooling is probably the **Bag of Visual Words model** (BoW) [30, 129, 98]. In this model, a codebook c of N_{bow} visual words is first generated by vector quantizing the LIDS of a training set of images (encoding step), by using various clustering techniques such as (kmeans [30], random forests [122], lattice-based [185]. For a new image, the final N_{bow}-dimensional signature is then obtained by approximating each LID to the closest visual word, and then pooling them in the final BoW vector. This is achieved by, for example, counting the resulting visual words occurrences, either without retaining spatial knowledge [30], or by building a spatial pyramid [96], or using methods like linear coordinate coding [208], Sparse Coding [206] and Local Coordinate Coding [194].

Given the partition of the k-dimensional space determined by the universal codebook c, the BoW signature approximates $p(x^{(I)})$, the joint PDF of the LIDs in a new image, with their joint probability given the codebook c, namely $p_{bow}(x^{(I)}) = p(x^{(I)}|c)$)see Fig. 4.4 (e)). A particular type of Bag of Words is the (Fig. 4.4 (d)) **Lattice- based BoW** [185]. This approach builds a vocabulary of k-dimensional hypercubes generated through the *monodimensional* quantization of each dimension of the LID (without involving therefore any clustering or operation in the high-dimensional space) in a fixed number of N_{lat} bins, and then reduce such vocabulary c_{lat} according to the informativeness of the resulting codewords. Even if this approach does not involve clustering, the size of the resulting codebook is exponential with the number of LID components ($O(N_{lat}^k)$), implying therefore to expensive searches and storage costs for the training phase.

A generative-discriminative approach for LID aggregation is the one of Fisher Vectors [79]. In this approach, the distribution of the image LIDs is approximated by first estimating the global LIDs PDF with a Gaussian Mixture Model [135] arising from the LIDs of a training set (encoding step), as shown in Fig. 4.4 (f). For each image, the gradient of the log likelihood of the set of image LIDs with respect to the parameters λ of the GMM is then computed using Fisher Kernels [76]. Finally the concatenation of the resulting partial derivatives is sorted in the final signatures (the Fisher Vectors) that model the probability of the image LIDs given the GMM parameters, namely $p_{fv}(x^{(I)}) = \bigtriangledown_\lambda \log p(x^{(I)}|\lambda)$.

As said, both approaches (BoW, Fisher Vectors) represent the joint probability of the image LIDs indirectly: they describe the behavior of the LIDs in an image given a universal model of the global LIDs space, obtained through operations (generally very expensive) in the k-dimensional space, such as clustering or mixture modeling. Despite its proved accuracy, this type of representation leads to a lack of discriminative power for complex classification tasks, and to high computational complexity in the training phase.

4.2 Our Approach: Marginal Modeling for Feature Aggregation

Given the properties and the issues of the most common feature aggregators, we propose in this Chapter a set of LIDs aggregators that differ significantly from the most common methods for feature pooling. We present three alternative and complementary approaches that achieve a global representation of the image LIDs behavior based on the **marginal** analysis. In contrast to the tendency of LIDs aggregators to use **multivariate** analysis approaches for the joint PDF approximation of the image LIDs $p(x^{(I)})$, in our models [146, 149] we exploit pure *monovariate* analysis for both unidimensional and multidimensional probability estimation, leading to more efficient and effective aggregated signatures for image categorization.

We can see our three contributions as three stages towards the complete accurate modeling of the multivariate LIDs PDF. Based on monovariate analysis, we first build a feature approximating the *marginal* PDF of the image keypoints, then we model their *independent joint probability* based on marginals, and finally their *multivariate probability* based Copulae structures. As a matter of fact, our methods for feature encoding and pooling can be summarized as follows (see Fig. 4.1).

1. We first introduce the **MEDA [146] signature** (Marginal Estimations for Descriptors Aggregation). In this approach, the shared model u is a set of n unidimensional bins ("letters") per dimension, obtained by quantizing the marginal distribution of each component of the LID The final image representation is a $k \times n$ histogram collecting the occurrences of such letters at each dimension. The MEDA signature represents therefore a concatenation of the approximated marginal distributions of the image LIDs components.
   This approach is very efficient, because it performs the vector quantization in a 1-d space, eliminating the correlation between the LID components by analyzing their distributions independently. However, by doing so, MEDA brakes the relations between the LID components, losing a lot of useful information regarding their multidimensional bounds.

2. In order to partially recover from this loss, in [149] we introduce the **Multi-MEDA kernel**, that represents a first attempt to improve the MEDA analysis by adding some multivariate information. The idea is that, if we assume that the LIDs components are independent, we can estimate their joint probability by multiplying their marginals (see Eq. 4.1). Since the actual Cartesian product of the marginal approximation would be computationally infeasible, we embed this process in a kernel for Support Vector Machines, that we name MultiMEDA. By a quick mathematical formulation, we kernelize the multiplication of the marginal values, generating a multidimensional probability out of the MEDA marginal approximations directly in the learning step.
   Even if MultiMEDA improves the MEDA discriminative power, it is still based on the assumption that the LIDs components are independent and that their marginals are uncorrelated. We need therefore to find a way to recover the

actual multidimensional information arising from the relations between the LIDS, and model the real multivariate PDF without using a universal multivariate model.

3. Given these observations, we present the **COMS** signature, (COpulae and Marginals). Our idea is to use Copula Theory [127] to build a complete multivariate analysis of the LID space and generate a feature vector out of such analysis. Why Copulae? Copulae are statistical tools for linking the marginals of the variables in a random vector with their multivariate joint distribution, modeling separately marginal distributions and their dependence structure (the Copula). We use Copulae to analyze the LIDs multivariate density by using marginal distributions only, in an efficient and statistically meaningful way. Given the marginals approximations in MEDA, for each image we model the corresponding Copula structure, that we store in the **Co-MEDA** vector, and we couple it with its MEDA descriptor, generating the COMS signature. With COMS, we finally achieve a complete representation of the real distribution of the image LIDS, based on pure marginal analysis, without referring to any $k-$dimensional universal models such as codebooks or Gaussian mixtures.

We test our proposed feature aggregators with MMIR system for scene recognition and video retrieval, and we show that such methods are not only very efficient, but they also lead to very discriminative image signatures. Moreover, given that the type of probabilistic analysis we perform on the LIDs differs significantly from the existing methods for feature pooling, when we combine our methods with BoW or Fisher kernel we introduce a new, complementary type of information regarding the keypoints distribution, achieving great improvements on the global performances of the visual analysis systems considered.

4.3 Marginal Modeling: Visual Alphabets for Descriptor Aggregation (MEDA) Model

In this Section, we describe our first idea for LIDs aggregation based on marginal analysis.

We present here a simple, fast and effective algorithm for local feature quantization that we name **MEDA** (Marginal Estimation for Descriptors Aggregation). This approach provides a different way of aggregating local descriptors that does not involve any clustering or operation in the high-dimensional space, leading to an image signature that requires much less computation and provides better accuracy compared to traditional BOW models.

Similar to the BOW model, a k-dimensional local invariant descriptors are used to describe a set of interest points in the image. While the general approach is to then perform LID quantization in the k-dimensional space, the basic idea of our approach is to model the k-dimensional space by the k marginal distributions approximations.

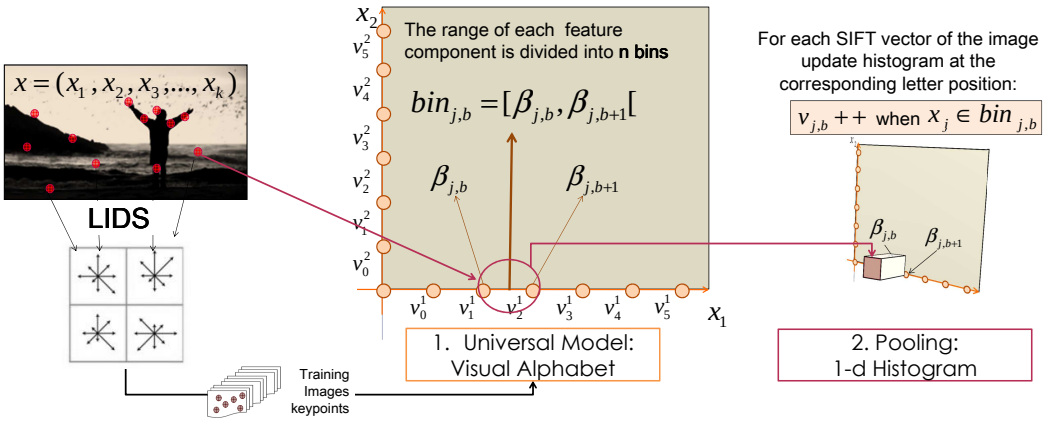


Figure 4.3: MEDA: a histogram representing the component distribution over unidimensional bins.

This is obtained with an aggregation process that involves (see Fig. 4.3 for a visual explanation) two steps:

1. **(encoding)**we quantize the range of each dimension of the LID into n bins, defining a reduced set of possible values that each of the k components of a described point can take , leading to a universal codebook $c_{meda} = \{c^j_{meda}\}^k_{j=1}$ for each dimension composed of $1-d$ words (letters) obtained through one-dimensional marginal quantization.

2. **(pooling process)** given an image and its set of descriptors, we count the frequencies of the computed bin values, and we collect them in a $k \times n$ histogram, i.e. the MEDA image signature. Therefore, the marginal, i.e. the probability distribution of each component of the LID of a new image is approximated by a histogram representing its frequency over unidimensional bins. The resulting MEDA vector represents the concatenation of approximations of the marginal distributions: $p_{meda}(x^{(I)}) = \cup^k_{j=1} p_j(x^{(I)}_j | c^j_{meda})$, see Fig. 4.4 (a).

Following the textual metaphor of the BOW, our method defines, for each component of the LID, a set of possible 1-d visual *letters*, namely the bin values; the collection of such *letters* is a *visual alphabet* that allows the mapping of an image into a fixed-length attribute vector. We present and compare three different methods to define the values in the visual *alphabet*, based on different types of range quantization, namely uniform quantization, quantile-based quantization and an entropy-based quantization we perform using a decision tree.

MEDA vectors carry therefore pure marginal information regarding the LIDs distribution, in contrast with existing approaches that model pure multivariate information. With MEDA, we therefore bring new knowledge about the LIDs distribution in the local image analysis, estimating a different, complementary probability, $p_{meda} \neq p_{bow}, p_{fv}$. We indeed tested the performances of MEDA for semantic analysis not only by considering it as a stand-alone descriptor, but also by combining

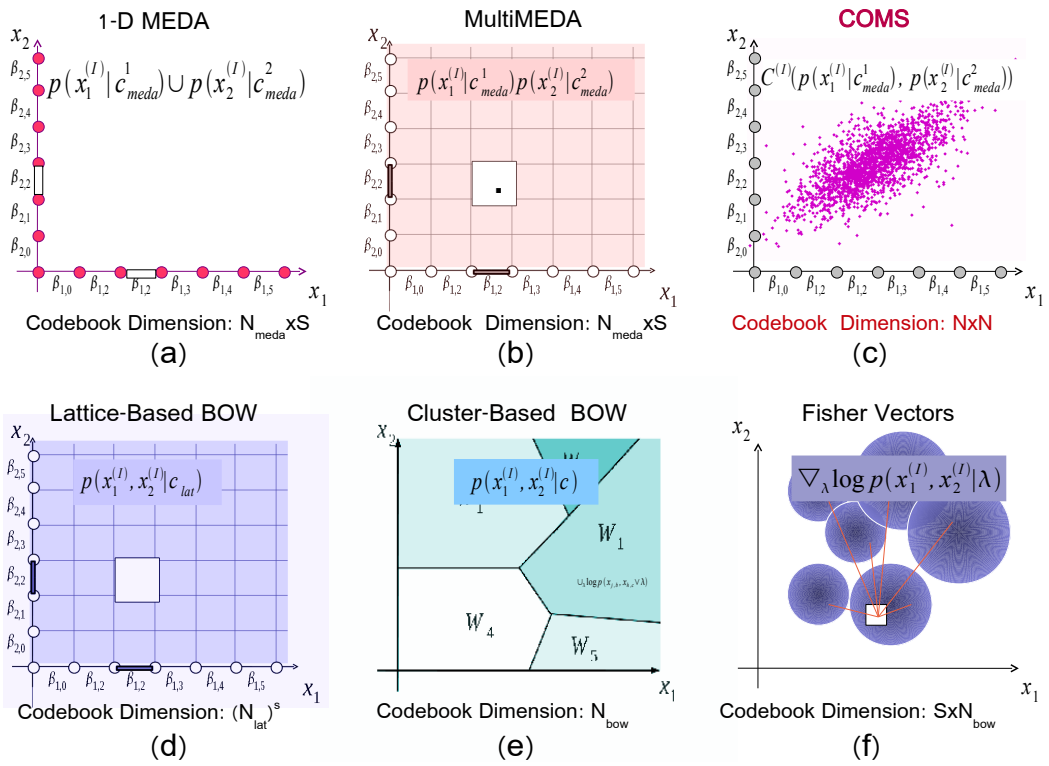


Figure 4.4: Comparison between the existing LID aggregators and our Copula-Based approach, based on the type of probabilistic analysis they perform on the image LIDs.

it with multivariate LIDs aggregators such as BoW, showing its effectiveness and complementarity for a variety MMIR tasks.

In the following, we will first present a detailed implementation of the MEDA model (Sec. 4.3.1) for local features quantization; we then explain, in Sec. 4.3.2, a variety of methods to approximate the range of the components. Finally, in Sec. 4.3.3 we will test the MEDA model for semantic analysis by embedding it into MMIR systems for scene recognition and concept detection for video retrieval.

4.3.1 The Signature: Marginals Estimation for Descriptors Aggregation

We propose an image representation that collects in a histogram the frequency of each component of the locally extracted vectors. While the BOW model quantizes the local features in a multi-dimensional space (*words*) determined by the descriptor length, here the quantization is performed in a 1-d space, for each component (*letter*) of the LID.

As in BoW, we start our aggregation process with an encoding step. We consider a set of LIDs each image I of a training set, namely $x^{(I)} = (x_1^i, \ldots, x_k^i)$, where each

element x_j^i represents the value of the descriptor x^i at position j, $j = 1, \ldots, k$ in image I.

After normalization, each element x_j^i can take a value in the finite interval $R = [-1, 1]$, which covers a very large set of possible discrete values $a_1, a_2, \ldots, a_m$. The idea here is to quantize R by mapping it into a smaller set of n discrete values

$$\beta_{j,b} \in R, b{=}0, \ldots, n{-}1, n{<}m$$

corresponding to a set of bins $c_{meda}^j = bin_{j,1}, \ldots, bin_{j,n-1}$. Such values correspond to our *alphabet*, defined for each dimension of the LID:

$$bin_{j,b}{=}[\beta_{j,b}, \beta_{j,b+1}[\tag{4.2}$$

(+1 is added to the last bin).

By doing so, each element in an image can be represented by the index of corresponding bin $bin_{j,b} : \beta_{j,b} \leq x_j^i < \beta_{j,b+1}$. The choice of the bin boundaries values will be discussed in the next Section.

We have therefore defined a set of shared visual *letters* (our universal model) that can be used to approximate the marginal distribution of the jth element of the descriptors in the image. We can now perform the pooling, namely represent the image as the collection of the number of elements x_j^i, $\forall x_j^i \in I$ that fall into each of the identified bins .

The resulting signature for the image I is a vector

$$v^{(I)} = (v_{1,1}, v_{1,2}, \ldots, v_{2,1}, v_{2,2}, \ldots, v_{k,n}) \tag{4.3}$$

with $v_{j,b} = \#\{x^i : x_j^i \in bin_{j,b}\}$, $\forall x_j^i \in I$,[1]

The MEDA vector in Eq. (4.3) can be seen as a concatenation of k n-dimensional vectors $\{v_{j,\cdot}\} = \{p(x_{j,1}^{(I)}), \ldots, p(x_{j,n}^{(I)})\}$. Each $v_{j,\cdot}$ represents the approximation of the marginal $p(x_j^{(I)})$ of the jth component of the LIDs in image I, given the probability $p(x_{j,b}^{(I)})$ at each bin $bin_{j,b}$.

The dimension of the MEDA signature is therefore $n \times k$.

4.3.2 Alphabet Construction

How to define the boundaries of such bins, our *letters*, so that the marginal of the jth component, $\forall j$ is properly estimated? In this section we tackle this issue using three different approaches, namely:

1. **uniform quantization**: the range is divided into n equally spaced bins, see Sec. 4.3.2.1

2. **quantile-based quantization**: the range is divided so that the probability of a sample to fall into a bin is equal for all the n bins in the quantized space,

[1] $\#\{\cdot\}$ is a function that counts the number of the elements that satisfy the condition in brackets.

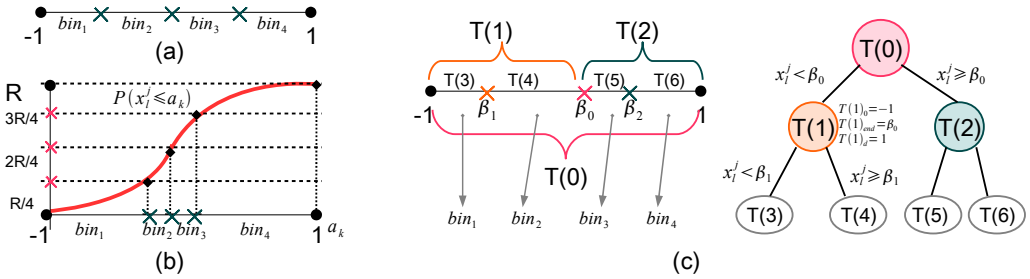


Figure 4.5: Three versions of the MEDA alphabet: (a) uniform, (b) quantile-based, (c) entropy based.

see Sec. 4.3.2.2

3. **tree-based quantization**: for each bin, the boundaries are learnt by minimizing the overall entropy given a progressively smaller interval of R, as shown in Sec. 4.3.2.3. Each of these methods leads to a different version of the MEDA histogram, that will be evaluated in Sec 4.3.3.

4.3.2.1 Uniform Bins

The simplest approach to define the bin boundaries over the data range is the uniform quantization. The advantage of such a simple approach is that it does not require prior knowledge of the marginal distribution of the components. As every component of the LID can take values in the same interval [-1,1], the resulting *alphabet* is an identical set of *letters* for every j. The range R is divided into n equal intervals of length $2/n$, and the set of bins valid for every component (see Fig. 4.5(a)) is defined as:

$$bin_b = [-1 + \frac{2b}{n}, -1 + \frac{2(b+1)}{n}[$$

4.3.2.2 Quantile-Based Bins

Here we try to adapt the width of each bin to the probability distribution of the component over the data range. When computing quantile-based bins, we generate a simple universal model that takes into account the general marginal behavior of each component of the LID. We will then use such general model to define the specific behavior of the image LID.

This process requires a learning phase in which we identify the probability of the jth component of the descriptors to take the value a_r, $r = 1, \ldots, \nu$, with $\nu >> n$ in the range R, (see Fig. 4.5(b)).

We need a dataset of N images over which we collect W described points x^l, $l = 1, \ldots, W$; we can then define the marginal:

$$p(a_r^j) = \#\{x^l : x_j^l = a_r\}$$

and the cumulative probability

$$P(x_j^l \leq a_r) = \sum_{s=1}^{r} p(a_s^j)$$

of each component given all the LIDs in the training set.

We want now each component x_j^l to be equally probable for all the bins in the range: we need therefore to find those values in the interval for which $p(x_j^l \in bin_{j,b}) = W/n$ for all b, being $\sum p(a_r^j) = W$. The final set of bins is defined as:

$$bin_{j,b} = [a_r^j : P(x_j^l \leq a_r) = \frac{bW}{n}, a_r^j : P(x_j^l \leq a_r) = \frac{(b+1)W}{n}[$$

4.3.2.3 Entropy-Based Bins

We propose a partition of the data range into a set of unbalanced bins, selected based on the minimization of the overall entropy. Here again we need a learning phase on a training set of N images and a total of W keypoints x^l, $l = 1, \ldots, W$ but in this case, we perform a *supervised* search in order to find the best boundaries for our bins.

We build, for each position j of the LID, a decision tree T^j, with n splits built in n iterations, that progressively learns the boundaries of each $bin_{j,b}$.

Each node $T^j(t)$, at depth $T_d^j(t)$ of the tree considers the set of x_j^l that take values between $T_0^j(t)$ and $T_{end}^j(t)$. The tree growing starts from the root node $T^j(0)$, corresponding to the whole set of $x_j^l \in R$ and, at each step, finds the value θ_t^j in R for which the resulting partition of the data has the minimum entropy, i.e. the optimum bin boundary.

If we assume the dataset is categorized in c classes $y_1, .., y_c$, the general entropy of the data for a split $a_r \in R$ is:

$$\begin{aligned} H(y|a_k) = &-p(x_j^l < a_r) \sum_{p=1}^{c} p(y_p|x_j^l < a_r) log(p(y_p|x_j^l < a_r)) \\ &-p(x_j^l \geq a_r) \sum_{p=1}^{c} p(y_p|x_j^l \geq a_r) log(p(y_p|x_j^l \geq a_r)) \end{aligned}$$

with $p(y_p|x_j^l < a_r)$ and $p(y_p|x_j^l \geq a_r)$ being the probability of a component belonging to an image labeled with category y_p to fall into the low/high bin generated by the split.[2]

[2] $p(y_p|x_j^l < a_r) = \frac{\#\{x_j^l : x_j^l < a \in y_p\}}{\#\{x_j^l : x_j^l < a_r\}}$; $p(y_p|x_j^l \geq a_r) = \frac{\#\{x_j^l : x_j^l \geq a \in y_p\}}{\#\{x_j^l : x_j^l \geq a_r\}}$

The following is the pseudo-code that summarizes how to grow a decision tree to learn the *alphabet* for the jth component:

Grow_Tree
$T^j(0) = \{root\}$
repeat
  choose unmarked leaf $T^j(t)$
  find $\theta_t^j = \arg\min_{a_r} H_t(y|a_r)$, $T_0^j(t) \leq x_j^l < T_e^j nd(t)$
  if $T(t)_d^j < max_depth$ **then**
    $T_0^j(t+1) \leftarrow T_0^j(t)$, $T_{end}^j(t+1) \leftarrow \theta_t^j$ {left child}
    $T_0^j(t+2) \leftarrow \theta_t^j$, $T_{end}^j(t+2) \leftarrow T_{end}^j(t)$ {right child}
  else
    mark $T^j(t)$
until all leaves are marked
$\beta_{j,b} \leftarrow$ in order tree walk on θ_t^j

Two child nodes $T^j(1)$ and $T^j(2)$ are created as the result of the split at the first iteration (see Fig. 4.5 (c)); at the second iteration, $T^j(1)$ will find the best split for the set of elements for which holds $-1 \leq x_j^l < \theta_0$, while $T^j(2)$ will consider those x_j^l that lie between θ_0^j and 1. The process is iterated until the maximum depth (max_depth) required to identify n bins is reached. Finally, the set of boundaries θ_t^j found is sorted and the bin values are assigned according to Eq. (4.2).

4.3.3 Experimental Validation

This Section presents an evaluation of the different versions of MEDA as described in the previous Section. In order to test the discriminative power of our descriptor, we built an MMIR system based on the MEDA signature in its different versions for a variety of challenging tasks. We compare accuracy and computational efficiency of MEDA and BOW on two datasets (indoor and outdoor) for scene recognition. We then test the effectiveness of the two approaches for concept detection in a video retrieval system built for the TrecVID 2010 [168] database.

4.3.3.1 Scene Recognition Task

We evaluate the performances of our model for image recognition into two challenging datasets, for indoor and outdoor scene categorization.

First, we extract the image local descriptors using the PCA-SIFT method described in [87], which reduces the dimensionality ($d = 36$) of the original SIFT (as proposed in [105], with $d = 128$) by applying PCA on the gradient image around the salient point. Once the local descriptors are extracted, we aggregate them using both BOW, by clustering a subset of training images using a standard k-means algorithm, and MEDA model (we implement the three different versions of

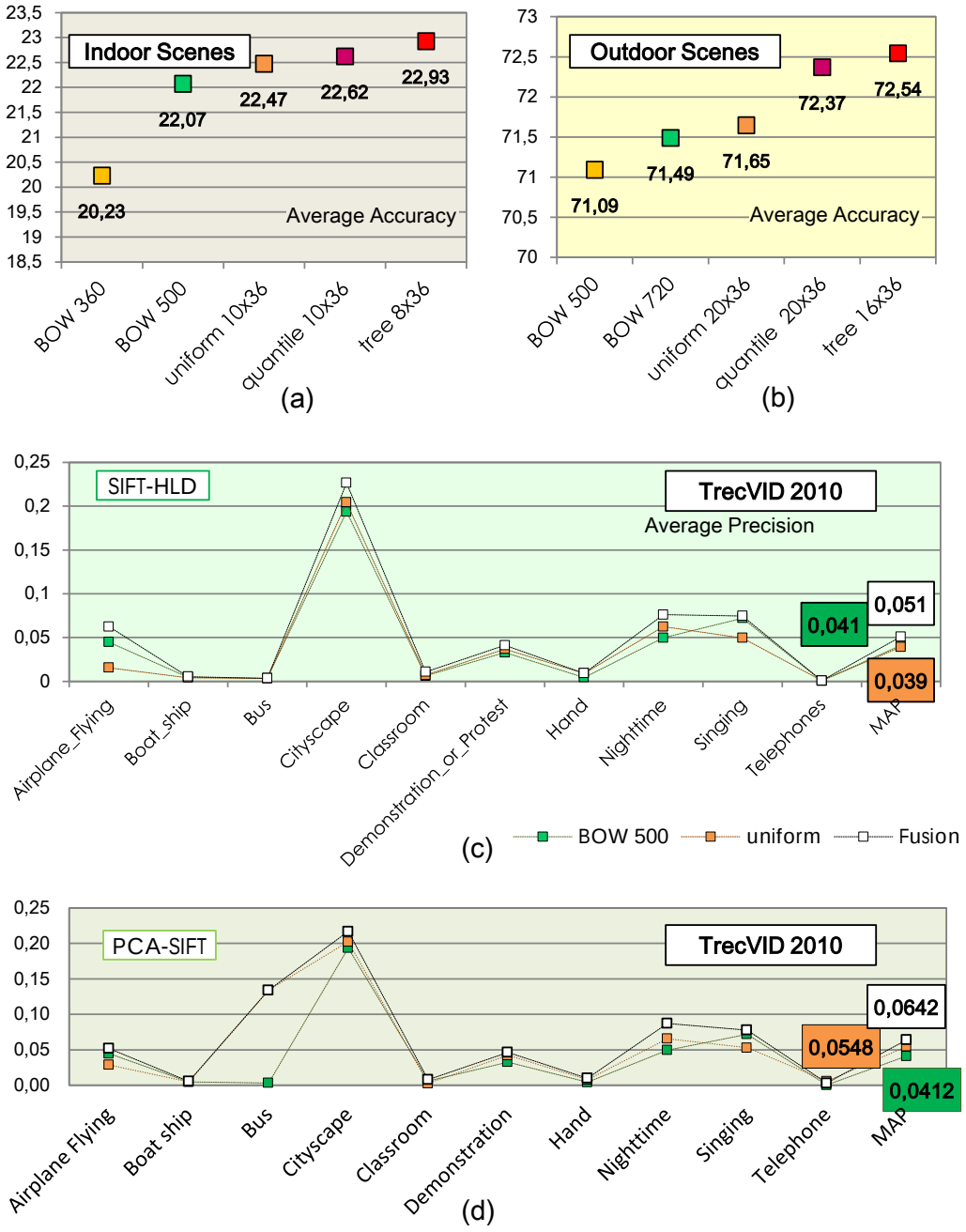


Figure 4.6: Comparison of performances of MMIR systems based on MEDA and BOW for (a,b) indoor and outdoor scene recognition and (c) video retrieval with input SIFT HLD (d) video retrieval with input PCA-SIFT

the MEDA model according to the methods in Sec. 3). For both MEDA and BoW, we finally use a one vs. all SVM with chi-2 kernel to separate each class from the

Indoor 67 Dataset			Outdoor 8 Dataset		
BOW	360	62503	BOW	500	10027
	500	86685		720	14429
MEDA	uniform (10×36)	135	MEDA	uniform (20×36)	55
	quantile (10×36)	163		quantile (20×36)	23
	tree (8×36)	401		quantile (20×36)	23

Table 4.1: Comparison of the computational times (in seconds) for computing BOW and MEDA in the training phase, for Indoor and Outdoor scenes dataset.

others,. For all the features and datasets, we use the training/test subset separation of the corresponding baselines. The number of bins of the MEDA descriptor for each dataset is optimized in the learning phase. As evaluation measure, we use the average multiclass prediction accuracy.

Experimental Setup

The first MMIR task for which we test the performances of our new MEDA model is the **Outdoor Scenes Dataset**.

We use as input to learning machines our 3 different versions of the MEDA signature.

- MEDA with uniform quantization, with 20 bins per dimension, resulting in a signature with $20 \times 36 = 720$ dimensions (*uniform* 20×36 in Fig. 4.6 (b)).
- Quantile-based MEDA, with 20 bins per dimension, resulting in a 720-dimensional signature (*quantile* 20×36 in Fig. 4.6 (b)).
- Entropy-based MEDA, with 16 bins per dimension, leading to a 576-dimensional signature, marked as *tree* 16×36 *in Fig. 4.6* (b).

In order to compare the performances with MEDA, we build BOW signatures of comparable dimensionality, creating a set of visual dictionaries with 500/720 visual words (*BOW 500/BOW 720* in Fig. 4.6 (b)).
For the second group of scene recognition experiments, we classify the images of the **Indoor Scenes Dataset** using MEDA and BOW. We define the following signatures for this experiment (see Fig. 4.6 (a)):

- *uniform* 10×36, MEDA with uniform quantization and 10 bins ;
- *quantile* 10×36, quantile-based MEDA with 10 bins
- *tree* 8×36, tree-based MEDA with 8 bins.

In order to compare our features with a traditional descriptor for multivariate LID analysis, we compute BoW signatures with similar dimensionality, based on dictionaries of 360/500 visual words ($BOW360, BOW500$ in Fig. 4.6 (a)).

Results Discussion

As we can see from Table 4.1, the most complex version ($tree \cdot \times 36$) of the MEDA model is more than 150 times less computationally expensive compared to the BOW model corresponding to the same feature size.

Moreover, we show in Fig 4.6(a-b) that MEDA is not only efficient, but it outperforms in accuracy the BOW model by 10% for the Indoor Scenes Dataset and 3% for the Outdoor Scenes.

Despite its simplicity, the BoW model with k-means is an example of multidimensional modeling of Local Image Descriptors. With our results we show that, for relatively small numbers of visual words, the approximation generated by the multivariate analysis and the related vector quantization deals to an information loss and lacks of properly modeling the correlation between the LID components. As a matter of fact, a simple descriptor such as MEDA, that involves uni-dimensional quantization only, and that is based on pure marginal approximation, outperforms BoW for a number of words comparable to the MEDA dimensionality.

4.3.3.2 Video Retrieval Task

For the Light semantic Indexing Task of TrecVID (SIN), participants are required to build a retrieval system that produces a ranked list of relevant shots ten semantic concepts. In this Section, we test and compare the MEDA descriptor for this task.

Experimental Setup

For our experiments, we extract SIFT-HLD [105] descriptors with $k = 128$ and PCA-SIFT descriptors with $k = 36$ from the video keyframes (see Chapter2 for baselines explanation) and quantize them using BOW with 500 words (*BOW 500* in Fig. 4.6 (c-d)), as in [153], and MEDA with uniform quantization (*uniform* in Fig. 4.6 (c-d)). The number of MEDA bins is optimized per concept in the training phase. For both descriptors, a set of SVM-based classifiers is trained with chi-2 kernel to detect the concept presence, for each concept. The partitioning of the datasets in training and test subset is exactly as explained in Chapter 2 for the baselines.

The concepts score of MEDA and BOW are then linearly fused (*fusion* in Figg. 4.6(c-d)) to evaluate the effectiveness of the combination of the two approaches. As evaluation measure, we use the Mean Average Precision.

Results Discussion

Despite its simplicity and efficiency, MEDA achieves retrieval results comparable with the BOW model, as shown in Fig.4.6(c-d) for traditional SIFT-HLD modeling. When dealing with PCA-SIFT, where the dimensions represent the projection

of the SIFT dimensions on orthogonal components, and therefore they are more informative, MEDA clearly outperforms BOW for the semantic Indexing Task.

The most interesting result here is the improvement we obtain on the BOW-based retrieval (+25% on the final MAP for 128-d LIDs, +50% for 36-d LIDs) by combining it with the MEDA-based retrieval. Our technique for descriptor aggregation brings new, complementary information to a traditional BOW model. As a matter of fact, MEDA calculates the frequency of each component (visual *letter*), while BOW calculate the frequency for the whole vector (visual *word*). MEDA gathers therefore a more holistic property of the image LIDs, counting the occurrences of each specific characteristic of the LIDs over all the image, without considering the LID vector as a whole chunk of locally extracted information. This holistic property of MEDA makes it therefore more suitable for scene and global concept recognition, rather than local object recognition, justifying therefore the poor results we obtain with MEDA for some local concepts e.g. Airplane_Flying, compared to the improvement brought for concepts such as Nighttime.

4.4 Multidimensional Modeling of Marginals: Multi-MEDA Kenrel

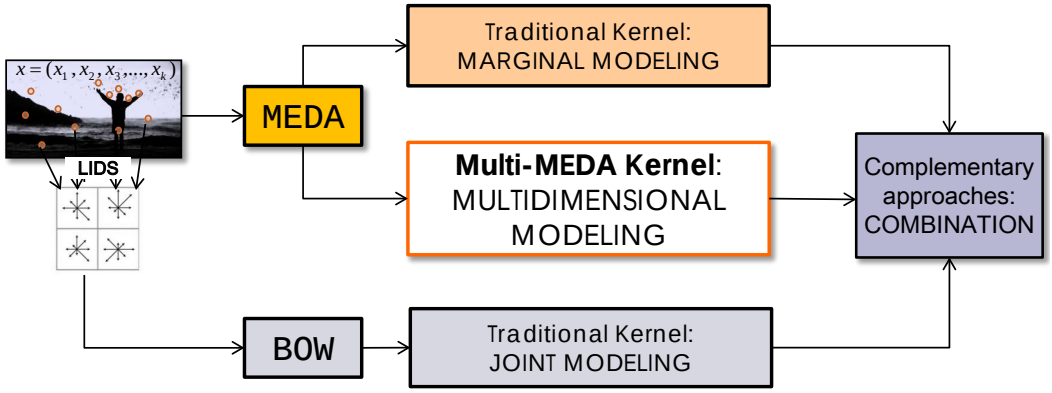


Figure 4.7: Multi-MEDA: a kernel over MEDA descriptors for multivariate probability modeling from marginals

As shown in Sec 4.3, the MEDA approach is very efficient, because it performs independent vector quantization in a 1-d space, forgetting the correlation between the LID dimensions analyzing their monovariate distributions only. Nevertheless, by doing so, MEDA brakes the relationships between the LID components, losing a lot of useful information for image representation. However, LID vectors arise from the analysis of an entire image region, and each element in a LID is crucial to define the surroundings of an interest point. It is therefore important to analyze the real multivariate information that characterizes those vectors.

Given these observations, we present here a first attempt to compensate the loss of information caused by the marginal analysis, by building **Multi-MEDA**, namely a new kernel function designed for the MEDA signature, that allows to model the joint contribution of the LID components in an efficient way. Our Multi-MEDA kernel models in a *linear* time the k-dimensional LID space, by deriving a multivariate probability from the k marginal approximations. With our approach, we keep as input to the kernel machine the classic, marginal-based MEDA signature, but we increase its discriminative power by analyzing it under a multidimensional perspective through the kernel formulation.

The main idea behind the Multi-MEDA kernel is that, since MEDA considers each dimension of the LID as an independent variable, we can approximate the joint distribution of the LID components by multiplying their marginal distributions, according to Eq. 4.1. However, an image signature supporting such model would require a k-fold Cartesian product of n-dimensional vectors, namely the multiplication of the approximations concatenated in the MEDA signature. This would lead again to an exponentially complex problem ($O(n^k)$), with a codebook of n^k elements and an extremely high-dimensional feature. For this reason, the key aspect of our approach is that *we do not compute explicitly the image signature nor the visual dictionary*, and instead we shift the computation of the multivariate probability inside the kernel machine. As a matter of fact, Multi-MEDA is a **shift-invariant kernel that embeds the marginals multiplication**, i.e. the Cartesian product of the marginal approximations. The most important property of our kernel is that it does not require exponential time to achieve the multidimensional modeling. We indeed show that the cost of computing the k-dimensional joint probability with the Multi-MEDA kernel becomes linear with the dimension of the LID and the number of letters in the MEDA codebook ($(O(nk)$). Therefore, although MEDA is built to describe marginal 1-d probabilities, when placing the Multi-MEDA kernel on top of MEDA signatures, we can reconstruct a model of the LID space that is based on a $k-d$ multi-variate probability, without needing to quantize the $k-d$ space.

Compared to the models generated by traditional SVM kernels over MEDA signatures, Multi-MEDA represents the LID space under a new, complementary point of view, as shown in Figg. 4.7 and 4.4. Multi-MEDA allows to explore two spaces (marginal-based and multidimensional) with the same feature (MEDA). Moreover, both the MEDA model and Multi-MEDA model are in turn different from the joint distribution approximation generated by traditional BoW approaches. By introducing Multi-MEDA, we therefore introduce a new discriminative source of information regarding the LID distribution, that can be combined with the MEDA and BoW models, leading to a significant increase (+50 %) of the MMIR performances, without requiring the computation of new LIDs, and without introducing exponential complexity.

In Sec. 4.4.1 we will look at the novelty introduced by MultiMEDA, explaining the technical and statistical differences with existing methods. We will then recall some MEDA principles and look at our marginal-based descriptor from a kernel perspective (see Sec. 4.4.2). Finally, in Sec. 4.4.3 we will show our kernelized

solution for marginal multiplication, namely the MultiMEDA kernel, and in Sec. 4.4.4 validate our theory with a set of experimental results.

4.4.1 Peculiarities of The MultiMEDA Kernel

We summarize here the novelty of MultiMEDA compared to similar approaches, from both a technical and a statistical point of view.

As said, eliminating the correlation between the LID elements with visual alphabets can cause losses of precious information for image description. This issue motivates us to perform a kernel-based analysis on the MEDA signature, that allows to learn a multivariate model, and that therefore takes into account the relations between the LIDs components. To our knowledge, the Multi-MEDA approach is one of the first attempts to improve the MEDA model by focusing on the kernel properties.

As a matter of fact, the cooperation between LID aggregators and kernels has mainly been investigated for extending the traditional BoW model. Various steps of the BoW approach has been improved through the interaction with kernels: a better codebook generation is achieved in [203] by using the Histogram Intersection Kernel in an unsupervised manner, while in [51] codebooks are used as free parameters of a Multiple Kernel Learning-based learning algorithm. maji2008classification swain1991color Lazebnik et al. in [97] use Spatial Pyramid Kernels to add the spatial information in the BoW model learning. The learning step is also improved in [12] by mapping the image LIDs into a low dimensional feature space, then averaging such vectors to obtain a set-level feature, and finally using a linear classifier to model the resulting vectors.

Our approach is different from the mentioned approaches because, first of all, we do not analyze the BoW model, but we extend instead the MEDA model to a multi-dimensional model through a kernel-based learning. Moreover, although we generate a model that works on a multivariate probability, the space that we explore through the Multi-MEDA kernel is statistically different from the space determined by vector quantization in BoW. What are the reasons of these differences?

MultiMEDA allows for multidimensional probability estimation (Fig. 4.4 (b)) by performing a kernelized Cartesian product of the marginal approximations in MEDA, assuming independence between LIDs components. The probability generated by MultiMEDA is therefore $p_{Mmeda}(x^{(I)}) = \prod_{j=1}^{k} p_j(x_j^{(I)}|c_{meda}^j)$. Since the computation of this multivariate distribution is performed inside the Multi-MEDA kernel, in our approach we do not need to compute a new visual signature or express explicitly the shared codebook, and we use instead as input the traditional MEDA vector.

p_{meda} and p_{Mmeda} are therefore generated using the same feature vector, but analyzing it with different kernels (traditional RBF or linear in the first case, Multi-MEDA in the second case). However, while the first one is a 1-dimensional *marginal* probability, the second is an actual *multivariate* probability distribution. Therefore, with our approach, we allow to construct two different models of the LID space

using the same input vector. The two models generated represent different sources of information regarding the position of the examples in the feature space. In this way, we "feed two birds with one seed": we explore two, complementary, probability distributions using one single descriptor.

Moreover, even if p_{bow}, p_{fv} and p_{Mmeda} are all multi-dimensional approximations of the LID distribution, p_{bow}, p_{fv} represent an estimation of the *real* joint probability, because they do not assume component independence. On the other hand, p_{Mmeda} is a k-d probability inferred from the set of k monodimensional probabilities in p_{meda}, assuming, as in MEDA, that the LID components are independent. We can therefore say that $p_{bow}, p_{fv} \neq p_{Mmeda}$. MultiMEDA and BoW allow to learn the LID space with different, complementary approaches. We will verify such complementarity in our experimental results.

4.4.2 MEDA from a Kernel Perspective

In order to understand the Multi-MEDA approach, we detail in this Section the kernel perspective of the MEDA signatures, namely how the kernel function is formulated when evaluating MEDA vectors.

What does "kernelized" mean? In MMIR frameworks, kernel machines are used at Level 3 to learn the input space using as input visual descriptors such as MEDA, SM, etc. (see Chapter 5 for further details). In the learning phase, the machine learns how to separate the feature space into two classes.

In order to do so, kernels are used to evaluate similarities between such features and define an optimal decision boundary, namely a hyperplane in the feature space.

Among the many kernel functions used to model the feature space (e.g. chi-square, polynomial), the Radial Basis Function (RBF) kernel has been shown to perform well for image retrieval applications [210].

For two input vectors v and w, the RBF kernel has equation

$$k(v, w) = \exp(-\lambda||v - w||^2).$$

When MEDA is used in conjunction with a RBF-based classifier, the kernel function evaluates the differences between the letters frequencies for each pair of training images I and J. In order to show this behavior, We will use the following notation:

- for image I, the LIDs are in the set $x^{(I)} = \{x_j^i\}$ and the MEDA signature is $v = \{v_{j,b}\}$
- for image J, the LIDs are in the set $y^{(J)} = \{y_j^i\}$ and the MEDA signature is $w = \{w_{j,b}\}$

In order to understand the kernel view of MEDA, in Figure 4.8, we propose a 2-d representation (namely a scenario where the LID has dimension $k = 2$) of the MEDA-based feature space. The MEDA vector in Eq. 4.3 can be seen as a set of k n-dimensional vectors $\{v_j\} = \{p(x_{j,1}), \ldots, p(x_{j,n})\}$, $j = 1, \ldots, n$. Each v_j represents

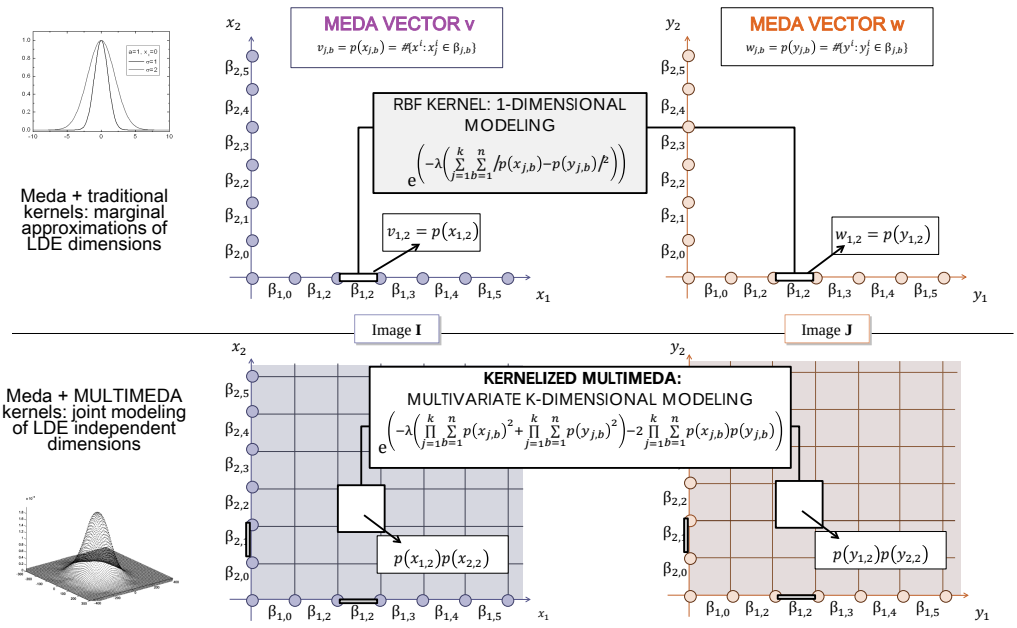


Figure 4.8: Placing kernels on top of MEDA: marginal (RBF/traditional kernels) and multidimensional (Multi-MEDA kernel) approaches

the approximation of the marginal $p(x_j)$ of the jth component of the LEDs in image I. In the 2-dimensional case, the kernel function of MEDA signatures becomes:

$$\begin{aligned} k(v,w) &= \exp(-\lambda(\sum_{b=1}^{n}|v_{1,b}-w_{1,b}|^2+\sum_{b=1}^{n}|v_{2,b}-w_{2,b}|^2) \\ &= \exp(-\lambda(\sum_{b=1}^{n}|p(x_{1,b})-p(y_{1,b})|^2+ \\ &\quad +\sum_{b=1}^{n}|p(x_{2,b})-p(y_{2,b})|^2)), \end{aligned} \tag{4.4}$$

i.e. for each dimension j, the sum over n bins of the squared differences between the signature values at each bin b.
It is therefore straight-forward to extend such kernel view and consider the real case, i.e. when $k >> 2$. In this scenario, the kernel evaluates the marginal contribution of all dimensions ($j = 1, \ldots, k$) and the previous equation becomes:

$$k(v,w) = \exp(-\lambda(\sum_{j=1}^{k}\sum_{b=1}^{n}|p(x_{j,b})-p(y_{j,b})|^2)) \tag{4.5}$$

As confirmed by the summation of Eq. (4.5) the current formulation of the

MEDA signature analyzes the marginal distribution of each dimension of the LID independently, without taking into account the interactions between the components in the k-dimensional space.

4.4.3 Kernelized Multi-MEDA: Multidimensional Probability Estimation From Marginals

As explained before, the MEDA modeling generates a 1-dimensional probability, while a model based on LID k-dimensional vectors should exploit a multivariate probability. In the Multi-MEDA model, we derive a k-dimensional probability from the marginal (1-d) probabilities computed for each dimension of the LID. Since the computation of such signature would result in an extremely high-dimensional vector, we shift the multidimensional modeling at a kernel level, embedding the k-d evaluation of MEDA in a RBF kernel. In this Section, we start from the MEDA formulation and show its multi-dimensional extension, that we then kernelize to build the Multi-MEDA model.

Recall that MEDA vector in Eq. 4.3 can be seen as a set of k n-dimensional vectors $\{v_{j,\cdot}\} = \{p(x_{j,1}^{(I)}), \ldots, p(x_{j,n}^{(I)})\}$, $j = 1, \ldots, n$. Each $v_{j,\cdot}$ representing the approximation of the marginal $p(x_j^{(I)})$ of the jth component of the LIDs in image I.

Having a MEDA vector for each image, we want now to derive a joint probability by exploiting the combination of the occurrences of all the dimensions. Since MEDA analyzes each dimension independently, in order to estimate the joint probability, we can multiply the contribution of the marginals of all components.[3]

For image I, this would result in a k-dimensional vector determined by the k-fold Cartesian product of all vectors v_j, $\forall j$. The model codebook would be the Cartesian product of all the k scalar alphabets, namely the set of hypercubes:

$$\begin{aligned} c_{(1,2,\ldots,k)} &= c_1 \times c_2 \times \ldots, \times c_k = \\ &= \{(\beta_{1,1}, \ldots, \beta_{k,1}), \ldots, (\beta_{1,n}, \ldots, \beta_{k,n})\} = \\ . &= \{(\beta_{1,b}, \beta_{2,d}, \ldots, \beta_{k,e}\},\ b, d, e = 1, \ldots, n \end{aligned}$$

Each value of the n^k-dimensional signature would be therefore the product of the occurrences of the unidimensional bins that concur in generating each hypercube:

$$v_{(1,b),(2,d),\ldots,(k,e)} = p(x_{1,b}) \cdot p(x_{2,d}) \cdot \ldots \cdot p(x_{k,e}). \tag{4.6}$$

The number of hypercubes to consider in such multidimensional formulation of the MEDA signature is exponential with the number of dimensions of the LID, which is typically 128 for traditional SIFT [105] vectors or 36 for PCA-SIFT [87]. Treating such high-dimensional feature, even with a small number of training samples, becomes impractical with traditional kernel machines. This motivates us to shift

[3] ($P(A, B) = P(A) \cdot P(B)$ if A, B are independent, see Eq. 4.1)

this multivariate probability computation inside an RBF-like kernel, and create the Multi-MEDA kernel.

As proposed for the previous analysis, we start with the 2-d example ($k = 2$, see Figure 4.8) and we then extend it to the more realistic k-d case.

When we want to take the Cartesian product of the marginals (as in Eq. (4.6) when $k = 2$) inside an RBF-like kernel, for the two images I and J the formulation in Eq. 4.5 becomes

$$k(v,w) = \exp(-\lambda(\sum_{b=1,c=1}^{n} |p(x_{1,b}) \cdot p(x_{2,c}) - p(y_{1,b}) \cdot p(y_{2,c})|^2)) \tag{4.7}$$

Developing the power in Eq. (4.7), we obtain:

$$\begin{aligned} k(v,w) &= \exp(-\lambda(\sum_{b=1,c=1}^{n} p(x_{1,b})^2 \cdot p(x_{2,c})^2 + p(y_{1,b})^2 \cdot p(y_{2,c})^2 - \\ &\quad -2(p(x_{1,b}) \cdot (y_{1,b}) \cdot (x_{2,b}) \cdot (y_{2,c})))) \\ &= \exp(-\lambda(\sum_b p(x_{1,b})^2 \sum_c p(x_{2,c})^2 + \sum_b p(y_{1,b})^2 \sum_c p(y_{1,c})^2 \\ &\quad -2 \sum_b p(x_{1,b}) p(y_{1,b}) \sum_c p(x_{2,c}) p(y_{2,c}))). \end{aligned} \tag{4.8}$$

The trick that allows us to compute Multi-MEDA in a linear time is that, when extending Eq. 4.8 to the k-dimensional space, the squares of the MEDA elements are multiplied over all dimensions independently, and the previous Equation becomes:

$$\begin{aligned} k(v,w) &= \exp(-\lambda(\prod_{j=1}^{k} \sum_{b=1}^{n} p(x_{j,b})^2 + \prod_{j=1}^{k} \sum_{b=1}^{n} p(y_{j,b})^2 - \\ &\quad -2 \prod_{j=1}^{k} \sum_{b=1}^{n} p(x_{j,b}) p(y_{j,b}))), \end{aligned} \tag{4.9}$$

which has a complexity linear in k and n $O(kn)$.

This allows us to use directly the original MEDA vectors as input to the kernel-based classifier, without pre-computing the dictionary hypercubes and the multidimensional MEDA (Eq.(4.6)). Moreover, unlike [185], this product-based formulation allows us to increase both the number of letters in the 1-d alphabets and the LID dimension without exponential increase of computation.

4.4.4 Experimental Validation

We test the effectiveness of the kernelized Multi-MEDA on two semantic analysis tasks, namely scene categorization and video retrieval. We compute MEDA (learnt

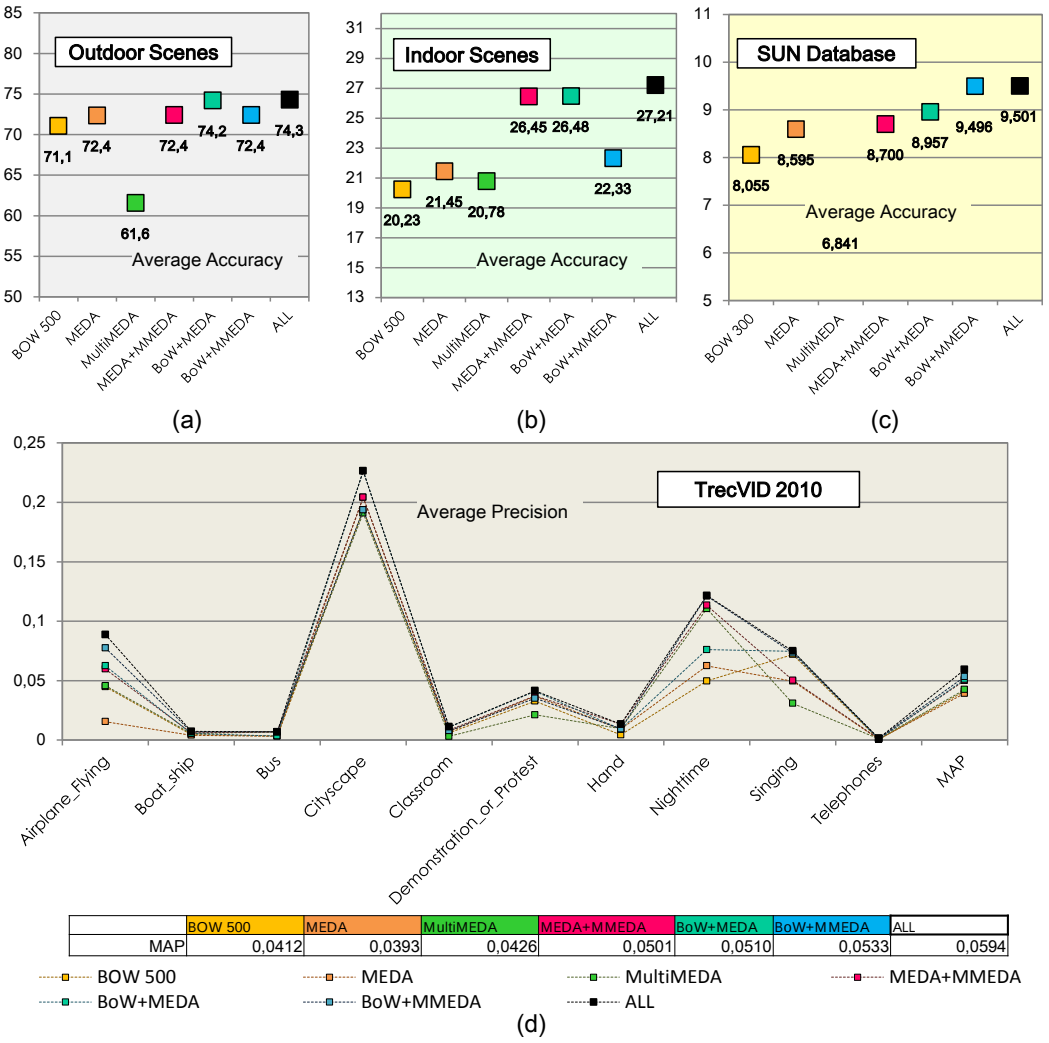


Figure 4.9: Comparison of performances of MMIR systems based on MEDA, MultiMEDA, BOW and their combination for (a,b,c) indoor, outdoor and large scale scene recognition and (d) video retrieval

with traditional kernels), BoW and Multi-MEDA models on the input images and we compare their performances. We use them as stand-alone descriptors and we then analyze the effects of their combinations, on the two given tasks. In this Section, we show that our proposed multidimensional modeling achieves good performances in both the mentioned tasks, comparable with both MEDA and BoW. Moreover, when we combine Multi-MEDA with the other LID aggregators, we show that it actually provides complementary information, as hypothesized in Sec.4.4.1, bringing a significant improvement in our experimental results.

MEDA, BoW, and Multi-MEDA share the same input seed. Therefore, the first step of our experiments is the extraction of a set of SIFT keypoints. We choose

for this purpose to extract SIFT-HLD keypoints as defined in the baselines. We then aggregate them using both BoW, by clustering a subset of training images using a standard k-means algorithm, and MEDA models (using the percentile-based technique proposed in [146]). We then learn a model based on each signature using SVMs with chi-square kernels.

In order to compute the kernelized the MULTI-MEDA, we use the MEDA signatures as input for our RBF-based multidimensional kernel in Eq. (4.9). One major issue is that the MEDA values are not normalized. Therefore, the products over k dimensions in Eq. (4.9) result in very high values. This values become the negative exponent of the RBF kernel, and $k(v, w)$ becomes close to zero. The similarity between the two vectors cannot be estimated reliably without normalization. In order to cope with this issue, we normalize the MEDA signature by m/n inside the kernel formulation. This is because each element in the MEDA vector represents a fraction (approximately $1/n$) of the total number of vectors (m), namely the one that take a given value in a given dimension. Moreover, instead of taking the product of such small values, that would bring the exponent to zero, we compute the sum of the log of those terms. Equation (4.9) becomes therefore:

$$\begin{aligned} k(v,w) = \exp(-\lambda(\sum_{j=1}^{k} log((\frac{n}{m})^2 \sum_{b=1}^{n} p(x_{j,b})^2) + \\ + \sum_{j=1}^{k} log((\frac{n}{m})^2 \sum_{b=1}^{n} p(y_{j,b})^2) \\ -2\sum_{j=1}^{k} log((\frac{n}{m})^2 \sum_{b=1}^{n} p(x_{j,b})p(y_{j,b})))). \end{aligned} \tag{4.10}$$

In the following experiments, other parameters or vector quantization models can be used, but given the statistical difference between the three approaches, the performances. of the stand-alone models and their combined contributions would not change significantly.

4.4.4.1 Scene Categorization

For the task of scene categorization, we choose three different datasets, namely the Indoor scenes database [143], the outdoor scenes database [133] and the SUN database [205] For every database, for every feature we select an experimental setup similar to the corresponding baseline , and we look at the experimental results.

Experimental Setup

For the two small-scale databases, **Indoor and Outdoor scenes**, we compute quantile-based MEDA and BoW signatures as in Sec. 4.3.3.

For the large-scale **SUN** database for scene understanding, we compute MEDA with 20 uniform bins, and BOW with 300 visual words.

We then use the MEDA vectors as input for 1-vs-all SVM with Multi-MEDA kernel to compare and combine the performances of the three descriptors. The predictions resulting from the models based on each of the three features are then combined with weighted linear fusion.

Experimental Results

Results on scene recognition show that actually the Multi-MEDA kernel models the LID space in a meaningful and effective way: Multi-MEDA achieves substantially good results for scene categorization.

Moreover, we can see here some evidences of the complementarity of the MEDA and MultiMEDA approaches, and their effectiveness for semantic analysis: the combination of MEDA and Multi-MEDA gives an improvement of about 2% for Outdoor Scenes, 30 % for Indoor Scenes, and 8 % for the SUN database, compared to BoW-only based classification, without involving any clustering, parameter tuning, or operation in the high-dimensional space.

Finally, we can observe that MEDA, MultiMEDA and BoW are mutually complementary, by looking at the performances of the three descriptors combined together: +3% for Outdoor Scenes, +34 % for Indoor Scenes, and 17 % for the SUN database, compared to BoW-only classification.

4.4.5 Video Retrieval

We use the TrecVID 2010 dataset to test the effectiveness of our proposed approach in a video retrieval task. In particular, we focus on the challenging Light semantic Indexing Task (SIN), of TrecVID [168] 2010.

In our framework, we extract 128-length SIFT features extracted from interest points based on Harris Laplace point detector (SIFT-HLD). From such points we extract the following signatures:

- BoW with 500 words
- MEDA with a number of bins per dimension that have been adapted for each concept (typically 10), as in Sec. 4.3.3

We learn models based on BoW and MEDA descriptors using a chi-square kernel. We then apply the Multi-MEDA kernel on top of the MEDA signatures and compare results with Mean Average Precision. For all the features, we use the same experimental setup as our baselines.

Results in Fig. 4.9 (d) shows that the kernelized solution that we propose in this Chapter is a good source of information for semantic MMIR. Multi-MEDA, as a stand-alone model, brings an improvement of around 13% to both traditional MEDA and BoW models. The concepts for which MEDA was not performing as good as BoW (e.g. Bus, Telephones, Airplane_Flying) benefit from the multidimensional modeling in the learning phase.

In the TrecVID results we can also clearly notice the complementarity of the kernelized multidimensional modeling that we propose in this Chapter with respect to the existing approaches. As a matter of fact, the combination of just two out of the three models considered for this task gives an average improvement of 30% compared to using the traditional BoW model only. Moreover, when we fuse the contribution of MEDA, Multi-MEDA and BoW together we obtain a prediction on the test set that is 50% more precise compared to traditional aggregators alone.

4.5 Multivariate Modeling of Marginals: Copula Signatures

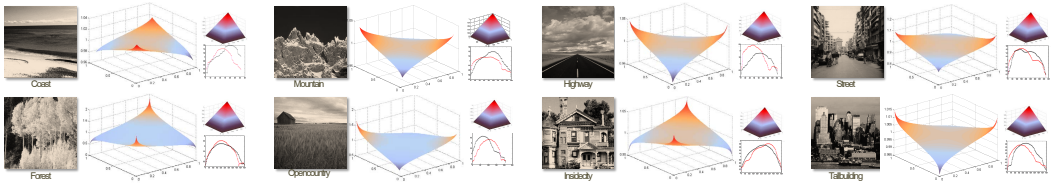


Figure 4.10: The shape, for different classes, of Gaussian Copula PDF (big plot), CDF and marginals (small plot) arising from the first two dimensions, , i.e. the most informative, of the set of PCA-SIFT [87] extracted from the image.

Despite the improvements brought by MultiMEDA towards the complete multivariate modeling of the LIDs probability based on marginal analysis, one major issue with our new kernel is that MultiMEDA stems from the assumption of component independence. However, we know that the components of traditional LIDs are actually correlated and the main aim is to model the real multivariate distribution. Our idea is to build a LID-based feature vector based on MEDA that can compensate this loss of information and finally reconstructs the LIDs joint density.

How can we model a multivariate distribution given pure marginal information only? We find here a solution to our problem in the **Copula theory**. This theory [166] tells us that *marginals can actually play an important role* in multivariate modeling. According to Copula theory, the PDF of a k-dimensional random vector x can be decomposed into k marginal distributions and one Copula function. While the **marginals** describe the *probability of each variable of the random vector*, the **Copula function** represents the *dependencies between the marginals*, and defines the probability of the vector by mapping the marginal PDF of the variables to their joint PDF. Such mapping is either pre-defined or calculated based on the marginal values, without therefore involving computationally expensive multidimensional searches. For this reason, Copulae are generally employed as efficient tools for multivariate modeling, and widely adopted in financial and medical data analysis. Here, we apply Copulae to MMIR and LID-based analysis. The main intuition is that, for an image I, we can fit a Copula with the marginals of the LIDs in I, and

then describe I according to the resulting PDF shape. Following Copula Theory, in order to build such representation, we should study separately the LIDs marginals and their dependencies.

Given these observations, we present **COMS** (COpulae and Marginals Signature): a Copula-inspired *extension of MEDA* that, by using Copulae, allows for efficient multivariate analysis of image LIDs using pure marginal information. COMS combines the MEDA vector with its complementary feature, that we name **CoMEDA** - Copula over MEDA. While MEDA models the pure monovariate information of the marginal *distributions*, CoMEDA represents the Copula structure: the marginal *dependencies*, namely the mapping between the image LIDs marginal values and the image LIDs joint density. The resulting COMS feature (MEDA+CoMEDA) reflects directly the PDF of the LIDs in an image, without involving the estimation of a global k-dimensional LID model such as visual codebooks. COMS is therefore much more discriminative and much faster in the training phase compared to both Fisher Vectors and BoW.

In the following, we will first give a general overview of our approach and detail its statistical peculiarities compared to traditional feature aggregators, see Sec. 4.5.1. We will then give some notions on Copula theory in Sec. 4.5.2, and then apply it to LID aggregation in Sec. 4.5.3. Finally, we will look at the performances of MMIR systems for semantic analysis based on COMS, and compare it with MEDA, BoW and Fisher Vectors in Sec. 4.5.4

4.5.1 COpulae and Marginal Signatures: an Overview

We explain in the following the general approach of COMS and its peculiar way of aggregating local image descriptors, different from the majority of the existing approaches.

How do we model the Copula-based feature? In our approach, we focus on a particular type of Copula, the Gaussian Copula C_Σ . This function describes the CDF (Cumulative Distribution Function[4]) of a random vector through the shape of a multivariate Gaussian CDF with the following properties:

1. Its variables are the Gaussian inverse of the marginals of the vector p_j^{-1}
2. its covariance matrix is the correlation matrix between the marginal inverses,
3. its mean is zero.

The Gaussian Copula function depends on one parameter only, namely its covariance/correlation matrix, corresponding to the dependencies between the inverse of the marginals.

We therefore fit a Gaussian Copula with the image LIDs, and store in CoMEDA the values of correlation coefficients of the marginal inverses in Σ directly, giving

[4]As we will see later, the Copula-based PDF is easily inferable from the equation of a Copula CDF

$p_{CoMeda}(x^{(I)}) = corr(p_1^{-1}(x_1^{(I)}|c_{meda}^1), \ldots, p_k^{-1}(x_k^{(I)}|c_{meda}^k))$. By doing so, we represent in a single feature the marginal dependencies determining the Copula structure. We then match the CoMEDA features using traditional kernel machines such as Support Vector Machines.

Despite the accuracy of CoMEDA as a stand-alone descriptor for MMIR, we know from Copula Theory that we can achieve a complete representation of the image PDF only when we combine the marginals and their Copula together: CoMEDA represents the multivariate complement of the monovariate MEDA vector, since it stores the dependencies between the marginal distributions. We therefore concatenate MEDA and CoMEDA in a single, very discriminative, Copula-inspired image descriptor, COMS, which we then use as input for the learning system. Therefore in COMS, MEDA+CoMEDA, namely the union of the two fundamental element of the LIDs density according to Copula theory, we model a complete Copula-based distribution, $p_{COMS}(x^{(I)}) = C_\Sigma(p_1(x_1^{(I)}|c_{meda}^1), \ldots, p_k(x_k^{(I)}|c_{meda}^k))$.

COMS is statistically different from the space determined by BoW, MEDA and Fisher Vectors (see Fig. 4.4 (c)). First, $p_{COMS} \neq p_{meda}, p_{Mmeda}$ because we are not analyzing the independent marginal behavior, but we are instead trying to estimate the multivariate density of the image LIDs through Copulae. Moreover, COMS. despite the underlying marginal analysis, does not assume independence between the LID components, but models instead a real joint PDF based on the marginal dependencies.

Despite its multivariate nature, we can also say that $p_{COMS} \neq p_{bow}, p_{fv}$ because, while BoW approximates the joint LIDs distribution through Vector Quantization given a global codebook, and while Fisher Vectors store the results of parameter adaptation for GMM fitting, COMS *directly* stores the parameter of the image joint PDF, leading to a more informative image feature modeling the real joint PDF based on the marginal dependencies. Since both MEDA and CoMEDA arise directly from the analysis of the image LIDs marginals, COMS does not require to build a universal model using unsupervised search on a training set in the k-dimensional space such as GMM, k-means or hypercube exploration to define the global LID density, saving a lot of computational time on training.

4.5.2 Copulae: Linking Marginals with Joint Distributions

In this Section, we give an overview of the Copula theory, highlighting the notions that we retain more useful for our proposed approach.

Given a 2-dimensional random vector $x = \{x_1, x_2\}$, we define $u = \Pi_1(x_1) = [P(x_1 \leq X_1)]$, $v = \Pi_2(x_2) = [P(x_2 \leq X_2)]$ as the marginal *cumulative* distribution functions (CDFs) of x_1 and x_2 respectively, and $\Pi(x_1, x_2) = P[x_1 \leq X_1, x_2 \leq X_2]$ as the vector cumulative joint distribution. For ease of understanding, we first consider a bivariate case.

In order to be defined as a two-dimensional Copula, C needs to fulfill the following requirements (see [127]):

- It is defined over the interval $[0, 1]$

- $\forall t \in [0,1]$, then $C(t,0) = C(0,t) = 0$ and $C(t,1) = C(1,t) = 1$
- $\forall u_1, u_2, v_1, v_2 \in [0,1]$, with $u_1 \leq u_2$ and $v_1 \leq v_2$, $C(u_2, v_2) - C(u_1, v_2) - C(u_2, v_1) + C(u_1, v_1) \geq 0$

As said, a Copula C, is defined as a unique mapping that assigns the joint CDF of X given each ordered pair of values of its marginals, namely:

$$\Pi(x_1, x_2) = C(\Pi_1(x_1), \Pi_2(x_2)) = C(u, v),$$

and, following Sklar's theorem and assuming that Π_1, Π_2 are continuous:

$$C(u, v) = \Pi(x_1, x_2) = \pi(\pi_1^{-1}(u), \pi_2^{-1}(v)). \tag{4.11}$$

Where π is a given multivariate distribution function, and $\pi_{\cdot}^{-1}(cdot)$ is the inverse of the corresponding univariate distribution. Eq. 4.11 allows to construct a Copula from a given multivariate distribution function π, that in our case will be the Gaussian distribution.

The Copula function by itself describes the vector CDF. However, we might want to represent the vector in terms of probability density function (PDF), i.e. $p(x) = p(x_1, x_2) = P[x_1 = X_1, x_2 = X_2]$. In order to obtain $p(x_1, x_2)$ we have to compute *copula density*, namely the CDF derivative, i.e., following Eq. (4.11) :

$$p(x_1, x_2) = \frac{\delta^2 C(u, v)}{\delta u, \delta v} = \frac{p^*(\pi^{-1}(u), \pi^{-1}(v))}{p^*(\pi^{-1}(u)), p^*(\pi^{-1}(v))},$$

where p^* is the PDF corresponding to π, e.g. Gaussian CDF π, Gaussian PDF p^*.

The Copula describes therefore the dependence between the components of a random vector, no matter the function describing their marginal distributions: if we know the mapping C, the joint density $p(x_1, x_2)$ can be inferred from the marginal CDFs u and v.

4.5.2.1 Gaussian Copulae

A particular type of Copulae is the Gaussian Copula, which belongs to the class of Elliptical Copulae (i.e. Copulae following Elliptical distributions such as Laplacian, T-Student, etc..). The Gaussian Copula structure is a multivariate normal distribution: in this model, π corresponds to the multivariate Gaussian CDF, while π^{-1} corresponds to the inverse of the univariate normal CDF.

A Gaussian Copula C_Σ is then defined for the two-dimensional random vector x as (following Eq. (4.11)):

$$C_\Sigma(u, v) = \theta_\Sigma(\theta^{-1}(u), \theta^{-1}(v)), \tag{4.12}$$

being $\theta^{-1}(\cdot)$ the inverse of the univariate normal CDF, and θ_Σ the bivariate (or

multivariate, when $s > 2$) standard with mean zero and covariance Σ, giving

$$C_\Sigma = \frac{1}{\sqrt{det(\Sigma)}} exp\left(-\frac{1}{2} \cdot \begin{pmatrix} \theta^{-1}(u) \\ \theta^{-1}(v) \end{pmatrix}^T \Sigma^{-1} \mathbf{I} \begin{pmatrix} \theta^{-1}(u) \\ \theta^{-1}(v) \end{pmatrix} \right), \tag{4.13}$$

How to find the covariance matrix Σ? When dealing with normal distributions, the correlation values between two variables fully define their dependencies. In Gaussian Copulae, Σ corresponds therefore to the correlation matrix between the inverse standard univariate normal CDF

$$\Sigma(\theta^{-1}(u), \theta^{-1}(v)) = \frac{cov(\theta^{-1}(u), \theta^{-1}(v))}{\sigma(\theta^{-1}(u))\sigma(\theta^{-1}(v))} \tag{4.14}$$

4.5.2.2 Why Gaussian Copulae?

As said, the Gaussian Copula function arises from pure marginal analysis: both the variables (inverse normal of marginal CDFs) and the parameter (correlation between the inverse marginal CDFs in Eq. (4.14) are constructed by manipulating the marginal distributions with simple operations ($O(k)$ for $\theta^{-1}(\cdot)$, and $O(k^2)$ for Σ). Gaussian Copulae represent therefore an efficient way to estimate the joint PDF of vectors with the following properties

- they have a small dimensionality, namely a low value of s and
- they have marginals that can be easily modeled.

In fact, local image descriptors satisfy conditions (1) and (2). The dimensionality of LIDs is generally $k \leq 128$. Moreover, it exists a descriptor for LID marginal approximation, MEDA, which have been proved to effectively model the univariate distributions of the LID components. Gaussian Copulae can be therefore very efficient tools to estimate the joint PDF of LIDs.

Moreover, a Gaussian Copula C_Σ depends on one parameter only, namely the covariance-correlation matrix Σ, whose computational time that is quadratic with s, making it easy to characterize an image through its Copula shape. Furthermore, various fast implementations are available to easily and quickly treat with multivariate normal densities, due to their popularity, making the computation of this Copula very easy. This motivates us to use Gaussian Copulae to efficiently and effectively approximate the distributions of the LIDs in an image and generate an image signature out of it.

4.5.3 COMS: Multivariate LID Analysis from Marginal Values

In this Section, we show how to exploit Copulae Theory to aggregate LIDs and build effective and efficient compact image signatures based on local descriptors.

In order to perform LID-based analysis, for each image I, we first extract m salient points and describe them using a k-dimensional normalized SIFT [105] descriptor $x^{(I)} = (x_1^i, \ldots, x_k^i)$, $i = 1, \ldots, m$. Recall that, for an image I, we define

$p_j(x_j^{(I)})$, $j = 1, \ldots, s$ as the marginal distribution of the j_{th} component of the image LIDs, and $p(x^{(I)})$ as their joint density.

The main idea is that, similar to Copula Theory, we can approximate $p^{(I)}(x)$ for an image I by extracting:

A its set of marginals $p_j(x_j^{(I)})$ and

B a Gaussian Copula Function

and use it as a discriminative image signature for MMIR purposes.

While it already exists a feature (A) approximating the marginals (i.e. MEDA), we are missing (B) a feature to represent the Copula structure. We therefore design CoMEDA for this purpose (See Fig. 4.11 for a visual explanation of our approach).

Therefore, we first (A) extract from image I the MEDA vector $v^{(I)}$ containing the LIDs marginals approximations.

We then (B) use them, as we will show in Sec.4.5.3.1, to fit an image-specific Gaussian Copula $C_\Sigma^{(I)}$, that defines an approximation $p_C(x^{(I)})$ of the joint distribution of the image LIDs. We characterize the image I with the Copula structure of its LIDs by storing in the CoMEDA feature the values of the image-specific covariance matrix $\Sigma^{(I)}$, namely the unique parameter of the resulting Copula-based PDF.

Finally, we achieve a complete model of the LID density by combining the CoMEDA feature of an image I with its marginal counterpart, i.e. the MEDA vector for image I, into a final image signature, namely COMS.

4.5.3.1 Fitting a Copula with the Image LIDs

Once we have extracted the marginal information from the Image LIDs, we can then use it to calculate the corresponding Gaussian Copula. This will allow us to characterize each image with the distribution of its LIDs (using the parameters of the Copula-based density as signature). First, for each dimension of the LID, for each of the k marginals $\tilde{p_j}(x_j^{(I)})$ that we obtain with the MEDA histogramming[5], we compute the corresponding k univariate CDFs $u^{(I)}(1) = \pi_j(x_1^{(I)}), \ldots u^{(I)}(k) = \pi_k(x_k^{(I)})$, normalized in the interval $[0, 1]$. According to the Gaussian Copula theory, we then compute the inverse of the normal CDF, namely

$$\theta^{-1}(u^{(I)}(1)), \ldots, \theta^{-1}(u^{(I)}(k)). \tag{4.15}$$

If we now want to define a Gaussian Copula $C_\Sigma^{(I)}$ representing the CDF of the LIDs for image I, we should extend the multivariate Gaussian in Eq. (4.12), for SIFT vector analysis with $k \gg 2$, giving, for image I,

$$C_\Sigma^{(I)}(u^{(I)}(1), \ldots, u^{(I)}(k)) = \theta_\Sigma^{(I)}(\theta^{-1}(u^{(I)}(1)), \ldots, \theta^{-1}(u^{(I)}(k))). \tag{4.16}$$

[5]In practice, we will use for our experiments a more refined way to estimate the marginal distribution shape, namely a kernel density estimator [165]

and from the Copula theory, we know that $\Sigma^{(I)}$ can be computed as the correlation matrix between the inverse of the LID marginals, namely:

$$\Sigma^{(I)}(a,b) = \frac{cov(\theta^{-1}(u^{(I)}(a)), \theta^{-1}(u^{(I)}(b)))}{\sigma(\theta^{-1}(u^{(I)}(a)))\sigma(\theta^{-1}(u^{(I)}(b)))} \tag{4.17}$$

where $a, b = 1, \ldots, j$, $cov(\cdot, *)$ corresponds to the covariance between $(\cdot)$ and $(*)$, and $\sigma(\cdot)$ is the standard deviation of variable $(\cdot)$.

4.5.3.2 The CoMEDA Vector

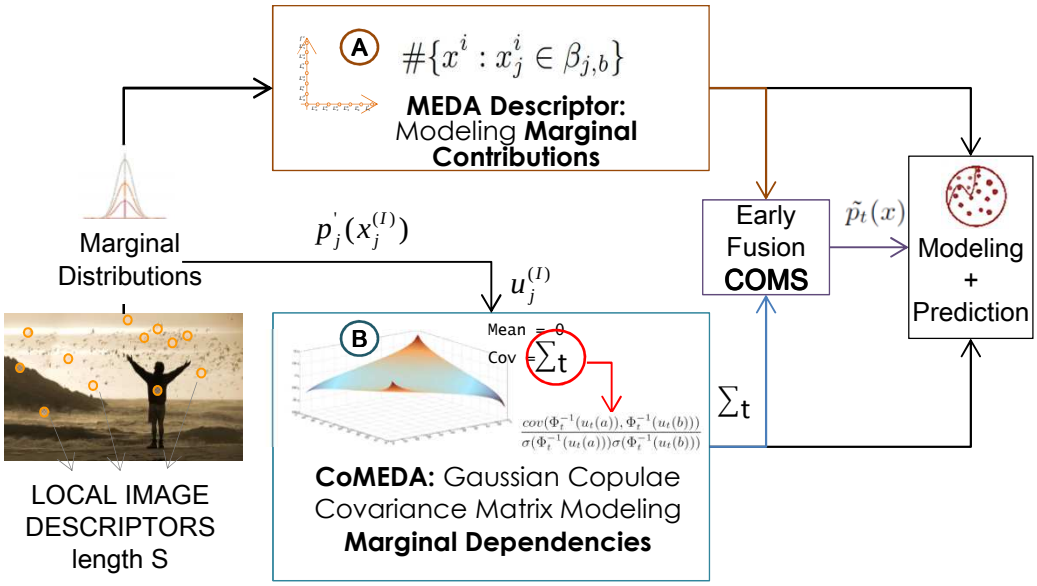


Figure 4.11: Our Copula-based analysis on the local image descriptors.

How can we capture the behavior of the Copula structure we just described, and store it into a single, effective feature? As we can observe, Eq. (4.16), has only one parameter, the covariance matrix $\Sigma^{(I)}$. Such covariance matrix describes the dependencies structure between the LIDs marginals and determines the equation of the multivariate distribution.

We therefore fill the CoMEDA vector $\mu^{(I)}$ for an Image I with the values corresponding to $\Sigma^{(I)}$, namely the correlation coefficients of the inverse marginal approximations of the LIDs in the image. The complexity of CoMEDA is quadratic with the number of dimensions of the LIDs, and its dimensionality is $\frac{k \times k}{2}$, being $\Sigma^{(I)}$ typically a symmetric matrix. CoMEDA does not imply therefore exponential computation or multidimensional vector quantization for multivariate LID representation. This low dimensional feature (we will select $k = 36$) can be easily then used as input for discriminative classifiers, that will learn a model of the LIDs space based on the CoMEDA feature representation.

4.5.3.3 COMS: MEDA + CoMEDA

CoMEDA gathers the main element of the Copula structure: it is the representation of the LIDs multidimensional information arising from the dependencies between marginal distributions.

However, we can observe that the shape of Eq. (4.12) is determined both by $\Sigma^{(I)}$ and by the behavior of the LIDs marginal distributions, specific of the image I. Recall that, as a matter of fact, Copula theory states that the joint distribution of a random vector can be represented by its marginal distributions and a multivariate Copula structure. This suggests us that, in order to have a complete representation of the LID space, we should combine the CoMEDA feature of image I with a descriptor approximating the marginal behavior of I, e.g. MEDA. Therefore, for each image, we concatenate these two types of information regarding the LID distribution, MEDA and CoMEDA, both very discriminative features, into a single image descriptor COMS $h^{(I)} = \{v^{(I)}, \mu^{(I)}\}$. By doing so, we enrich the representation of the LID space, and determine a good approximation of the LID joint distribution.

4.5.4 Experimental Validation

In this Section we will show the performances of our Copula-based approach, comparing it with the most effective LID aggregators available in literature.

We test the effectiveness of our approach for two, challenging Multimedia Retrieval tasks, namely video retrieval and scene recognition. Since all the descriptors work over the same input, namely local image descriptors, the first step of our experiments is to compute the image LIDs.

Since we want to keep the dimensionality low, from all the images/keyframes in our datasets we compute PCA-SIFT [87] ($s = 36$) as described in Chapter 2. We then aggregate them using the following approaches for comparison (See Fig. 4.12):

1. *bow*, the Bag of Words Model computed, as in [30], through a codebook built with k-means clustering

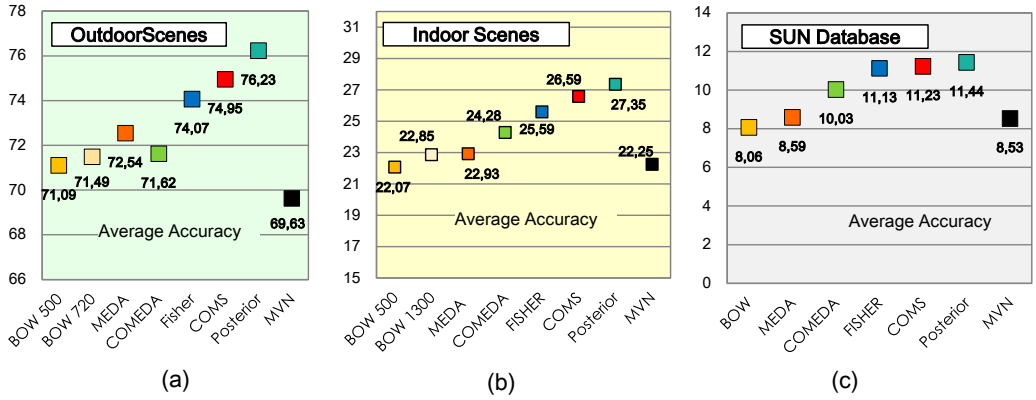
2. *Meda*, the marginal-based descriptor in [146] described in Sec. 4.3

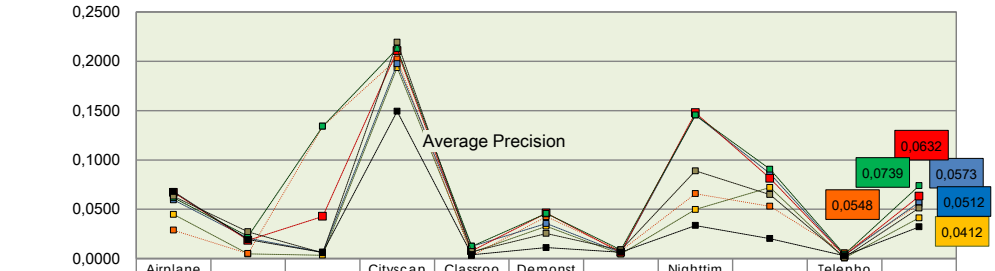
3. *Fisher*, the Fisher Vectors approach, computed using and adapting the implementation in [79]

4. *CoMeda*, our Copula-based descriptor, i.e. the values of the correlation coefficients of the inverse of the marginals

5. *COMS*, the early combination of MEDA and CoMEDA

Moreover, in order to prove the reasonableness of our Copulae-based LID processing, we compute another feature, that we call MVN (Multivariate Normal), that stores the values of the mean and covariance matrix of the image LIDs vectors (different from CoMEDA, that treats with LIDs marginals). The difference of effectiveness





	Airplane Flying	Boat ship	Bus	Cityscape	Classroom	Demonstration	Hand	Nighttime	Singing	Telephone	MAP
Bow	0,0448	0,0047	0,0034	0,1936	0,0061	0,0328	0,0044	0,0497	0,0720	0,0004	0,0412
MEDA_36	0,0288	0,0050	0,1341	0,2024	0,0029	0,0425	0,0079	0,0658	0,0530	0,0059	0,0548
Co-Meda	0,0595	0,0206	0,0065	0,1975	0,0124	0,0355	0,0061	0,1455	0,0862	0,0034	0,0573
COMS	0,0672	0,0185	0,0428	0,2108	0,0079	0,0461	0,0063	0,1479	0,0813	0,0030	0,0632
Posterior	0,0617	0,0225	0,1341	0,2129	0,0124	0,0457	0,0089	0,1455	0,0906	0,0044	0,0739
Fisher	0,0632	0,0272	0,0058	0,2194	0,0070	0,0255	0,0087	0,0890	0,0649	0,0012	0,0512
MVN	0,0679	0,0192	0,0062	0,1493	0,0038	0,0112	0,0064	0,0335	0,0202	0,0028	0,0321

(d)

Figure 4.12: Comparison of performances of MMIR systems based on MEDA, BOW and COMS for (a,b,c) indoor, outdoor and large scale scene recognition and (d) video retrieval

between COMS (or other multivariate approaches) and MVN will show the discriminative value added by treating the LIDs with models more complex than a simple multivariate Gaussian PDF.

Then, we use the computed descriptors as input to Support Vector Machines (SVM) with Chi-square kernels, to build models able to predict the image category, or the presence of a given concept (in the case of Video Retrieval).

Finally, in order to further prove the effectiveness of the combination of MEDA and CoMEDA, we combine and weigh the predictions coming from the MEDA-only model and the CoMEDA-only model, and we name this class of experiments *Posterior*, see Fig. 4.12.

For all the features and datasets, we follow the training/test set splitting and training used by our baselines.

We show that our approach outperforms the other methods in all the databases

	BOW	MEDA	Co-MEDA	Posterior	Early	Fisher	MVN
Outdoor	71,489175	71,64705	71,624425	76,2299375	74,9513375	74,0664625	69,6271
Indoor	22,85358485	22,92517727	24,2833197	27,34952576	26,58517879	25,59342576	22,245
SUN	8,0554	8,5945	10,0302	11,4408	11,2343	11,1285	8,534

Table 4.2: Seconds of computation for various LIDs aggregators on the training phase

considered for scene recognition and for video retrieval. Overall, we can say that posterior fusion of MEDA and CoMEDA is slightly more effective than *COMS*, because we add one parameter to weigh the contribution of the two descriptors. We can also observe that the simple MVN descriptor has a weaker discriminative power compared to all the other descriptors, suggesting that adding complexity in the LID modeling actually is useful for MMIR performances improvement.

Regarding computational costs, as we can see from Table. 4.2 (c), the time to compute CoMEDA, for the training set, has the same order of magnitude as the MEDA feature, because it does not require to estimate a universal model such as the BoW codebook.

4.5.4.1 Scene Recognition

In this Section we present the results of our experiments for small scale (indoor/outdoor) and large scale scene recognition. The goal for this task is to build a model able to classify test images with the correct class, selected out of a set of pre-defined mutually exclusive categories.

We achieve this goal by learning our features with a one-vs.-all multiclass SVM, and assigning the image category according to the classifier that outputs the highest score. The typical evaluation measure for this task is the average accuracy on the test set.

In the following we will see the experimental setup and results for the various datasets considered. A visual representation of the results can be found in Fig. 4.12(a-b).

Small Scale Scene Recognition

For the **Outdoor Scenes**, we compare the mentioned LID aggregators, computing them as follows:

- *Bow* with 500 and 720 visual words;
- *Meda* with quantile-based quantization, as shown in Sec. 4.3.3
- *Fisher* with 64 Gaussians in the mixture (final dimension is 2304),
- *CoMeda* (dimensionality 1296)
- *COMS*, with 1656 components,
- *MVN* with 36*36(covariance)+36(mean)=1332 dimensions

Our results (see Fig. 4.12 (a)) show that, even if *CoMeda* by itself does not outperform *Meda*, when they are combined together with early (*COMS*) and *Posterior* fusion, namely when we follow the Copula Theory approach, the resulting model is much more effective than both Bag of Words and Fisher Vectors.

For **Indoor Scenes** dataset [133], we follow a similar experimental setup. The details of the features that we compare follow:

- *bow* with 1300 visual words;
- *Meda* with percentile quantization, as proposed in [146], with 10 bins per dimension (resulting in a feature with 288 components), see Sec. 4.3.3
- *Fisher* with 32 Gaussians in the mixture (final dimension is 1152),
- *CoMeda* (dimensionality 1296),
- *COMS*, with 1656 components,
- *MVN* with 36*36(covariance)+36(mean)=1332 dimensions.

Results for indoor scenes (see Fig. 4.12 (b)) show a similar trend as the experiments on the outdoor scenes datasets. The CoMEDA feature used as a stand-alone descriptor is actually more performing (+6%) than BoW, and it is improved by its combination with the MEDA descriptor (+ 16% of *COMS* and +20% of *Posterior* over *bow*), with a great improvement, = 6% over the Fisher Vectors-based classification.

Large Scale Scene Recognition

The LIDs aggregators that we compute for the SUN database [205] are as follows:

- *bow* with 500 visual words;
- *Meda* with uniform quantization, as proposed in [146], with 10 bins per dimension (resulting in a feature with 360 components),
- *Fisher* with 32 Gaussians in the mixture (final dimension is 2304),
- *CoMeda* (dimensionality 1296),
- *COMS*, with 1584 components,
- *MVN* with 36*36(covariance)+36(mean)=1332 dimensions.

In the results for this dataset (see Fig. 4.12 (c)), we can see a homogeneous accuracy score obtained the COMS/*Posterior*/*Fisher* descriptor, all outperforming by around 40% the simpler approaches such as MEDA and BoW.

4.5.4.2 Video Retrieval

For this task, we focus on the challenging TrecVID 2010 [168] light semantic Indexing Task, comparing the results in terms of Mean Average Precision. Here, we compute the following descriptors for comparison:

- MEDA with fixed quantization (with a number of bins tuned, as in [146], for each concept),
- *bow* with 500 visual words,
- *Fisher* with 32 Gaussians in the mixture (final dimension is 2304),
- *CoMeda* (dimensionality 1296), and
- *COMS*, with 1584 components, and finally *MVN* with 36*36(covariance)+36(mean)=1332 dimensions.

As shown in Fig. 4.12(d), the effectiveness of our method is even more clear for this challenging task: while *COMS* outperforms *bow* by more than 50% and *Fisher* by 23%, the posterior fusion of MEDA and CoMEDA is further improving the performances of our proposed method for video retrieval, with an increase of around 78 % over BoW and 44% over the Fisher Vector-based retrieval.

4.6 Summary and Future Work

We presented a set of methods for local image descriptors encoding and pooling. While the majority of the existing methods for feature pooling aggregate the LIDs based on their multivariate distribution, the key aspect of all the techniques proposed in this Chapter is that the aggregation of the LIDs is based on the *approximation of their marginal distribution*. This peculiarity of our methods not only leads to very efficient and effective techniques for feature aggregation, but allows to introduce a new cue for LID-based image analysis, namely the information arising from the univariate distribution of the LID components.

The three methods correspond to three steps towards the complete modeling of the multivariate LIDs distribution based on marginal analysis: we have first built an image descriptor, that we named MEDA, that approximates the LIDs marginals by quantizing each dimension of the local descriptors. We have then built a multidimensional model by constructing a kernel over MEDA signatures that allows to generate a multivariate probability out of the MEDA descriptors by multiplying their marginal approximations. Finally, we coupled the MEDA descriptor with CO-MEDA, namely a Copula-based signature describing the link between the LID marginal and their joint distribution, leading to an image signature that fully describe the LIDs multivariate distribution.

Possible tracks for future work include the construction of MEDA, MultiMEDA and COMS signatures using as input LIDs local descriptors other than SIFT, for

example SURF [6] or HOG [31]. Another possible development for our techniques is to concentrate the marginal analysis over specific regions, in order to embed some spatial information into the LID aggregation. Finally, given the good performances of our 1-d quantization technique when combined with k-d quantization, we could explore the possibility of 2 or 3-d quantization and build multi-dimensional visual dictionaries. Similarly, we could explore the possibility of learning with MultiMEDA, the joint contributions of the independent marginals over hypercubes of dimensions $l < k$ (e.g. considering the interactions between components 2 by 2, 3 by 3 etc.), building a complete model with various level of multivariate analysis.

While we have already sufficiently extended MEDA and MultiMEDA, COMS represents one of the first attempts in literature to introduce Copulae for LID pooling. It opens therefore a wide range of possibilities for future work in this direction. For example, the COMS idea can be extended by finding more effective kernels for Copula-based signature matching, such as kernels based on Bhattacharyya distance or Kullback-Leibler divergence. Moreover, we could use different Copula structures, such as Clayton or T-student Copulae, or build an ad-hoc Copula formulation to better model the LIDs multivariate distribution.

Chapter 5

Level 2: A Multimedia Retrieval Framework Based on Automatic Relevance Judgments

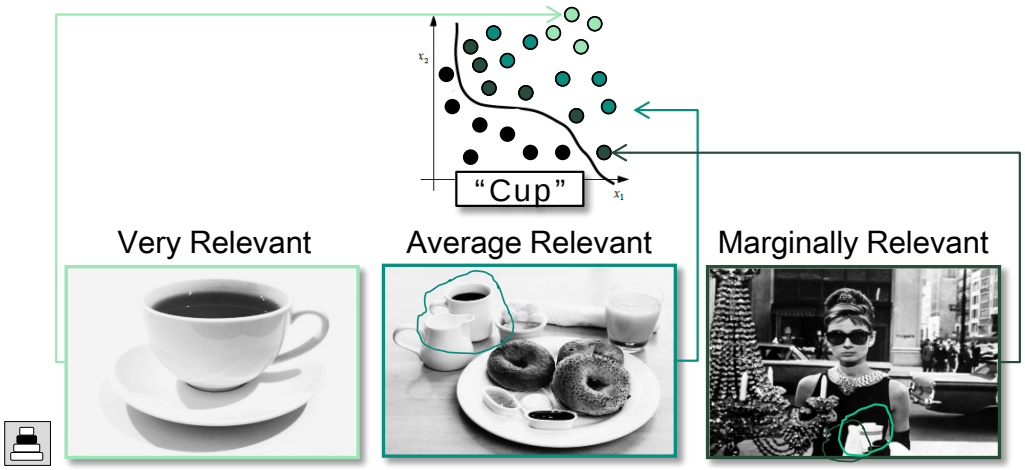


Figure 5.1: Introducing graded-relevance into multimedia retrieval.

Traditional Content Based Multimedia Retrieval (CBMR) systems measure the relevance of visual samples using a ***binary scale*** *(Relevant/Non Relevant). However, a picture can be relevant to a semantic category with different degrees, depending on the way such concept is represented in the image. In this Chapter, we build a CBMR framework that supports graded relevance judgments. In order to quickly build graded ground truths, we propose a measure to reassess binary-labeled databases without involving manual effort: we automatically assign a reliable* ***relevance degree*** *(Non, Weakly, Average, Very Relevant) to each sample, based on its position with respect to the hyperplane drawn by Support Vector Machines in the feature space. We test the effectiveness of our system on a large-scale database for video retrieval, and we show that our approach outperforms the traditional binary relevance-based frameworks for this task.*

In the previous Sections, we looked at global and pooled descriptors and applied them to MMIR for semantic analysis. Features are very important to describe in a few values the salient properties of visual content, and in order to build a complete system for visual analysis, we need to **learn** the links between the features and the semantics of the image.

In this Chapter, we aim at improving the learning level of the MMIR pyramid for semantic analysis. We will apply the same techniques for Aesthetic analysis in Chapter 5.

The core of the “intelligence" of an automatic image analysis system lies in the learning framework used for classification. Supervised learning systems such as Support Vector Machines [17], Neural Networks [173], Decision trees [158] learn, from a set of input data samples with a variety characteristics, a model able to separate data based on some common patterns. A learning system learns from **training examples**, and infers rules and partitions of the input space.

A training example is generally composed of a feature vector associated with a corresponding *annotation*, representing the *label*, or the category, associated to the input sample. When given a new sample, the learning framework will be able, given the knowledge extracted in the training phase, to predict the sample position in the input space, and therefore its category and characteristics.

Learning in Semantic Multimedia Information Retrieval

In semantic analysis systems, we generally work with supervised learning frameworks that receive as input either pixel-level features (Level 0 of the pyramid) or pooled (Level 1 of the pyramid) features from a set of training images, together with some *semantic annotations*. Such annotations are generally related to semantic properties of the images: which is the content depicted in the image? Which semantic concepts is the image *relevant* to?

In order to build discriminative models, supervised learning techniques require therefore manually-assessed ground truth annotations associated with the images in the training set. When labeling a dataset, real assessors are asked to categorize an image or a shot according to its topical relevance with respect to a given concept. The learning framework then learns how to separate the input data based on its feature and the corresponding annotations, building a model able to distinguish between images that are relevant and image that are non-relevant to a given semantic concept, and to classify (in case of image categorization) and rank (in case of retrieval) new images accordingly.

Given the features from the lower levels, at this level of the image analysis chain, the general performances of the system mainly depend on two main elements:

1. The **quality of the annotations**, namely how well they indicate the presence of the content represented in the image, i.e. the relevance of the image with respect to a given topic

2. The **generalization ability of the learning algorithm**, namely how well the framework perform accurate analysis on new, unseen examples

Our observation is that the notion of “relevance” used by general learning frameworks in MMIR might cause lack of accuracy in both the mentioned elements.

As a matter of fact, In most cases (e.g. the TrecVID collaborative annotation [5]), the notion of relevance is measured using a **binary scale**: a visual input is

either “positive” or “negative” for the concept considered. Likewise, in MMIR, we mainly use fast, effective learning systems, such as SVMs, that work on *binary separations* of the input space only, without allowing for intermediate degrees of annotations.

Our Contribution: Graded Relevance vs. Binary Relevance

One major issue regarding this approach is that this type of assessment assumes that all the relevant elements are identically relevant and that all the irrelevant samples are equally non-relevant. However, a picture can be relevant to a semantic category with different degrees (see Fig. 5.1), depending on the way such concept is represented in the image, and the binary learning can cause then inconsistencies when predicting the presence of semantic concepts of new samples, due to the variety of the forms in which a concept can appear in a picture. We would need therefore ground truth annotations reflecting *non-binary* relevance judgments in order to better represent the diversity of the semantic compositions of natural images. Moreover, we would need a learning framework that is able to deal with such multiple degrees of annotations.

A similar issue has been arisen for Web Information Retrieval in [55], where Gordon et al. exposed the need for information retrieval systems able to distinguish relevant documents from marginally relevant ones. This issue was then partially addressed by the TREC editions [65] and improved by Sorumen [171] by formalizing the concept of **degree of relevance**, namely “the potential usefulness of documents for a reader trying to learn about a topic" [171] and re-assessing the TREC corpus allowing for non-binary relevance judgements.

Given these observations, in our work we introduce the notion of **graded relevance** learning for MMIR: we re-design the concept of traditional learning for semantic analysis, by building a learning framework for image and video retrieval that supports graded relevance judgments. In order to quickly build graded ground truths, we propose a measure to reassess binary-labeled databases without involving manual effort: we automatically assign a reliable relevance degree (Non, Weakly, Average, Very Relevant) to each sample, based on its position with respect to the hyperplane drawn by Support Vector Machines in the feature space. We test the effectiveness of our system on a large-scale database, and we show that our approach outperforms the traditional binary relevance-based frameworks for video retrieval. Moreover, we show another application of our graded relevance degree assignment method, that we re-use for reducing the annotation noise in the TrecVID data.

In the remainder of this Chapter, we will first strongly motivate the need of graded relevance learning systems for image categorization and retrieval (see Sec. 5.1), and give a broad overview of our approach and its novelty. We then recall in Sec. 5.2 some principles of standard techniques for learning in visual data analysis. We will then show our graded-relevance extension in Sec. 5.3 and test its effectiveness in Sec. 5.4 for video retrieval.

5.1 Graded Relevance for Visual analysis: Motivations and Contributions

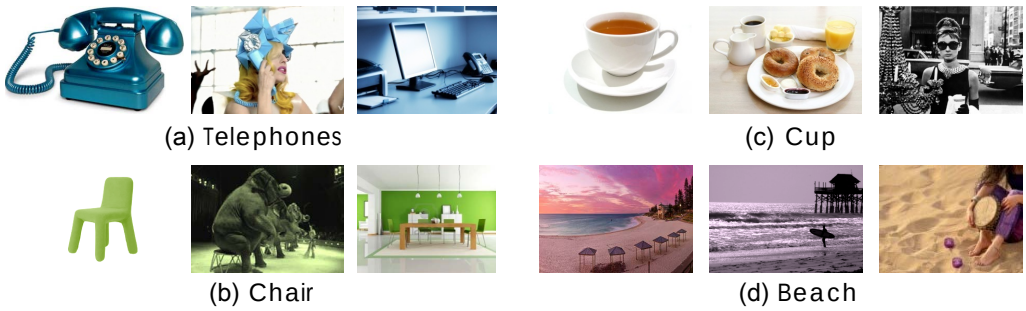


Figure 5.2: Relevance is a relative notion: images labeled as positive for (a) "telephone" (b) "chair" (c) "cup" (d) "beach" actually have different visual evidences.

In this Section, we explain in detail the reasons of choosing graded relevance assignments and learning for image and video retrieval and categorization. We will then give a brief explanation of our contribution, and showing the novelty that it brings compared to traditional systems for multimedia retrieval.

5.1.1 Why Graded Relevance?

"As all human notions, relevance is messy and not necessarily perfectly rational"[159].

Each group in Fig. 5.2 shows a set of images that would be annotated as positive for the same corresponding concept: even if we can acknowledge that all the images are relevant with respect to the group label (e.g. images in group *a* contain the concept "Telephones"), the global semantic content of each image differs. *Intuitively*, we would say that each image is relevant for the associated concept with a different degree (for example, similar to web search engines, labels or grades such as "weakly relevant" or "very relevant" could be assigned). A distribution of relevance inferences over a graded scale would reflect better the human way of understanding concepts. From a *learning system* point of view, binary judgments imply that both marginally-relevant samples and very representative samples are treated equally when modeling the concept feature space: this might cause inconsistencies in the classification process. In a *multimedia retrieval framework*, concept models might be therefore less effective due to the contrast between the intra-class diversity and the binary relevance judgment.

Relevance is a fundamental notion for information retrieval: as pointed out in [159], while traditional bibliographic and classification frameworks aim to describe/categorize samples, retrieving information involves, besides description and categorization, the need for *searching*, and "searching is about relevance". Graded

relevance-based learning methods first appeared for real Web search engines, where pages cannot be simply categorized as relative/non relative, but they need a multi-level relevance assignment. Several algorithms have been proposed to learn ranking functions from relative relevance judgments, like RankNet[179], based on neural networks, RankBoost[48], or the regression-based learning proposed in [211] by Zheng et al. How are these "grades" assigned? Generally, in traditional information retrieval such reassessment is done manually, either using real expert assessors [171], or using Amazon MechanicalTurk [170]. For web-based searches, the relevance judgment can be inferred in an automatic way, using the users' clickthroughs (see [89] for an overview of implicit relevance feedback method).

In the image analysis and video retrieval field, graded relevance has been rarely explored. As a matter of fact, traditional multimedia retrieval systems (see, for example, [153]) generally rely on binary-labeled keyframes or images. However, it was recently shown [38] that a video retrieval framework benefits from a graded-relevance annotated training set: in [38] the development set is reassessed by assigning, for each generally "relevant" frame, a degree of relevance from Somehow Relevant to Highly relevant. Three new training sets are then created based on different combinations of the relevance-based partitions.

5.1.2 Automatic Graded Relevance Assignments for Multimedia Retrieval

Our work is one of the first attempts in literature to build a complete automatic graded relevance system for Content-based retrieval.

When building a graded-relevance framework for feature learning, the first step is to reassess the training samples, labeled as positive/negative, by assigning a "degree" of relevance. Generally [171] [38], the level of relevance of each sample is labeled manually. However, when dealing with large collections of visual data, e.g. the 400 hours of training videos for TrecVID [168] 2011, such re-assessment becomes time-consuming and practically unfeasible.

We propose here an effective automatic graded-relevance based framework for image recognition and video retrieval. With our system, we can treat noisy and marginally relevant samples with less importance, achieving a better usage of our training set, thus improving the performances of traditional binary-relevance systems. Moreover, the key aspect of our framework is that, unlikely [38], *the relevance degree of a training sample is assessed automatically*: we assign to each sample a reliable and realistic relevance judgment, without involving any manual effort.

To auto-reannotate each training sample in the database according to a non-binary relevance scale, we find a measure that first assigns a fuzzy membership judgment (i.e. how much a sample is representative/positive for a given concept). The idea is to exploit the learning methods traditionally used in video retrieval frameworks: the SVMs. We are inspired by few works in machine learning literature that reassess the samples in a binary-labeled training set based on the learnt feature space, looking at the distribution of the features given the class. Generally,

they assign to the samples automatically a fuzzy membership score, namely a value representing their relevance for a given class. For example [101] defines an automatic membership measure as a function of the mean and radius of each class; this work is then extended by Lin et al in [102], that uses an heuristic strategy based on a confidence factor and a trashy factor in training data to automatically assign a score to each sample.

Given these promising works, we therefore choose in our approach to assign the image relevance degree based on the position of the sample with respect to the hyperplane drawn by a Support Vector Machine (SVM) [17] in the feature space.

The second step is the *discrete category assignment*. Based on the computed relevance score, we re-categorize the training dataset into 4 groups for every concept: Very Relevant, Average Relevant, and Weakly Relevant and Non Relevant samples. By training the system on such multiple repartitions, we then build a multi-level model for each semantic concept considered. When assigning labels to a new sample, the system outputs a set of concept prediction scores (one for each relevance-based level of the model), that we weigh and combine to obtain a final label. Moreover, we will also use this fuzzy score to identify the samples that have been wrongly annotated, and discover new positives for the training set (See Sec.5.4 for further details).

Our framework is somehow similar to the framework presented in [38]; however, in their work, the manual database re-assessment involves a lot of human effort and might increase the labeling noise. In this Chapter we automate this process by automatically assigning a class membership degree to each sample.

An example of using automatic relevance assignment for image recognition is represented by the work of Ji et al. [80], where, to solve a face gender classification problem, the distance to the SVM hyperplane is used to measure the importance of each sample in a dataset for a given class. Another example can be found in [139], where the confidence of an image region label is again derived from the sample distance from the hyperplane. Similar to the work in [80], we use a SVM-based measure to identify a fuzzy relevance score for each class, that we then discretize, in order to label our training sets with three relevance degrees. However, instead of using the raw distance value, we prefer to use a calibrated, thresholded value, that still depends on the distance to the hyperplane, but it is expressed with the probability of a given sample to be positive with respect to a concept.

We test the effectiveness of our system by comparing it with traditional binary-relevance frameworks for video retrieval. We consider a large scale, noisy, database of internet archive video, namely the TrecVID database, and we show that traditional categorization systems and features benefit from our automatic graded relevance-based multi-level model when retrieving this kind of biased data based on visual appearance.

5.2 The Baseline: Binary-Relevance Learning Frameworks

In order to understand our approach, we need to deeply understand the key elements of the learning based on binary annotations.

We recall here some properties of the most commonly used learning machine for image analysis, namely the Support Vector Machine, and show how this tool is generally used for multimedia categorization frameworks. Traditional multimedia categorization systems associate a set of images or videos with a semantic label given a low-dimensional description of the input, namely a feature vector. Multimedia retrieval systems use categorization frameworks to build lists of pictures/shots ranked according to their pertinence with respect to a semantic concept or query. In both cases, the core of the system is composed by a set of SVMs, namely supervised learning techniques that build models able to predict the presence of a given object or concept in the visual input.

In order to build such system, a set of training samples $v_1, v_2, \ldots, v_n$ is required, where v_i, $i = 1 \ldots n$ are the feature vectors extracted from the visual input data, such as MEDA or Saliency moments. For a set of concepts or categories $\{c_1, c_2, \ldots . c_p\}$ (e.g. "Telephones","Cup"), each sample in the training set v_i is labeled either as "positive", $y_{il} = +1$, $l = 1, \ldots, p$, (the concept is present in the visual input represented by v_i) or "negative", $y_{il} = -1$ (no visual trace of the concept is found in v_i).

A set of SVM-based classifiers, one for each concept/category, is used to learn the training feature space and then to label new samples according to the same scheme. The idea behind the SVM is to find a hyperplane that separates the two classes in training the feature space, given the distribution of the relative and non-relative samples with respect to a given concept. Such hyperplane satisfies the equation $w_l^T v - b_l = 0$, where $w_l = \sum_i \alpha_{il} y_{il} v_i$ has been proved in [18] to be the linear combination of the support vectors (i.e. the samples v_i for which the corresponding Lagrangian multiplier α_i is non-zero).

When a new point z needs to be categorized, the system assigns the corresponding label y_{zl} based on the side of the hyperplane where z falls after computing the dot product-based decision function $f_l(z)$:

$$y_{zl} = \begin{cases} -1 & \text{if } f_l z < 0 \\ +1 & otherwise \end{cases}$$

For a retrieval framework, see, for example [153], a concept score $p(y_{zl} = 1|z)$ is obtained for sample z based on decision function values; generally, various features are used to infer such scores, that are then combined and sorted in order to finally rank the results according to their pertinence with respect to a given concept.

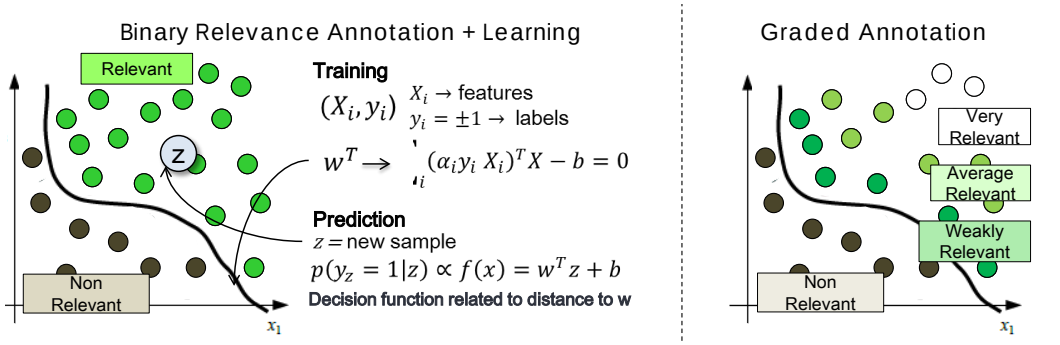


Figure 5.3: Differences and analogies between binary and non-binary systems.

5.3 Our Approach: A Graded-Relevance Learning Framework

As showed in Sec. 5.2, a SVM separates the feature space so that, at the top of the MMIR pyramid, our application will be able to distinguish between positive and negative new samples for each given concept. This boundary is found based on a binary relevance judgment, y_{il}, i.e. the groundtruth annotation. As discussed before, such binary division might be too restrictive compared to the range of possible instances of a semantic concept in the visual input. In order to allow a better usage of our data, we go here beyond the Relevant/Irrelevant subdivision, by reassessing our binary-relevance based training set with graded relevance judgments: in the new training set, a keyframe can be either Irrelevant (negative), Weak/Marginally Relevant, Average Relevant or Very Relevant to a given category. We then integrate the inferred relevance degree in a multi-level concept classifier.

The proposed framework works as follows (see Fig. 5.4):

1. The features extracted from the training samples are processed by a set of binary p SVM-based classifiers (one for each concept). According to such models, we analyze the position of each training sample v_k with respect to the hyperplane, using a calibrated decision value, and extract, for each concept c_l, a fuzzy membership score σ_{kl}. This is a continuous value representing how much a given sample is representative for a semantic concept (see Sec. 5.3.1 for more details).

2. As shown in Section 5.3.2, for each concept, we sort the positive training samples according to their fuzzy relevance scores and we set two thresholds so that we are able to re-categorize the samples using discrete relevance degrees. We obtain three subsets of Strongly, Average and Weakly Relevant training samples. All the negatives are equally labeled as Non Relevant samples.

3. Similar to [38], we then build a multi-level model by training the system on three different, relevance-based training sets. Then, as presented in Sec. 5.3.3,

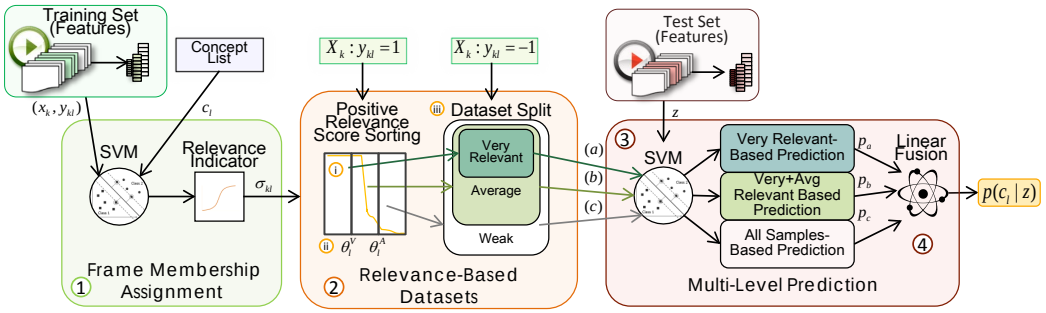


Figure 5.4: Visual representation of our Relevance Based Framework

given a new test image, for all c_l we obtain from the multi-level model three different concept prediction scores, that we then combine with weighted linear fusion to obtain one single output score. Such output score is then used for ranking and thresholded to determine the image label.

5.3.1 Decision Values as Relevance Indicators

As any traditional learning system, we start from an annotated training set of images/keyframes represented using low level features, namely our labeled samples. Given a set of non-negative samples, how to automatically define the fuzzy degree of relevance σ_{kl} of each sample with respect to a semantic concept? We tackle this problem by exploiting the SVM decision values of the training set.

The idea is that if, for a concept c_l, we are able to define how "positive" the sample is, given its position with respect to the hyperplane, we can have a good estimation of its relevance degree for that given concept. As a matter of fact, various works [80, 101, 139] showed that there is a correlation between the distance to the hyperplane (or the distance to the class center) and how much each sample is representative for a given class (the bigger its distance from the boundary, the higher its relevance with respect to the positive/negative category).

In our approach, we use as a fuzzy membership measure for a training sample a thresholded version of the decision function, according to the solution proposed in [141] to translate the uncalibrated decision value into a probabilistic output.

First, we calculate $f_l(v_k)$, namely the decision value for concept c_l, $\forall v_k, k = 1, \ldots, n$ in the training set samples. We then estimate the membership assignment as the positive class posterior probability $\sigma_{kl} = p(y_{kl} = 1 | f_l(v_k))$ with a parametric model based on fitting a sigmoid function:

$$\sigma_{kl} = \frac{1}{1 + \exp(A f_l(v_k) + B)}, \tag{5.1}$$

Where A and B are parameters adapted in the training phase to give the best probability estimates.

5.3.2 A Multi-Level Training Set with Different Relevance Levels

Once the *continuous* value σ_{kl} is computed for each training sample v_k, the next step is to build a graded relevance learning framework. In order to achieve this goal, we need to have a *discrete* relevance degree for each training sample, so that we are able to perform a relevance-based split of the training set into smaller, consistent subsets with different degrees of relevance with respect to a concept c_l.

As pointed out in [88], there is no universal rule to define such number of relevance degrees in a graded system. However, as shown in Sec 5.4, our experimental results suggest to set to 4 the number of relevance levels considered (i.e. Very Relevant, Average Relevant, Marginally Relevant, Non Relevant).

We therefore separate, for each concept, the positive/relevant training samples into three groups: Very Relevant Samples, that represent the most representative images/keyframes for a given class, Average Relevant Samples, and Weakly Relevant Samples; all the negatives are equally labeled as Non Relevant samples. We then generate three repartitions of our training database, based on which a multi-level model will be learnt (see Sec. 4.3). Having the fuzzy membership score σ_{kl} for each relevant sample, the discretization procedure is very simple:

i For each c_l, we take the *positive* ($v_k : y_{kl} = 1$) training samples and sort them according to their corresponding σ_{kl}, in decreasing order.

ii We now want to find a partition of the positive samples in three classes, according to the relevance scale selected. Based on the shape of the curve drawn by the sorted fuzzy relevance scores, we identify two thresholds, θ_l^V and θ_l^A. We use and test three different approaches to choose such thresholds:

(ii.a) we split the curve into equally spaced intervals,

(ii.b) we choose the thresholds manually such that, intuitively, the intra-partition variance of the scores value is minimized

(ii.c) we choose the values corresponding to 1/3 and 2/3 of the maximum membership score for the concept considered .

For each concept c_l, the Very Relevant samples are then defined as the positive $v_k : 1 < \sigma_{kl} < \theta_l^V | y_{kl} = 1$; the Average Relevant samples as $v_k : \theta_l^V < \sigma_{kl} < \theta_l^A | y_{kl} = 1$; the Weakly Relevant as $v_k : \theta_l^A < \sigma_{kl} < 0 | y_{kl} = 1$.

iii Finally, similar to [38] we create three new training sets: (a) merges the Very Relevant Samples with all the Non Relevant (i.e. our *negatives*, $v_k : y_{kl} = -1$), (b) merges (a) with the Average Relevant Samples, and (c) considers all positives and negatives samples.

5.3.3 Multi-Level Prediction and Fusion

Once we have created the three concept-specific training subsets, for each concept we build our multi-level model: it consists of three different SVM-based models, each of them learning a partition (a), (b), (c). Each level of the model separates the feature

	BOW (HLD)	*BH Manual*	*BH Equal*	***BH Max***	*BOW* (DOG)	*BD Manual*	*BD Equal*	***BD Max***	*CM*	*CM Manual*	***CM Equal***	*CM Max*
Airplane_Flying	0,045	0,047	0,048	**0,047**	0,018	0,018	0,018	**0, 019**	0,016	0,017	**0,017**	0,016
Boat_Ship	0,005	0,005	0,006	**0,006**	0,008	0,010	0,013	**0,012**	0,006	0,008	**0,011**	0,006
Bus	0,003	0,003	0,004	**0,003**	0,004	0,012	0,006	**0,021**	0,001	0,002	**0,001**	0,001
Cityscape	0,194	0,196	0,195	**0,194**	0,201	0,208	0,201	**0,204**	0,152	0,152	**0,152**	0,152
Classroom	0,006	0,007	0,008	**0,011**	0,003	0,003	0,004	**0,005**	0,001	0,003	**0,001**	0,001
Demonstration	0,033	0,035	0,033	**0,033**	0,037	0,035	0,037	**0,037**	0,010	0,010	**0,010**	0,010
Hand	0,004	0,004	0,005	**0,005**	0,008	0,009	0,009	**0,009**	0,004	0,006	**0,004**	0,00
Nighttime	0,050	0,051	0,050	**0,051**	0,039	0,040	0,044	**0,042**	0,027	0,027	**0,028**	0,029
Singing	0,072	0,073	0,074	**0,074**	0,078	0,082	0,082	**0,089**	0,074	0,077	**0,074**	0,076
Telephones	0,000	0,001	0,001	**0,001**	0,004	0,038	0,013	**0,022**	0,001	0,001	**0,001**	0,001
MAP	0,041	0,042	0,042	**0,042**	0,040	0,046	0,043	**0,046**	0,029	0,030	**0,030**	0,030

	WF	*WF Manual*	*WF Equal*	***WF Max***	*EDGE*	*E Manual*	*E Equal*	***E Max***
Airplane_Flying	0,028	0,028	0,028	**0,029**	0,028	0,028	0,028	**0,028**
Boat_Ship	0,002	0,003	0,002	**0,002**	0,015	0,018	0,017	**0,017**
Bus	0,001	0,002	0,002	**0,002**	0,001	0,001	0,001	**0,001**
Cityscape	0,107	0,111	0,118	**0,112**	0,193	0,203	0,198	**0,203**
Classroom	0,001	0,001	0,001	**0,002**	0,010	0,012	0,012	**0,010**
Demonstration	0,002	0,002	0,002	**0,002**	0,010	0,011	0,011	**0,011**
Hand	0,002	0,003	0,002	**0,002**	0,005	0,007	0,007	**0,009**
Nighttime	0,008	0,008	0,011	**0,024**	0,026	0,037	0,045	**0,060**
Singing	0,024	0,031	0,028	**0,027**	0,033	0,028	0,038	**0,038**
Telephones	0,001	0,001	0,001	**0,001**	0,002	0,002	0,002	**0,002**
MAP	0,018	0,019	0,020	**0,020**	0,032	0,035	0,036	**0,038**

Table 5.1: Applying graded relevance for video retrieval (TrecVID 2010 semantic Indexing Task): per-feature results for binary relevance learning, graded relevance with manual threshold tuning, graded relevance with equally spaced thresholds, graded relevance with threshold values proportional to the maximum concept score value

space in a different way, according to the annotations of the subset considered. When a new test sample z needs to be classified, we compute, using probabilistic SVM, three prediction scores for each concept (each of them is generated by a level of the model). We therefore obtain , $\forall l$, $p_a(y_{zl} = 1|z), p_b(y_{zl} = 1|z), p_c(y_{zl} = 1|z)$.

Each of these predictions is generated by a different relevance-based partition, which gives a different, complementary type of information regarding the relevance degree of the new sample to be classified. In order to exploit such different cues and obtain a single output, we then merge the three outputs using weighted linear fusion, as follows:

$$p_{zl} = p(y_{zl} = 1|z) = \sum_t w_t p_t(y_{zl} = 1|z), \tag{5.2}$$

$t = a, b, c, \forall l$, where w_t is a concept-specific weight learnt with cross-validation on development data.

For retrieval purposes, we then rank, for each query l the test samples according to p_{zl} in decreasing score, while for image categorization, the final label y_{zl} is assigned according to the following scheme:

$$y_{zl} = \begin{cases} -1 & \text{if } p_{zl} < 0.5 \\ +1 & otherwise \end{cases}$$

5.4 Experimental Validation

In this Section, we use our proposed graded relevance learning framework for the semantic video retrieval MMIR task. We compare the graded relevance framework with the classical binary-relevance systems (our baselines) for this task. Moreover,

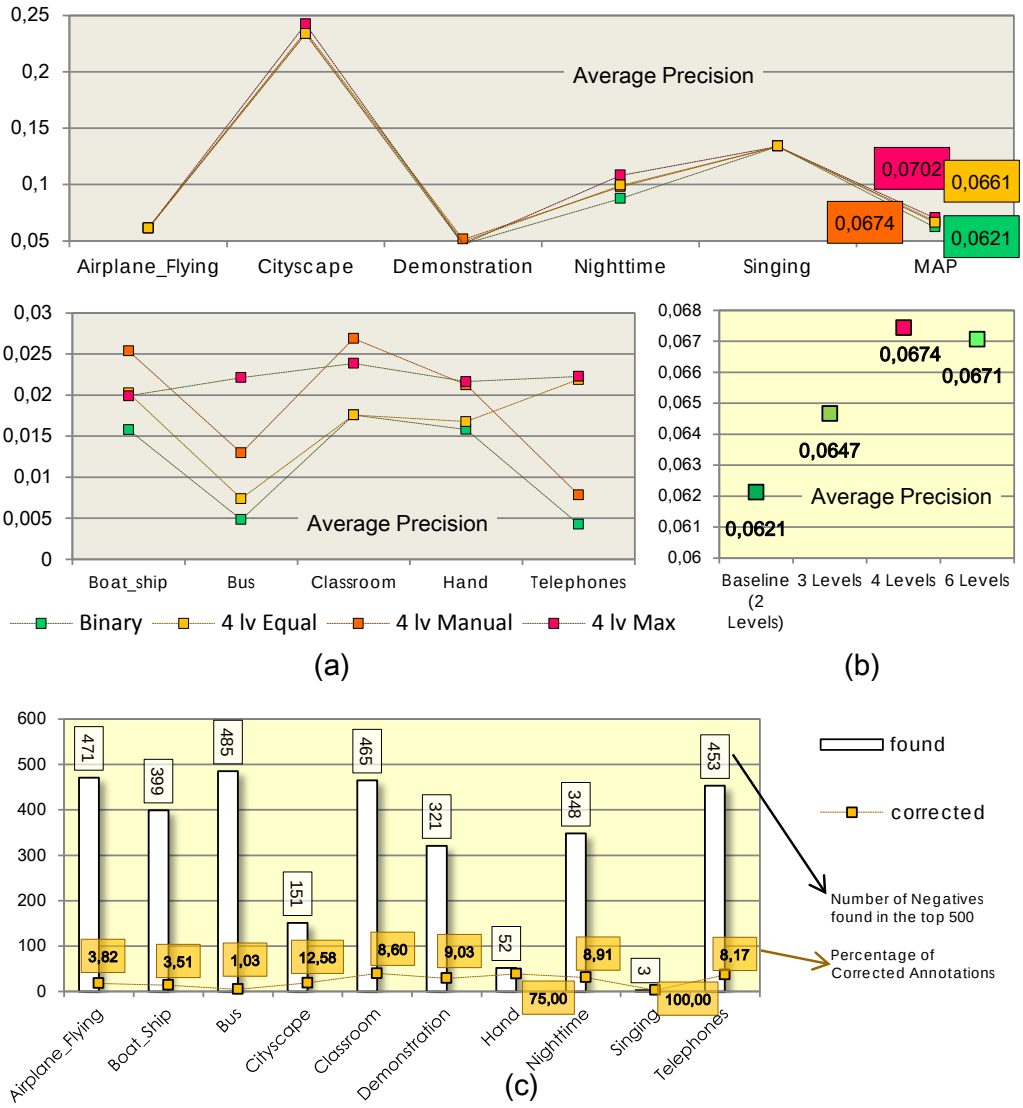


Figure 5.5: Our experimental results on the TrecVID 2010 Dataset. (a) Graded Relevance MMIR system with 4 levels, vs. traditional binary learning system. (b) Performances of the graded relevance dataset given the different relevance scales. Results are shown in terms of Average Precision. (c) percentages of re-assessed annotations, given the samples that has been labeled as negative but are detected as very relevant by our system



Figure 5.6: Automatic relevance-based reassessment: for given semantic concepts, examples from the three relevance-based categories are shown.

we re-use our automatic way for sample relevance re-assessment to partly remove the labeling noise of the dataset we consider.

5.4.0.1 Graded Relevance Framework Setup: Scale Selection and Relevance Visual Results

In order to test the effectiveness of our approach, we compare it with our baseline run for TrecVID 2010 [153], namely a pool of visual features (Sift DOG[105]+BoW, Sift HLD+BoW, Color Moments [175], a Wavelet Feature [187], and the MPEG7 edge histogram [200]) learnt by binary concept-specific SVMs.

Our Graded Relevance frameworks are built on top of such baselines. Having already computed all the features for all the images for the binary relevance MMIR system, the next step is to reassess the binary annotations and re-learn a set of models based on the resulting graded relevance degrees.

As we already have binary annotated datasets, we need to (1) add a fuzzy membership score to each frame, (2) find proper thresholds to obtain a discrete relevance category assignment, and (3) build a multi-level model as described in Sec. 5.3. We therefore proceed as follows:

1. For each feature f, we can re-use the model built in the baseline to estimate the fuzzy membership score σ_{kl}^{f} of a keyframe/image in the training set v_k for a concept/category c_l. Instead of using directly feature-based membership scores, that might supply incomplete information (e.g. the most relevant samples given the color or the edge distribution only), we combine them to obtain one single σ_{kl} for each sample.

2. Now that we have a fuzzy score, how to select the number of discrete levels that we will use to re-categorize the training set? As shown in Fig. 5.5 (b), we experimented with different subdivisions of the relevant samples of the

training set and tested their respective performances on the video retrieval task. Results shown in Mean Average Precision yield to the selection of a 4-level graded scale (namely Highly, Average and Weakly Relevant, and the Non Relevant label assigned to all the negatives) to reassess the training set. Given the trend of the fuzzy membership score curve, we manually select the thresholds θ^V and θ^A, as the two values, that for each concept minimize the score variance without reducing too much the number of positive samples inside the resulting partitions.

3. Finally, for every feature and every concept, given the new training set repartitions, three models are created and then used to predict the presence of the concept, combining the three outputs as shown in Sec. 5.3.2. At the end of this step we will have, for a new sample z, a concept score p_{zl}^f for each feature. Such feature-specific concept scores are then fused with linear fusion, similar to the binary baseline.

5.4.0.2 Experimental Results

For the semantic Indexing Task of TrecVID 2010, we present the results of both systems in terms of Mean Average Precision, the standard evaluation measure used for TrecVID assessments. We first look at the per-feature results. As we can see from Table 5.4 that the weaker features (e.g. Edge Histogram, +20% and Wavelet Feature, + 10%) benefit from our graded system. Moreover, we can see from Fig. 5.5 that the overall MAP increases of about 9%, when considering the ensemble of features combined together, with some peaks for those concepts for which the binary system was less performing, e.g. Classroom +53%, Telephones +83%, Bus +167% and BoatShip +60%, probably due to the variety in the visual appearance of relevant samples.

5.4.0.3 Qualitative Results and Binary Annotations Correction

Is the subdivision that we automatically obtain by re-ordering their samples based on their position to the hyperplane reliable? Fig. 5.6 shows examples from the three relevance-based classes: as we can see, our proposed method actually separates samples according to their relevance with respect to the given category or query, and in some cases, among the "Weakly Relevant" samples we can even find *wrongly annotated images*.

This suggests us to explore further the possibilities of these methods, in order to see if the case of wrongly annotated examples is a localized phenomenon, or if the TrecVID database is actually noisy in terms of groundtruth annotations. In order to check for noisy candidates, we take the list of training shots ranked according to their σ_{kl} . We then take the top 500 shots of this list and isolate the shots that have been annotated as *negative*. These should represent the samples that, despite their negative annotation, are detected to be very relevant by the learning framework. In

fact, they are supposed to be the ones for which the system would more likely assign a positive label.

We manually checked all such noisy candidates, and found out that actually many of them (around 10% average) have been originally wrongly annotated as negative examples. Even if they were labeled as negative, some of these shots are practically identical to frames annotated as positive, introducing noise in the training set and making the modeling more difficult.

5.5 Summary and Future Work

We presented a multimedia categorization and retrieval framework based on automatic graded relevance annotations. We automatically reassessed binary-labeled databases by assigning a degree of relevance to each sample based on its position with respect to the SVM hyperplane, and build an effective graded-relevance based CBMR system. We showed that our system, by allowing different degrees of relevance, outperforms the traditional binary-based frameworks for both image recognition and video retrieval.

Our simple approach can be improved in various ways. First, the automatic relevance fuzzy score assignment can be refined by using more complex machine learning-based measures, or by considering the combination of the relevance scores of a sample with respect to different concepts. Moreover, we can automate the discretization procedure (from fuzzy to discrete relevance degrees) by designing a measure that infers the best thresholds from the shape of the positive membership scores curve. Finally, while in our framework, similar to traditional CBMR systems, we use simple SVM classifiers for ranking, we could explore the learning methods used for web page ranking (e.g. [211]), that are designed to support graded-relevance, achieving a higher discriminative power.

Moreover, we have presented a first attempt of using the sample distance to the hyperplane to remove the labeling noise of the positive examples of a very challenging dataset. This approach can be extended and re-used for the denoising of various datasets annotations, in order to re-evaluate both the negative and the positive examples, by building a more formal or automatic model for the correction of the sample labels.

Chapter 6

Level 3: Beyond Pure semantics: the Synergy with aesthetic analysis

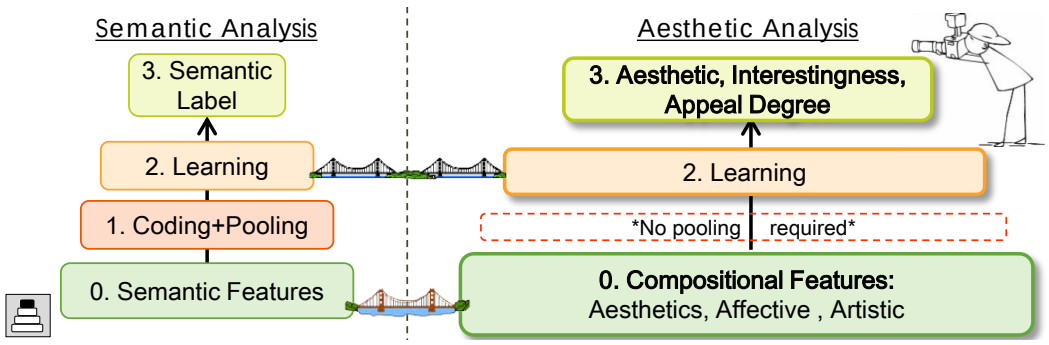


Figure 6.1: aesthetic analysis and semantic analysis are two closely related applications within the MMIR fields

MMIR for aesthetic analysis provides a set of techniques to automatically assign a beauty degree to a given image. In this Chapter, we re-use many of the techniques built in this thesis for semantic analysis, and apply it to aesthetic analysis. We assess the image appeal degree by using semantic features together with aesthetic features, and learn models based on graded relevance. Moreover, we also explore the other way around: are aesthetic analysis tools useful for semantic analysis? We determine here the importance of aesthetic features for semantic prediction. Overall, in this chapter we enrich both types of visual analysis by exploring the synergy between semantic and aesthetics. We show the benefits and the limits of this synergy, and propose some improvements in this direction.

We have seen in the previous chapters a substantial set of contributions at all levels of the MMIR chain, and applied them to semantic analysis by testing their performance on benchmarking datasets. We will now see how to re-use and mix together some of the technologies presented to go **beyond semantics**.

When reaching the upper level of the MMIR pyramid, namely the **application level**, the whole process, from low-level feature extraction to model learning, is enclosed in an intelligent "black box" and embedded into a real user-oriented application. Based on the variety of the underlying features, one or more of these "black boxes" can be used for the prediction of the image labels. The contribution of the different predictions can be then combined into a single framework for MMIR.

At this level of the chain, the system generally takes as input an image and gives as output a label related to some characteristic of the image, given its features and the underlying model. The nature of the label can be related to various peculiarities of the image: the objects, scenes, events, or, for example, the level of beauty or appeal of the image. Besides the efficiency, at the top of the pyramid there is one single visible parameter: the **accuracy** of the application, namely how accurate the predictions that the underlying image analysis system generates are.

Until now, following the traditional track of MMIR technologies, we mainly concentrated our efforts and techniques to build automatic *semantic analysis*(SA), whose task is to automatically predict the objects depicted in an image, the scenes where the pictures have been shot, or general semantic concepts such as actions or events. However, in the recent years, a new application for automatic image analysis has attracted the attention of MMIR researchers: *aesthetic analysis* (AA), namely a set of techniques to automatically assess the image beauty and appeal. This Chapter will explore possible solutions for this application.

Automatic Semantic Analysis and Aesthetic Analysis: Similarities and Boundaries

While with our traditional MMIR systems we predict the presence of given semantics in an image, aesthetic analysis frameworks predict the aesthetic degree of its visual content. semantic analysis techniques are generally more focused on the analysis of the *content* of the image: they learn models based on *semantic* features, such as the Saliency Moments [148], namely descriptions of the image content. On the other hand, aesthetic analysis frameworks [32, 130] learn models able to predict image beauty based on *compositional* features, that describe how much an image is following given photographic rules [32], and what the general arrangement and layout of the image is.

While semantic features such as MEDA, Saliency Moments, etc. give information about the content, AA features collect the attributes related to the shooting process and the image composition.

Despite their different applications and underlying features, semantic and aesthetic analysis systems are closely related fields, both from a technical and a perceptive point of view.

- From a *technical* point of view, they share the same learning framework, adopted by AA systems from SI. In both cases, a model (level 2, Chapter 5) is learnt on annotated (with content labels or aesthetic degree) training data (namely semantic or compositional features, level 0 and 1) through machine learning techniques, and then used to label (with object/scenes categories or beauty degree) a test image.
  But analogies between SA and AA are not limited to their implementations.

- Content and aesthetics are closely related in natural images also from a *perceptive* point of view. First, as proved in [35] the type of objects depicted in



Figure 6.2: Similar images, similar aesthetics.

an image can strongly influence the aesthetic judgment (e.g. people, animals, faces). Moreover, it is well known in photographic theory [91] that the image shooting process and its composition technique, as well as the emotion vehiculated and the degree of visual appeal, change according to the content to be depicted. Content is therefore important to determine the image composition, and, subsequently, its aesthetic degree.

Given this relation, we could also assume that the other way around is equally valid, given the image aesthetics and compositional rules, we can infer some image semantics. Groups of semantically similar images can share the same compositional attributes, making compositional-aesthetic information an additional cue for semantic analysis.

Our Contribution: Semantics at the Service of Aesthetic Analysis, and Vice versa

These observations regarding the junctions between these two fields suggest us that, by merging and combining AA and SA, and their underlying systems, we can enrich the overall visual analysis and obtain higher accuracy and better performances at an application level in both types of analysis. Our idea is that the synergy between semantic analysis and aesthetic analysis can help both image category and aesthetic degree prediction. Our aim is therefore to merge the intelligent systems underlying the two applications and improve the performances on top of the pyramid.

The first step towards the complete understanding of the synergy between aesthetics and semantics is to explore the contribution that semantic analysis tools for MMIR brings to image appeal assessment frameworks. In this Chapter, we will therefore first address one main question: **How is semantic analysis influencing aesthetic prediction?**

In order to answer the first question, we re-use many of the new rules and features in the previous chapters and our strong background in semantic analysis for aesthetic prediction. We employ *semantic features* such as the Saliency Moments, that we originally created for SA problems, that we combine it with a new *compositional* feature vector we build for this purpose containing artistic, affective and aesthetic features. We then learn a model based on such different types of features through a graded relevance learning systems of Chapter 5. We use as annotations for this purpose, the "interestingness" degrees assigned by the Flickr crowd to the images of the popular online photo management service. We then apply the resulting model for the **prediction of the appeal** of images and videos.

The key aspect of our approach is that it strongly relies on the *interaction* between different sources of information. Not only we combine semantic and computational aesthetics information, but we also build a **compositional** feature vector collecting existing and new features from three, closely related fields: aesthetic image analysis, affective image analysis and artistic image analysis. While in aesthetic analysis literature [36] compositional attributes are generally related to the simple image layout (**aesthetic** attributes, e.g. rules of thirds), here we extend this definition to include **affective** (emotional) and **artistic** attributes that can help characterizing the "intent" [47] of the photographer when composing a given picture.

The fusion of such different, discriminative and complementary sources of information about the scene attributes, together with semantic features, brings a substantial improvement on interestingness prediction performances, compared to systems based on aesthetic features only.

Moreover, we also investigate the other way around: **How is aesthetic analysis information affecting semantic prediction?** In order to explore this possibility, we look at the prediction improvements on semantic annotation of scenes, obtained by adding our *compositional (aesthetic, affective, artistic) feature vector* to a classic SA framework for **scene recognition** based on the Saliency Moments descriptor (see Chapter 3, and [148]).

Overall, in this Chapter we investigate the intersections between semantics and aesthetics with the aim of improving the global effectiveness of AA and SA systems. Since the general frameworks for AA and SA share the same pyramidal structure of MMIR systems (see Fig. 6.1), it is practically straight forward to combine the knowledge of these two conjoint fields. Content and aesthetics are complementary sources of information regarding the image depicted, and we can exploit their combination to enrich both the aesthetic and semantic learning.

In the following we will first identify similar works in the field, and highlight the peculiarities of our approaches in Sec. 6.1, and then describe in detail our compositional descriptor gathering aesthetic, affective and artistic features (see Sec. 6.2). We will then show in Sec. 6.3 how we employ it together with semantic features under a graded-relevance learning framework for the prediction of the interestingness degree of images and videos. Finally, in Sec. 6.4 we will combine it with semantic features for the improvement of a scene categorization system.

6.1 Related Work

We show here the novelty introduced by our work, one of the first attempt to improve semantic analysis and aesthetic analysis through their interaction.

Existing aesthetic image analysis frameworks automatically define the beauty degree of an image, generally by using learning systems trained on compositional features. Datta et al. in their pioneer work [32] learn features that model photography rules, and Wong et Al improve it in [201] by adding saliency information in the prediction framework. Here, we go beyond the pure compositional analysis by extending the pool of features used for aesthetic prediction, embedding semantic features in the AA framework. Some work has been done in this direction by Obrador et Al. [130], that build different aesthetic models for different image categories, using pre-defined manually labeled image categories. The use of semantic features for aesthetic prediction has been explored in [35], where semantic concepts such as animals, scenes, people, are detected and the probability of their presence is used as an attribute to predict image aesthetics. Our work differs from the one in [35] because we do not train any concept model (in order to avoid complexity and prediction noise generated by the low precision of semantic analysis systems), but we instead use the semantic features in an unsupervised way, and predict the aesthetics of an image given its semantic content without explicitly labeling it. Moreover, we also improve the AA learning framework by using a graded relevance semantic analysis system, previously used for video retrieval [150].

On the other hand, semantic analysis works generally by building MMIR frameworks for scene categorization (using holistic features [148, 205]), object recognition (using local features [30]), or concept detection for video retrieval [168]. Generally, such systems use local or global visual features that represent the pure image content, without considering all the information coming from the image composition, layout and shooting style. However, are compositional features useful for semantic analysis? In our work, we address this question by creating a scene categorization system that embeds some compositional features. To our knowledge, the work that appears to be more similar to ours is the one presented by Van Gemert [188], that incorporates into the spatial pyramid descriptor some style attributes for object recognition. Our work differs from [188] first because of the final application (scene vs object recognition), and second because we directly apply the compositional features for semantic analysis rather than using composition to extend an existing algorithm.

6.2 A New Set of Compositional Features Modeling the Photographer's "Intent"

The image aesthetics is strongly influenced by the image layout and arrangement, namely its *composition*. Previous works in computational image composition [130, 36] understands **composition** as a set of objective rules for constructing the image

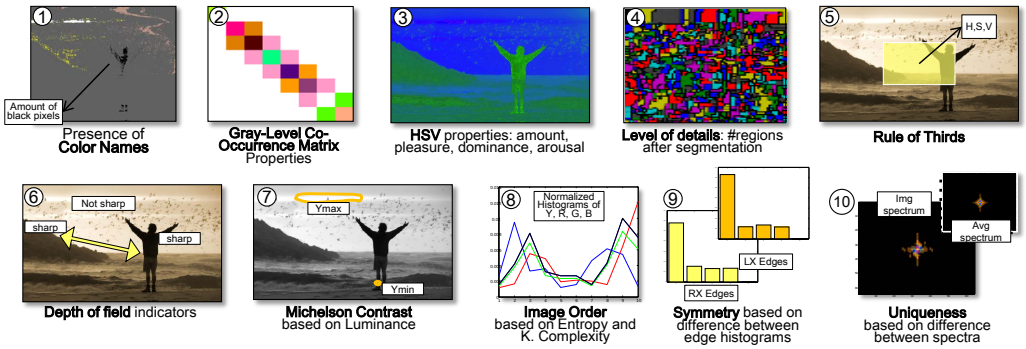


Figure 6.3: The compositional features we extract for aesthetic and semantic image analysis.

layout. For example, compositional attributes have been defined for aesthetic scene analysis as *"characteristics related to the layout of an image that indicate how closely the image follows photographic rules of composition"* [36]. This is the reason why many beauty predictors use mainly features related to the image composition in order to assess the image aesthetic degree.

However, arranging pictures is not only about applying objective rules, but it is also about following an artistic, intuitive process and convey intentions, meanings and emotions [47]. Therefore, here, we extend this concept to include features describing image emotional and artistic traits.

As Freeman states in [47] *"So far we have been concerned with the vocabulary of grammar and composition, but the process usually begins with purpose - a general or specific idea of what kind of image a photographer wants"*. In order to model the photographer's "intent" as defined by Freeman, we summarize the image composition using, besides aesthetic-compositional features, also affective attributes, that describe the emotions that a given image arouses through affective measures, and artistic attributes, that determine, for example, the "order" of a given image. In addition to existing features, e.g. low depth of field indicators [32], or color names [110], we implement two new compositional features: our own version of "image uniqueness", namely a measure evaluating the novelty of the image content, and our own formula to determine image "symmetry".

Many of the features we extract have been proved to be discriminative in their respective domains, but here, we test their discriminative ability for interestingness prediction and scene classification.

In order to properly describe the image composition, we therefore extract a set of features from three closely related domains, (aesthetic analysis [32, 130] , affective image analysis [110] and artwork analysis [155]), and collect them into a single *compositional descriptor*.

aesthetic image analysis aims at building systems that automatically define the beauty degree of an image: for example, Datta et al. in [32] extract features

that model photography rules using a computational approach to predict subjective aesthetic scores for photographs; such model is improved in [202] by adding saliency information in the aesthetic degree prediction framework.

In **affective image analysis**, the aim is to automatically define the type of emotions that a given image arouses: in [196], specific color-based features are designed for affective analysis and in [110], a pool of features arising from psychology and art, and related to the image composition, is proposed to infer the emotions generated by digital images.

In **art image analysis**, specific computational features (e.g. complexity, shape of segments) are designed to investigate patterns in paintings [155] or to assess artwork quality [100].

We therefore design a compositional descriptor of 43 features coming from emotion-based image recognition, aesthetic analysis, and painting analysis. For each image/frame, we extract our compositional 43-d feature vector $a = \{a(i)\}_{i=1}^{43}$, by gathering the following features (see Fig. 6.2):

1. **Color names**, a(1-9). Similar to [110] we count the amount of 9 different common colors ('black','blue','green','flesh','magenta','purple','red','white','yellow') in the image: different color combinations are used from artists/photographers to arouse different emotions.

2. **GLCM properties**, a(10-19). Gray-level co-occurrence matrices [61] are efficient ways to infer the image texture properties, because they describe the distribution of similar image values given a distance offset. Texture properties are of crucial importance to determine the affective content of a given image. Here, similar to [110], we fill our feature vector with the properties of correlation, homogeneity, energy, entropy and dissimilarity inferred from the GLCM matrix of a given image.

3. **HSV features**, a(20-25). After transforming the image into HSV space, we take the mean of hue, saturation and brightness, and compute *pleasure, arousal* and *dominance* features according to the values assigned to the affective features in [136] and then [110]. [1]

4. **Level of detail**, a(26). We measure image homogeneity from [110] based on the number of segments resulting after waterfall segmentation [7].

5. **Rule of thirds**, a(27-29). The rule of thirds in photography states that the most relevant subjects in an image should be placed along the horizontal/vertical lines intersection resulting from dividing an image in a 3×3 grid.

[1] Pleasure $= 0,69Y + 0,22S$
Arousal $= -0,31Y + 0,60S$
Dominance $= 0,76Y + 0,32S$

We evaluate how much the image follows the photography rule of thirds by taking the mean of Hue, Saturation and Brightness of the image inner rectangle, as in [32].

6. **Low depth of field**, a(30-38). The depth of field measures the ranges of distances from the observer that appear acceptably sharp in a scene. We extract low DoF indicators using wavelet coefficients as described in [32].

7. **Contrast**, a(39). As in [34], we extract the contrast Michelson measure [116]. The Michelson contrast is defined as the ratio between the difference and the sum of maximum and minimum luminance of the image.

8. **Symmetry**, a(40). Image Symmetry is a very important element to define the image layout. We define our own symmetry feature: we extract the Edge Histogram Descriptor [200] on both the left half and the right half of the image (but inverting major and minor diagonals in the right half), and retain the difference between the resulting histograms as the amount of symmetry in the image.

9. **Image Order**, a(41,42). According to Birkhoff [11], image beauty can be found in the ratio between order and complexity. Following this theory, image (in particular, arts and painting) order is computed in [155] using an information theory approach. We compute here the image order using first a Shannon Entropy approach and then a Kologomorov Complexity approach, as proposed in [155]. The first measure is computed based on the difference of the maximum entropy for an RGB image and the entropy of an image given its color palette (computed based on the entropy of the normalized histogram of the luminance values). The second value is computed with the differences between the sum of the entropy values of the normalized histograms of the three color channels and the complexity (1-compression ratio) of the image.

10. **Uniqueness**, a(43). How much an image represents a novelty compared to known information, how much is an image unique, i.e. it differs from the common image behavior? this variable can tell much about the artistic content of an image. We propose a new solution to address this question. We define the common image behavior according to the "1/f law" [157], saying that the average amplitude spectrum of a set of images obeys a 1/f distribution. We measure the uniqueness by computing the Euclidean distance between the average spectrum of the images in the database and the spectrum of each image (See Fig. 6.4 for visual examples).

We finally normalize all the features in the range [0,1] and combine them into our compositional vector. While feature 1-7 and 9 are *existing* descriptors from affective (1-4), aesthetics (5-7) and artistic (9) image analysis, the aesthetic feature of **symmetry** (8) and the artistic feature of **uniqueness** represent our *novel* contribution.

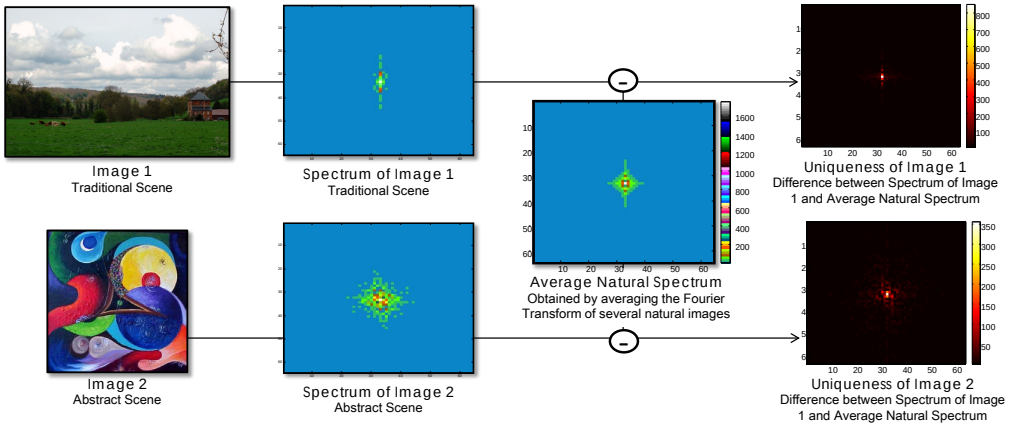


Figure 6.4: Visual explanation of the Uniqueness feature. The average Fourier spectrum of natural images is subtracted from the image spectrum, resulting in a uniqueness features with higher values for original pictures

6.3 Retrieving Appealing Images and Videos by Learning Flickr-based Graded Judgments

Our first work that investigates the benefit of the combination aesthetics-semantics is a MMIR framework that predicts the image "interestingness", typically related to the beauty of the scene depicted (see [151] for details). Our aim here is to build a system that, given an image (or a video sequence), can output a value corresponding to the appeal of its visual content.

Given the huge volumes of visual information surrounding us, automatic assessment of image beauty and appeal is becoming of crucial importance for the development of effective user centered visual applications.

For example, one of the tasks set for the ACM Multimedia Grand Challenge 2012 by the Japanese national public broadcasting channel NHK, is named "Where is the beauty?". In this track, participants are provided with a set of broadcasting videos (from the NHK channel) describing Japanese famous landscapes and touristic sites. The task is to automatically extract the beautiful scenes in such corpus and rank them in terms of beauty.

We participated to this task building a novel system for image appeal prediction based on semantic features, compositional features and graded relevance. We define a peculiar notion of beauty and use it to create an external training database with graded judgments. Based on this data we train a graded relevance framework (i.e. a CBMR system able to deal with multiple degrees of annotations), that will provide a list of beautiful scenes extracted from the NHK videos.

Our *novel contributions* can be summarized in the flow of our proposed approach:

1. **Notion of Beauty.** In the NHK challenge, the notion of beauty, and the subsequent evaluation and annotation of the training images, is entrusted to



Figure 6.5: Examples of *interesting* (top) and *non interesting* (bottom) Flickr images from Japan.

the participants. However, there is no universal agreement on an objective measure of beauty. It is therefore difficult to ask users to define the aesthetic degree of images and build an annotated dataset out of it. Moreover, mere scenic beauty might be of limited importance when describing broadcasting data, that is typically edited to "attract" the passive TV user. We therefore choose to model and automatically predict the *more informative property of "interestingness"*, namely an indicator representing how much the visual content is appealing for the audience. How do we quantify and use this attribute to build a training set for NHK corpus evaluation?

2. **Training Database.** Since the NHK videos come without aesthetic labels, we decide to train our system on an external database of images annotated with interestingness degrees, reliably computed from the preferences of the large Flickr audience. As a matter of fact, we exploit the *Flickr "interestingness" criteria* and build our *training set* by downloading a set of "interesting", "average interesting" and "non interesting" (3-level annotations) Japan-related pictures.

3. **Feature Extraction.** From both the Flickr-based and the NHK databases we extract a set of compositional and semantic features, including the two new features we design to define image *"symmetry" and "uniqueness"*.

4. **Two Learning Frameworks.** We first train a traditional binary CBMR system using a simple binary interesting/non interesting partition of the training images. We then learn a *graded relevance framework* similar to the one presented in Chapter 5, that can deal with multiple degree of annotations. Both systems, given a new image, are able to output a score reflecting its appeal/interestingness degree.

5. **Testing and Ranking.** Flickr data is made of still images, while our test data is composed of video sequences. Therefore, in order to transfer the knowledge from the training set to the test set, we first take a set of still frames per shot in the NHK corpus. We then predict the interestingness score for each

frame, and we rank shots according to the highest score obtained by its frames. For the final submission, we *combine two lists of shots*, one based on the binary retrieval and the other based on graded annotations. We also present some results on Flickr data to show the improvements brought by our choices, compared to AA-only based analysis.

In the following, we will first analyze in detail our notion of beauty and the resulting training database from Flickr images, see Sec. 6.3.2. We then explain in Sec. 6.3.2.1 the set of features used and the two proposed learning frameworks based on binary and graded relevance. Finally, we will look at the effectiveness of our choices for interestingness prediction of the Flickr dataset, see Sec. 6.3.3.

6.3.1 Training and Test Datasets: from Flickr Interestingness to Video Appeal

For the 2012 NHK Grand Challenge we are provided with a database of 10 videos depicting beautiful scenes in Japan touristic places. Shot boundaries and video sources are given, but no annotation is provided. However, since we address the challenge with a supervised learning framework, we require positive/negative labels associated with the data, in order to feed the supervised learning techniques beneath our system. To obtain such annotations, we should ask a substantial amount of users to give their preferences regarding the beauty of the NHK scenes. However, this turned out to be infeasible due to time and resources constraints. In order to have reliable, non-subjective, average judgments about the video beauty, we would need a large number of diverse human annotators. And even in this scenario, the notion of beauty of such users could be biased by the education level, the gender, nationality, etc. Moreover, we would like to retain all the scenes in the challenge videos as a test set, in order to maximize the probability of finding beautiful shots for the final submission.

Given these observations, we decided to train the whole system on external annotated data, and then apply the learnt models to retrieve the whole NHK corpus. How to define such external dataset? How to find a properly annotated training set, that can be suitable to rank the video scenes we are provided with?

First of all, we clarify our definition of beauty. Since we are dealing with broadcasting data, that is typically edited to *attract* the final user, we assume that the scenes to be shown to the final users need to be appealing, besides being beautiful: they need therefore to be "interesting". It was proved indeed [160] that beauty and interestingness are closely related, and that the second property includes the former one: precisely, beauty is said to be the "first derivative" of interestingness. An interestingness-based annotated dataset would give us not only an aesthetic judgment on the scenes, but also an evaluation of how much appealing they are, and how much curiosity they arouse.

But how can visual content be judged for its interesting qualities without involving human subjectivity? Flickr Interestingness criteria is the answer to this issue: a value representing the appeal of each photo in the Flickr collection, determined

by number of views, comments, bookmarks, citations in discussion etc. As we can see from Fig. 6.5, thanks to the "crowd wisdom", the most interesting images for a given query are high-quality, beautiful pictures, while the least interesting (second row) are more anonymous, less stunning pictures.

How do we use Flickr interestingness to create a training set suitable for our test data? Our approach is as follows (see Fig. 6.6).

1. First, since the NHK videos are shot in popular places in Japan, we select the 20 most touristic Japanese landmarks, helped by Japan guidebooks.

2. We then crawl the Flickr website and download 200 images per landmark, namely the 100 most interesting and the 100 less interesting pictures tagged with the proposed landmarks.

3. We then use the set of most interesting pictures as positives and the less interesting as negatives for our annotated training set, and build our models based on such data. However, a binary partition, based on positive/negative annotations only can be a bit too restrictive for the type of information we want to infer.

4. An image can be appealing for a user with different degrees, depending on the way the image is composed. We therefore download additional 100 images for each query, that we define as "average interesting", namely, the pictures that, for each query, lie in the middle of the returned sorted list. We then train a graded relevance system (see [150]) based on training set sub-partitions.

5. Finally, we sample the NHK videos, extracting 16 frames per shot, and apply the learnt models to such test dataset.

6.3.2 Our Proposed System

As shown in Fig. 6.6, in order to produce the final appealing scenes list we proceed as follows. From the Flickr-based and the NHK databases (i.e. the still frames extracted from the videos) we extract a set of compositional and semantic features (see Sec 6.3.2.1). We then use the training features with their corresponding interesting/non interesting annotations to train a *binary* learning system, as shown in Sec 6.3.2.2. We then extend the training set to allow for non-binary annotations and feed a graded relevance interestingness-based retrieval system (see Sec. 6.3.2.1). We use both systems to rank the shots in the test NHK set according to their interestingness, as shown in Sec. 6.3.2.3. Finally, we combine the outputs of both systems to compose the final list of beautiful shots, as shown in Sec 6.3.2.4.

6.3.2.1 Feature Extraction

As a first step, we extract a set of discriminative image features coming from both the AA domain and the SA field.

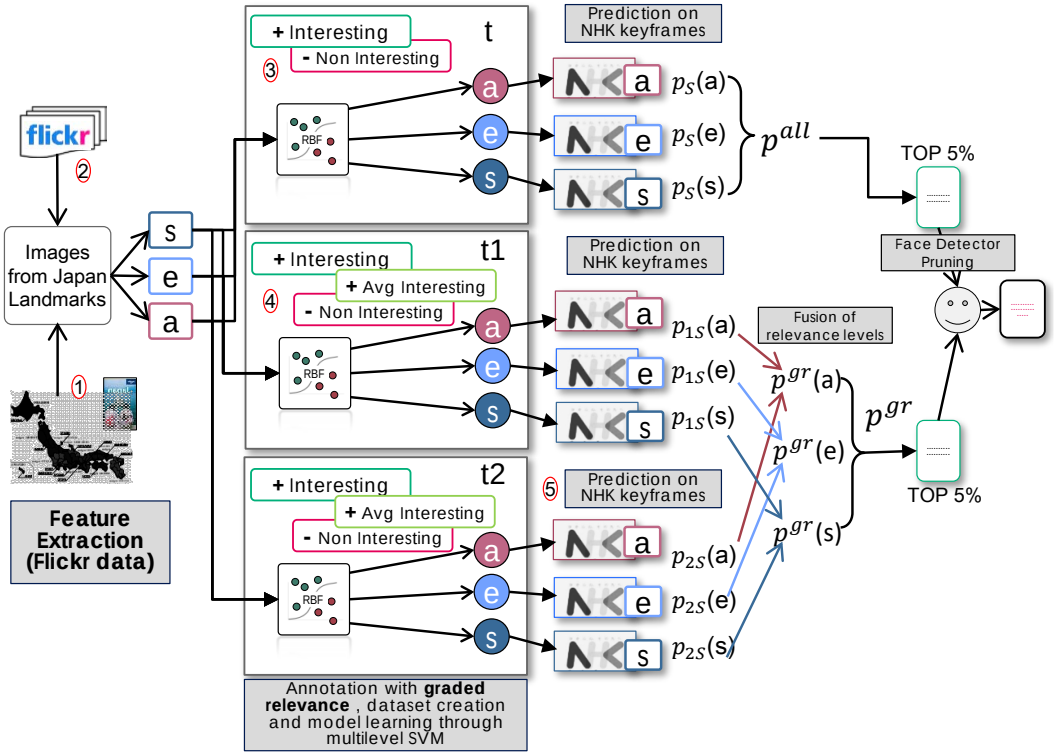


Figure 6.6: The flow of the proposed system for video scene appeal evaluation

We first extract a set of **affective, artistic and aesthetical features**, according to the compositional descriptor presented in 6.2, obtaining the feature vector $a = \{a(i)\}_{i=1}^{43}$.

In order to enrich the visual analysis, we then compute a set of **semantic features**.

As proved in [35], and confirmed by our results, the semantic content of an image plays an important role in determining its interestingness degree. In our system, we extract two semantic features (whose contributions will be combined with the compositional features in Sec. 6.3.2.2 and 6.3.2.3), namely the MPEG7 **Edge Histogram Descriptor** (EHD) [200], that generates a 80-dimensional vector, namely $e = \{e(j)\}_{j=1}^{80}$ and the **Saliency Moments Descriptor** (SM) [148], a 462-d feature vector $s = \{s(l)\}_{l=1}^{462}$. We chose the former one because it is a holistic image representation highlighting the image composition, which is typically very important to define the aesthetic degree of an image. Similarly, the SM descriptor is based on the shape of the salient region, and it is proved in [201] that visual attention information is closely related to image aesthetics.

6.3.2.2 Binary Relevance System

The first system we use for ranking NHK video scenes is a traditional learning framework. We extract from the training Flickr images the feature vectors a, e, s in Sec. 6.3.2.1. We then label each feature vector with a positive/negative label according to the interestingness of its corresponding image, and we use them as input to feed a set (one per feature) of Support Vector Machines (SVM) with RBF Kernel. We have now three feature-specific models able to distinguish between appealing/non appealing images.

On the NHK test set, we extract 16 frames per shot, for the 10 videos provided. We then classify the resulting frames with the feature-specific models: by doing so, we obtain, for each frame F an interestingness score $p_F(f) = p(int|f)$, $f = a, e, s$, corresponding to the output probability of the f-specific SVM. Since we want to determine the beauty of an entire shot S, and we have several frames per shot, we retain as interestingness score for a given shot the maximum of the scores of the frames belonging to that shot, namely $p_S(f) = max(p_{F \in S}(f))$. Finally, we compute the output interestingness score for each shot by linearly combining the output of the three feature-specific predictors: $p_S^{(all)} = \sum_f w_f \cdot p_S(f)$. We then rank the results according to p_S^{all} and the top 5% shots are retained for the final run.

6.3.2.3 Graded Relevance System

In our second framework, similar to the graded relevance system of Chapter 5 [150], we use 3 levels of labels(interesting, average interesting and non-interesting) and build a graded relevance retrieval system able to deal with multiple degrees of annotations.

First, we extend our training set by adding the images (and their corresponding features) that have been ranked by Flickr as being “average interesting". We then create two training subset: $(t1)$ considers as positives the "interesting" images and as negative the “average interesting" plus the “non interesting" images; $(t2)$ consider as positives the “interesting" and “average interesting" images, and as negative the “non interesting" ones. We then build a 2-layer model by training the system, for each type of feature, with both training subsets. Each level of the multi-layer model is a model generated by a different relevance-based partition of the training set, and it will therefore provide complementary information regarding the interestingness degree of a new sample to be ranked.

As a matter of fact, when we test on the NHK database, for each shot (processed as in the Binary Relevance System) we obtain two, complementary, feature-based interestingness score, namely $p1_S(f) = p(int|f, t1)$ and $p2_S(f) = p(int|f, t2)$, that we linearly combine into $p_S^{gr}(f) = \lambda(p1_S) + (1 - \lambda)(p2_S)$ order to retain the information coming from both levels of the multi-layer mode. Finally, feature-based scores are combined as in the binary system, results are ranked and the top 5% shots are retained for the final run.

	(s)	(e)	(a)	a+e+s
Binary	0,16646227	0,11358656	0,17658801	0,18944645
Graded	0,17648103	0,12364157	0,20284673	0,21484038

Table 6.1: MAP for the top 10% results of each system in our framework.

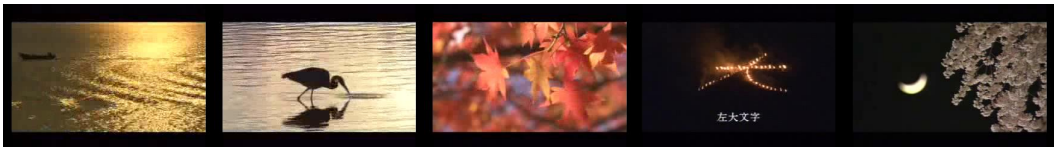


Figure 6.7: Snapshots the selected appealing scenes

6.3.2.4 The Submitted Run

We compose our final run by combining the top 5% of each list (namely, top 5% of Binary System and top 5% of Graded Relevance System). As we observed on our training data, the ranked lists of the two systems are very different, and they are more likely to show more interesting examples on the top positions.

Moreover, we run a face detector on the final list, that identifies the size and the number of the human faces represented in each listed frame. We then eliminate those frames for which the detector identifies a face with size $\geq 1/4$ of the frame surface, since such frames are very likely not to contain sceneries or landscapes.

Given the resulting keyframes and shots detected by our system as being the more "interesting" of the NhK corpus, our system has been selected as one of the finalists of the MM Grand Challenge 2013.

6.3.3 Evaluation

In this Section, we present some results on our development data, namely the Flickr annotated data, that proves the effectiveness of our solutions.

We split the images of Flickr dataset into a train and a test subset, in order to learn the various parameters of our framework (i.e. the RBF parameters, w_f in Sec 6.3.2.2, λ in Sec 6.3.2.3). We learn a model for each system (binary, graded) on the Flickr training subset, then we test on the Flickr test subset and evaluate the results on the using Mean Average Precision (MAP) over the list of the images that have been ranked in the top 10 %.

Results in Table 1 show the performances of the three features (aesthetics, a, EHD, e, SMD, s) used alone and their improvement after posterior fusion (+ 7% over a-based classification only, which show the importance of semantic features in the interestingness-based retrieval) in the binary-relevance system. Moreover, we show the MAP results for the graded relevance system, that improves the results of the previous system by 6% for $s-$based retrieval, 8% for $e-$based retrieval, 15% for $a-$based retrieval, and by 14% when we consider the contribution of the three features together.

6.4 Enhancing Semantic Features with Compositional Analysis for Scene Recognition

The second work we propose to investigate the relations between aesthetics and semantics is a MMIR system for the semantic scene categorization task, namely the automatic prediction of the image scene category (where was the image taken?) based on a pre-defined set of scene classes.

In scene recognition literature, semantic features are extracted to analyze the image content using either local analysis, based on local interest point descriptors (see Chapter 4) aggregated into a compact image representation, or global analysis (see Chapter 5), where general properties of the image, such as color or texture distribution, are summarized into a single descriptor.

Semantic information is without discussion the primary cue for scene identification. However, as mentioned, there exists another important source of information regarding the image scene, namely its *composition*, that not only is helpful to assess the image beauty, but it also could be an important cue to recognize the scene category. We understand here as image composition a combination of aesthetic, affective and artistic components that concur in creating its photographic style, intent [47] and layout, as shown in Sec. 6.2.

How is this related to scene identification? For example, intuitively it is more likely than an image with a high level of symmetry depicts a non-natural scene (e.g. a building), or that a picture with high level of detail comes from indoor environments. Moreover, as proved in [189], groups of semantically similar images can share the same compositional attributes (e.g. same point of view and depth of field for buildings or sport fields, same color contrast for natural outdoor scenes, see Fig. 6), as also studied in Photogrpahy Theory [47].

Given these observations, we explore here the role of compositional attributes for scene recognition using a computational approach. We design a categorization system that incorporates affective, aesthetic and artistic features, and combines them with traditional semantic descriptors for scene classification. The fusion of such different, discriminative and complementary sources of information about the scene attributes brings a substantial improvement of the scene categorization performances, compared to systems based on semantic features only.

We test the effectiveness of our *compositional descriptor* for scene classification using a variety of challenging datasets [143, 133, 205], including the SUN [205] dataset, that contains around 400 categories of very diverse scenes. We first use our compositional vector of Sec. 6.2 as a stand-alone descriptor, and we verify that compositional features carry discriminative power for scene categorization. Moreover, we show that, by summarizing the image layout properties into an image descriptor for classification, we introduce a new, complementary source of information regarding the scene characteristics. Therefore, when we combine our descriptor with traditional semantic features in a complete scene categorization system, we increase the classification accuracy of a semantic feature-only system by 13-15% for

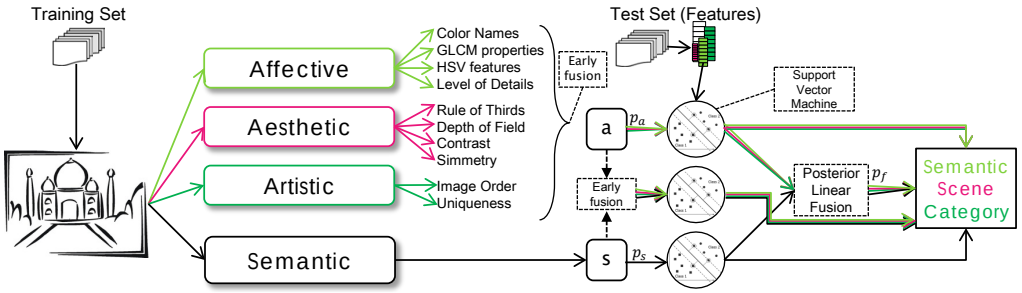


Figure 6.8: Combining compositional and semantic attributes for scene recognition

both small-scale [143, 133] and large-scale [205] scene understanding datasets.

In the following, we will first see a broad overview of our framework and then verify the discriminative power and the complementarity of compositional features for scene analysis given our experiments.

6.4.1 Analyzing Compositional Attributes for Scene Recognition

Scene recognition systems automatically categorize a given image into a pre-defined set of semantic classes corresponding to different scenery situations. In our approach, we exploit for this purpose the informativeness regarding image composition and photographic style typical of aesthetic, artistic and affective image features. We then combine them with the discriminative traditional semantic features in a complete scene categorization system that predicts an image class based on such diverse sources of information.

While we use traditional learning frameworks for classification, the peculiarity of our system is the choice of particular, discriminative image features that go beyond the traditional semantic descriptors for scene categorization by evaluating not only the content but also the compositional style of the image.

The set of features we combine is as follows:

- The core of the discriminative power of our scene recognition system is the set of semantic features for categorization. Here, we select to compute a powerful global feature for scene recognition, namely the Saliency Moments (SM) descriptor, see Chapter 3, $s = \{s(i)\}_{i=1}^{462}$.
- We then extract our **compositional feature vector** $a = \{a(i)\}_{i=1}^{43}$ gathering artistic, affective and aesthetic features (see Fig. 6.2).

Our general framework is basically a traditional image categorization/retrieval framework (see Fig. 6.8): based on compositional image features, for each category, we learn a model from the training images with Support Vector Machines (SVMs). Similarly, we train a set of SVMs (one for each class) using a set of semantic features. In the test phase, for a new image, given both compositional and semantic features and the models previously computed, we obtain, for each category c, $p_a(c)$ i.e.

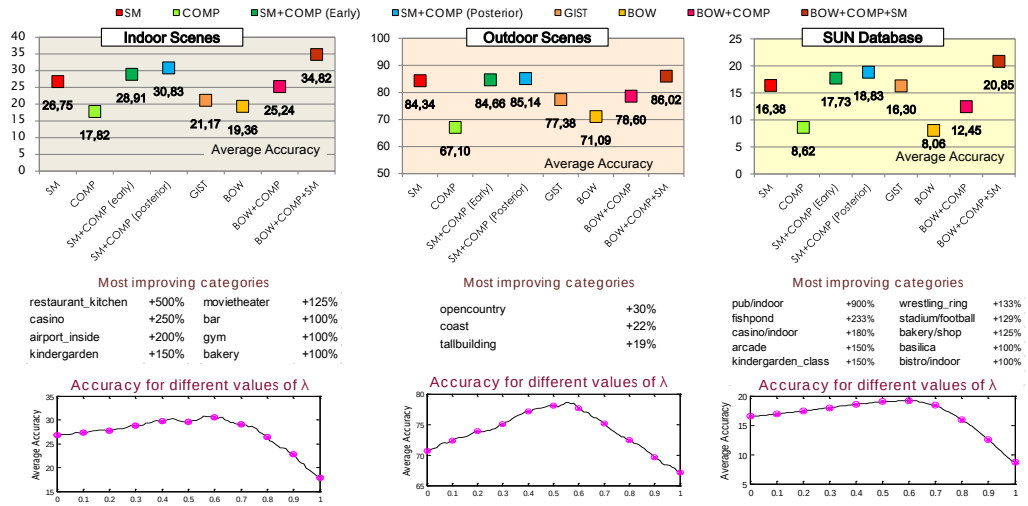


Figure 6.9: Results of large scale and small scale scene recognition

the category score given compositional features, and $p_s(c)$, i.e. the category score given semantic features. We retain the prediction from each model to test the discriminative ability of each feature, and we assign the category as $\arg\max_c p_x(c)$, being $x = a, s$.

In order to explore the complementarity of our features, we use two types of combination for the prediction of the scene category based on both compositional and semantic attributes.

- **Early Fusion:** We combine the vectors a and s in a single vector $as = \{as(i)\}_{i=1}^{505}$ gathering the information coming from such diverse set of sources. We then build a model based on the early-combined feature vector that will be able, in the test phase, to predict a category score $p_{as}(c)$.

- **Posterior Fusion:** We then combine the prediction scores with weighted linear fusion, namely $p_f(c) = \lambda(p_a(c)) + (1 - \lambda)(p_s(c))$, where λ is a value learnt during training. The final image category is assigned according to the resulting category scores after fusion.

6.4.2 Experimental Results

In order to test the effectiveness of the proposed approach, and verify the usefulness of aesthetic and affective features for semantic analysis, we use our framework for two scene recognition tasks: small scale categorization and large scale categorization. For the first task, we use two very popular benchmarking datasets for indoor [143] and outdoor [133] scene recognition, while for large scale scene recognition, we test our system on the challenging SUN database [205].

For each database, we first compute the classification accuracy given the model built using each semantic feature (i.e. "COMP" or "SM" in Fig. 6.9). We then

look at the classification performances resulting from using our compositional feature ("COMP") as a stand-alone descriptor. Furthermore, we show the effectiveness of the combination of aesthetic and compositional features by first fusing semantic and aesthetic features in a single, early fused descriptor (e.g. "SM+COMP (early)"). Finally, we combine the predictions of the single-descriptor-based models with posterior linear fusion. We fix the parameter λ for fusion and show the resulting, improved, performances (e.g. "SM+COMP (posterior)" in Fig. 6.9). For all descriptors and datasets proposed, we learn the feature space through a multi-class SVM with Radial Basis Function Kernel and we evaluate the performances by average multiclass accuracy, using training/tests subdivisions similar to our baselines for such datasets.

In order to further enrich our analysis, besides the SM descriptor we also compute a more purely semantic feature, namely the BOW signature aggregating PCA-SIFT [87] descriptors ("BOW"). We use the BOW feature as input for multiclass SVM with chi-squared kernel, and combine the resulting predictions with the compositional descriptor using posterior fusion.

6.4.2.1 Small Scale Scene Recognition

Automatic classification of images into scene categories is performed here using the proposed framework over two small scale dataset for indoor and outdoor scene recognition.

In all cases, we see from Fig. 6.9 that the combination between the two sources of information is bringing substantial improvement to the final performances, and that both sources of information (aesthetic, semantic) concur in discriminating the scenes. As a matter of fact, we can observe that the average accuracy reaches its maximum value when the weight λ, representing the importance of the aesthetic information for semantic classification, is around 0.5. This means that, when combining aesthetic and semantic prediction, the MMIR system will give equal importance to the aesthetic-based and to the semantic-based classification, suggesting that compositional analysis is a promising cue for semantic MMIR.

Outdoor Scenes

Results show that, by combining aesthetic, affective and artistic features in our compositional descriptor ("COMP") we obtain an effective descriptor (68% of accuracy VS 12.5% of a random classifiers) for outdoor scene recognition. Moreover, we can see that, while its combination with the SM descriptor does not bring much improvement, its fusion with the BOW features increases the performances of the BOW-only classification by 11%. This is because SM is an extremely effective descriptor by itself for outdoor scenes, and because it contains already some compositional information related to saliency, while the BoW feature arises from local analysis of the shapes and contours, leading to a more accurate description of the pure content of the image.

Indoor Scenes

Results in this task clearly highlight the effectiveness of compositional features for scene recognition: while the accuracy of the compositional descriptor alone is not as good as semantic features (around 17% vs. 26% of SM), but still more than 10 times better than a random classifier ($\sim$ 1,4%), the scenario changes when we combine it with traditional semantic features. As a matter of fact, both the early (+ 8%) and the posterior (+ 15%) fusion with the Saliency Moment descriptor successfully enhance the final scene recognition performances. Similar, more evident behavior when we combine the compositional features with the BOW descriptor: such fusion brings an improvement of 30 % compared to BOW-only classification. Being BOW and SM complementary, and being both complementary to compositional features, we also tried to combine the predictions resulting from the three stand-alone models using posterior linear fusion. The improvement over the classification based on SM (i.e. the most performing stand-alone descriptor) in this case is more than 20%, suggesting that introducing compositional features in the pool of existing semantic features is a promising cue for indoor scene recognition.

Large Scale Scene Recognition

Finally, we present our results for large scale scene recognition over the challenging SUN database. Results on this dataset follow the same pattern of the previously analyzed experiments: the combination of the SM with aesthetic/affective features brings an improvement of 8% with early fusion and 13% with late fusion compared to the SM-only classification, thus confirming the discriminative ability and the complementarity of aesthetic and compositional features for scene recognition even on a large scale.

6.5 Summary and Future Work

In this Chapter, we have investigated the possibility to apply the MMIR techniques for semantic analysis to another MMIR application, namely aesthetic analysis, to automatically assess the image beauty and appeal. Given the set of features commonly used for aesthetic analysis, we also explored the possibility of using aesthetic and compositional cues for semantic analysis, by embedding aesthetic information into a complete MMIR system for Scene Recognition.

Given our studies, we have verified that both applications, semantic and aesthetic analysis, benefit from the combination and integration of the respective underlying MMIR systems. Can we further investigate the synergy between those two applications? Two main tracks can be followed for our future work.

Improving AA with SA. As said, content plays an important role for aesthetic prediction, and different contents will generally show different compositional arrangements. We therefore aim to build a content-aware aesthetic framework with multiple aesthetic models, each one built according to the characteristics of a group of visually similar images. Some work has been done in this direction by Obrador

et Al. [130], that build different aesthetic models for different image categories, using pre-defined manually labeled image categories. However, the relevance of an image to one category is not always binary, as shown in [150], thus changing the compositional rules and the aesthetic appreciation. Moreover, even if extended with automatic classification, such work would be strongly dependent on the classifier performances. Our idea is to perform an unsupervised pre-grouping of the training images, by automatically defining a set of appearance-based clusters based on semantic features. We could then infer an aesthetic model for each "semantic" cluster, and then predict the aesthetic degree of the image according to its group and to its aesthetic features.

Improving SA with AA. On the other hand, we can improve the scene categorization system by looking at the compositional features that are more useful to distinguish each class from the others. For example, symmetry might be more useful to identify a skyscraper scene, rather than contrast. For each classifier, we could design a set of category-specific compositional vector, which can be constructed based on the discriminative ability of each feature for the class.

Chapter 7

Conclusions and Future Perspectives

Multimedia Information Retrieval is a complex research discipline. Its intrinsic complexity is due to its global goal: building machines that can see what we see and understand what we understand. Moreover, the research in the field is very diverse, since several computer science domains are involved in the creation of a MMIR system, from low-level signal processing, to statistic and probabilistic modeling, to machine learning techniques. Due to the complexity of this task, computer vision systems still perform an order of magnitude worse compared to their biological counterparts, i.e. the human vision systems. A huge gap exists between what humans see and understand, and what computer can infer from the image pixels. In this thesis we presented a set of techniques to reduce this gap.

As we showed in this manuscript, MMIR follows a complex chain of stages that *translate the image pixel values into intelligible concepts*: first, the image pixel are processed and their properties are summarized into **low-level features**. Such features are then either **pooled** into a compact image descriptor, or directly used as input for **learning machines**. Learning frameworks build models of the feature space by linking the feature values to the image properties. Such properties are expressed as **labels**, and can represent the type of **objects** depicted in the image, the **scene** and **location** where a given image has been shot, the **aesthetic** degree of the image, or the **emotions** it arouses. At the top level of the pyramid, i.e. the application level, the MMIR system is used for label **prediction**, given new, unknown images.

7.1 Our contributions From a Multidisciplinary Point of View: the Lessons Learnt

In this thesis, we presented a set of solutions for each level of processing of the MMIR chain: from the feature point of view, we developed **low-level features based on saliency** and several methods for **feature pooling based on marginal analysis**. From the learning point of view, we build a **graded-relevance learning framework** for video retrieval, and finally, at an application level, we applied the lessons learnt to **predict the image interestingness based on aesthetics and semantics**.

The common property of all the methods we proposed is that they tend to be **multidisciplinary**: we design our techniques by borrowing technologies and

principles from a wide variety of different sources that are not directly related with computer vision. We believe that the discovery of new cues for image analysis can bring not only substantial improvements with respect to existing techniques, but it also generates techniques providing complementary information regarding the image properties, and that can be therefore used in **combination** with existing techniques.

For example, by using the biological visual principles of visual attention and gist perception in Chapter 3, we created a very powerful low-level feature that can be compared for efficiency to the low-level global features, and for accuracy to the compact descriptors generated by aggregating local image descriptors. This suggests us that human visual systems and their understanding can be a useful source of inspiration for the development of effective computational vision systems. By analyzing how we process the real word scenes and objects in their early vision stage, and by carefully studying the recent developments in **neurobiology**, we can build more discriminative image features and learning frameworks.

Another example of our multidisciplinary approach is given by the techniques we design at Level 1 of the image analysis chain (Chapter 4) for local feature aggregation. Here, we borrow from **economic statistics** the theory of Copulae, and we successfully apply it for marginal-based multivariate modeling of the local image keypoints. The resulting compact feature is much more effective for semantic analysis compared to the complex traditional models such as Bag of Words, without involving any universal model construction. This work represents a first, successful attempt of using Copulae for image statistics. But Copula theory is a whole branch of statistics that can provide several tools to solve computer vision problem, and that we wish to explore in our future work.

At the learning level (Chapter 5), we built a graded-relevance learning framework. The concept of graded relevance is not new in the **web information retrieval** field. Web pages ranking are based on a non-binary scale of relevance, and ad-hoc learning and ranking techniques have been proposed for this purpose. Such techniques can be re-used for Multimedia Information Retrieval, and deep, interesting studies can be done in how to adapt such ranking methods to the video retrieval field.

Finally, at an application level, in Chapter 6 we explore the synergy of many domains within the MMIR field: **semantic analysis, aesthetic analysis, affective Analysis, artistic analysis**. The combination of such variety of cues for image analysis enriches the global visual analysis and brings substantial improvements to the performances of scene recognition and image interestingness prediction. In particular, one of the key elements of the MMIR system we build for interestingness prediction is that the groundtruth information is not generated manually, but by using a non-standard procedure in traditional MMIR systems, namely the **crowd sourcing**.

7.2 Future Perspectives: Does Content matter?

The work we outlined in Chapter 6 represents our first attempt to use *contextual* information for MMIR. Contextual information is a precious source of information regarding the image properties, since it is composed of textual captions and descriptions directly input by the image users. With the explosion of the on-line photo management and sharing tools, we have available an incredible amount of metadata that we can use to improve our aesthetic and semantic prediction.

It was pointed out [167] that visual-based analysis might not be even actually useful for multimedia-based applications, and that *contextual information* could be enough to perform automatic image analysis. It is true that visual analysis research, constrained by benchmarking experimental datasets, tend to rely on the mere *content* (compositional, semantic) analysis of the images, losing precious information coming from the media *context* (metadata, user characteristics). Semantic and Aesthetic MMIR systems also lack of considering real-world applications, limiting the experiments to the pure prediction of image characteristics (image category or image beauty). As a matter of fact, despite their effectiveness for image properties prediction tested on benchmarking datasets, pure content-based techniques have several performances limitations when coming to practical uses.

However, living in a world of user-generated content, we are learning that we must consider what the user needs, analyze her behavior, her way of producing media, and all the information she is giving as input about her visual content. We can say that *content analysis is not over*, but we have to find its right place, specific applications that need its informativeness, merge it with other sources to enrich the media analysis. *Content matters*, and, in the following, we can see a possible future track to follow to improve the prediction of the image interestingness and appeal based on content and context. The global intention is to follow the idea of synergy between different sources initiated in the last chapter of this thesis

With the wide diffusion of picture management tools such as Flickr, Picasa, Facebook, the need of automatic assessment of image beauty, interestingness and appeal becomes clearer day by day. Instead of trying to model interestingness using pure visual cues, we should consider several factors that might be involved in aesthetic image judgment on online photo services, for example:

1. The position of the user in the **social network**: how much is she appreciated? How many contacts does she have? How interesting were her previous pictures for the crowd?

2. The **content** of the image: there are some subjects that are more likely to attract the crowd than others, depending on the culture and nationality, the current news, the general tendency of humans to focus on certain type of objects and scenes etc.. Important is also the **location** where a given image has been shot.

3. The **aesthetics** of the image: the way the image is composed, arranged, the

photography rules that have been used to enhance its beauty, the type of camera that is used for shooting, the emotions vehiculated etc..

If we are able to infer this kind of information regarding the user and its media, we could then translate such elements into numerical features and model (learning the importance of each feature) the image interestingness using as groundtruth a large number of Flickr images annotated with their corresponding Flickr interestingness values [1].

How to extract such features? **Social Network analysis** (see, for example, [24]), very popular in the recent years, can help us understanding the social position and profile of the user (1). Regarding the content and location (2), besides the **tag and metadata** that the user provides with her media, we could also rely on pure content-based techniques such as the ones we presented in this thesis. Similarly, aesthetic and compositional features (3) can be calculated by modeling photographic rules, emotions, and artistic traits using a computational approach, and we can enrich this pool of features with some **forensic analysis** for camera identification [108]. Moreover, there will be other factors impacting interestingness that we can discover thanks to **user perception** studies.

Similarly, we could help affective and artistic image analysis with similar frameworks, and perhaps re-use similar techniques.

As a conclusion, we showed with this thesis that the synergy between different, related or unrelated field can bring substantial improvements towards the real modeling of a computer vision system. Perhaps, by increasing the interaction of such variegate cues, and exploiting the new technologies that the digital visual world is providing us, we could soon infer, in a reliable way, at least 500 of those thousand words an image is worth.

[1] http://www.flickr.com/explore/interesting/

Bibliography

[1] Vireo group in http://vireo.cs.cityu.edu.hk/links.html. (Cited on page 27.)

[2] F. S. Abas and K. Martinez. Classification of painting cracks for content-based analysis. In *Electronic Imaging 2003*, pages 149–160. International Society for Optics and Photonics, 2003. (Cited on page 26.)

[3] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1597–1604. IEEE, 2009. (Cited on pages 39, 43 and 49.)

[4] S. Arya and D. Mount. Algorithms for fast vector quantization. In *Data Compression Conference, 1993. DCC'93.*, pages 381–390. IEEE, 1993. (Cited on page 64.)

[5] S. Ayache and G. Quénot. Trecvid 2007 collaborative annotation using active learning. In *In Proceedings of the TRECVID 2007 Workshop*, 2007. (Cited on page 102.)

[6] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *Computer Vision–ECCV 2006*, pages 404–417, 2006. (Cited on pages 6, 16, 36, 61, 62, 64, 100 and 168.)

[7] S. Beucher. Watershed, hierarchical segmentation and waterfall algorithm. *Mathematical morphology and its applications to image processing*, pages 69–76, 1994. (Cited on page 123.)

[8] S. Bhattacharya, R. Sukthankar, and M. Shah. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *Proceedings of the international conference on Multimedia*, pages 271–280. ACM, 2010. (Cited on page 17.)

[9] N. Bianchi-Berthouze. K-dime: an affective image filtering system. *Multimedia, IEEE*, 10(3):103–106, 2003. (Cited on page 25.)

[10] I. Biederman. Visual object recognition. In *Readings in philosophy and cognitive science*, pages 9–21. MIT Press, 1993. (Cited on page 15.)

[11] G. Birkhoff. Aesthetic measure, 1933. (Cited on page 124.)

[12] L. Bo and C. Sminchisescu. Efficient match kernel between sets of features for visual recognition. *Advances in Neural Information Processing Systems*, 2(3), 2009. (Cited on pages 23 and 80.)

[13] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. (Cited on page 21.)

[14] A. Borji, D. Sihite, and L. Itti. Salient object detection: A benchmark. (Cited on page 38.)

[15] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via plsa. *Computer Vision–ECCV 2006*, pages 517–530, 2006. (Cited on page 21.)

[16] A. Bosch, A. Zisserman, and X. Muoz. Image classification using random forests and ferns. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007. (Cited on page 21.)

[17] B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992. (Cited on pages 8, 28, 102, 106 and 170.)

[18] B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992. (Cited on page 107.)

[19] Y. Boureau, N. Le Roux, F. Bach, J. Ponce, and Y. LeCun. Ask the locals: multi-way local pooling for image recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2651–2658. IEEE, 2011. (Cited on page 19.)

[20] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2559–2566. IEEE, 2010. (Cited on page 18.)

[21] E. Bursztein, M. Martin, and J. Mitchell. Text-based captcha strengths and weaknesses. In *Proceedings of the 18th ACM conference on Computer and communications security*, pages 125–138. ACM, 2011. (Cited on page 25.)

[22] W. M. Campbell, D. E. Sturim, and D. A. Reynolds. Support vector machines using gmm supervectors for speaker verification. *Signal Processing Letters, IEEE*, 13(5):308–311, 2006. (Cited on page 19.)

[23] M. Castelhano and J. Henderson. The influence of color on the perception of scene gist. *Journal of Experimental Psychology*, 34(3):660–675, 2008. (Cited on page 51.)

[24] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web*, pages 721–730. ACM, 2009. (Cited on pages 142 and 179.)

[25] Y. Chen and J. Z. Wang. Image categorization by learning and reasoning with regions. *The Journal of Machine Learning Research*, 5:913–939, 2004. (Cited on page 22.)

[26] Y. Chen, J. Z. Wang, and R. Krovetz. Clue: Cluster-based retrieval of images by unsupervised learning. *Image Processing, IEEE Transactions on*, 14(8):1187–1201, 2005. (Cited on page 20.)

[27] Y. Chen, X. S. Zhou, and T. S. Huang. One-class svm for learning in image retrieval. In *Image Processing, 2001. Proceedings. 2001 International Conference on*, volume 1, pages 34–37. IEEE, 2001. (Cited on page 22.)

[28] M. Cohen, G. Alvarez, and K. Nakayama. Gist perception requires attention. *Journal of Vision*, 10(7):187, 2010. (Cited on page 51.)

[29] C. Colombo, A. Del Bimbo, and P. Pala. Semantics in visual information retrieval. *Multimedia, IEEE*, 6(3):38–53, 1999. (Cited on page 25.)

[30] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, page 22. Citeseer, 2004. (Cited on pages 10, 17, 18, 20, 28, 36, 62, 63, 66, 95 and 121.)

[31] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005. (Cited on pages 6, 36, 61, 62, 64, 100 and 168.)

[32] R. Datta, D. Joshi, J. Li, and J. Wang. Studying aesthetics in photographic images using a computational approach. *Computer Vision–ECCV 2006*, pages 288–301, 2006. (Cited on pages 6, 16, 17, 22, 25, 118, 121, 122, 124 and 168.)

[33] R. Datta, J. Li, and J. Z. Wang. Imagination: a robust image-based captcha generation system. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 331–334. ACM, 2005. (Cited on page 25.)

[34] M. Desnoyer and D. Wettergreen. Aesthetic image classification for autonomous agents. In *Proc. ICPR*. Citeseer, 2010. (Cited on page 124.)

[35] S. Dhar, V. Ordonez, and T. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1657–1664. IEEE, 2011. (Cited on pages 25, 118, 121 and 129.)

[36] S. Dhar, V. Ordonez, and T. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1657–1664. IEEE, 2011. (Cited on pages 120, 121 and 122.)

[37] C.-J. Du and D.-W. Sun. Learning techniques used in computer vision for food quality evaluation: a review. *Journal of Food Engineering*, 72(1):39–55, 2006. (Cited on page 25.)

[38] N. Elleuch, M. Zarka, I. Feki, A. Ben Ammar, and A. Alimi. Regimvid at trecvid 2010: Semantic indexing. In *In Proceedings of the TRECVID 2010 Workshop*, 2010. (Cited on pages 105, 106, 108 and 110.)

[39] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. (Cited on pages 8, 24 and 171.)

[40] J. Farquhar, S. Szedmak, H. Meng, and J. Shawe-Taylor. Improving bag-of-keypoints image categorisation: Generative models and pdf-kernels. 2005. (Cited on pages 18 and 23.)

[41] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007. (Cited on pages 8, 24, 30, 46, 56 and 171.)

[42] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 524–531. Ieee, 2005. (Cited on pages 15, 36 and 62.)

[43] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 524–531. IEEE, 2005. (Cited on page 21.)

[44] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–264. IEEE, 2003. (Cited on pages 15, 36 and 62.)

[45] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–264. IEEE, 2003. (Cited on page 20.)

[46] P. Forssén, D. Meger, K. Lai, S. Helmer, J. Little, and D. Lowe. Informed visual search: Combining attention and object recognition. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pages 935–942. IEEE, 2008. (Cited on page 37.)

[47] M. Freeman. *The photographer's eye: composition and design for better digital photos*. Focal Pr, 2007. (Cited on pages 120, 122 and 132.)

[48] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.*, 4:933–969, December 2003. (Cited on page 105.)

[49] D. Gabor. Theory of communication. part 1: The analysis of information. *Electrical Engineers-Part III: Radio and Communication Engineering, Journal of the Institution of*, 93(26):429–441, 1946. (Cited on page 54.)

[50] B. Gao, T.-Y. Liu, T. Qin, X. Zheng, Q.-S. Cheng, and W.-Y. Ma. Web image clustering by consistent utilization of visual features and surrounding texts. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 112–121. ACM, 2005. (Cited on page 20.)

[51] P. Gehler and S. Nowozin. Let the kernel figure it out; principled learning of preprocessing for kernel classifiers. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2836–2843. IEEE, 2009. (Cited on page 80.)

[52] H. Goëau, P. Bonnet, A. Joly, N. Boujemaa, D. Barthélémy, J.-F. Molino, P. Birnbaum, E. Mouysset, M. Picard, et al. The imageclef 2011 plant images classification task. In *ImageCLEF 2011*, 2011. (Cited on page 25.)

[53] K.-S. Goh, E. Y. Chang, and W.-C. Lai. Multimodal concept-dependent active learning for image retrieval. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 564–571. ACM, 2004. (Cited on page 22.)

[54] V. Gopalakrishnan, Y. Hu, and D. Rajan. Random walks on graphs to model saliency in images. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1698–1705. IEEE, 2009. (Cited on page 39.)

[55] M. Gordon and P. Pathak. Finding information on the world wide web: the retrieval effectiveness of search engines. *Information Processing & Management*, 35(2):141–180, 1999. (Cited on page 103.)

[56] S. Gordon, H. Greenspan, and J. Goldberger. Applying the information bottleneck principle to unsupervised clustering of discrete and continuous image representations. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 370–377. IEEE, 2003. (Cited on page 20.)

[57] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007. (Cited on page 24.)

[58] J. Han and K. Ma. Fuzzy color histogram and its use in color image retrieval. *Image Processing, IEEE Transactions on*, 11(8):944–952, 2002. (Cited on pages 14 and 41.)

[59] A. Hanjalic and L.-Q. Xu. Affective video content representation and modeling. *Multimedia, IEEE Transactions on*, 7(1):143–154, 2005. (Cited on page 26.)

[60] R. M. Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):786–804, 1979. (Cited on page 15.)

[61] R. M. Haralick and L. G. Shapiro. *Computer and Robot Vision*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition, 1992. (Cited on page 123.)

[62] S. Harding, M. Cooke, and P. Konig. Auditory gist perception: an alternative to attentional selection of auditory streams? *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint*, pages 399–416, 2007. (Cited on page 51.)

[63] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. *Advances in neural information processing systems*, 19:545, 2007. (Cited on pages 39 and 49.)

[64] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, page 50. Manchester, UK, 1988. (Cited on page 15.)

[65] D. Hawking. Overview of the trec-9 web track. In *Proceedings of TREC*, volume 9, pages 500–249, 2000. (Cited on page 103.)

[66] X. He, R. S. Zemel, and M. A. Carreira-Perpinán. Multiscale conditional random fields for image labeling. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–695. IEEE, 2004. (Cited on page 21.)

[67] P. Hong, Q. Tian, and T. S. Huang. Incorporate support vector machines to content-based image retrieval with relevance feedback. In *Image Processing, 2000. Proceedings. 2000 International Conference on*, volume 3, pages 750–753. IEEE, 2000. (Cited on page 22.)

[68] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. Ieee, 2007. (Cited on pages 37, 39, 43, 44, 46, 50, 52, 53 and 58.)

[69] C.-W. Hsu and C.-J. Lin. A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425, 2002. (Cited on page 22.)

[70] J. Huang, S. Kumar, M. Mitra, and W. Zhu. Image indexing using color correlograms, 2001. US Patent 6,246,790. (Cited on page 36.)

[71] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Image indexing using color correlograms. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 762–768. IEEE, 1997. (Cited on page 14.)

[72] N. Inoue, T. Saito, K. Shinoda, and S. Furui. High-level feature extraction using sift gmms and audio models. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3220–3223. IEEE, 2010. (Cited on pages 18 and 19.)

[73] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254–1259, 2002. (Cited on pages 39, 43, 49 and 58.)

[74] K. Ivanova and P. Stanchev. Color harmonies and contrasts search in art image collections. In *Advances in Multimedia, 2009. MMEDIA'09. First International Conference on*, pages 180–187. IEEE, 2009. (Cited on page 26.)

[75] K. Ivanova, P. Stanchev, E. Velikova, K. Vanhoof, B. Depaire, R. Kannan, I. Mitov, and K. Markov. Features for art painting classification based on vector quantization of mpeg-7 descriptors. *Data Engineering and Management*, pages 146–153, 2012. (Cited on page 26.)

[76] T. Jaakkola, D. Haussler, et al. Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, pages 487–493, 1999. (Cited on page 66.)

[77] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. *Computer Vision–ECCV 2008*, pages 304–317, 2008. (Cited on page 19.)

[78] H. Jegou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3304–3311. IEEE, 2010. (Cited on pages 19 and 63.)

[79] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011. (Cited on pages 10, 18, 19, 28, 62, 63, 66 and 95.)

[80] Z. Ji and B. Lu. Gender classification based on support vector machine with automatic confidence. In *Neural Information Processing*, pages 685–692. Springer, 2009. (Cited on pages 106 and 109.)

[81] J. Jia, N. Yu, and X.-S. Hua. Annotating personal albums via web mining. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 459–468. ACM, 2008. (Cited on page 20.)

[82] F. Jing, C. Wang, Y. Yao, K. Deng, L. Zhang, and W.-Y. Ma. Igroup: web image search results clustering. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 377–384. ACM, 2006. (Cited on page 20.)

[83] C. R. Johnson, E. Hendriks, I. J. Berezhnoy, E. Brevdo, S. M. Hughes, I. Daubechies, J. Li, E. Postma, and J. Z. Wang. Image processing for artist identification. *Signal Processing Magazine, IEEE*, 25(4):37–48, 2008. (Cited on page 26.)

[84] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2005. (Cited on pages 16 and 54.)

[85] J. Jones and L. Palmer. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233, 1987. (Cited on page 54.)

[86] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on*, pages 2106–2113. IEEE, 2009. (Cited on page 39.)

[87] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. 2004. (Cited on pages 16, 28, 74, 83, 88, 95 and 135.)

[88] J. Kekäläinen. Binary and graded relevance in ir evaluations–comparison of the effects on ranking of ir systems. *Information processing & management*, 41(5):1019–1033, 2005. (Cited on page 110.)

[89] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37:18–28, September 2003. (Cited on page 105.)

[90] J. J. Kivinen, E. B. Sudderth, and M. I. Jordan. Learning multiscale representations of natural scenes using dirichlet processes. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007. (Cited on page 21.)

[91] B. Krages. *Photography: the art of composition*. Allworth Press, 2005. (Cited on page 119.)

[92] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1106–1114, 2012. (Cited on page 22.)

[93] M. Lades, J. C. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Wurtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *Computers, IEEE Transactions on*, 42(3):300–311, 1993. (Cited on page 21.)

[94] M. Larson, M. Soleymani, P. Serdyukov, V. Murdock, and G. Jones. Working notes proceedings of the mediaeval 2010 workshop. *Pisa, Italy*, 2010. (Cited on page 24.)

[95] S. Lazebnik and M. Raginsky. Supervised learning of quantizer codebooks by information loss minimization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(7):1294–1309, 2009. (Cited on page 18.)

[96] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006. (Cited on pages 19 and 66.)

[97] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. Ieee, 2006. (Cited on pages 23 and 80.)

[98] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision*, 77(1):259–289, 2008. (Cited on pages 18, 62 and 66.)

[99] C. Leistner, H. Grabner, and H. Bischof. Semi-supervised boosting using visual similarity learning. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. (Cited on page 20.)

[100] C. Li and T. Chen. Aesthetic visual quality assessment of paintings. *Selected Topics in Signal Processing, IEEE Journal of*, 3(2):236–252, 2009. (Cited on pages 17, 26 and 123.)

[101] C. Lin and S. Wang. Fuzzy support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):464–471, 2002. (Cited on pages 106 and 109.)

[102] C. Lin and S. Wang. Training algorithms for fuzzy support vector machines with noisy data. *Pattern Recognition Letters*, 25(14):1647–1656, 2004. (Cited on page 106.)

[103] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum. Learning to detect a salient object. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007. (Cited on page 39.)

[104] Y. Liu, D. Zhang, and G. Lu. Region-based image retrieval with high-level semantics using decision tree learning. *Pattern Recognition*, 41(8):2554–2570, 2008. (Cited on page 21.)

[105] D. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. (Cited on pages 6, 27, 36, 61, 62, 64, 74, 77, 83, 92, 113 and 168.)

[106] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1150–1157. Ieee, 1999. (Cited on pages 15 and 16.)

[107] X. Lu, P. Suryanarayan, R. B. Adams Jr, J. L. M. G. Newman, and J. Z. Wang. On shape and the computability of emotions. 2012. (Cited on page 26.)

[108] J. Lukas, J. Fridrich, and M. Goljan. Digital camera identification from sensor pattern noise. *Information Forensics and Security, IEEE Transactions on*, 1(2):205–214, 2006. (Cited on page 142.)

[109] Y. Ma and H. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *2003 multimedia conference: proceedings of the 11th ACM international conference on multimedia, Berkeley, California, USA, November 4-6, 2003*, page 374. Association for Computing Machinery, 2003. (Cited on page 39.)

[110] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the international conference on Multimedia*, pages 83–92. ACM, 2010. (Cited on pages 6, 22, 26, 122, 123 and 168.)

[111] F. Mahmoudi, J. Shanbehzadeh, A.-M. Eftekhari-Moghadam, and H. Soltanian-Zadeh. Image retrieval based on shape similarity by edge orientation autocorrelogram. *Pattern recognition*, 36(8):1725–1736, 2003. (Cited on page 15.)

[112] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. *arXiv preprint arXiv:0809.3083*, 2008. (Cited on page 18.)

[113] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. (Cited on page 22.)

[114] P. K. Mallapragada, R. Jin, A. K. Jain, and Y. Liu. Semiboost: Boosting for semi-supervised learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(11):2000–2014, 2009. (Cited on page 20.)

[115] B. S. Manjunath and W.-Y. Ma. Texture features for browsing and retrieval of image data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(8):837–842, 1996. (Cited on page 15.)

[116] A. Michelson. *Studies in optics*. Dover Pubns, 1995. (Cited on pages 6, 124 and 168.)

[117] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *ICCV 2001, Eighth IEEE International Conference on*, volume 1, pages 525–531. IEEE, 2001. (Cited on page 15.)

[118] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. *Computer Vision, ECCV 2002*, pages 128–142, 2002. (Cited on page 15.)

[119] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International journal of computer vision*, 60(1):63–86, 2004. (Cited on pages 16 and 27.)

[120] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, pages 1615–1630, 2005. (Cited on page 16.)

[121] F. Moosmann, D. Larlus, and F. Jurie. Learning saliency maps for object categorization. In *ECCV International Workshop on The Representation and Use of Prior Knowledge in Vision*. Citeseer, 2006. (Cited on page 37.)

[122] F. Moosmann, B. Triggs, F. Jurie, et al. Fast discriminative visual codebooks using randomized clustering forests. *Advances in Neural Information Processing Systems 19*, pages 985–992, 2007. (Cited on pages 18, 20 and 66.)

[123] H. Müller, P. Clough, T. Deselaers, and B. Caputo. *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*, volume 32. Springer, 2010. (Cited on page 24.)

[124] H. Müller, T. Deselaers, T. Deserno, J. Kalpathy-Cramer, E. Kim, and W. Hersh. Overview of the imageclefmed 2007 medical retrieval and medical annotation tasks. *Advances in Multilingual and Multimodal Information Retrieval*, pages 472–491, 2008. (Cited on page 25.)

[125] N. M. Nasrabadi and W. Li. Object recognition by a hopfield neural network. *Systems, Man and Cybernetics, IEEE Transactions on*, 21(6):1523–1535, 1991. (Cited on page 21.)

[126] D. Navon. Forest before trees: The precedence of global features in visual perception* 1. *Cognitive psychology*, 9(3):353–383, 1977. (Cited on pages 49 and 50.)

[127] R. Nelsen. *An introduction to copulas*. Springer Verlag, 2006. (Cited on pages 68 and 90.)

[128] U. Niaz, M. Redi, C. Tanase, and B. Merialdo. EURECOM at TrecVid 2012: The light semantic indexing task. In *TRECVID 2012, 16th International Workshop on Video Retrieval Evaluation, October 29, 2012, National Institute of Standards and Technology, Gaithersburg, USA*, Gaithersburg, ÉTATS-UNIS, 10 2012. (Cited on pages 14 and 57.)

[129] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2161–2168. Ieee, 2006. (Cited on pages 18, 20, 62 and 66.)

[130] P. Obrador, M. Saad, P. Suryanarayan, and N. Oliver. Towards category-based aesthetic models of photographs. *Advances in Multimedia Modeling*, pages 63–76, 2012. (Cited on pages 25, 118, 121, 122 and 137.)

[131] P. Obrador, L. Schmidt-Hackenberg, and N. Oliver. The role of image composition in image aesthetics. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 3185–3188. IEEE, 2010. (Cited on pages 17 and 25.)

[132] A. Oliva and P. Schyns. Diagnostic Colors Mediate Scene Recognition. *Cognitive Psychology*, 41(2):176–210, 2000. (Cited on pages 51 and 56.)

[133] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3), 2001. (Cited on pages 8, 10, 15, 24, 27, 29, 36, 46, 49, 50, 54, 56, 57, 86, 98, 132, 133, 134 and 171.)

[134] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006. (Cited on pages 49, 51 and 52.)

[135] D. Ormoneit and V. Tresp. Averaging, maximum penalized likelihood and bayesian estimation for improving gaussian mixture probability density estimates. *Neural Networks, IEEE Transactions on*, 9(4):639–650, 1998. (Cited on pages 18, 64 and 66.)

[136] C. Osgood, G. Suci, and P. Tannenbaum. *The measurement of meaning*, volume 47. University of Illinois Press, 1967. (Cited on page 123.)

[137] S. Palmer. *Vision science: Photons to phenomenology*, volume 1. MIT press Cambridge, MA, 1999. (Cited on page 38.)

[138] G. Pass, R. Zabih, and J. Miller. Comparing images using color coherence vectors. In *Proceedings of the fourth ACM international conference on Multimedia*, pages 65–73. ACM, 1997. (Cited on page 14.)

[139] M. Paterno, F. Lim, and W. Leow. Fuzzy semantic labeling for image retrieval. In *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, volume 2, pages 767–770. IEEE. (Cited on pages 106 and 109.)

[140] H. Permuter, J. Francos, and I. Jermyn. A study of gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recognition*, 39(4):695–706, 2006. (Cited on page 21.)

[141] J. Platt. Probabilistic outputs for support vector machines. *Bartlett P. Schoelkopf B. Schurmans D. Smola, AJ, editor, Advances in Large Margin Classifiers*, pages 61–74. (Cited on page 109.)

[142] A. Pronobis, B. Caputo, P. Jensfelt, and H. I. Christensen. A discriminative approach to robust visual place recognition. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 3829–3836. IEEE, 2006. (Cited on page 14.)

[143] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 413–420. IEEE, 2009. (Cited on pages 8, 24, 29, 56, 57, 86, 132, 133, 134 and 171.)

[144] B. Rafkind and S.-F. Chang. Angular radial edge histogram. 2008. (Cited on page 15.)

[145] A. Rao, R. K. Srihari, and Z. Zhang. Spatial color histograms for content-based image retrieval. In *Tools with Artificial Intelligence, 1999. Proceedings. 11th IEEE International Conference on*, pages 183–186. IEEE, 1999. (Cited on page 14.)

[146] M. Redi and B. Merialdo. Marginal-based visual alphabets for local image descriptors aggregation. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 1429–1432. ACM, 2011. (Cited on pages 36, 63, 67, 86, 95, 98 and 99.)

[147] M. Redi and B. Merialdo. Saliency-aware color moments features for image categorization and retrieval. In *Content-Based Multimedia Indexing (CBMI), 2011 9th International Workshop on*, pages 199–204. IEEE, 2011. (Cited on pages 37 and 58.)

[148] M. Redi and B. Merialdo. Saliency moments for image categorization. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 39. ACM, 2011. (Cited on pages 22, 38, 58, 118, 120, 121 and 129.)

[149] M. Redi and B. Merialdo. Exploring two spaces with one feature: kernelized multidimensional modeling of visual alphabets. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, page 20. ACM, 2012. (Cited on pages 63 and 67.)

[150] M. Redi and B. Merialdo. A multimedia retrieval framework based on automatic graded relevance judgments. *Advances in Multimedia Modeling*, pages 300–311, 2012. (Cited on pages 121, 128, 130 and 137.)

[151] M. Redi and B. Merialdo. Where is the interestingness?: retrieving appealing videoscenes by learning flickr-based graded judgments. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1363–1364. ACM, 2012. (Cited on page 125.)

[152] M. Redi and B. Merialdo. Direct modeling of image keypoints distribution through copula-based image signatures. In *Proceedings of the 3rd ACM International Conference on Multimedia Retrieval*, page 20. ACM, 2013. (Cited on page 63.)

[153] M. Redi, B. Merialdo, and F. Wang. Eurecom and ecnu at trecvid 2010: The semantic indexing task. In *In Proceedings of the TRECVID 2010 Workshop*, 2010. (Cited on pages 14, 15, 28, 56, 77, 105, 107 and 113.)

[154] L. Renninger and J. Malik. When is scene identification just texture recognition? *Vision Research*, 44(19):2301–2311, 2004. (Cited on page 49.)

[155] J. Rigau, M. Feixas, and M. Sbert. Conceptualizing birkhoff's aesthetic measure using shannon entropy and kolmogorov complexity. *Computational Aesthetics in Graphics, Visualization, and Imaging*, 2007. (Cited on pages 122, 123 and 124.)

[156] C. Rosenberg, M. Hebert, and H. Schneiderman. Semi-supervised self-training of object detection models. 2005. (Cited on page 20.)

[157] D. Ruderman. The statistics of natural images. *Network: computation in neural systems*, 5(4):517–548, 1994. (Cited on page 124.)

[158] S. R. Safavian and D. Landgrebe. A survey of decision tree classifier methodology. *Systems, Man and Cybernetics, IEEE Transactions on*, 21(3):660–674, 1991. (Cited on pages 21 and 102.)

[159] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part iii: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, 58(13):2126–2144, 2007. (Cited on page 104.)

[160] J. Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. *Anticipatory Behavior in Adaptive Learning Systems*, pages 48–76, 2009. (Cited on page 127.)

[161] P. Schyns and A. Oliva. From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. *Psychological Science*, 5(4):195, 1994. (Cited on pages 15, 50 and 52.)

[162] N. Sebe, Q. Tian, E. Loupias, M. Lew, and T. Huang. Color indexing using wavelet-based salient points. In *Content-based Access of Image and Video Libraries, 2000. Proceedings. IEEE Workshop on*, pages 15–19. IEEE, 2002. (Cited on page 42.)

[163] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(3):411–426, 2007. (Cited on page 54.)

[164] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 300–312, 2007. (Cited on pages 37, 50, 51 and 55.)

[165] B. Silverman. *Density estimation for statistics and data analysis*, volume 26. Chapman & Hall/CRC, 1986. (Cited on pages 65 and 93.)

[166] A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8(1):11, 1959. (Cited on pages 10, 88, 160 and 173.)

[167] M. Slaney. Web-scale multimedia analysis: Does content matter? *MultiMedia, IEEE*, 18(2):12 –15, feb. 2011. (Cited on pages 141 and 178.)

[168] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press. (Cited on pages 8, 22, 24, 56, 74, 87, 99, 105, 121 and 171.)

[169] J. R. Smith and S.-F. Chang. Visualseek: a fully automated content-based image query system. In *Proceedings of the fourth ACM international conference on Multimedia*, pages 87–98. ACM, 1997. (Cited on page 14.)

[170] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics, 2008. (Cited on page 105.)

[171] E. Sormunen. Liberal relevance criteria of trec-: counting on negligible documents? In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 324–330. ACM, 2002. (Cited on pages 103 and 105.)

[172] C. Spampinato, V. Mezaris, and J. van Ossenbruggen. Multimedia analysis for ecological data. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1507–1508. ACM, 2012. (Cited on page 25.)

[173] D. F. Specht. Probabilistic neural networks. *Neural networks*, 3(1):109–118, 1990. (Cited on page 102.)

[174] M. Stricker and A. Dimai. Color indexing with weak spatial constraints. *Storage and Retrieval for Image and Video Databases IV*, 2670, 1996. (Cited on page 42.)

[175] M. Stricker and M. Orengo. Similarity of color images. In *Proceedings of SPIE*, volume 2420, page 381, 1995. (Cited on pages 6, 14, 27, 31, 36, 41, 43, 55, 56, 113 and 168.)

[176] H. Sturges. The choice of a class interval. *Journal of the American Statistical Association*, 21(153):65–66, 1926. (Cited on page 65.)

[177] C. Suchy-Dicey. What the Gist? A Case Study in Perception and Attention. (Cited on page 51.)

[178] J. Sun, X. Zhang, J. Cui, and L. Zhou. Image retrieval based on color distribution entropy. *Pattern Recognition Letters*, 27(10):1122–1126, 2006. (Cited on page 14.)

[179] K. Svore, L. Vanderwende, and C. Burges. Enhancing single-document summarization by combining ranknet and third-party sources. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 448–457, 2007. (Cited on page 105.)

[180] M. Swain and D. Ballard. Color indexing. *International journal of computer vision*, 7(1):11–32, 1991. (Cited on pages 14, 22 and 41.)

[181] M. Szummer and R. W. Picard. Indoor-outdoor image classification. In *Content-Based Access of Image and Video Database, 1998. Proceedings., 1998 IEEE International Workshop on*, pages 42–51. IEEE, 1998. (Cited on page 14.)

[182] H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *Systems, Man and Cybernetics, IEEE Transactions on*, 8(6):460–473, 1978. (Cited on page 14.)

[183] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118. ACM, 2001. (Cited on page 22.)

[184] A. Torralba, A. Oliva, M. Castelhano, and J. Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, 113(4):766–786, 2006. (Cited on pages 37 and 39.)

[185] T. Tuytelaars and C. Schmid. Vector quantizing feature space with a regular lattice. In *11th IEEE International Conference on Computer Vision (ICCV '07)*, pages 1–8, Rio de Janeiro, Brazil, 2007. IEEE Computer Society. (Cited on pages 18, 66 and 84.)

[186] T. Tuytelaars and L. Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. In *British Machine Vision Conference*, volume 2, page 4, 2000. (Cited on page 15.)

[187] M. Unser. Texture classification and segmentation using wavelet frames. *Image Processing, IEEE Transactions on*, 4(11):1549–1560, 1995. (Cited on pages 6, 15, 27, 31, 36, 56, 58, 113 and 168.)

[188] J. van Gemert. Exploiting photographic style for category-level image classification by generalizing the spatial pyramid. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 14. ACM, 2011. (Cited on page 121.)

[189] J. van Gemert. Exploiting photographic style for category-level image classification by generalizing the spatial pyramid. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 14. ACM, 2011. (Cited on page 132.)

[190] J. van Gemert, C. Veenman, A. Smeulders, and J. Geusebroek. Visual word ambiguity. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(7):1271–1283, 2010. (Cited on pages 18 and 66.)

[191] R. H. van Leuken, L. Garcia, X. Olivares, and R. van Zwol. Visual diversification of image search results. In *Proceedings of the 18th international conference on World wide web*, pages 341–350. ACM, 2009. (Cited on page 20.)

[192] L. Viet Tran. *Efficient image retrieval with statistical color descriptors*. PhD thesis, Linköping, 2003. (Cited on page 14.)

[193] D. Walther, U. Rutishauser, C. Koch, and P. Perona. On the usefulness of attention for object recognition. In *Workshop on Attention and Performance in Computational Vision at ECCV*, pages 96–103. Citeseer, 2004. (Cited on page 37.)

[194] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3360–3367. IEEE, 2010. (Cited on pages 19 and 66.)

[195] L. Wang, K. Chan, and Z. Zhang. Bootstrapping svm active learning by incorporating unlabelled images for image retrieval. In *IEEE computer society conference on computer vision and pattern recognition*, volume 1. IEEE Computer Society; 1999, 2003. (Cited on page 22.)

[196] W. Wang and Y. Yu. Image emotional semantic query based on color semantic description. In *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, volume 7, pages 4571–4576. IEEE, 2005. (Cited on page 123.)

[197] W. Wei-ning, Y. Ying-lin, and J. Sheng-ming. Image retrieval by emotional semantics: A study of emotional space and feature extraction. In *Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on*, volume 4, pages 3534–3539. IEEE, 2006. (Cited on page 26.)

[198] G. G. Wilkinson. Results and implications of a study of fifteen years of satellite image classification experiments. *Geoscience and Remote Sensing, IEEE Transactions on*, 43(3):433–440, 2005. (Cited on page 25.)

[199] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1800–1807. IEEE, 2005. (Cited on page 18.)

[200] C. Won, D. Park, and S. Park. Efficient use of MPEG-7 edge histogram descriptor. *Etri Journal*, 24(1):23–30, 2002. (Cited on pages 15, 27, 31, 36, 56, 58, 113, 124 and 129.)

[201] L. Wong and K. Low. Saliency-enhanced image aesthetics class prediction. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 997–1000. Ieee, 2009. (Cited on pages 25, 121 and 129.)

[202] L. Wong and K. Low. Saliency-enhanced image aesthetics class prediction. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*. Ieee, 2009. (Cited on page 123.)

[203] J. Wu and J. Rehg. Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 630–637. IEEE, 2009. (Cited on page 80.)

[204] Q. Wu, C. Zhou, and C. Wang. Content-based affective image classification and retrieval using support vector machines. *Affective Computing and Intelligent Interaction*, pages 239–247, 2005. (Cited on page 25.)

[205] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE, 2010. (Cited on pages 8, 16, 22, 24, 29, 86, 98, 121, 132, 133, 134 and 171.)

[206] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition,*

2009. CVPR 2009. IEEE Conference on, pages 1794–1801. IEEE, 2009. (Cited on pages 18 and 66.)

[207] A. Yarbus, B. Haigh, and L. Rigss. *Eye movements and vision*, volume 2. Plenum press New York, 1967. (Cited on page 38.)

[208] K. Yu, T. Zhang, and Y. Gong. Nonlinear learning using local coordinate coding. *Advances in Neural Information Processing Systems*, 22:2223–2231, 2009. (Cited on page 66.)

[209] H. Zhang, E. Augilius, T. Honkela, J. Laaksonen, H. Gamper, and H. Alene. Analyzing emotional semantics of abstract art using low-level image features. *Advances in Intelligent Data Analysis X*, pages 413–423, 2011. (Cited on page 26.)

[210] L. Zhang, F. Lin, and B. Zhang. Support vector machine learning for image retrieval. In *Image Processing, 2001. Proceedings. 2001 International Conference on*, volume 2, pages 721–724. IEEE, 2001. (Cited on pages 22, 28 and 81.)

[211] Z. Zheng, K. Chen, G. Sun, and H. Zha. A regression framework for learning ranking functions using relative relevance judgments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 287–294. ACM, 2007. (Cited on pages 105 and 115.)

[212] Z.-H. Zhou, K.-J. Chen, and Y. Jiang. Exploiting unlabeled data in content-based image retrieval. *Machine Learning: ECML 2004*, pages 525–536, 2004. (Cited on page 20.)

Appendix A

Nouvelles Méthodes pour la Recherche Sémantique et Esthétique d'Informations Multimédia

A.1 Résumé

Comment pouvons-nous permettre aux ordinateurs de voir le monde comme nous le faisons, et de déduire des milliers de vrais mots à partir d'une image numérique? Comment pouvons-nous fabriquer des machines qui transforment *pixels* en *sémantique* et en autres informations pertinentes concernant le contenu de l'image? Un ensemble de solutions est proposé par une discipline de récherche appellée **Multimedia Information Retrieval** (MMIR), qui permet de combler l' écart entre les valeurs au niveau du pixel et la compréhension sémantique de l'image.

Les techniques MMIR visent à extraire automatiquement des informations sur les objets, les scènes, les émotions représentées dans l'image, sur la base de l'analyse de son *contenu* visuel. Les chercheurs MMIR conçoivent des systemès qui apprennent comment lier les valeurs de pixels, qui sont résumées dans des *charactéristiques visuelles* non-redondantes, à un ensemble de *concepts intelligibles*.

Un système MMIR peut automatiquement **classer** ou **récuperer** une image sur la base de son aspect visuel, en **reconnaissent** automatiquement le contenu et les propriétés de l'image. Pour une image donnée, un cadre de classification délivre un ensemble de courtes descriptions de son contenu, que nous appelons *labels* ou *annotations, étiquette*. Les étiquettes peuvent être considérées comme des jugements positifs ou négatifs sur la **pertinence** d'une image par rapport à un concept donné (par exemple "il y a un chat", "il n'y a pas de souris"). Compte tenu des annotations assignées automatiquement, les systèmes MMIR peuvent aussi aller au-delà de classification, permettant **la recherche des images par leur contenu**.

L' *objectif* d'un cadre de récupération multimédia dépend de la nature des annotations assignées à l'image. Unensem système MMIR peut classer ou récupérer automatiquement des images non seulement en fonction de leur *sémantique*, mais aussi en se basant sur les *émotions* qu'elle suscite, le degré de *beauté* de l'image donnée ou l'interêt de l'image.

Comment est-ce-que ça fonctionne? Dans ce qui suit, nous allons illustrer un système MMIR avec une simple structure pyramidale ayant plusieurs *niveau d' ab-*

stractions, où à la base il se déroule le traitement au niveau du pixel pour l'extraction de caractéristiques, et où le sommet correspond au niveau de l'application, à savoir l'attribution de l'étiquette.).

Tout commence par une **groundtruth**, un ensemble d' images annotées avec leurs étiquettes correspondantes, à savoir des annotations attribuées manuellement reflétant la présence/absence de concepts donnés.

Compte tenu de ce groundtruth, les systèmes MMIR suivent un proces standard:

1. Ils extraient en premier lieu un ensemble de caractéristiques (niveau 0), à savoir un ensemble restreint de valeurs très instructifs concernants l'aspect visuel de l'image (par exemple,quelle est la couleur la plus dominante dans l'image?).

2. Ils utilisent ensuite ces caractéristiques, ou leur version réduite (niveau 1), avec les étiquettes de l'image, en input pour des machines d'apprentissage, qui apprennent un modèle (niveau 2) capable d'associer les valeurs des caractéristiques a la présence/absence d'un concept donné.

3. Un tel système intelligent sera (niveau 3) alors capable d'annoter automatiquement les nouvelles, *inconnues*, images de test compte tenu de leurs caractéristiques et du modèle calculé.

Les systèmes MMIR sont donc des cadres complexes dont les performances dépendent de nombreux facteurs: le type de caractéristiques visuelles utilisées, le cadre d'apprentissage, les techniques de réduction de la redondance, la qualité des annotations, l'application, etc. Inspiré par des disciplines très diverses, dans cette thèse, nous allons discuter de plusieurs contributions différentes qui améliorent la qualité des systèmes MMIR.

Pour chacun des facteurs qui jouent un rôle important dans un cadre MMIR, nous présentons de nouvelles techniques qui améliorent les *performances* globales d'un système de recherche multimédia, à savoir l' *exactitude* des étiquettes prédites, ou la précision des résultats récupérés. Nous allons apprendre comment ameéliorer les performances des systèmes MMIR pour l'analyse sémantique en construisant de nouvelles fonctionnalités de bas niveau, de nouveaux agrégateurs de caracteristiques, et un cadre d'apprentissage particulier. Nous allons nous concentrer principalement sur deux types d'applications pour nos études MMIR: l' **analyse sémantique**, à savoir l'extraction automatique des étiquettes decrivantes les objets, les scènes ou de concepts généraux, et l'**analyse esthétique**, à savoir la prédiction automatique du degré de beauté de l'image.

Pour les niveaux 0, 1 et 2 (extraction de caractéristiques, regroupement et apprentissage), nous concevons un ensemble de nouvelles techniques que nous testons sur des bases de données générales pour l'analyse comparative de techniques de classification/récupération sémantique. Dans nos premières études, l'application principale de nos techniques est donc principalement l'analyse sémantique. Au niveau 3, nous allons au-delà des applications sémantiques pures et nous construisons un

système de prédiction de l'esthetique de l'image et de sa beauté, en intégrant la plupart des techniques que nous avons proposées dans les niveaux inférieurs dans un système MMIR que nous construisons pour l'analyse esthétique. Dans ce qui suit, nous allons voir un aperçu de nos contributions.

- **Niveau 0: Caractéristiques Hybrides à base de saillance pour catégorisation de l'image (Chapitre 3)** Les caractéristiques de bas niveau sont les racines de tous les systèmes intelligents d'analyse de l'image. Les caractéristiques ou *signatures, descripteurs* sont des ensembles descriptifs de chiffres résumantes les propriétés visuelles importantes de l'image.
  Dans notre contribution a ce niveau, nous constructisons un ensemble de nouvelles fonctionnalités de bas niveau, très discriminants pour l'analyse sémantique, inspirées par la théorie de la perception visuelle.
  Les caracteristiques sémantiques peuvent être classés en deux groupes: *locales* et *globales*. Alors que les premiéres sont très instructives sur les détails et les contours d'images et robustes aux transformations, les caractéristiques globales decrivent le comportement général de l'image, au détriment d'une certaine précision lorsque les conditions d'éclairage / de rotation changent. Malgré leur efficacité, l'inconvénient majeur des caractéristiques locales est leur coût de calcul et l'obligation au regroupement. D'autre part, les caractéristiques globales sont de faible dimensions et rapides à calculer.
  Nos caractéristiques se situent à un point intermédiaire entre les deux approches mentionnées: nous concevons un ensemble de descripteurs **hybrides**, à savoir des caractéristiques *globales* de basse dimensionnalité qui intègrent certaines informations localement analysées. L'information locale est integrée dans un descripteur global en utilisnt des **cartes de saillance de l'image**, à savoir des matrices avec des valeurs d'intensité plus élevée correspondant aux régions qui ont la plus grande probabilité d'attirer l'attention de l'observeur de l'image. Nous présentons deux caractéristiques suivantes cet approche hybride: le descripteur **Saliency-Aware Color Moments**, à savoir une amélioration d'une caractéristique de couleur a travers des techniques de saillance, et le **Saliency Moments**, à savoir un descripteur basé sur le traitement de la distribution locale de la saillance.
  Nous testons l'efficacité de nos descripteurs pour la catégorisation d'objets, la reconnaissance de scènes et la récupération de la vidéo, et nous montrons que non seulement ils surpassent les fonctionnalités existantes pour MMIR, mais que aussi la saillance apporte d'information *complémentaire* aux descripteurs existants généralement utilisés dans MMIR.

- **Niveau 1: Agrégation des caractéristiques locales à travers l'analyse marginale et les copulae (Chapitre 4)** Dans certains cas, les fonctions de bas niveau ne peuvent pas être utilisées directement comme input pour le troisième niveau de la pyramide, à savoir le cadre d'apprentissage, mais elles ont besoin de passer par une étape intermédiaire qui les agrège en une signature visuelle compacte. La solution générale est donc de *regrouper* tous les

descripteurs des images locales dans une signature de longueur fixe représentant le comportement des points-clés de l'image. Ce résultat est obtenu avec un premièr **codage** d'un ensemble d'entraînement des descripteurs locaux en un nombre des valeurs plus petit, à savoir un code universel partagé avec un nombre donné de "mots visuels". Pour obtenir une nouvelle image, la quantité variable de descripteurs est ensuite **groupée** en une nouvelle signature de l'image qui agrége leurs propriétés en regardant leur distribution, compte-tenu du dictionnaire commun, ayant donc une dimension fixe égal au nombre de mots visuels.
Pour ce niveau, nous étudions le regroupement et l'encodage, en proposant un ensemble de nouvelles techniques pour l'agrégation des descripteurs locaux rapide et efficace, inspiré par des techniques de modélisation provenant du statistique économique.
Notre observation est la suivante: les techniques traditionnelles pour l'agrégation des descripteurs ont besoin de procédures coûteuses pour les techniques de codage. Il a a été prouvé que cette approche est très efficace pour les applications MMIR. Cependant, un des inconvénients majeurs est son intrinsèque coût de calcul et stockage, ainsi que la perte d'information due à l'étape de regroupement.
Les solutions que nous proposons diffèrent sensiblement de l'approche traditionnelle, en améliorent l'efficacité et la précision des méthodes traditionnelles de regroupement. Tout d'abord, nous concevons le **MEDA** descripteur, qui décrit le comportement des descripteurs locaux basé sur les approximations de leur distribution marginale, conduisant à une signature de l'image extrêmement léger a calculer mais qui garde une grande précision pour la classification et l'extraction. Nous améliorons cette méthode en construisant **Multimeda**, un noyau pour le Support Vector Machines qui est capable d'extraire une probabilité multidimensionnelle des descripteurs des images en calculant le produit des approximations marginales stockées dans MEDA. Enfin, nous utilisons la théorie des Copulae [166] pour calculer la probabilité réelle conjointe des descripteurs locaux, basé sur l'information marginale pure stockée dans MEDA. Nous modélisons la distribution multivariée des points-clés de l'image sans impliquer un processus de codage dans l'espace multidimensionnel. Le vecteur **COMS** résultant s'avère être beaucoup plus efficace que les techniques de l'état de l'art pour l'agrégation locale appliquée à la reconnaissance des scènes et à la récupération de la vidéo.

- **Niveau 2: Un cadre d'extraction multimédia basé sur des jugements automatiques de pertinence gradué (Chapitre 5)** Aprés avoir obtenu la signature de longueur fixe, la prochaine étape vers la compréhension automatique des caractéristiques de l'image est l'étape d'**apprentissage**. A ce niveau, nous utilisons des cadres d'apprentissage supervisé qui *apprennent* comment faire la distinction entre les images contenant des caractéristiques différentes (i.e. contenu ou esthétique différente), puis *prédisent* les étiquettes correspon-

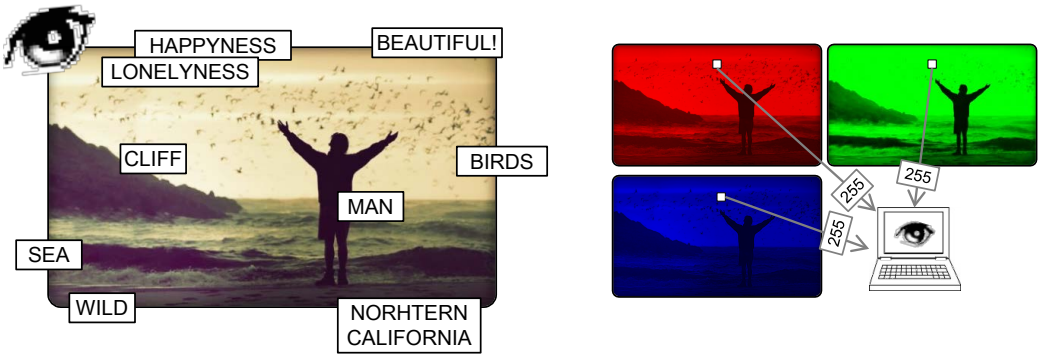
dant aux nouvelles images.
Nous allons nous concentrer sur la conception d'un cadre d'apprentissage utilisé pour la recuperation video, en introduisant dans MMIR certains concepts de la recherche d'information web.
La principale observation est que, dans l'étape d'apprentissage des systèmes MMIR, les étiquettes d'entraînement sont attribuées sur un échelle binaire (pertinent / non pertinent). Cela signifie que les annotations générées par l'utilisateur identifient la simple présence ou absence d'un concept dans les données visuelles, sans tenir compte des options intermédiaires. Cependant, une image peut être pertinente pour une catégorie sémantique avec des degrés divers, en fonction de la façon dont ce concept est représenté dans l'image.
Différent des cadres les plus courants, dans le Chapitre 5, nous construisons un cadre d'apprentissage qui est capable de gérer des **jugements de pertinence classés**, à savoir de degrés d'annotations differents, reflétant les différents niveaux de pertinence d'un image par rapport à un concept donné. Étant donné que l'annotation manuelle est un processus coûteux et imprécis, afin de construire rapidement des groundtruth graduées, nous proposons une façon de réévaluer les bases de données binaires sans impliquer un effort manuel: nous attribuons automatiquement un degré de pertinence gradé fiable (non, faiblement, moyen, très pertinent) à chaque image, en fonction de sa position par rapport à l'hyperplan dessiné par une Support Vector Machines dans l'espace des caractéristiques.
Nous testons l'efficacité de notre système sur deux bases de données à grande échelle, et nous montrons que notre approche surpasse les cadres binaires traditionnelles pour la récupération de la vidéo.

- **Niveau 3: Au-delà de la sémantique pure: la synergie avec l'analyse esthétique (Chapitre 6)** Que reste-t-il dans la chaîne de recherche d'information multimédia? Il nous manque l'étape de prédiction, à savoir l'attribution automatique des étiquettes sémantiques/esthétiques à des nouvelles images inédites. A ce niveau de la chaîne, nous **testons** les performances de notre modèle de calcul de base en calculant la précision des prédictions sur les nouvelles images.
A ce niveau de la chaîne, un aspect important est le type d'application pour lequel le système est conçu. Traditionnellement, les cadres MMIR sont conçus par des chercheurs pour **l' analyse sémantique générale**, c'est à dire la reconnaissance des objets et de scènes. Mais les techniques d'extraction et de classification peuvent aussi aller au-delà de la prédiction sémantique générale, en appliquant, par exemple, les techniques d'analyse sémantique à des domaines plus étroits, **domaines spécifiques** qui bénéficient de l'automatisation apportée par MMIR (par exemple, les images médicales, les vidéos de cuisine, les images satellites . . .). En outre, les techniques MMIR peuvent être utilisées pour prédire de l'information qui n'est pas strictement liée à les objets de l'image et les scènes, abandonnant la sémantique et se con-

centrant sur les émotions, la valeur **artistique** ou **esthétique**.
Dans le dernier Chapitre technique, nous étudions une nouvelle application émergente dans le domaine de la récupération Multimédia: l'analyse esthétique. Nous appliquons au problème de l'évaluation automatique de l'interet de l'image des nouvelles techniques de MMIR, aussi que certaines existantes, traditionnellement utilisées pour des tâches sémantiques. Nous réutilisons beaucoup de leçons apprises dans les Chapitres précédents et nous lesappliquons à la prédiction du degré esthétique du contenu visuel. Nous ajoutons des informations sémantiques à des cadres d'analyse esthétique, et nous observons l'amélioration apportée par les caractéristiques basées sur le contenu de l'image a la prédiction de la beauté globale.
Par ailleurs, nous explorons également dans l'autre sens: est-que les outils d'analyse esthétiques sont utiles pour l'analyse sémantique? Nous étudions l'importance de l'analyse esthétique pour les applications sémantiques, en testant l'efficacité des caractéristiques esthétiques d'un cadre MMIR de reconnaissance de scène.
Dans notre contribution au niveau de l'application, nous enrichissons donc l'analyse visuelle sémantique et esthétique en explorant la synergie de ces deux applications pour MMIR. L'idée principale est que l'analyse sémantique et esthétique sont deux applications étroitement liées dans la récupération multimédia. Nous montrons les avantages et les limites de cette synergie, et proposons quelques améliorations allant dans ce sens.

La propriété commune de toutes les méthodes que nous allons proposer, est leur **multidisciplinarité**: nous concevons nos techniques en empruntant des technologies et des principes d'une grande variété de sources qui ne sont pas directement liées à la vision par ordinateur. Nous croyons que la découverte de nouvelles ressources pour l'analyse d'image peut apporter non seulement des améliorations substantielles par rapport aux techniques existantes, mais elle peut génèrer également des techniques qui fournissent des informations complémentaires sur les propriétés d'image, et qui peuvent être donc utilisées en **combinaison** avec des techniques existantes.

A.2 Introduction



"Une image vaut mille mots". Considération incontestable, lorsqu'il s'agit d'êtres humains et leur *système de vision humain*. Quand on regarde une image, nous identifions des objets, des actions, des scènes, des repères, mais nous lions le contenu de l'image avec nos souvenirs, avec un ensemble d'émotions, de films, de livres, de sensations .... Toutefois, lorsqu'il s'agit de *systèmes de vision artificiels*, le scénario change. Pour un système de vision artificielle, la même image vaut tout simplement mille ou plus *pixels*, à savoir triplets de nombres discrets représentant le montant de vert, de rouge et de bleu dans un point de l'image.

Comment pouvons-nous permettre aux ordinateurs de voir le monde comme nous le faisons, et en déduire des milliers de vrais mots a partir d'une image numérique? Comment pouvons-nous fabriquer des machines qui transforment *pixels* en *sémantique* et d'autres informations pertinentes concernant le contenu de l'image? Un ensemble de solutions est proposé par **Multimedia Information Retrieval** (MMIR), qui permet de combler l' écart entre les valeurs au niveau du pixel et la compréhension sémantique de l'image.

Les techniques MMIR visent à extraire automatiquement des informations sur les objets, les scènes, les émotions représentées dans l'image, sur la base de l'analyse de son *contenu* visuel. Les chercheurs MMIR conçoivent des systemès qui apprennent comment lier les valeurs de pixels, qui sont résumées dans des *charactéristiques visuelles* non-redondantes, à un ensemble de *concepts intelligibles*.

Un système MMIR peut automatiquement **classer** ou **récuperer** une image sur la base de son aspect visuel, en **reconnaissent** automatiquement le contenu et les propriétés de l'image. Compte tenu d'une image, un cadre de classification délivre un ensemble de courtes descriptions de son contenu, que nous appelons *labels* ou *annotations, étiquette*. Les étiquettes peuvent être considérées comme des jugements positifs ou négatifs sur la **pertinence** d'une image par rapport à un concept donné (par exemple "il y a un chat", "il n'y a pas une souris"). Compte tenu des annotations assignées automatiquement, les systèmes MMIR peuvent aussi aller au-delà de classification, permettant **la recherche des images par le contenu**.

L' *objectif* d'un cadre de récupération multimédia, dépend de la nature des

annotations assignées à l'image. Un système MMIR peut classer automatiquement ou récupérer des images non seulement en fonction de leur *sémantique*, mais aussi basé sur les *émotions* qu'il suscite, le degré de *beauté* de l'image donnée ou combien l'image est *intéressante*.

Les systèmes MMIR sont donc des cadres complexes dont les performances dépendent de nombreux facteurs: le type de caractéristiques visuelles utilisées, le cadre d'apprentissage, les techniques de réduction de la redondance, la qualité des annotations, l'application, etc. Inspiré par des disciplines très diverses, dans cette thèse, nous allons discuter de plusieurs contributions différentes qui l'améliorent la qualité des systèmes MMIR, par exemple des caractéristiques globales fondées sur la saillance, des cadres d'apprentissage à pertinence graduée. Pour chacun des facteurs qui jouent un rôle important dans un cadre MMIR, nous présentons des nouvelles techniques qui améliorent les *performances* globales d'un système de recherche multimédia, à savoir la *exactitude* des étiquettes prédit, ou la précision des résultats récupérés. Nous allons nous concentrer principalement sur deux types d'applications pour nos études MMIR: l' **analyse sémantique**, à savoir l'extraction automatique des étiquettes decrivantes les objets, les scènes ou de concepts généraux, et l'**analyse esthétique**, à savoir la prédiction automatique du degré de beauté de l'image.

Dans le reste de ce Chapitre introductif, nous allons donner un aperçu général de la motivation et de la structure de cette thèse. Nous allons d'abord expliquer en détail l'importance (voir la Section A.2.1.) Du MMIR. Nous donnerons ensuite un aperçu des processus et les étapes clés qu'un système MMIR necessite pour assigner automatiquement des annotations aux images, et enfin nous mettrons en évidence nos contributions dans le domaine (Secc. A.2.2 et A.2.3) .

A.2.1 L'importance de la Recherche d'Information Multimédia Aujourd'hui

En 2012, les utilisateurs de Flickr [1] ont téléchargé sur le site de gestion de photos une quantité impressionnante de fichiers, 517.863.947 ajouts. Et Flickr est juste un des nombreux services en ligne qui permettent de partager du contenu audiovisuel numérique.

Nous vivons dans un monde visuel numérique : actualités, films, photos, images et vidéos générées par l'utilisateur ... Avec la large diffusion des appareils portables et des connexions internet à haut débit, nous produisons, modifions et partageons des quantités énormes des images et de vidéos presque *instantanément*, permettant à autres utilisateurs d'accéder au contenu visuel, a partir de tous les points du monde interconnecté au en même temps. Accèder à ce contenu signifie avoir un oeil sur le monde: nous développons des idées et des concepts en recevant et transmettent des informations visuelles, nous partageons des émotions et des souvenirs à travers des images parce que "Une image vaut mille mots".

Des millions d'utilisateurs chaque jour explorent cet espace multimédia, en cherchant dans les collections multimédias les éléments multimédias qui mieux s'adaptent

[1] www.flickr.com

à leurs besoins. Et Multimedia Information Retrieval aide les utilisateurs à explorer cet espace.

En effet, afin de mieux organiser, rechercher et sélectionner des parties de cette énorme quantité des informations visuelles, nous avons besoin d'outils efficaces et efficients pour l'indexer et récupérer les éléments de ces collections de données. Une des façons les plus intuitives pour explorer l'information visuelle est de rechercher et de sélectionner les images en fonction de leur *contenu*, en recherchent de données visuelles qui contiennent des *sémantique* spécifiques, par exemple, "images avec des chats". On pourrait aussi réchercher des images selon son *degré de beauté* ou selon les émotions que les images génèrent, par exemple, "images effrayantes".

Cependant, en pratique, l'exploration sémantique, esthétique et affective n'est pas si simple. Une image vaut mille mots, mais ces mots sont dans nos yeux humains, tandis que des milliers de pixels formant une image n'ont pas de sens pour une machine, si la machine ne sait pas comment les comprendre. Le monde visuel numérique est très difficile d'ordonner et d'organiser: il nous faut donc des procédures automatiques pour soutenir une telle exploration.

Bien que les documents textuels peuvent être indexés et récupérés, par exemple, en comptant les fréquences des mots, c'est une autre histoire lorsqu'il s'agit des images numériques, car dans le contenu visuel il n'y a pas de " mots" réels à indexer. Une solution serait d'ordonner les données visuelles manuellement, en attribuant un ensemble de mots à chaque image décrivant leurs caractéristiques, puis les classer et récupérer sur la base de ces descriptions. Toutefois, compte tenu du volume de données multimédia que nous traitons, il est pratiquement impossible de demander aux humains d'étiqueter tout le monde visuel numérique à la main.

La solution générale, adoptée par les services en ligne pour la recherche des images, est de travailler avec des informations textuelles relatives aux données visuelles. Pour une requête textuelle, tels systèmes recherchent des images pertinentes compte tenu de leur informations *contextuelle*, à savoir tout le texte liée à l'image, mais provenant de sources externes, puis classent les images en fonction de leur pertinence par rapport aux concepts exprimés dans la requête. Par exemple, dans les moteurs de recherche des images Web populaires, tels que Yahoo! ou Google, les sémantiques des images sont déduites étant donné les concepts intelligibles inférables du texte entourant l'image dans les pages Web liées à l'image. Une autre approche populaire, utilisé par les collections multimédias en ligne plus courants tels que Flickr, YouTube ou Facebook, est de récupérer et organiser les images fournies par les "tags", courtes étiquettes textuelles qui sont generées par les propriétaires de l'image et qui en quelque sorte reflètent certaines propriétés visuelles telles que le contenu, la localisation, les émotions, etc.

Les deux approches présentent plusieurs inconvénients pratiques. Tout d'abord, les concepts infèreés du texte environnant de l'image sont souvent peu fiables lorsqu'il s'agit de sémantique complexe, et en particulier avec les émotions. Alors que les tags générées par l'utilisateur peuvent être moins bruyant, l' étiquetage manuel sémantique des données visuelles prend du temps et peut être souvent incomplète, ou la plupart du temps totalement absent. En outre, les informations contextuelles

discriminative n'apparaît que dans les images web, tandis que pour les collections des images hors ligne la seule source d'information qui peut aider à la catégorisation de l'image est dans ses pixels. Afin de déduire les milliers de mots exprimées par

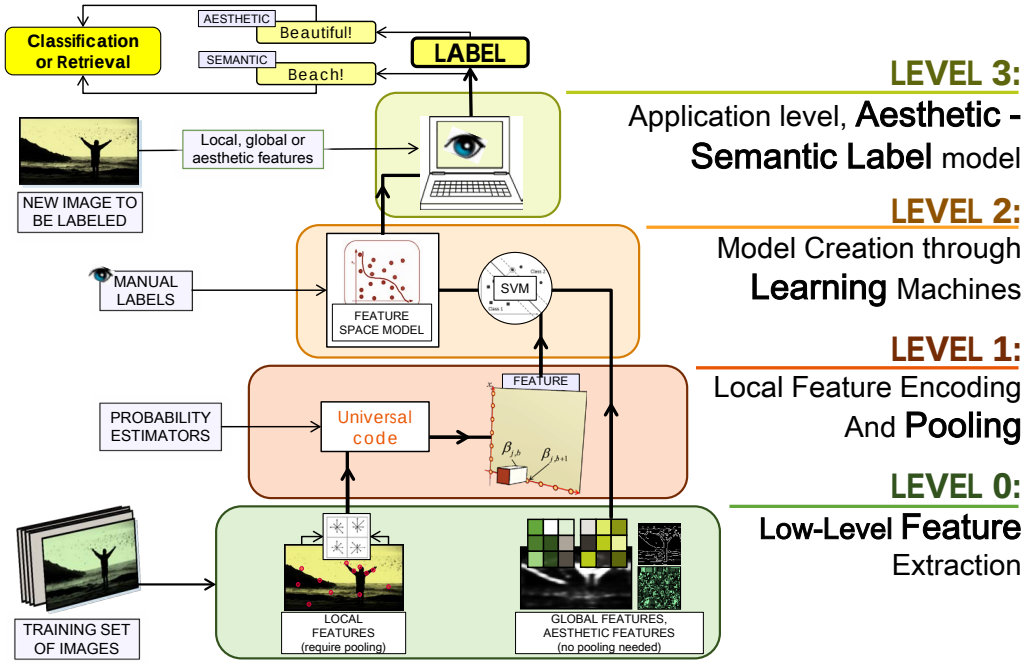


Figure A.1: Nous représentons un cadre de récupération multimédia en utilisant une structure pyramidale: de l'extraction de caractéristiques a la prédiction de l'étiquette et le classement de l'image

une image, nous aurions besoin donc des outils *intelligents* qui puissent automatiquement déduire le contenu d'un image a partir de son aspect visuel: nous avons besoin de cadres qui modèlent le système de vision humain, et qui puissent *reconnaître* les objets perceptibles, les scènes, l'esthétique et les émotions représentées dans l'image.

Multimedia Information Retrieval est une discipline de recherche portant sur ces questions. MMIR étudie en effet la façon de lier les valeurs d'intensité de pixels d'une image avec son sens général. Les systèmes de récupération Multimédia étiquettent automatiquement les images avec des informations sur leur contenu, en utilisant des machines d'apprentissage qui "voient" la sémantique, l'esthétique et les émotions générées par l'image, et qui les traduisent en mots intelligibles.

Comment ça marche? Dans ce qui suit, voir Sec. A.2.2, nous allons illustrer un système MMIR avec une simple structure pyramidale ayant des *niveau des abstractions*, où la base il y le traitement au niveau du pixel pour l'extraction de caractéristiques, et où le sommet correspond au niveau de l'application, à savoir l'attribution de l'étiquette (voir figure A.1.). Tout commence par un **groundtruth**, une formation ensemble des images annotées avec leurs étiquettes correspondantes, à savoir

des annotations attribuées manuellement reflétant la présence/absence de concepts donnés. Compte tenu de ce groundtruth, les systèmes MMIR extrait en premier un ensemble de caractéristiques (niveau 0), à savoir un petit ensemble de valeurs très instructifs concernant l'aspect visuel de l'image (par exemple, c'est laquelle la couleur la plus dominante dans l'image?). Ils utilisent ensuite ces caractéristiques, ou leur version réduite (niveau 1), avec les étiquettes de l'image, en input pour des machines d'apprentissage, qui apprennent un modèle (niveau 2) capable d'associer les valeurs des caractéristiques a la présence/absence d'un concept donné. Un tel système intelligent sera (niveau 3) alors capable d'annoter automatiquement les nouvelles, *inconnues*, images de test compte tenu de leurs caractéristiques et du modèle calculé.

Pour chaque niveau de la pyramide MMIR, dans cette thèse, nous donnons un aperçu général des techniques et nous concevons un ou plusieurs solutions novatrices qui visent à enrichir l'analyse visuelle globale. Nous allons apprendre comment augmenter les performances des systèmes MMIR pour l'analyse sémantique en construisant de nouvelles fonctionnalités de bas niveau, nouveaux agrégateurs des caracteristiques, et un cadre d'apprentissage particulier. Nous allons ensuite réutiliser les leçons apprises dans nos études d'analyse sémantique pour construire un système d'archivage multimédia pour l'analyse esthétique. Nous allons voir un aperçu de nos contributions à la Sec. A.2.3.

A.2.2 Comment ça Marche? La Pyramide des Analyse des Images

Dans cette thèse, nous exprimons un système de recherche multimédia come un cadre pyramidal stratifié. Dans le **pyramide MMIR**, chaque niveau correspond à une étape différente du processus de transformation de l'information visuelle. Chaque niveau re-traite les signaux sortent de niveaux plus bas, des pixels de l'image brute à la compréhension automatique du sens de l'image. Plus la couche est élevé, plus haut le niveau d'abstraction, des valeurs entières de pixels signification discrets, à l'étiquette de l'image sémantique/esthétique intelligible pour l'homme. Plus le niveau est élevé, moins grande est la quantité d'information traitée, à partir de la carte complète de pixel à l'étiquette de l'image simple. Dans ce qui suit, nous allons jeter un oeil aux caractéristiques de chaque couche.

Niveau 0: Les caractéristiques de bas niveau

Les caractéristiques de bas niveau sont les racines de tous les systèmes intelligents d'analyse de l'image. Les caractéristiques ou *signatures, descripteurs* sont des ensembles descriptives des chiffres résumant les propriétés visuelles importantes de l'image. Cela signifie que les images avec sémantique ou esthétique similaires devraient avoir des caractéristiques de bas niveau similaires. Les caractéristiques donc aident la machine à percevoir la similitude ou la dissemblance entre les images comme les humains font.

A ce stade du processus MMIR chaque pixel de l'image est examiné, et les informations pertinentes sont résumées dans un petit ensemble de nombres, stockée

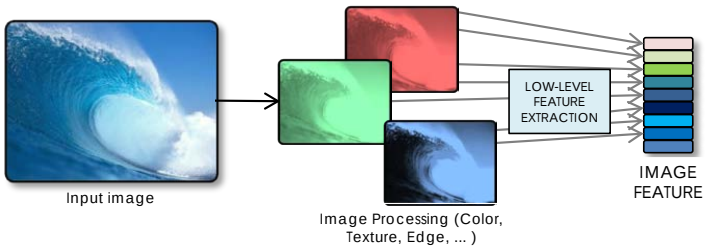


Figure A.2: Extraction de caractéristiques: un petit ensemble de nombres décrivant les propriétés de l'image sont extraites des pixels de l'image

dans la signature de l'image finale. Les techniques d'extraction de caractéristiques peuvent etre variées: de la simple moyenne de pixels ou points, à la détection de bords, de coins et de propriétés de texture. En général, les types de caractéristiques extraites changent avec l'application du système MMIR. Les informations pixel peuvent être "pertinent" pour une tâche avec des degrés différents: par exemple, la valeur de la *contraste* d'une image donnée nous en dit plus sur les caractéristiques esthétiques de l'image plutôt que sur les objets représentés.

Mais quelles caractéristiques sont importantes pour la sémantique et l' esthétique? Les caractéristiques sémantiques sont généralment conçues pour décrire le *contenu* de l'image: les objets et leur emplacement dans la scène. Dans la littérature de MMIR pour l'analyse sémantique, nous pouvons trouver deux approches opposées pour l'extraction de caractéristiques: **caractéristiques locales**, comme SIFT [105] ou SURF [6], à savoir descriptions statistiques de la distribution des bords autour des points d'intérêt locaux, et **caractéristiques globales**, qui résument les propriétés générales de l'image en un seul descripteur, comme la couleur [175] ou la texture [187]. Les **caractéristiques esthétiques** sont plutôt conçues pour décrire la *composition* de l'image ou son *style*, comme son contraste [116], son niveau de détail [110] ou sa profondeur de champs [32].

Les caractéristiques de bas niveau sont le sous-sol de chaque système MMIR et leur *caractére informatif*, ou *capacité discriminative*, à savoir la quantité d'informations fiables sur le contenu de l'image/esthétique qu'ils transportent, est crucial pour le développement de cadres de MMIR efficaces pour l'analyse de l'image.

Niveau 1: Codage des descripteurs locaux et regroupement

Dans certains cas, les fonctions de bas niveau ne peuvent pas être utilisées directement comme input pour le troisième niveau de la pyramide, à savoir le cadre d'apprentissage, mais ils ont besoin de passer par une étape intermédiaire qui les agrège en une signature visuelle compactet. C'est le cas de descripteurs locaux telles que HoG [31], SIFT [105] et SURF [6] .

La raison de ce problème est que le cadre d'apprentissage nécessite en input une signature avec une *longueur fixe* et un petit nombre de dimensions, tandis que, dans

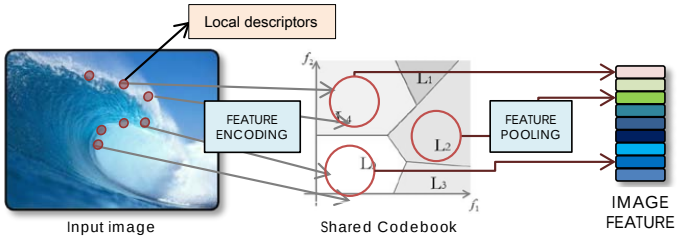


Figure A.3: Regroupement: les caractéristiques locales sont regroupées en un descripteur compact de l'image.

l'extraction des caractéristiques locales, un *montant variable* des descripteurs locaux est extrait de chaque image, ce qui rend l'extraction directe des signatures avec la même dimension pour toutes les images impossible.

La solution générale est donc de *regrouper* tous les descripteurs des images locaux dans une signature de longueur fixe représentant le comportement des points-clés de l'image. Ce résultat est obtenu avec un premièr **codage** d'un ensemble d'entraînement des descripteurs locaux en un nombre des valeurs plus petit, à savoir un code universel partagé avec un nombre donné de "mots visuels". Pour obtenir une nouvelle image, la quantité variable de descripteurs est ensuite **groupée** en une nouvelle signature de l'image qui agrége leurs propriétés en regardant leur distribution, compte-tenu du dictionnaire commun, ayant donc une dimension fixe égal au nombre de mots visuels.

Le regroupement est un autre important niveau de traitement dans MMIR: le processus d'agrégation est généralement réalisé par quantification vectorielle et le regroupement des descripteurs de grande dimension peut causer des pertes d'informations sur la répartition des descripteurs locaux, réduisant ainsi le caractère informatif globale des caractéristiques. La signature de l'image finale doit conserver les informations d'origine autant que possible, tout en gardant la dimension bas et égal pour toutes les images.

Niveau 2: Modèle d'apprentissage

Une fois obtenue une signature de longueur fixe, la prochaine étape vers la compréhension automatique des caractéristiques de l'image est l'étape d'**apprentissage**. A ce niveau, nous utilisons des cadres d'apprentissage supervisé que *apprennent* comment faire la distinction entre les images contenant des caractéristiques différentes (i.e. contenu ou esthétique différente), puis *prédire* les étiquettes correspondant aux nouvelles images.

Semblable à des cerveaux humains, qui reconnaissent au monde basé sur l'association avec leurs souvenirs, l'apprentissage artificiel supervisé nécessite un groundtruth, à savoir un ensemble de **entraînement** des images dont les étiquettes correspondantes qui devront être prévues sont connues. Ces annotations sont généralement attribuées précédemment à la main, en demandant à l'homme d'

indiquer la présence ou l'absence d'un concept sémantique, d'une émotion générée, ou d'un degré esthétique.

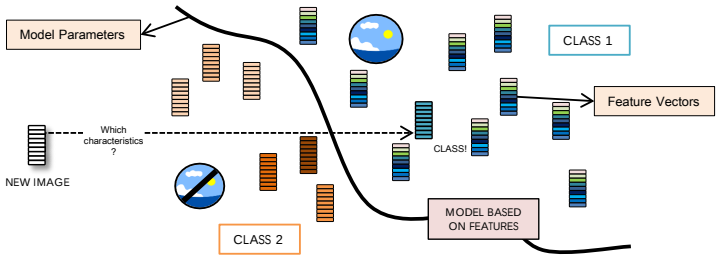


Figure A.4: Modèle d'apprentissage: sur la base des valeurs de caractéristiques, les cadres d'apprentissage apprennent les paramètres pour distinguer entre les différentes catégories de l'image.

En général, chaque propriété à prédire, c'est à dire chaque étiquette, représente un problème d'apprentissage distinct. En traitant le groundtruth, le cadre d'apprentissage détermine les liens entre les valeurs de caractéristiques et de l'étiquette (par exemple, une valeur élevée de couleur bleue est susceptible d'être corrélée avec la présence de l'eau, et peu susceptible de représenter la présence de l'herbe). Dans le cas de la prédiction esthétique, le cadre d'apprentissage apprend à distinguer entre les images attrayantes et non attrayantes, compte tenu des valeurs de caractéristiques esthétiques.

La sortie de cette étape est un ensemble de *modèles*, un pour chaque étiquette à prédire, qui contiennent toutes les équations et les paramètres pour partitionner l'espace de fonction dans des groupes ayant des étiquettes similaires (par exemple degré esthétique similaire, présence du même objet . . .).

La précision des partitions dans l'étape d'apprentissage dépend souvent de la qualité de la base de données de entraînement: les annotations de groundtruth et les signatures sont cruciaux pour la construction d'un modèle significatif. En outre, la précision du modèle peut varier en fonction du type d'algorithme d'apprentissage utilisé, et de comment il trouve des motifs communs entre les caractéristiques. En général, pour les systèmes MMIR, nous utilisons des Support Vector Machines [17], des simples cadres d'apprentissage binaires basées sur des mesures de similarité du noyau.

Niveau 3: tâches et les applications

Qui reste-t-il dans la chaîne de recherche d'information multimédia? Il nous manque l'étape de prédiction, à savoir l'attribution automatique d'étiquettes sémantiques/esthétique à des nouvelles images inédites. A ce niveau de la chaîne, nous **testons** les performances de notre modèle de calcul base en calculant la précision des prédictions sur les nouvelles images.

Au sommet de la pyramide, nous ne traitons pas des caractéristiques à grande échelle: pour une nouvelle image, le système MMIR *prédit* une ou plusieurs éti-

quettes, étant donné les modèles des concepts spécifiques construites dans le niveau inférieur, et les valeurs des caractéristiques de la nouvelle image. Ces étiquettes sont prévues avec un certain degré de *confiance*, à savoir une valeur représentant la fiabilité des labels attribuées automatiquement.

Selon la **tâche** pour laquelle le système est conçu, les systèmes MMIR peuvent ensuite présenter les résultats de deux manières différentes. Lorsque l'objectif est de *classer* un ensemble des images en catégories exclusives pré-définies (par exemple, des scènes ou des objets), le système MMIR, étant donné une image, envoie sa catégorie correspondante, et les performances sont évaluées par la **exactitude** globale, à savoir le nombre des images correctement classées dans le nombre total des images. Lorsque nous traitons avec la *recherche des images*, le système MMIR présente les résultats d'une donnée *query* textuelle a partir d'un ensemble prédéfini d'étiquettes de l'image. Compte tenu de l'indice de confiance attribué par le prédicteur, le système MMIR *récupère* une liste des images classées en fonction de leur pertinence par rapport à la requête. Une façon courante pour évaluer la qualité de ces listes de classement est la Mean Average Precision (précision moyenne arithmétique), à savoir une mesure qui tient compte du montant des images pertinentes qui sont récupérées compte tenu de la requête, et l'ordre dans lequel ils apparaissent dans les listes de classement des résultats retournés .

A ce niveau de la chaîne, un aspect important est le type d'application pour laquelle le système est conçu. Traditionnellement, les cadres MMIR sont conçues par des chercheurs pour **l' analyse sémantique générale**, c'est à dire la reconnaissance des objets et de scènes. Compte tenu de l'importance reconnue de l'analyse sémantique automatique au fil des ans, l'élément crucial pour déterminer la qualité d'un système MMIR sémantique est l'étape d'**évaluation**. L'évaluation est généralement réalisée en utilisant des bases de données comparatives disponibles publiquement (principalement développé pour la reconnaissance des scènes [143, 133, 205] ou des objet [41]), construits pour comparer les performances des différentes techniques en utilisant un groundtruth commun, ou par la partecipation aux défis internationaux et aux campagnes d'évaluation telles que TRECVID [168] ou Pascal VOC [39] .

Mais les techniques d'extraction et de classification peuvent aussi aller au-delà de la prédiction sémantique générale, en appliquant, par exemple, les techniques d'analyse sémantique pour des domaines plus étroites, **domaines spécifiques** qui bénéficient de l'automatisation apportée par MMIR (par exemple, les images médicales, des vidéos de cuisine, images satellites . . .). En outre, les techniques MMIR peuvent être utilisés pour prédire de l'information qui n'est pas strictement liée à l'objet de l'image et des scènes, abandonnant la sémantique et en se concentrant sur les émotions, la valeur **artistique** ou **esthétique**.

A.2.3 Nos Contributions à Chaque Niveau de la Pyramide

Compte tenu des propriétés et des éléments essentiels des systèmes MMIR, dans cette thèse, nous proposons un ensemble de nouvelles contributions à tous les niveaux de

la chaîne MMIR, avec l'objectif d'améliorer l'analyse visuelle globale et la précision des systèmes MMIR construit pour l'analyse sémantique et esthétique .

La particularité de notre travail est son intrinsèque multidisciplinairté: nous empruntons, étudions, étendons et réutilisons les techniques de domaines qui sont traditionnellement externe ou non directement liés à la récupération multimédia. En introduisant ces nouveaux signaux dans les systèmes MMIR, nous construisons des solutions qui sont très efficace en termes de reconnaissance, et généralement aussi complémentaires aux approches existantes.

Pour le niveau 0, 1 et 2 (extraction de caractéristiques, regroupement et l'apprentissage), nous concevons un ensemble de nouvelles techniques qui nous testons sur des ensembles de données générales pour l'analyse comparative de classification/récupération sémantique. Dans nos premières études, l'application principale de nos techniques est donc principalement l'analyse sémantique. Au niveau 3, nous allons au-delà des applications sémantiques pures et nous construisons un système de prédiction de l'esthetique de l'image et de sa beauté, en intégrant la plupart des techniques que nous avons proposées dans les niveaux inférieurs dans un système MMIR que nous construisons pour l'analyse esthétique. Dans ce qui suit, nous allons voir un aperçu de nos contributions.

Niveau 0: Caractéristiques Hybrides à base de saillance pour catégorisation de l'image (Chapitre 3)

Dans le Chapitre 3, nous opérons directement au niveau des pixels. Nous constructisons un ensemble de nouvelles fonctionnalités de bas niveau, très discriminants pour l'analyse sémantique, et inspirées par la théorie de la perception visuelle.

Comme nous avons vu, les caracteristiques sémantiques peuvent être classés en deux groupes: *locales* et *globales*. Alors que les premiéres sont très instructives sur les détails et les contours image et invariants aux transformations, les caractéristiques globales decrivent le comportement générale de l'image, en perdent une certaine précision lorsque les conditions d'éclairage / de rotation changent. Malgré leur efficacité, l'inconvénient majeur des caractéristiques locales est leur coût de calcul et l'obligation regroupement. D'autre part, les caractéristiques globales sont extrêmement faibles en dimensions et rapide à calculer.

Nos caractéristiques se situent dans un point intermédiaire entre les deux approches mentionnées: nous concevons un ensemble de descripteurs **hybrides**, à savoir caractéristiques *globales* de basse dimensionnalité qui intègrent certaines informations localement analysées. L'information locale est integrée dans un descripteur global en utilisent des **images cartes de saillance**, à savoir matrices avec des valeurs d'intensité plus élevée correspondant aux régions qui plus probablement attirent les fixations de l'homme dans l'image. Nous présentons deux caractéristiques suivantes cet approche hybride: le descripteur **Saliency-Aware Color Moments**, à savoir une amélioration d'une caractéristique de couleur a travers des techniques de saillance, et le **Saliency Moments**, à savoir un descripteur basée sur le traitement de la distribution locale de saillance.

Nous testons l'efficacité de nos descripteurs pour la catégorisation d'objets, la reconnaissance de scènes et la récupération de la vidéo, et nous montrons que non seulement ils surpassent les fonctionnalités existantes pour MMIR, mais que la saillance apporte aussi de l'information *complémentaire* aux descripteurs existants généralement utilisé dans MMIR.

Niveau 1: Agrégation des caractéristiques locales à travers l'analyse marginale et des copulae (Chapitre 4)

Dans le Chapitre 4, nous étudions le regroupement et l'encodage, en proposant un ensemble de nouvelles techniques pour l'agrégation des descripteurs locaux rapide et efficace, inspiré par des techniques de modélisation statistique économique.

Notre observation est que les techniques traditionnelles pour l'agrégation des descripteurs ont besoin de procédures coûteuses pour les techniques de codage. Pr evaluer la loi de probabilité multidimensionnelle des composants de descripteurs locaux, ils ont besoin d'un ensemble d'entraînement afin de construire un dictionnaire commun qui permettra pour produire une représentation de l'image compacte. La représentation de l'image est alors calculée en calculant la distribution multidimensionnelle des descripteurs locaux donné le livre de code global. Cette approche a été prouvé être très efficace pour les applications MMIR. Cependant, un des inconvénients majeurs est son stockage intrinsèque et son coût de calcul, ainsi que la perte d'information due à l'étape de regroupement.

Les solutions que nous proposons diffère sensiblement de l'approche traditionnelle, en améliorent l'efficacité et la précision du méthodes traditionnelles de regroupement. Tout d'abord, nous concevons le **MEDA** descripteur, qui décrit le comportement des descripteurs locaux basés sur les approximations de leur **d**istribution marginale, conduisant à une signature de l'image qui est extrêmement léger a calculer mais qui garde une grande précision pour la classification et l'extraction. Nous améliorons cette méthode en construisant **Multimeda**, un noyau pour le support Vector Machines qui est capable d'extraire une probabilité multidimensionnelle des descripteurs des images en calculant le produit des approximations marginaux stockés dans MEDA. Enfin, nous utilisons la théorie de Copula [166] pour calculer la probabilité réelle conjointe des descripteurs locaux, basé sur l'information marginal pure stockée dans MEDA. Nous modélisons la distribution multivariée des points-clés de l'image sans impliquer un processus de codage dans l'espace multidimensionnel. Le vecteur **COMS** résultant s'avère être beaucoup plus efficace que les techniques de l'état de l'art pour l'agrégation locale appliquée à la reconnaissance de scène et la récupération de la vidéo.

Niveau 2: Un cadre d'extraction multimédia basé sur des jugements automatiques de pertinence gradues (Chapitre 5)

Dans le Chapitre 5, nous nous concentrons sur la conception d'un cadre d'apprentissage utilisé pour la recuperation video, en introduisant dans MMIR certains concepts de la recherche d'information web.

La principale observation est que, dans l'étape d'apprentissage des systèmes MMIR, les étiquettes d'entraînement sont attribuées sur un échelle binaire (pertinent / non pertinent). Cela signifie que les annotations générées par l'utilisateur identifient la simple présence ou absence d'un concept dans les données visuelles, sans tenir compte des options intermédiaires. Cependant, une image peut être pertinente pour une catégorie sémantique avec des degrés divers, en fonction de la façon dont ce concept est représenté dans l'image.

Différent de cadres les plus courantes, dans le Chapitre 5, nous construisons un cadre d'apprentissage qui est capable de g'erer des **jugements de pertinence classés**, à savoir de degrés de annotations differents, reflétant les différents niveaux de la pertinence d'un image par rapport à un concept donné. Étant donné que l'annotation manuelle est un processus coûteux et imprécis, afin de construire rapidement des groundtruth graduées, nous proposons une façon de réévaluer les bases de données binaires sans impliquer un effort manuel: nous attribuons automatiquement un degré de pertinence fiable (non, faiblement, moyen, très pertinent) à chaque image, en fonction de sa position par rapport à l'hyperplan dessiné par une Support Vector Machines dans l'espace des caractéristiques.

Nous testons l'efficacité de notre système sur deux bases de données à grande échelle, et nous montrons que notre approche surpasse les cadres binaires traditionnelles pour la récupération de la vidéo.

Niveau 3: Au-delà de la sémantique pure: la synergie avec l'analyse esthétique (Chapitre 6)

Dans le dernier Chapitre technique, nous étudions une nouvelle application émergente dans le domaine de la récupération Multimédia: l'analyse esthétique. Nous appliquons des techniques de MMIR nouveaux et existants, traditionnellement utilisés pour des tâches sémantiques, au problème de l'évaluation automatique d'appel de l'image. Nous réutilisons beaucoup de leçons apprises dans les Chapitres précédents et l'appliquons à la prédiction du degré esthétique du contenu visuel. Nous ajoutons des informations sémantiques à des cadres d'analyse esthétique, et nous voions l'amélioration apportée par les caractéristiques basées sur le contenu de l'image et a la prédiction de la beauté globale.

Par ailleurs, nous explorons également dans l'autre sens: sont les outils d'analyse esthétiques utile pour l'analyse sémantique? Nous étudions l'importance de l'analyse esthétique pour les applications sémantiques, en testant l'efficacité des caractéristiques esthétiques d'un cadre MMIR de reconnaissance de scène.

Dans notre contribution au niveau de l'application, nous enrichissons donc l'analyse visuelle sémantique et esthétique en explorant la synergie de ces deux applications pour MMIR. L'idée principale est que l'analyse sémantique et esthétique sont deux applications étroitement liées dans la récupération multimédia. Nous montrons les avantages et les limites de cette synergie, et proposons quelques améliorations dans ce sens.

Ce manuscrit est structuré en suivant la structure pyramidale nous avons illustré. Tout d'abord, dans le Chapitre 2, nous donnons un aperçu détaillé de l'état de l'art des techniques pour les systèmes MMIR appliquées à l'analyse sémantique et esthétique. Les chapitres 3-6 expliquent nos contributions à chaque niveau de la pyramide MMIR, à partir de l'extraction de caractéristiques au niveau de l'application. Enfin, au Chapitre 7, nous tirons des conclusions sur le travail effectué pour cette thèse.

A.3 Conclusions et perspectives d'avenir

Le Multimedia Information Retrieval est une discipline de recherche. Sa complexité intrinsèque est due à son objectif global: la construction de machines qui peuvent voir ce que nous voyons et comprendre ce que nous comprenons. En outre, la recherche dans le domaine est très variée, plusieurs domaines de l'informatique sont impliqués dans la création d'un système MMIR, à partir du traitement du signal de bas niveau, à la statistique et la modélisation probabiliste, à des techniques d'apprentissage automatique. En raison de la complexité de cette tâche, les systèmes de vision par ordinateur encore performent un ordre de grandeur pire par rapport à leurs homologues biologiques, les systèmes de vision de l'homme. Un é énorme existe entre ce que les humains voient et comprennent, et ce que l'ordinateur peut déduire de pixels de l'image. Dans cette thèse, nous avons présenté un ensemble de techniques visant à réduire cet écart.

Comme nous avons montré dans ce manuscrit, MMIR suit une chaîne complexe d'étapes qui *traduisent les valeurs de pixel d'image en concepts intelligibles*: d'abord, les pixels d'image sont traitées et leurs propriétés sont résumées dans **caractéristiques de bas niveau**. Ces éléments sont ensuite soit **regroupées** dans un descripteur d'image compact, ou directement utilisé comme entrée pour des **machines d'apprentissage**. Ensuite, des cadres d'apprentissage construisent des modèles de l'espace des descripteurs en reliant sur les valeurs caractéristiques des propriétés de l'image. Ces propriétés sont exprimées en **étiquettes**, et peuvent représenter le type de **objets** représentés dans l'image, la **scène** et la **location** où une image donnée a été prise, le degré **esthétique** de l'image, ou les émotions qu'elle suscite. Au niveau supérieur de la pyramide, c'est à dire le niveau de l'application, le système MMIR est utilisé pour ls **prédiction** des labels pour les nouvelles images, inconnues.

A.3.1 Les Contributions d'un Point de vue Multidisciplinaire: les Leçons Apprises

Dans cette thèse, nous avons présenté un ensemble de solutions pour chaque niveau de traitement de la chaîne MMIR: du point de vue de la fonctionnalité, nous avons développé **caractéristiques de bas niveau basé sur la saillance** et plusieurs méthodes pour **le regroupement des descripteurs basé sur l'analyse marginale**. Du point de vue de l'apprentisage, nous avons construit un **cadre à pertinence graduée** pour la récupération de la vidéo, et enfin, au niveau de l'application, nous avons appliqué les leçons apprises pour la **prédiction de l'intérêt d'image basée sur l'esthétique et sémantique** .

La propriété commune de toutes les méthodes que nous avons proposées, c'est qu'ils ont **multidisciplinaires**: nous concevons nos techniques en empruntent des technologies et des principes d'une grande variété de sources qui ne sont pas directement liées à la vision par ordinateur. Nous croyons que la découverte de nouveaux ressources pour l'analyse de l'image peut apporter non seulement des améliorations

substantielles par rapport aux techniques existantes, mais il peut génèrer également des techniques qui fournissent des informations complémentaires sur les propriétés d'image, et qui peuvent être donc utilisées en **combinaison** avec existant techniques.

Par exemple, en utilisant les principes visuels biologiques de l'attention visuelle dans le Chapitre 3, nous avons créé un descripteur très puissant de bas niveau qui peut être comparé pour son efficacité à des caractéristiques globales de bas niveau, et pour son exactitude à des descripteurs compacts générés en agrégeant les descripteurs d'images locaux. Cela nous suggère que les systèmes visuels de l'homme et leur compréhension peut être une source d'inspiration utile pour le développement de systèmes de vision de calcul efficaces. En analysant la façon dont nous traitons les scènes et les objects réels pendant le stade de la vision précoce, et en étudiant attentivement les développements récents dans la **neurobiologie**, on peut construire des caractéristiques de l'image plus discriminantes.

Un autre exemple de notre approche multidisciplinaire est donné par les techniques que nous concevons au niveau 1 de la chaîne d'analyse d'image (Chapitre 4) pour la fonction d'agrégation locale. Ici, nous empruntons du domaine de **statistiques économiques** la théorie des Copules, et nous appliquons avec succès pour la modélisation multidimensionnelle basée sur l'information marginale des points-clés de l'image. La caractéristique résultant est beaucoup plus efficace pour l'analyse sémantique par rapport aux modèles traditionnels complexes tels que sac de mots, sans impliquer la construction d'un modèle universel. Ce travail constitue une première tentative réussie d'introduire les copulae pour les statistiques d'image. Mais la théorie Copula est toute une branche des statistiques qui peuvent fournir plusieurs outils pour résoudre le problème de la vision par ordinateur, et que nous souhaitons explorer dans nos travaux futurs.

Au niveau d'apprentissage (Chapitre 5), nous avons construit un cadre d'apprentissage à pertinence graduée. La notion de pertinence graduée n'est pas nouvelle dans le domaine de **la recherche d'information web**. Les classements des pages Web sont basés sur une échelle non binaire de pertinence, et les techniques d'apprentissage et de classement ont été proposées ad hoc à cet effet. Ces techniques peuvent être réutilisés pour des problémes de Multimedia Information Retrieval et profondes études peuvent être faites sur la façon d'adapter ces méthodes de classement dans le domaine de l'extraction vidéo.

Enfin, au niveau de l'application, dans le Chapitre 6, nous explorons la synergie de nombreux domaines dans le champ MMIR: **l'analyse sémantique, l'analyse esthétique, l'analyse affective, l'analyse artistique**. La combinaison de ces divers indices d'analyse d'image enrichit l'analyse visuelle globale et apporte des améliorations substantielles aux performances de reconnaissance de scènes et de la prédiction de l'intérêt de l'image. En particulier, l'un des éléments clés du système MMIR nous construisons pour la prédiction de l'intérêt est que l'information de groundtruth n'est pas généré manuellement, mais en utilisant une procédure non standard dans les systèmes MMIR traditionnels, à savoir le **crowd sourcing**.

A.3.2 Perspectives d'Avenir: une Question de Contenu?

Le travail presenté dans le Chapitre 6 représente notre premier tentative d'utilisation des informations contextuelles pour MMIR. L'information contextuelle est une source précieuse d'informations sur les propriétés de l'image, car il est composé de sous-titres et des descriptions textuelles entrées directement par les utilisateurs d'image. Avec l'explosion de la gestion de photos en ligne et des outils de partage, nous avons à disposition une quantité incroyable de métadonnées que nous pouvons utiliser pour améliorer notre prédiction esthétique et sémantique.

Il à été souligné [167] que l'analyse visuelle peut-être pas plus encore réellement utile pour les applications multimédia, et que *l'information contextuelle* pourrait être suffisant pour effectuer une analyse automatique de l'image. Il est vrai que la recherche sur l'analyse visuelle, contraint par l'analyse comparative sur des ensembles de données expérimentales, à tendance à se reposer sur l'analyse des images basee sur le simple *contenu* (composition, sémantique), perdant de précieuses informations provenant du media *contexte* (métadonnées, caractéristiques de l'utilisateur). Les systèmes de MMIR sémantique et esthétique manquent également d'envisager des applications dans le monde réel. En fait, en dépit de leur efficacité pour la prédiction des propriétés de l'image pour des bases de données comparatives, les techniques basées sur le contenu pur ont plusieurs limitations de performances en venant à des utilisations pratiques.

Cependant, en vivent dans un monde de contenu généré par l'utilisateur, nous apprenons que nous devons considérer les besoins des utilisateurs, analyser son comportement, sa façon de produire des médias, et toutes les informations qu'elle donne en entrée avec son contenu visuel. Nous pouvons dire que *l'analyse de contenu n'est pas fini*, mais nous devons trouver sa place, des applications spécifiques qui ont besoin de son caractère informatif, le fusionner avec d'autres sources afin d'enrichir l'analyse des médias. Nous allons présenter une possible voie de l'avenir à suivre pour améliorer la prédiction de l'intérêt de l'image et basé sur le contenu et le contexte. L'objectif global est de suivre l'idée de synergie entre les différentes sources initiées dans le dernier Chapitre de cette thèse

Avec la large diffusion des outils de gestion d'image tels que Flickr, Picasa, Facebook, la nécessité de l'évaluation automatique de la beauté et de l'intérêt de l'image devient plus clair de jour en jour. Au lieu d'essayer de modéliser l'intérêt utilisant la seule infromation visuel purs, nous devons tenir en compte plusieurs facteurs qui pourraient être impliqués dans le jugement esthétique des images sur les services de photo en ligne, par exemple:

1. La position de l'utilisateur dans le **réseau social**: combien est-elle appréciée? Combien de contacts at-elle? Combien intéressant étaient ses images précédentes pour la foule?

2. **content** de l'image: il y a des sujets qui sont plus susceptibles d'attirer la foule que d'autres, en fonction de la culture et de nationalité, les nouvelles courantes, la tendance générale de l'être humain de se concentrer sur certains

types d'objets et des scènes etc. Important est aussi la **location** où une image donnée à été prise.

3. **esthétique de l'image**: la façon dont l'image est composé, arrangé, les règles de la photographie qui ont été utilisées pour améliorer sa beauté, le type d'appareil qui est utilisé pour la prise, les émotions véhiculées etc.

Si nous sommes capables de conclure ce genre d'information concernant l'utilisateur et de ses médias, nous pourrions alors traduire ces éléments dans les caractéristiques numériques et modeler l'intérêt image en utilisant comme groundtruth un grand nombre d'images Flickr annotée avec leurs valeurs de Flickr interestingness correspondants. [2]

Comment extraire ces caractéristiques? **Social Network Analysis** (voir, par exemple, [24]), très populaire dans les dernières années, peut nous aider à comprendre la position sociale et le profil de l'utilisateur (1). En ce qui concerne le contenu et l'emplacement (2), en plus des **tag et métadonnées** que l'utilisateur fournit avec ses médias, nous pourrions également compter sur des techniques basées sur le contenu purs tels que ceux que nous avons présenté dans cette thèse. De même, les caractéristiques esthétiques et de composition (3) peuvent être calculées en modélisant les règles photographiques, des émotions, et des traits artistiques en utilisant une approche computationnelle.

En conclusion, nous avons montré avec cette thèse que la synergie entre les différents domaines, liés ou non peut apporter des améliorations considérables vers la véritable modélisation d'un système de vision par ordinateur. Peut-être, en augmentant l'interaction de ces ressources variegates, et l'exploitation des nouvelles technologies que le monde visuel numérique nous fournit, nous pourrions bientôt en déduire, de manière fiable, au moins 500 de ces mille mots q'une image vaut.

[2] http://www.flickr.com/explore/interesting/.