THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR DE L'ISAE-ENSMA

ÉCOLE NATIONALE SUPÉRIEURE DE MÉCANIQUE ET D'AÉROTECHNIQUE

(Diplôme National – Arrêté du 25 mai 2016 Modifié par l'Arrêté du 26 Août 2022)

ÉCOLE DOCTORALE n° 651 MIMME

MATHÉMATIQUES, INFORMATIQUE, MATÉRIAUX, MÉCANIQUE, ÉNERGÉTIQUE

SECTEUR DE RECHERCHE :

INFORMATIQUE ET VISION PAR ORDINATEUR

Par **David TRAPARIC**

Vers une stratégie basée sur l'intelligence artificielle pour la création de contextes immersifs pour le cinéma

Soutenue le 18/12/2025

Membres du jury :

Mme. Isabelle WATTIAU	Professeure des Universités	ESSEC	Rapporteure
M. Frédéric MORAIN-NICOLIER	Professeur des Universités	Université de Reims	Rapporteur
M. Yacine GHAMRI-DOUDANE	Professeur des Universités	La Rochelle Université	Examinateur
M. Chaker LARABI	Professeur des Universités	Univ. de Poitiers	Directeur de thèse
M. Ladjel BELLATRECHE	Professeur des Universités	ISAE-ENSMA	Codirecteur de thèse
M. Thomas TAFFOREAU	Responsable Postprod. ICE	CGR Cinémas	Invité

REMERCIEMENTS

Bordel... C'est pas trop tôt.

Tychus Findlay

J'aimerais citer Chaker et reprendre ses mots au début de la thèse : « Tu verras, à la fin, tous mes doctorants me détestent et c'est tout à fait normal dans une thèse ». Je comprends aujourd'hui ô combien il avait raison, et je comprends également que les échantillons présentent parfois tous un biais commun. André approuvera.

Je dédie cette thèse au concert de *Ours et al.* et à ses autres membres (Maugan De Murcia, Arthur Cavalier, Nicolas Prouteau, Clément Joubert, Guillaume Crashbomb Deau), le plus bel accomplissement qui soit arrivé pendant cette période. Je la dédie également à Mathieu Tews Traparic qui m'a fait découvrir le goût pour la programmation, l'esprit critique et scientifique mais aussi la musique, dès mon plus jeune âge et qui me façonnent encore aujourd'hui.

Je remercie tout particulièrement Thomas Tafforeau, pour son suivi et son aide tout au long de la thèse, son enthousiasme ainsi que sa disponibilité malgré les périodes intensives de postproduction. Merci également à Evaëlle Renoux pour ses informations sur la postproduction, son envie de faire avancer ICE, et pour cette collaboration entre son expertise graphique et mes recherches qui ont donné lieu à la documentation interne du fonctionnement ICE la plus complète et précise à ce jour.

Merci à mes amis d'ENSEIRB-Matmeca, Antoine Siebert et Adrien Stutter Sergent, qui ont été mes senpai de mathématiques en école comme en BU, et qui sont restés des compagnons de vocal hors-pairs avec Matthieu Freez Vilain.

Merci aux aventuriers du royaume d'Andor d'avoir placé un Gor, Rémi Mousse Orveau, Maugan De Murcia, Kevin Pascreau et Prince Thorald.

Merci à tous les copains du laboratoire et merci à mes amis de La Rochelle, et tout particulèrement Aélia Renoux.

Merci à Anna Traparic, pour son soutien immense et indéféctible pendant cette épreuve psychologique.

Et enfin, merci aux personnes de l'ombre qui m'ont donné le goût de la pédagogie et du partage, les professeurs passionnés ainsi que les encadrants de stage bienveillants auprès de qui il était très agréable de travailler et d'apprendre. Antoine Marot et Stéphane Daguts au lycée, Bernadette Chaulet, François Louet et Clency Perrine en IUT, Romain Tajan, Pascal Vallet, Guillaume Ferré et Yannick Berthoumieu en école d'ingénieur, Nicolas Texier à Covéa et Christophe Engelvin au CEA.

Vers une stratégie basée sur l'intelligence artificielle pour la création de contextes immersifs pour le cinéma

Résumé :

Le format ICE (Immersive Cinema Experience) introduit par CGR Cinémas propose une extension spatiale du film grâce à des écrans LED latéraux, offrant une expérience de projection immersive. La création de ces contenus repose actuellement sur une postproduction manuelle, où les graphistes déterminent plan par plan les traitements visuels à appliquer aux bords de l'image. Cette thèse vise à modéliser et automatiser ce processus afin de formaliser les choix esthétiques et narratifs qui le regissent, tout en rendant la production de contenus immersifs plus cohérente et adaptable à grande échelle. Une première contribution consiste en la formalisation du processus décisionnel ICE à l'aide de graphes sémantiques et d'ontologies issues de l'analyse filmique. Cette modélisation symbolique structure les critères de création immersive et en souligne les limites pour l'inférence automatique, motivant le recours à des approches d'apprentissage profond. Ainsi, trois modèles ont été développés dans ce cadre : 1) R-ICE, basé sur un réseau convolutif 3D entrainé sur une base de bandes-annonces, atteint une précision de 72,71 % dans la prédiction des traitements à appliquer ; 2) ICE-AGE introduit un mécanisme d'attention multimodale exploitant la profondeur, la détection d'objets et l'échelle de plan pour affiner la décision ; et 3) ICE-Cubed élargit ces travaux à la reconstruction et à l'enrichissement visuel des contenus en combinant détection, inpainting et superrésolution afin de préserver la cohérence perceptuelle et l'immersion. Les modèles ont été évalués sur des bases de données construites en collaboration avec les experts du format ICE. Les résultats expérimentaux démontrent la performance du module de détection d'ICE-Cubed par rapport à l'état de l'art, et une étude perceptuelle menée sur des bandes-annonces confirme l'efficacité du système au niveau immersif.

Mots-clés : Vision par ordinateur, Affichage immersif, Cinéma, Apprentissage profond, Ontologie

---

Abstract :

The ICE (Immersive Cinema Experience) format, introduced by CGR Cinémas, extends the spatial boundaries of traditional film projection through lateral LED screens, offering an immersive viewing experience. The creation of ICE content currently relies on a manual postproduction process, in which graphic artists determine, shot by shot, the visual treatments to be applied to the edges of the image. This doctoral research aims to model and automate this process in order to formalize the underlying aesthetic and narrative decisions while making the production of immersive content more consistent and scalable. The first contribution consists in the formalization of the ICE decision-making process through semantic graphs and ontologies derived from film analysis. This symbolic modeling structures the aesthetic and narrative criteria guiding immersive creation, while revealing its limitations for automatic inference, thus motivating the use of deep learning approaches. Within this framework, three models have been developed : 1) R-ICE, based on a 3D convolutional neural network trained on a dataset of movie trailers, achieves an accuracy of 72.71% in predicting the appropriate visual treatment ; 2) ICE-AGE introduces a multimodal attention mechanism that leverages depth, object detection, and shot scale to refine decision-making ; and 3) ICE-Cubed extends these works toward visual reconstruction and enhancement by combining detection, inpainting, and super-resolution techniques to preserve perceptual coherence and improve immersion. The models were evaluated on datasets built in collaboration with ICE experts. Experimental results demonstrate the superior performance of the ICE-Cubed detection module compared to state-of-the-art methods, while a perceptual study conducted on movie trailers confirms the system's effectiveness in enhancing the immersive experience.

Keywords : Computer Vision, Immersive display, Cinema, Deep learning, Ontology

---

TABLE DES MATIÈRES

Remerciements **3**

Résumé — Abstract **5**

Table des matières **7**

Table des figures **11**

Liste des tableaux **15**

Liste des acronymes **17**

Traductions techniques **21**

Introduction **25**

1 Le cinéma immersif et l'IA **29**
1.1 Introduction 31
1.2 L'histoire de l'audiovisuel comme expérience immersive 31
1.2.1 Le début de la narration visuelle et le précinéma 31
1.2.2 Le cinéma sans caméra 34
1.2.3 Le cinéma photographique et numérique 35
1.2.4 L'évolution du son au cinéma comme modalité d'immersion 36
1.2.5 L'évolution des écrans de cinéma 37
1.2.6 L'évolution de la préproduction et de la postproduction 37
1.3 Le cinéma aujourd'hui : La réinvention de l'expérience face à la vidéo 38
1.3.1 L'apparition de la chronologie des médias en France 38
1.3.2 Le cinéma se réinvente face à la SVOD 39
1.3.3 Description du format ICE comme innovation immersive 41
1.4 Immersion 43
1.4.1 Immersion : définition et amélioration 43
1.4.2 L'immersion appliquée à ICE : facteurs disruptifs et actions permises 43
1.5 État de l'art par rapport à ICE 46
1.5.1 Les systèmes « Focus+Contexte » 46
1.5.2 Les systèmes « CAVE » et casques VR 48
1.6 L'apprentissage profond et sa place au sein de l'IA 49
1.6.1 Définitions 49
1.6.2 Le paradoxe de « l'effet IA » et redéfinitions 50
1.6.3 L'IA, de la fiction à aujourd'hui 53
1.6.4 La place de l'IA dans la société : Hivers de l'IA et luttes sociales . . 54

L'IA vue par les financements et le capital . . . 54
L'IA vue par le peuple . . . 55
1.6.5 Les rouages de l'apprentissage profond . . . 58
Le perceptron de Rosenblatt . . . 58
Fonctions d'activation modernes . . . 59
Le perceptron multicouche et la descente de gradient . . . 60
Couches entièrement connectées . . . 62
1.6.6 Réseaux de neurones convolutifs . . . 62
Convolution 1D . . . 62
Convolution 2D . . . 63
Sous-échantillonnage ou *pooling* . . . 64
Architecture globale des CNN . . . 64
1.7 Conclusion . . . 65

2 ICE-HOLSS : Une ontologie de la sémantique latente des scènes pour ICE 67
2.1 Introduction . . . 69
2.2 Postproduction et décision ICE . . . 70
2.2.1 Environnement de travail ICE . . . 70
2.2.2 Décisions et génération des contenus ICE . . . 70
WALL . . . 72
WALL EXPAND . . . 72
SEAM . . . 72
WALL EXPAND FH . . . 72
WALL COLOR . . . 72
2.2.3 Contraintes technologiques et de la technique actuelle . . . 72
Validation par les studios et obligation de cohérence . . . 73
Déphasage temporel . . . 74
Amplification de la vitesse . . . 74
2.2.4 Décisions des graphistes ICE . . . 75
Principe de base . . . 75
Cas remarquables . . . 76
Techniques avancées . . . 77
2.3 Formalisation de la décision ICE à l'aide d'ontologie . . . 78
2.3.1 Problématique . . . 78
2.3.2 État de l'art des ontologies du cinéma . . . 78
2.3.3 Qu'est-ce que l'ontologie, les graphes sémantiques et l'IA symbolique 80
2.3.4 Analyse filmique . . . 83
Échelle de plan . . . 84
Le cadrage . . . 86
Mouvement dans l'image et mouvement de caméra . . . 87
Image . . . 89
Son . . . 90
Profondeur . . . 91
Composition . . . 92
Lumière . . . 93
Montage . . . 94
2.4 Retours d'expérience sur la création d'une ontologie dans le cadre de ICE . 99
2.5 Conclusion . . . 101

3 R-ICE et ICE-AGE : Prédiction du niveau de détail adéquat 103
3.1 Introduction . . . 105
3.2 État de l'art sur l'analyse de vidéo . . . 105
La classification d'action . . . 106
Le résumé de vidéos . . . 106
Le Sous-titrage vidéo . . . 107
La détection d'actions . . . 108
La génération de propositions temporelles d'action (TAPG) . . . 108
Localisation d'action temporelle (TAL) . . . 108
Localisation spatiotemporelle d'action (STAL) . . . 109
« Compréhension de tâches humaines » et autres sous-genres . . . 109
3.3 Problématique et objectifs du projet . . . 110
3.4 Gestion des données . . . 110
3.4.1 Constitution de la base de données . . . 113
3.4.2 Méthodologie de stockage des données . . . 115
3.5 Méthodologie et appropriation d'architectures existantes . . . 116
3.5.1 Prétraitement des données . . . 117
3.5.2 Descripteurs issus de l'apprentissage profond . . . 117
3.5.3 Descripteurs sur-mesure . . . 119
3.5.4 Classification . . . 120
3.6 Expérimentations . . . 121
3.6.1 Jeu de données . . . 121
3.6.2 Détails d'implémentation . . . 123
3.6.3 Résultats et discussions . . . 124
3.6.4 Étude d'ablation . . . 127
3.7 ICE-AGE : Extension Graphique basée sur l'Attention . . . 129
3.7.1 Motivations . . . 129
3.7.2 Mécanisme d'attention . . . 130
3.7.3 Méthodologie et architecture . . . 136
3.8 Conclusion . . . 140

4 ICE-Cubed : Extension intelligente de contextes immersifs par inpainting 143
4.1 Introduction . . . 144
4.2 État de l'art . . . 145
4.3 Méthodologie . . . 150
4.3.1 Architecture du système . . . 150
4.3.2 Segmentation d'objets . . . 152
4.3.3 Création du logiciel libre d'annotation SSPICE et de la base de données 153
4.4 Résultats . . . 155
4.4.1 Configuration expérimentale . . . 155
4.4.2 Discussions sur la détection . . . 155
4.5 Validation perceptuelle . . . 157
4.5.1 Méthodologie de l'étude perceptuelle . . . 157
4.5.2 Résultats de l'étude perceptuelle . . . 159
4.6 Conclusion . . . 161

Conclusion 163
4.7 Perspectives . . . 165

Bibliographie **167**

Sources graphiques et multimédias **179**

A Annexes **181**

TABLE DES FIGURES

1.1 Trame de la bande-annonce de *Furiosa : Une saga Mad Max* lors d'une séance ICE 26

1.1 Frise chronologique de l'histoire du cinéma 31
1.2 Photo des premiers dispositifs de narration visuelle et des jouets optiques du précinéma comme précurseurs du cinéma d'aujourd'hui 32
1.3 Démonstration de l'effet phi et photographie d'une lanterne magique. 33
1.4 Gravures de dispositifs d'animation à usages individuels. 34
1.5 Le praxinoscope à projection et le Théatre optique des dessins animés d'Émile Reynaud 34
1.6 Le début du cinéma, avec les pellicules de Reynaud, d'Edison et des frères Lumière 35
1.7 Console d'orgue de cinéma 36
1.8 Le StageCraft, la salle de production virtuelle sur plateau qui a popularisé la technique 38
1.9 Illustration de la chronologie des médias 39
1.10 Photographies des équipements ICE 42
1.11 Processus de postproduction ICE pour un film 44
1.12 Division du champ de vision humain 45
1.13 Démonstration de l'acuité périphérique des lettres par l'œil humain 46
1.14 Schéma de la répartition des photorécepteurs dans la rétine humaine 47
1.15 Illustration de la loi de clôture de la théorie Gestaltique par l'exemple . . . 47
1.16 Réalisations commerciales ayant des points communs avec ICE, à savoir un dispositif d'affichage à but immersif mettant à profit la vision périphérique . 48
1.17 Réalisations scientifiques ayant des points communs avec ICE, à savoir un dispositif d'affichage à but immersif mettant à profit la vision périphérique . 49
1.18 Schéma de la catégorisation de l'IA la plus commune, jugée incomplète dans ce manuscrit, mais permettant une approche simplifiée pédagogique 50
1.19 L'abstraction des données allant de paire avec la spécialisation de l'IA vers l'apprentissage automatique puis l'apprentissage profond 51
1.20 Catégorisations de l'IA, corrigée en accord avec la proposition de définition de ce manuscrit 53
1.21 Similarité entre la fonction paramétrique et le perceptron 59
1.22 Les différentes fonctions d'activation utilisées depuis leur apparition 60
1.23 Perceptron multicouche 61
1.24 Explication visuelle d'une convolution 1D. Le fonctionnement peut être vu comme un glissement d'une fonction sur l'autre. Pour chaque position, la valeur de sortie est l'aire sous la courbe du produit des deux fonctions . . . 63

1.25 La convolution 2D 63
1.26 Schéma de l'architecture du VGG16 64

2.1 Capture d'écran du logiciel de travail Adobe Premiere Pro de l'équipe de postproduction ICE sur une bande-annonce de Buzz L'éclair 69
2.2 Zoom sur les pistes vidéos de l'outil Première Pro de l'équipe ICE 71
2.3 Illustration des différents prétraitements d'image disponibles à l'usage pour l'équipe ICE 71
2.4 Illustration de l'algorithme de Seam Carving pour l'agrandissement d'image 73
2.5 Illustration de la contrainte du déphasage temporel en ICE 74
2.6 Illustration de la contrainte d'amplification de la vitesse 75
2.7 Interface graphique de Protégé, le logiciel libre d'édition d'ontologie. 83
2.8 Exemples de plan débullé et son rôle 84
2.9 Schéma des différentes échelles de plan employés dans le cinéma 85
2.10 Illustration du décadrage 87
2.11 Exemples du clair-obscur et de l'éclairage franc 89
2.12 Exemple de cadre dans le cadre 93
2.13 Fonctionnement de base de notre ontologie représentée sous la forme d'un diagramme de classe. 96
2.14 Illustration des significations potentielles mais non certaines dans le diagramme de classe 97
2.15 Illustrations de l'héritage en diagramme de classe 98
2.16 Illustration de la combinaison de critères en diagramme de classe 99
2.17 Illustration d'une association de critères sous conditions en diagramme de classe 99
2.18 Aperçu des premiers niveaux des différentes entités de l'ontologie sur Protégé100

3.1 Arborescence des différentes sous-tâches qui composent la compréhension de vidéo 105
3.2 Exemple visuelle d'une action nécessitant la dimension temporelle pour être comprise, en l'occurence, contenu audiovisuel interactif où un personnage doit se projeter à l'aide de son marteau 106
3.3 Principe du résumé de vidéo 107
3.4 Illustration du sous-titrage vidéo 107
3.5 Différences entre les modèles TAPG d'approche basées ancre et d'approche sans ancre 108
3.6 Utilisation d'un modèle de localisation d'action spatiotemporelle (STAL) . . 109
3.7 Aspect d'un contenu vidéo de travail fourni par les studios de film 111
3.8 Illustrations de cartons divers 112
3.9 Origine et prise en compte des décalages temporels entre les fichiers vidéos des données de vérité terrain et les labels associés à ces données 112
3.10 Fondu enchainé et fusion entre les plans d'une piste vidéo. 114
3.11 Hiérarchie de la structure de données mise en place 115
3.12 Aperçu des données de film par plan cinématographique 115
3.13 Architecture de notre modèle R-ICE. 117
3.14 Structure du classifieur placé après l'extraction des caractéristiques par les différents backbones temporels. 120
3.15 Distribution des labels de notre base de données et fusion des labels. 122
3.16 Augmentation de données progressive au fil de l'entrainement 123

3.17 Matrices de confusion entre les étiquettes prédites et les annotations de référence . . . 125
3.18 Illustration des mauvaises prédictions de R-ICE lorsque celles-ci semblent tout à fait plausibles pour un œil non expert . . . 126
3.19 Utilité du bourrage à zéro . . . 127
3.20 Longueur des plans de notre jeu de données de bandes-annonces . . . 128
3.21 Plongement sémantique de mots . . . 131
3.22 Exemple de jetonisation d'une phrase . . . 132
3.23 Traduction mot à mot d'une phrase japonaise vers le français. Le texte traduit n'a plus de sens. . . . 132
3.24 Traduction efficace d'une phrase vers le français. Pour être traduite, la phrase doit prendre en considération tout son contexte, et les mots de la langue cible dépendent des mots qui apparaissent plus tôt ou plus tard dans la phrase de la langue d'origine. . . . 132
3.25 Construction des vecteurs contextuels comme première étape du mécanisme d'attention . . . 133
3.26 Mise en œuvre de matrices entrainables afin d'obtenir un poids d'attention efficace . . . 134
3.27 Mécanisme d'attention à tête multiple . . . 135
3.28 Exemple de carte de profondeur . . . 136
3.29 Illustration d'un flot optique . . . 136
3.30 Représentation de l'espace de couleur TSV par un cône et son utilité pour représenter un champ vectoriel de flux optique . . . 137
3.31 Module d'attention croisée . . . 137
3.32 Modules d'attention croisée venant de différentes modalités afin de classifier une trame. . . . 138
3.33 Architecture globale de ICE-AGE, avec une attention croisée par modalité et par trame. Le classifieur est un LSTM qui considère plusieurs trames successives en entrée pour inférer sa sortie. . . . 138
3.34 Graphiques de l'agorithme de séparation de grappe « Pic de densité » . . . 139
3.35 Test de « pic de densité » en guise de modèle de résumé de vidéo, afin d'extraire les trames principales des séquences . . . 140

4.1 Illustration de la nécessité de cohérence spatiale pour ICE. L'expansion de l'image doit concerner uniquement l'arrière-plan en raison de la configuration tridimensionnelle des écrans . . . 144
4.2 Similarités et comparaisons d'architectures entre les modèles profonds de compréhension et les modèles génératifs . . . 146
4.3 Exemples de tâches d'inpainting et d'outpainting réalisés par l'humain, tâches possibles bien avant l'apparition de l'intelligence artificielle . . . 147
4.4 Interpolation et extrapolation d'un signal 1D. . . . 147
4.5 Exemples d'inpainting avec la méthode InOut (2022), tirant profit d'un espace latent . . . 148
4.6 Exemples d'inpainting appliqué à nos données . . . 150
4.7 Schéma détaillé de l'architecture proposée pour générer des contextes périphériques sans élément, en tant qu'extension de trame . . . 150
4.8 Illustration de l'effet de subdivision « gauche » et « droite » dans l'architecture ICE-Cubed pour l'inpainting de trames. . . . 151

4.9 Effets de l'opération morphologique de dilatation et d'érosion sur un logotype, avec silhouette et lettres stylisées. 151
4.10 Démonstration de l'importance du module proposé de réduction d'artéfact . 152
4.11 SSPICE, le logiciel open source de détourage de trames vidéo ultrarapide, créé pour cette thèse 154
4.12 Comparaison visuelle et performances qualitatives de tous les modèles de segmentation utilisés 156
4.13 Comparaison de rendus en salle ICE, entre le visionnage d'un contenu d'origine et sa transformation une fois ICE-Cubed appliqué 159
4.14 Vingt-neuf contenus présentés aux cinq experts. Résultats de l'étude perceptuelle par rapport au contenu. 160
4.15 Vingt-neuf contenus présentés aux cinq experts. Résultats de l'étude perceptuelle par expert. 160

LISTE DES TABLEAUX

3.1 Comparaison des tailles d'origine des trames de chaque modèle 119
3.2 Précision (taux de prédiction correcte) des différents modèles d'extracteurs de caractéristiques temporels (squelette, ou dorsale), selon les différentes configurations de classifieurs possibles. 124
3.3 Précision (taux de prédiction correcte) des différents modèles d'extracteurs de caractéristiques temporels, selon les différentes combinaisons de configurations de classifieurs possibles. 125
3.4 Comparaison des performances de notre classifieur selon les différentes conditions d'image proposées dans la littérature (TSP, I3D, C3D) et celles conçues pour cette thèse. 129

4.1 Comparaison des performances entre différents modèles d'apprentissage profond et méthodes traditionnelles sur notre jeu de données annoté. La MAE doit être faible, tandis que la MF et la MS doivent être élevées. Les meilleurs scores pour chaque métrique sont soulignés. 157

LISTE DES ACRONYMES

BU *bibliothèque Universitaire.*

C3D *Convolutional 3D.* , 118, 119, 121, 123–125, 129
CA *Chiffre d'Affaires.* , 38, 40
CAVE *Cave Automatic Virtual Environment.* , 48
CEA *Commissariat à l'Énergie Atomique et aux Énergies Alternatives.* 3, , 49
CGR *Circuit Georges Raymond.* 5, , 25, 26, 42, 78, 98, 110, 121, 165
CIE *Commission Internationale de l'Éclairage.*
CNC *Centre National du Cinéma.* , 40
CNN *Convolution Neural Network.* , 55, 64, 135, 138
COD *Camouflaged Object Detection.* , 153
CSV *Comma-Separated Values.* , 112, 114

DCI *Digital Cinema Initiatives.*
DCP *Digital Cinema Package.*
DICE *Dataset of dIsturbing Cinematographic Elements*, jeu de données d'éléments cinématographiques perturbants. , 145, 164
DIS *Dichotomous Image Segmentation.*
DL *Deep Learning.*

FC *Fully Connected.* , 62, 120, 121
FCS *Fréquence Critique de Scintillement.* , 45
FHI *Future of Humanity Institute*, Institut pour l'Avenir de l'Humanité. , 52
FOV *Field of view*, champ de vision. , 49
fps *Frame per second.* , 118, 119
FVOD *Free Video On Demand.*

GAN *Generative Adversarial Networks*, ou réseaux antagonistes génératifs. , 148, 149
GPU *Graphical Processing Unit* ou processeur de calcul graphique. , 55, 123, 140, 155

HD *High Definition.* , 37
HFR *High Framerate.* , 41

HMD *Head-mounted Displays*. , 48, 49

HRSOD *High-Resolution Salient Object Detection*. , 153

HSL *Hue Saturation Lightness* ou Teinte Saturation Luminosité. , 122

HSV *Hue Saturation Value* ou Teinte Saturation Valeur.

HTML *HyperText Markup Language*. , 80

i/s *image par seconde*. , 41

I3D *Inflated 3D ConvNet*. , 118, 119, 123–126, 129

IA *Intelligence Artificielle*. , 48–50, 78, 80, 82, 83, 100, 101, 110, 123, 163, 164

ICE *Immersive Cinema Experience*. 5, 11, 12, , 25–27, 31, 41, 42, 44, 49, 65, 69–75, 77, 78, 99–102, 105, 110–114, 121, 126, 130, 137, 140, 141, 144, 161, 163–165

ICE-AGE ICE : Attention-based Graphical Extension. , 27, 105, 164

ICE-Cubed *Inpainting of Cinematographic Elements for Intelligent Context Expansion on ICE*, inpainting d'éléments cinématographiques pour l'extension intelligente de contextes sur ICE. , 27, 145, 158, 161, 164

ICE-HOLSS *Holistic Ontology of Latent Scene Semantics*, Ontologie Holistique de la Sémantique Latente des Scènes. , 70, 163, 164

ICP *ICE Cinema Package*.

IMDB *Internet Movie Database*, ou Base de données cinématographiques d'Internet. , 79

KDM *Key Delivery Message*.

LCP *LightVibes Cinema Package*.

LED *Light-Emitting Diode*. 5, , 25, 26, 37, 41, 42, 47, 49, 69, 72, 140

LLM *Large Langage Model*. , 48, 100, 117

LUT *Look Up Table*.

ML *Machine Learning*.

MLP *Multilayer Perceptron*. , 50, 60–62, 64, 120, 125, 127, 128

NLP *Natural Language Processing*. , 130

ORTF *Office de radiodiffusion-télévision française*. , 38

OSVP *On-Set Virtual Production* ou production virtuelle sur plateau. , 37

OWL *Web Ontology Langage*. , 82, 95–97

PAF *Paysage Audiovisuel Français*. , 40

PAR *Parabolic Alumnized Reflector (Can Light)*. , 41

PDG *Président-directeur général*. , 39

PLF *Premium Large Format*. , 25

PSNR *Peak Signal-to-Noise Ratio* ou le ratio du pic du signal à bruit. , 152

QC *Quality Check*. , 73

R-ICE *Robotic-ICE*. , 27, 105, 164

RAM *Random Access Memory* ou mémoire vive. , 19

RDF *Ressource Description Framework.* , 80–82, 96, 97
RDFS *Ressource Description Framework Schema.* , 81, 82
ReLU *Rectified Linear Unit.* , 60, 127
RGB *Red Green Blue.*
RVB *Rouge Vert Bleu.* , 47, 70, 116, 118

SI *Spatial Perceptual Information.* , 120
SITI *SI et TI.* , 126–128
SOD *Salient Object Detection.* , 153
SPARQL *SPARQL Protocol and RDF Query Language.* , 81
SSIM *Structural Similarity*, similarité de structure. , 152
SSPICE *Scene Segmentation Platform by ICE.* 14, , 145, 153, 154, 164
SVOD *Subscription Video On Demand.* , 31, 39, 40
SVR *Support Vector Machine for Regression.* , 127

TAPG *Temporal Action Proposal Generation* ou génération de propositions temporelles d'action. , 108
TI *Temporal Perceptual Information.* , 120, 136
TSP *Temporally-Sensitive Pretraining.* , 118, 119, 123–129
TV *Télévision.* , 37, 39

URI *Uniform Resource Identifier.* , 80, 81, 96, 97
URL *Uniform Resource Locator.* , 80, 81

VCR *Video Cassette Recording.* , 39
VFX *Video Effects.* , 73
VHS *Video Home System.* , 39
VOD *Video On Demand.*
VR *Virtual Reality.* , 41, 49
VRAM *Video RAM, Video Random Access Memory* ou mémoire vive vidéo. , 26, 123, 140, 155

W3C *World Wide Web Consortium.* , 80
WC *Wall Color.*
WEFH *Wall Expand Flou Horizontal.*
WWW *World Wide Web.* , 80

XML *Extensible Markup Language.* , 81, 82, 96, 97, 111, 113

TRADUCTIONS TECHNIQUES

La quasi-totalité du champ lexical de l'apprentissage profond, de la photographie ou encore du cinéma sont des termes provenant de l'anglais. Certaines traductions en français perdent totalement le sens d'origine, tandis que la population française utilise plutôt directement les termes anglais. Vu l'absence d'experts linguistiques ou scientifiques au sein de l'académie française montrant non seulement une incapacité à traduire de manière précise mais aussi à inventer de nouveaux mots (*streamer* traduit par « joueur en direct » alors qu'un *streamer* peut faire une revue de presse papier ou une performance musicale, *skin* par objet virtuel alors qu'on parle d'un cosmetique de personnage virtuel et pas une cryptomonnaie, *fatbike* traduit par vélo à pneus surdimensionnés), on ne saurait vers qui se tourner pour obtenir une traduction correcte. De plus, les dictionnaires ne reconnaissent pas tous la même liste de mots, et l'usage d'un mot dans la langue française précède toujours son entrée en dictionnaire, parfois même de 10 ans (*enjailler*, reconnu par Le Petit Larousse en 2022 alors que les apparitions remontent au moins à 2011 d'après trends.google.fr, ou encore « à priori », graphie moderne employée par Voltaire en 1759, proposée par l'académie française dans son dictionnaire de 1835 puis supprimée du dictionnaire de l'académie française en 1992 et pourtant celle recommandée par la réforme des rectifications orthographiques de 1990). À ce stade, à défaut de pouvoir utiliser une traduction reconnue et précise, il est logique que la population se tourne vers le mot d'origine afin de prioriser le but premier d'une langue, à savoir se faire comprendre précisément par son interlocuteur, plutôt que parler dans une seule langue mais de manière imprécise et favorisant la confusion. Dans la contrainte de devoir rédiger ce manuscrit entièrement en français, des choix ont nécessairement été faits par l'auteur sur la traduction des termes techniques, parfois uniquement entendus en anglais. Ces choix sont menés par la volonté d'allier compréhension des experts du domaine, des novices francophones et intégrité du sens d'origine dans la traduction.

Par ailleurs, toujours sur les choix linguistiques, la norme orthographique du manuscrit sera celle des rectifications orthographiques de 1990, orthographe recommandée depuis cette date et appliquée dans les manuels scolaires depuis 2017, qu'il est conseillé d'appliquer entièrement plutôt qu'à moitié. On écrira donc cofinancé, précinéma, postproduction, prémium, scénarios, évènement, à priori, plateforme, chaine, entrainement, artéfact, boite, cout, gout, naitre, à minima, kébab, contrejour, apparaitre, préentrainé, plutôt que co-financé, pré-cinéma, post-production, *premium*, *scenari*, événement, *a priori*, plate-forme, chaîne, entraînement, artefact, boîte, coût, goût, naître, *a minima*, kebab, contre-jour, apparaître, pré-entrainé, etc.

(cinematic) shot plan (cinématographique). , 121

action recognition, action classification Reconnaissance d'action, classification d'action. , 107

average pooling sous-échantillonnage par moyenne. , 64

back lights éclairage arrière. , 41, 70

batch size taille de lot, taille de batch. , 155

border matting *mattage de contours* (Le *matte painting*, *video matting*, *border matting* sont des termes complémentaires sur le *matte* qui est traduit par cache/contre-cache en français, alors que le terme anglais vient pourtant du français « mat » l'opposé du brillant, comme synonyme d'opaque. Si la traduction de *matte painting* par *peinture sur cache* a encore un sens, elle n'en a plus si on considère *matting* qui est le fait de déterminer un contour, une séparation entre une forme et une autre ou bien un sujet et un arrière-plan, alors traduit par détourage. À nouveau, détourage perd tout son sens si on essaie de l'appliquer au *border matting*, qui est le fait de déterminer, lorsqu'un pixel est un mélange colorimétrique de l'arrière-plan et du sujet, s'il convient mieux de le déterminer comme de l'arrière-plan ou du sujet. Traduire *video matting* par détourage vidéo mais *border matting* par détourage de contour est absurde. Faute de mot adéquat, le terme affinage de contour est parfois proposé.). , 153

chroma key compositing, ou chroma keying incrustation sur fond vert, incrustation chromatique. , 37, 38

cluster grappe.

clustering *délimitation de grappe* (calqué sur la traduction de « Computer cluster » par « Grappe de serveurs » publiée par la Commision d'enrichissement de la langue française dans le *Journal officiel* du 10 octobre 1998). , 139

compositing Composition d'images, ou incrustation d'image.

cut (film transition) Coupe de plan (transition entre deux plans cinématographiques). , 114, 152

Cut-off point Seuil de coupure. , 140

data set jeu de données. , 54

deep learning apprentissage profond. , 60, 105, 110

DevOps Intégration continue de développement. , 153

dummy data données factices. , 127

e-sport littéralement sport éléctronique, jeu vidéo à caractère compétitif et sportif. , 40

egocentric vision subjective. , 109

embedding plongement. , 130

epoch *époque* (Le terme anglais epoch est calqué sur la même étymologie grecque que notre mot français « époque », bien que l'homologue français ne soit jamais utilisé dans la littérature et les ressources en IA). , 123, 124, 127, 155

feature maps cartes de caractéristiques. , 119, 139

fine tuning spécialisation. , 118

frame trame. 13, 14, , 32, 34, 35, 42, 74, 106, 107, 109, 111, 112, 115, 117–119, 123, 127, 128, 130, 138–140, 153–156, 164

framework infrastructure logicielle, cadre logicielle, cadriciel, cadre méthodologique. , 110

fuzzy floue. , 101

hardware Matériel informatique ou éléctronique. , 72

high frame rate cadence d'image élevée, rafraichissement d'image élevée. , 41

High-key lighting éclairage franc. , 89

jittering tremblement. , 158

learning rate taux d'apprentissage. , 155

linear affine. , 131

live diffusion Diffusion en direct. , 40

loss function fonction de cout, fonction de perte. , 61

Low-key lighting éclairage discret. , 89

machine learning apprentissage automatique.

matte painting peinture sur caches. , 37

matting détourage, extraction de fond. , 37

max pooling sous-échantillonnage par maximum. , 64

MLOps Intégration continue d'apprentissage. , 153

model backbone *squelette de modèle,dorsale de modèle* (calqué sur la traduction de « Internet Backbone » par « Dorsale Internet » publiée par la Commision d'enrichissement de la langue française dans le *Journal officiel* du 16 mars 1999). , 118, 119, 123, 137

Moving head Spot motorisé, spot mobile, projecteur de lumière à tête motorisée. , 41

Multistream architecture *Architecture à voies multiples* (Le terme « voie » est calqué sur les voies I et Q de la modulation d'amplitude en quadrature.). , 137

one-hot *un* parmi *n*. , 131

Parabolic Alumnized Reflector Can Light Spot encastré à réflecteur parabolique aluminisé. , 41

patch imagette. , 141, 148

prompt instructions (textuelles). , 148

receptive field champs receptif. , 119

snippet bribe. , 117, 121, 127

speed dating Rencontre amoureuse expresse. , 40

Spot light *Projecteur de lumière collimatée, projecteur de lumière directionnelle* (Le terme « spot » est préféré car à la fois très court, très commun, et permet d'éviter la confusion entre projecteur vidéo, projecteur d'image et projecteur de lumière.).

storyboard scénarimage. , 37

Task localization Localisation de tâches. , 106
Timestamp Horodatage. , 111
travelling matte process extraction de fond en mouvement. , 37

unscripted non scriptées. , 109
untrimmed non tronquées. , 109

vanishing gradient phénomène de disparition du gradient. , 59
video captioning Sous-titrage vidéo. , 106
Video classification Classification vidéo. , 106
video frame interpolation interpolation temporelle de vidéo. , 117
Video Summarizing Résumé de vidéos. , 106
video understanding compréhension de vidéos. , 78, 105

zero padding bourrage à zéro. , 127, 128

INTRODUCTION

> Pour ceux qui viendront après, n'est-ce pas ?
>
> *Gustave Lumière*

Les travaux de recherche issus de cette thèse sont l'aboutissement d'un contrat de thèse CIFRE, cofinancé par l'Association Nationale de la Recherche et de la Technologie (ANRT), conclu entre les laboratoires Xlim CNRS de l'Université de Poitiers, le LIAS de l'ISAE-ENSMA ainsi que l'entreprise CGR Cinémas, entreprise d'exploitation de salles de cinéma.

Le cinéma, vecteur culturel et de sociabilisation, n'a jamais cessé de fonctionner ni d'évoluer depuis son invention et ses première représentations en 1892. La diversification des modes de consommations des films (télévision, cassette vidéo, internet avec *streaming* et piratage puis plateformes de vidéo à la demande comme Netflix) a amené le cinéma à chercher des alternatives afin de rester compétitif, créant ainsi des formats prémium, ou PLF, pour les plus grands complexes (ICE, IMAX, Barco Escape, Dolby Theater, ScreenX, 4DX), et de l'évènementiel pour les plus petits (ciné-débat, quiz concours).

ICE est la salle prémium conçue par CGR Cinémas en 2017. Elle présente plusieurs équipements, comme des écrans basse résolution placés sur les murs latéraux de la salle, des projecteurs PAR, des projecteurs à tête mobile, ainsi que de l'éclairage arrière. Les écrans LED, nommés panneaux ICE, sont placés dans la continuité de la toile de cinéma avec pour objectif d'améliorer l'immersion des spectateurs comme on peut le voir sur la figure 1.1.

Par défaut, les panneaux ICE viennent dupliquer les bords du films (360 px de gauche et 360 px de droite) afin de le prolonger horizontalement. Pour une sortie ICE, les graphistes doivent parcourir tout le film via leur logiciel de montage Adobe Premiere Pro, en défilant plan par plan. Pour chacun de ces plans, le graphiste va décider quel traitement d'image et de vidéo appliquer sur les bords de l'écran avant de les afficher dans les panneaux ICE. Cela peut signifier le floutage de l'image d'origine, le remplacement par un aplat de couleur, un ralentissement, un changement de la teinte dans un espace de couleur différent, un fondu enchainé entre deux options, etc. Bien que l'équipe converge vers un consensus commun, la décision est totalement subjective et dépend de l'effet procuré en salle de cinéma, pour savoir si un traitement d'image est perturbant en vision périphérique lointaine ou non. Ainsi, le raisonnement est une tâche de très haut niveau d'abstraction et ne peut que très difficilement (voire pas du tout) être formalisée.

On comprend alors qu'une des problématiques de ICE est le temps de postproduction conséquent : trois semaines sont nécessaires au développement d'un film pour ICE par une



FIGURE 1.1 – **Trame de la bande-annonce de *Furiosa : Une saga Mad Max* lors d'une séance ICE.** La duplication des bords du film mise en place par l'équipe de postproduction ICE chargée de choisir les contenus des écrans LED latéraux permet de prolonger l'image.

équipe de cinq graphistes de CGR avant la sortie du film. Dans le cas où l'on souhaite accélérer cette décision de manière informatique, un algorithme formel peu complexe suivant des règles, à la manière d'un système expert ou d'un graphe de décision ne pourra être qu'au maximum aussi efficace que le permet la qualité du formalisme mis en place. Compte tenu du fait que la difficulté de la tâche vient justement de ce formalisme à établir sur cette réflexion subjective, il convient beaucoup mieux de prendre la stratégie inverse. L'apprentissage profond permet d'inférer, à partir des entrées et des sorties correspondantes, les règles atomiques qui régissent la tâche. Ce dernier a maintes fois prouvé être plus efficace que les systèmes traditionnels en ce qui concerne les tâches à haute abstraction.

D'autre part, une seconde problématique vient du fait que la possibilité créative permise par la technique ICE est relativement restreinte, car tout contenu des panneaux ICE provient d'une duplication des bords du film. Là encore, les traitements d'image existants viennent d'algorithmes plus traditionnels, mais l'apprentissage profond peut être utilisé afin de créer un nouveau type de traitement d'image d'un haut niveau d'abstraction, comme par exemple « la reconstruction de l'arrière-plan de la scène » pour constituer l'image des panneaux ICE.

Ces dernières années, l'IA a permis de faire progresser de manière spectaculaire des domaines très différents, grâce à la modularité de l'apprentissage profond ainsi qu'à l'amélioration itérative qu'il facilite. De l'analyse d'images (VGG16, 2014) à l'analyse du langage (ChatGPT, 2022) en passant par la génération d'images (MidJourney 2022, Dall-E 2021), les bibliothèques informatiques destinées aux développeurs et les services commerciaux qui s'adressent au grand public se décuplent d'année en année. Cependant, plusieurs défis restent à relever pour la postproduction ICE à l'aide d'IA. En effet, l'apprentissage profond sur les données vidéos est plus complexe tant en termes d'architectures que de ressources computationnelles (mémoire VRAM, vitesse de calcul). De plus, très peu de travaux traitent de films ou du cinéma, et les autres travaux scientifiques qui traitent de vidéos utilisent souvent des bases de données relativement simples, avec des séquences vidéo controlées de quelques secondes uniquement. Pour finir, le dernier défi correspond à l'industrie du cinéma et du graphisme, qui a dans l'histoire été réfractaire à maintes reprises au progrès technique inexorable et impitoyable, et qui le sont davantage encore à l'IA. Il va donc falloir travailler sur un rendu irréprochable.

Ce manuscrit est organisé en quatres chapitres. Le premier traite de l'invention du cinéma, afin de mieux comprendre ce qui motive cette mouvance et pouvoir la faire évoluer dans le bon sens. Nous verrons que toutes les modalités du cinéma, qu'il soit question du film en lui-même, de son format d'image, de son accompagnement sonore ou d'autres expérimentations de modalités, toutes ces évolutions sont motivés par l'immersion recherchée chez le spectateur. Nous examinerons également les tensions entre l'industrie cinématographique et les autres modes de transmission du film, non pas dans une logique d'imitation, mais dans une perspective de différenciation, où chacun développe ses propres spécificités pour répondre à des besoins utilisateurs légèrement différents. Enfin, nous proposerons une définition de l'immersion dans le contexte d'une expérience audiovisuelle et discuterons des moyens de l'améliorer, en nous appuyant sur la littérature scientifique. Ce chapitre se conclura par une présentation des fondements de l'intelligence artificielle, comprenant une définition, un aperçu historique et une introduction à l'apprentissage profond.

Le deuxième chapitre sera consacré à la formalisation du processus de création ICE. Il décrira la manière dont les graphistes conçoivent les extensions latérales et les critères esthétiques, narratifs et perceptifs qui orientent leurs choix. Dans un second temps, une étude complète de l'analyse filmique sera menée afin d'identifier les consensus établis par les réalisateurs autour des effets recherchés sur le spectateur. Enfin, une modélisation de ces connaissances est réalisée sous la forme d'un graphe sémantique puis d'une ontologie ICE-HOLSS , permettant de structurer les relations entre les composantes filmiques, les intentions visuelles et les effets perceptifs. Cette étape constitue le socle nécessaire pour la conception d'une solution automatique proposée dans le chapitre suivant.

Le troisième chapitre présente la conception d'un réseau de neurones visant à automatiser le processus de décision des graphistes ICE. Il débute par un état de l'art sur l'analyse vidéo en apprentissage profond, afin d'identifier les approches les plus adaptées. Le modèle R-ICE , basé sur un réseau de neurones convolutif tridimensionnel, est ensuite décrit et entrainé sur les données issues des productions ICE depuis 2017. La constitution et la gestion de la base de données sont détaillées, ainsi qu'une augmentation de données, une étude d'ablation et la recherche du meilleur classifieur. Le chapitre introduit ensuite ICE-AGE , une version enrichie intégrant un mécanisme d'attention.

Dans le quatrième chapitre, la nécessité de cohérence spatiale entre la toile de cinéma et les panneaux ICE sera abordée. Celle-ci intervient dans le processus de création et restreint la créativité potentielle des graphistes ICE. Le modèle ICE-Cubed , incorporant plusieurs architectures de réseaux de neurones, sera proposé afin de résoudre ce problème. Concrètement, les éléments qui perturbent l'immersion lorsqu'ils sont situés dans la vision périphérique seront détectés puis *inpaintés* (gommés) par notre solution, permettant ainsi de lever ces obstacles aux graphistes. Pour ce faire, la création de SSPICE , un logiciel libre d'annotation de vidéo a été réalisée, ce qui a permis la conception de la base de données de films sur-mesure DICE .

Ces données ont été utilisées pour maximiser les performances de notre modèle unique en son genre. Enfin, une étude perceptuelle a été conduite en salle de cinéma ICE, avec les cinq experts de ICE pour vérifier la validité et l'utilité de la solution proposée.

En conclusion de ces travaux, nous récapitulerons les principaux résultats obtenus ainsi que les perspectives envisagées.

Publications et réalisations

Conférences :

- Traparic D., De Murcia M., Larabi C., Bellatreche L., 2025. ICE-Cubed : Inpainting of Cinematographic Elements for Intelligent Context Expansion. Dans : 2025 Springer Advanced Concepts for Intelligent Vision Systems (ACIVS).
- Traparic D., Larabi C., Bellatreche L., 2023. Towards Automatic Content Generation for Immersive Cinema Theater Based on Artificial Intelligence. Dans : 2025 IEEE 26th International Workshop on Multimedia Signal Processing (MMSP).
- Sendjasni A., Traparic D., Larabi C., 2022. Investigating normalization methods for CNN-based image quality assessment. Dans : 2022 IEEE International Conference on Image Processing (ICIP).
- Sendjasni A., Traparic D., Larabi C., 2023. Exploration de l'impact de la normalisation sur la performance de l'évaluation de la qualité basée sur des réseaux de neurones convolutionnels. Dans : 29ème Colloque Francophone de Traitement du Signal et des Images (GRETSI).

Codes :

- Traparic D., 2025, Scene Segmentation Platform by ICE, Apache License, `https://github.com/dtraparic/SSPICE`

CHAPITRE 1

LE CINÉMA IMMERSIF ET L'IA



Il est un fait patent, que ceux-là même qui ont le plus envie de gouverner les gens sont, *ipso facto*, les moins aptes à le faire.

Douglas Adams

Table des matières

1.1 Introduction 31
1.2 L'histoire de l'audiovisuel comme expérience immersive 31
1.2.1 Le début de la narration visuelle et le précinéma 31
1.2.2 Le cinéma sans caméra 34
1.2.3 Le cinéma photographique et numérique 35
1.2.4 L'évolution du son au cinéma comme modalité d'immersion 36
1.2.5 L'évolution des écrans de cinéma 37
1.2.6 L'évolution de la préproduction et de la postproduction 37
1.3 Le cinéma aujourd'hui : La réinvention de l'expérience face à la vidéo 38
1.3.1 L'apparition de la chronologie des médias en France 38
1.3.2 Le cinéma se réinvente face à la SVOD 39
1.3.3 Description du format ICE comme innovation immersive 41
1.4 Immersion 43
1.4.1 Immersion : définition et amélioration 43
1.4.2 L'immersion appliquée à ICE : facteurs disruptifs et actions permises 43
1.5 État de l'art par rapport à ICE 46
1.5.1 Les systèmes « Focus+Contexte » 46
1.5.2 Les systèmes « CAVE » et casques VR 48

1.6 L'apprentissage profond et sa place au sein de l'IA 49
1.6.1 Définitions 49
1.6.2 Le paradoxe de « l'effet IA » et redéfinitions 50
1.6.3 L'IA, de la fiction à aujourd'hui 53
1.6.4 La place de l'IA dans la société : Hivers de l'IA et luttes sociales . . 54
L'IA vue par les financements et le capital 54
L'IA vue par le peuple 55
1.6.5 Les rouages de l'apprentissage profond 58
Le perceptron de Rosenblatt 58
Fonctions d'activation modernes 59
Le perceptron multicouche et la descente de gradient 60
Couches entièrement connectées 62
1.6.6 Réseaux de neurones convolutifs 62
Convolution 1D 62
Convolution 2D 63
Sous-échantillonnage ou *pooling* 64
Architecture globale des CNN 64
1.7 Conclusion 65

1.1 Introduction

Bien que le cinéma évolue depuis sa création en 1892, la recherche d'une création audiovisuelle, sans but fonctionnel mais avec un objectif d'immersion et d'engagement remonte à bien plus longtemps que ce que l'on pourrait s'imaginer. Les peintures rupestres des grottes permettaient déjà la réalisation de spectacles narratifs devant un petit groupe d'individus.

Là où le cinéma des frères Lumière avait, à l'époque de sa création, le monopole de tous les types de consommation audiovisuels, ce n'est plus le cas de nos jours. L'arrivée de la télévision qui propose des films aux Français, puis de la cassette vidéo, ainsi qu'internet qui donne accès au *streaming* et au piratage, et à nos jours, les plateformes de SVOD, ont tous comme ancêtre commun le cinéma. Tous se partagent désormais des parts de plus en plus fines d'une même clientèle potentielle, et les salles perdent en compétitivité. Tous ces modes de transmission, bien qu'ils aient l'air de ponctionner la part de marché du cinéma, dépendent également de lui, puisque c'est en quasi-totalité de ce dernier que viennent les produits qu'ils transmettent.

Il convient de protéger le cinéma en revalorisant son intérêt. Une des manières d'y parvenir est de créer des salles de cinéma haut de gamme, comme ce que fait ICE , 4DX ou IMAX, afin d'améliorer l'immersion du spectateur par rapport à un cinéma classique ou un home-cinéma. Se pose alors la question de comment définir l'immersion, et comment l'améliorer dans un contexte comme celui-ci ? Peut-on l'améliorer grâce à la conception de réseaux de neurones ?

Ce chapitre traite d'abord de l'historique du cinéma ainsi que de l'évolution de chacune des modalités qui participe à l'immersion de l'expérience. Vient ensuite un état de l'art sur la littérature scientifique au sujet de l'immersion et des dispositifs d'affichage immersifs, avant de présenter l'intelligence artificielle, ce qu'elle est capable de faire et les architectures en réseaux de neurones.

1.2 L'histoire de l'audiovisuel comme expérience immersive

1.2.1 Le début de la narration visuelle et le précinéma

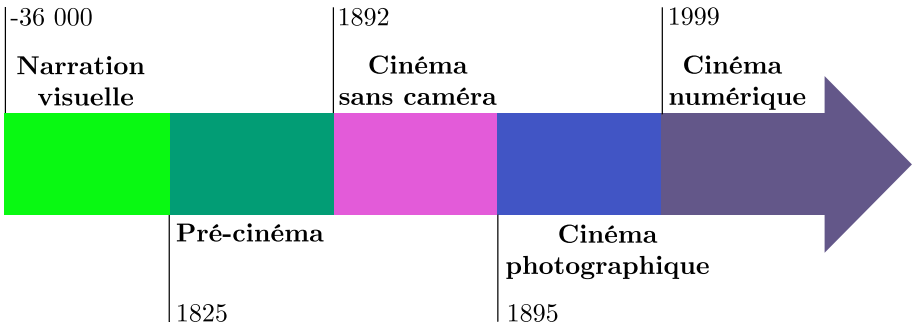


FIGURE 1.1 – **Frise chronologique de l'histoire du cinéma.** La Narration visuelle commence dès 36 000 av. J.-C., soit 26 000 avant même le début de la civilisation et l'agriculture. Le précinéma, le cinéma sans caméra, le cinéma photographique puis le cinéma photographique sont différents états que l'on peut distinguer de l'évolution du cinéma, à partir de 1825. (Crédit : auteur [M1])

Depuis son invention, le cinéma a suivi deux grandes dynamiques en termes d'évolution : une quête de réalisme sensoriel que l'on peut nommer « course à l'immersion » et un

développement continu de moyens techniques, ou « course à la technique ». Souvent, les avancées technologiques sont précisément motivées par la volonté de plonger davantage le spectateur dans l'image, d'amplifier l'effet de présence, ou de repousser les limites de la perception audiovisuelle. L'immersion devient ainsi un moteur narratif et perceptif, traversant les siècles sous différentes formes. La projection collective, c'est-à-dire celle destinée à être visionnée simultanément par un même public et non par un individu seul est ce qui définit le cinéma. C'est ainsi que nous pourrons le situer chronologiquement sur la figure 1.1.



a) b) c) d)

FIGURE 1.2 – **Photo des premiers dispositifs de narration visuelle et des jouets optiques du précinéma comme précurseurs du cinéma d'aujourd'hui**
a) Grotte de Chauvet b) Thaumatrope c) Phénakistiscope d) Zootrope

Néanmoins, la recherche de l'immersion pour des spectateurs commence bien avant comme avec les peintures rupestres de la grotte Chauvet (environ il y a 38 000 ans), qui exploitent la courbure des parois et les effets de torche pour suggérer du mouvement et une profondeur. On peut voir ces effets recherchés sur la figure 1.2-a, où certaines parties du cheval sont dessinées plusieurs fois pour symboliser leur mouvement. Ces peintures sont en effet un dispositif immersif primitif de narration visuelle, la première époque distincte sur notre figure 1.1, limité à un petit groupe de spectateurs [9] [10].

Bien plus tard, mais toujours en narration audiovisuelle, la lanterne magique (1659, voir figure 1.3) est le premier instrument de projection. Lors de spectacles, elle permet de projeter à un public élargi des images peintes sur verre, parfois accompagnées de commentaires ou de musique. La plaque de verre peut glisser devant l'objectif pour créer un mouvement dans le décor, certaines lanternes magiques ont également deux objectifs pour permettre un fondu enchainé entre deux plaques d'image différente. La lanterne magique introduit ainsi une première forme d'expérience collective combinant image et son, au service d'un récit.

Par la suite, la période du précinéma est marquée par divers dispositifs nommés « jouets optiques » qui ont été conçus pour créer l'illusion du mouvement. Ces derniers exploitent des phénomènes tels que la persistance rétinienne d'un côté ainsi que l'effet phi, représentant la sensation de mouvement lors de l'apparition successive d'images ou de différents motifs statiques. Devant ces différents états du contenu visuel analogue à des trames de vidéos numériques, le cerveau reconstitue le mouvement de transition entre ces différents états. Ce phénomène s'observe également sur des dispositifs lumineux basse résolution tels que les affichages de bus ou de métros, où les lettres semblent défiler alors qu'elles apparaissent

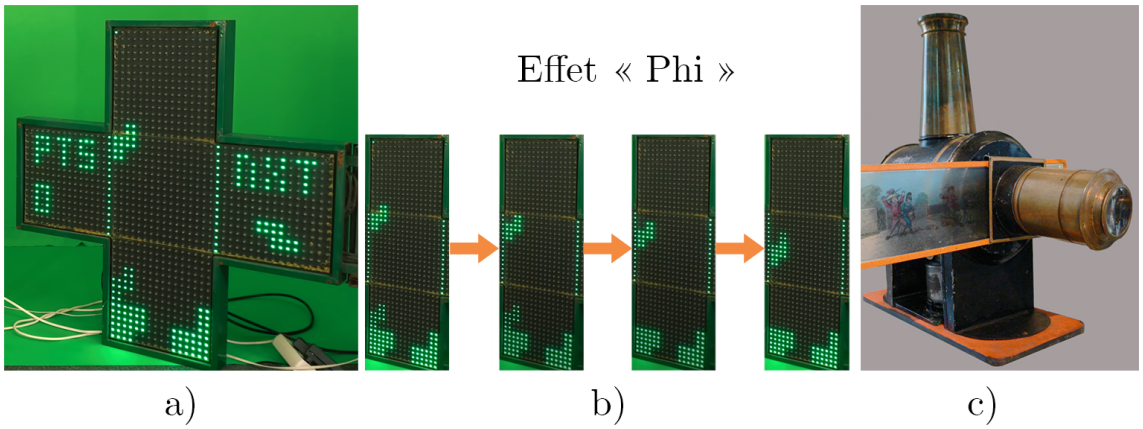


FIGURE 1.3 – **Démonstration de l'effet phi et photographie d'une lanterne magique.** Aucune élément ne bouge physiquement sur la croix de pharmacie, le mouvement est pourtant compris et ressenti par le cerveau humain. (Crédit : [M2])

de façon discrète [1] et séquentielle, comme illustré dans la figure 1.3.

Parmi ces inventions, le thaumatrope (1825), le phénakistiscope (1832), le zootrope (1834), le folioscope (ou *flipbook*, 1868) et le praxinoscope (1877) ont marqué l'histoire du précinéma. Ces appareils permettaient de visualiser des séquences d'images en boucle, offrant ainsi une première expérience de l'animation.

L'inventeur du phénakistiscope, Joseph Plateau, s'intéressait particulièrement à la persistance rétinienne et à ses implications sur la perception visuelle, allant jusqu'à étudier les effets de regarder directement le soleil pour comprendre la durée pendant laquelle une image reste sur la rétine. Plus tard en 1865, Antoine Claudet expérimentait avec un prototype de phénakistiscope destiné à la vision stéréoscopique, et c'est à travers ce dispositif qu'il fut amené à étudier les points fondamentaux de la persistance rétinienne, poursuivant ainsi les travaux de Plateau. Il montre notamment que le cerveau humain est capable de reconstruire les parties manquantes d'un mouvement ou d'une animation pour maintenir une sensation continue. Ses recherches mettent en évidence le lien étroit entre la perception physiologique de la rétine et la capacité cognitive à reconstituer le mouvement. Toutes ces investigations témoignent de la volonté des pionniers du précinéma de combiner compréhension scientifique et innovation technique pour enrichir l'expérience visuelle.

Inspirés par les jouets optiques de la période précinéma, le **Kinétoscope**, visible sur la figure 1.4, breveté par Thomas Edison et William Dickson en 1891 et accessible au public en 1893, et le **Mutoscope**, breveté par Herman Casler en 1894, représentent les premières expériences de films photographiques destinés à un visionnage individuel. Le Kinétoscope, utilise une bande animée passant devant un œilleton, tandis que le Mutoscope fonctionne comme un folioscope mécanique (voir figure 1.4-a) où les images imprimées sur des cartes sont actionnées par rotation manuelle. Bien que ces dispositifs permettent déjà de visualiser des photographies animées, ils restent néanmoins rattachés au précinéma, car ne pouvant pas assurer une projection collective. Cependant, des salles commerciales spécialisées ont

1. Discrète ici réfère aux mathématiques discrètes, par opposition à « continue ». Les textes numériques défilant ont un nombre fini de positions possibles, tandis qu'une écriture manuscrite sur un papier aurait un nombre infini de possibilités. Le même exemple décrit également la différence entre ce qui est analogique et numérique, tout ce qui est numérique étant nécessairement discret.

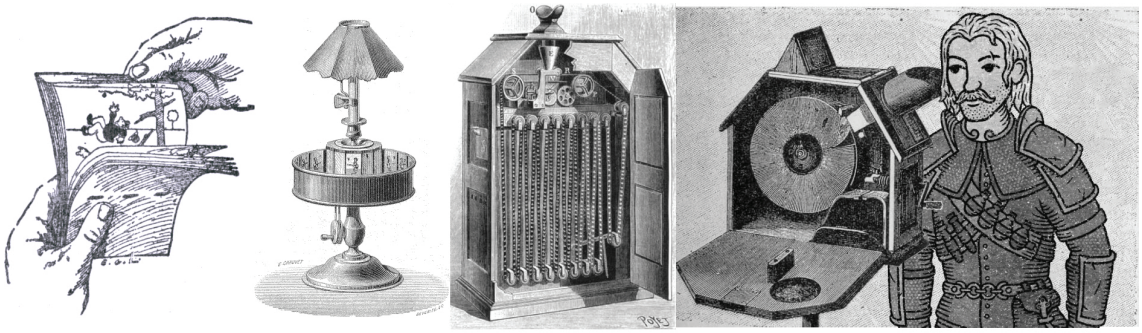


FIGURE 1.4 – **Gravures de dispositifs d'animation à usages individuels.** a) Folioscope b) Praxinoscope c) Kinétoscope d) Mutoscope. Le Kinétoscope et le Mutoscope sont commercialisés dans des salles prévues à cet effet et sont accessibles pour une visualisation individuelle. [M5]

proposé ces appareils à un public, offrant une expérience partagée et préparant le terrain pour la diffusion publique du cinéma photographique.

1.2.2 Le cinéma sans caméra

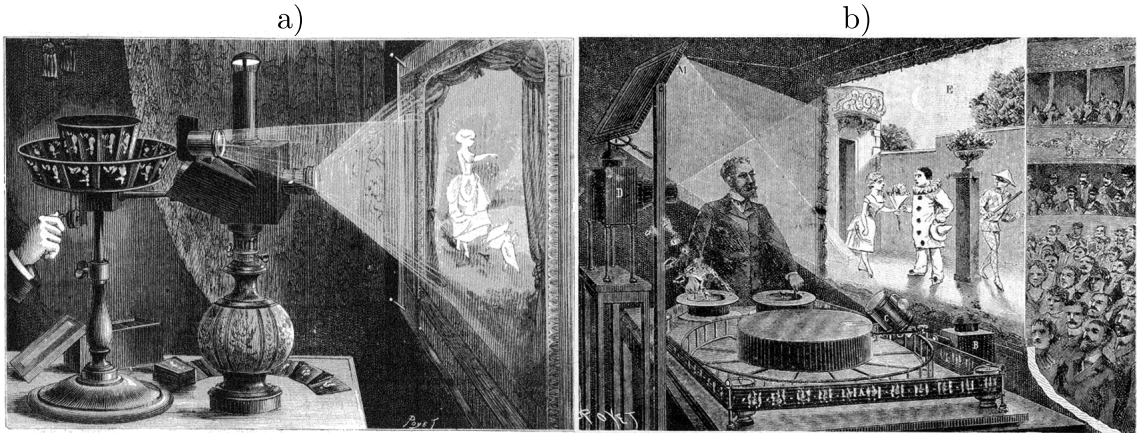


FIGURE 1.5 – **a) Praxinoscope à projection (1880) et b) Théatre optique des dessins animés d'Émile Reynaud (1888).** Le premier permet la diffusion à un public d'une boucle d'animation de quelques secondes sur un décor projeté indépendamment. Le second est la première diffusion publique de cinéma et dure plusieurs minutes. (Crédit : [M4])

Émile Reynaud, inventeur du praxinoscope de 1877, l'améliore pour créer le praxinoscope à projection en 1880 visible sur la figure 1.5-a qui permet ainsi un visionnage à un public à des boucles d'animation de quelques secondes, ce qui reste caractéristique du précinéma. Cependant, il présente en 1892 son théâtre optique, une avancée majeure dans le domaine de l'animation, marquant ainsi le début du cinéma avec des films de plusieurs minutes. Ce dispositif permettait de projeter sur grand écran des images peintes à la main, par dessus un décor fixe lui aussi projeté, donnant naissance aux premières projections publiques de dessins animés, telles que Pauvre Pierrot, dont la pellicule et la trame est visible en figure 1.6-a et 1.6-b. Ces projections accompagnées de musique en direct offraient une narration complète à une assemblée collective, seulement un an après le brevet du

Kinétoscope, mais surtout un an avant l'accessibilité de ce dernier au public et trois ans avant les films des frères Lumière. Dans Pauvre Pierrot, Reynaud intègre un système innovant de synchronisation sonore. Des tiges d'argent disposées sur la bande flexible à des moments clés actionnent un bruiteur qui, à l'aide d'un électro-aimant, produit des sons synchronisés avec l'image, tels que le bruit des coups portés sur Pierrot par Arlequin. Cette technique de synchronisation mécanique a d'ailleurs perduré bien au-delà des premières années du cinéma : jusqu'à la fin de l'utilisation de la pellicule 35 mm (2010 en France), les projecteurs utilisaient encore des repères métalliques adhésifs, surnommés « scotch magnétique », pour automatiser la transition entre les publicités, les bandes-annonces et le film principal au fur et à mesure du défilement de la pellicule.

1.2.3 Le cinéma photographique et numérique

Comme mentionné plus haut, le Kinétoscope marque le premier visionnage individuel d'un film photographique en 1893 et le Théatre optique marque le premier visionnage collective d'un film animé en 1892, que l'on considère comme le commencement du cinéma. Malgré les demandes de son associé, Edison refuse de faire étudier un appareil de projection sur écran pour une assemblée (affirmant que « l'on tuerait ainsi la poule aux œufs d'or » [11]).

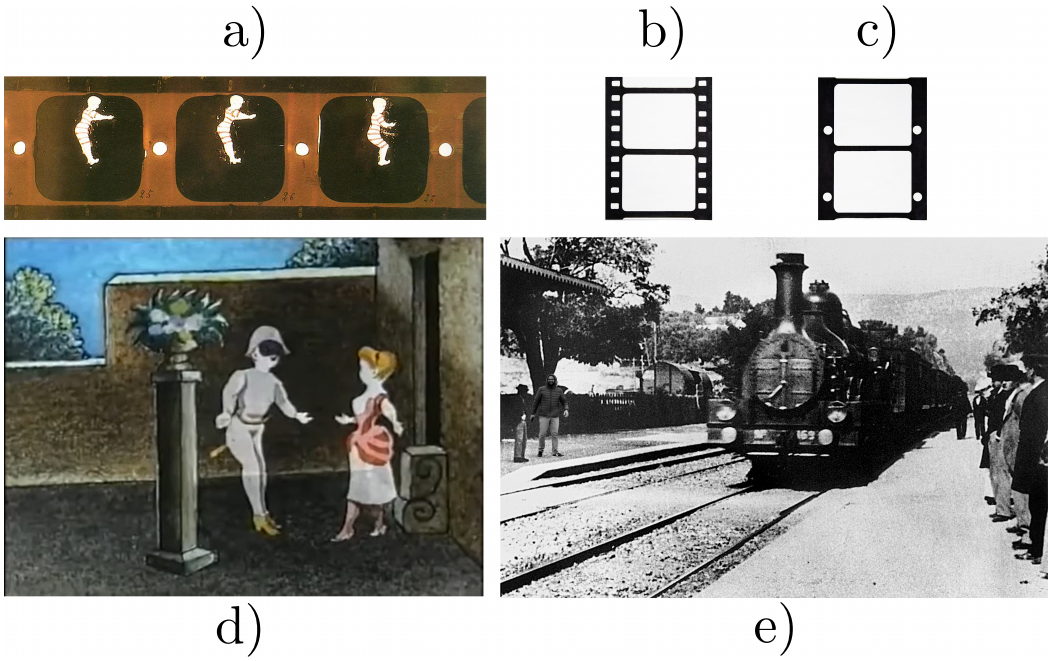


FIGURE 1.6 – **Le début du cinéma.** a) Pellicule de Reynaud b) Pellicule d'Edison c) Pellicule des frères Lumière d) Trame de « Pauvre Pierrot » e) Trame de « L'arrivée d'un train en gare de La Ciotat », l'une des premières diffusions photographiques animées publiques des Frères Lumière, les précurseurs du cinéma photographique.

Peu de temps après, les frères Lumière présentent en 1895 leurs premiers films au travers d'une projection publique, puis en 1896 « L'Arrivée d'un train en gare de La Ciotat ». Une trame du film est visible sur la figure 1.6, en plus d'une comparaison entre le format de pellicule des Frères Lumière, celui d'Edison, et celui de Reynaud. Si la célèbre panique du public pendant la projection, fréquemment relayée par la presse à posteriori relève davantage du mythe que du fait documenté, certains historiens doutent d'une telle

surréaction [12]. Malgré l'exagération, la scène illustre parfaitement la puissance immersive de l'image animée naissante.

Gunning introduit ainsi le concept du « cinéma des attractions », où l'accent est mis sur l'aspect spectaculaire de l'image plutôt que sur une immersion narrative.

1.2.4 L'évolution du son au cinéma comme modalité d'immersion



FIGURE 1.7 – **Console d'orgue de cinéma**, le point culminant de l'accompagnement sonore des cinémas avant qu'il soit remplacé par une bande sonore enregistrée. La performance de l'organiste qui en jouait durant les séances pouvaient recréer des musiques d'ambiances cohérentes ainsi qu'une grande palette de bruitages souhaités par le film.

La recherche de l'engagement et l'immersion du spectateur passe aussi par le son, qui a un grand rôle dans l'histoire du cinéma. D'après Azéma, des évidences montrent que l'aspect acoustique n'était déjà pas anodine devant les fresques du paléolithique [9], mais il a aussi son importance dès les projections des temps modernes. En effet, dès les premières projections publiques des vues photographiques animées de Louis Lumière en 1895, la musique accompagnait les films, souvent interprétée en direct par un pianiste ou un petit ensemble musical. Cette pratique visait à compenser l'absence de son enregistré et à renforcer l'engagement émotionnel du spectateur. À partir des années 1910, l'accompagnement musical se formalise davantage. Les maisons de production commencent à fournir des indications musicales spécifiques avec les bobines de films, orientant les musiciens sur les styles ou les morceaux appropriés pour différentes scènes. Parallèlement, l'orgue de cinéma, visible sur la figure 1.7, émerge comme une solution innovante pour enrichir l'expérience sonore. Inspiré de l'orgue orchestral conçu par Robert Hope-Jones vers 1890, cet instrument est conçu pour imiter les timbres des instruments d'orchestre et intégrer des effets sonores. Son adoption se généralise dans les grandes salles de cinéma à partir des années 1920, offrant une alternative économique et polyvalente aux orchestres traditionnels.

Cependant, l'avènement du cinéma parlant à la fin des années 1920, marqué par des films tels que Le Chanteur de jazz (1927) qui mèle voix, musique, et bruitages, rend progressivement obsolètes les ciné-concerts et les orgues de cinéma. La standardisation du son enregistré et les considérations économiques conduisent à leur déclin. Néanmoins,

ces pratiques connaissent des résurgences ponctuelles dans des contextes patrimoniaux ou artistiques, témoignant de leur valeur historique et esthétique.

1.2.5 L'évolution des écrans de cinéma

De la même manière que le son, la format de l'écran de projection a beaucoup évolué avant d'être stabilisé, et ce, conjointement avec les écrans d'ordinateur. L'histoire des formats d'image en cinéma reflète une coévolution constante entre contraintes techniques, volontés immersives et standards industriels. Les premiers films projetés, comme ceux des frères Lumière (1895) ou les productions Edison, étaient tournés en pellicule 35 mm avec un ratio proche du 1,33:1 (équivalent au 4/3), qui devient un standard en 1909. En 1932, l'introduction du son optique, qui consistait à inscrire sur la pellicule des variations lumineuses traduites ensuite en signal électrique par une cellule photoélectrique afin d'être restituées en son bouleverse le format de l'image. L'« Académie des arts et des sciences du cinéma » est ainsi conduite à normaliser un format légèrement élargi, le 1,37:1, nommé ratio académique, pour compenser la bande son ajoutée sur la pellicule. Ce format reste dominant jusqu'aux années 1950, où l'arrivée de la télévision en 4/3 pousse les studios à chercher une différenciation visuelle. C'est dans ce contexte que naissent les formats larges, dont le *CinemaScope* (2,35:1 puis 2,39:1). Dans les années 1990, le ratio 1,78:1 (ou 16/9) s'impose avec l'avènement de la télévision HD , facilitant l'adaptation des films de cinéma (2,35:1) aux écrans domestiques. Cette évolution conjointe des formats, tant en salle qu'à la maison, illustre la tension permanente entre immersion spatiale, fidélité technique et exigences industrielles.

1.2.6 L'évolution de la préproduction et de la postproduction

Côté technique, les évolutions de la préproduction et de la postproduction ne se font pas non plus attendre. Avant même les projections collectives, les réalisateurs de film expérimentent des procédés visuels pour dépasser la simple captation réaliste. Dès 1896, Georges Méliès introduit des trucages en caméra (Le Voyage dans la Lune, 1902), en jouant sur les fondus, les peintures sur caches (*matte painting*), et les surimpressions : les premières formes d'effets spéciaux en prise de vues. Dans les années 1930, l'apparition de la technicolor trichrome (notamment dans Le Magicien d'Oz, 1939) permet un nouveau traitement esthétique en postproduction, accompagné des premiers effets d'incrustation optique. Ce film constitue également un jalon pour l'incrustation de décors sur pellicule, grâce au procédé d'extraction de fond (*matting*). L'incrustation chromatique [13], issue du procédé par cache mobile (travelling matte process) développé dans les années 1940, se généralise dans les années 1950–1960 et devient un standard numérique à partir des années 1990. Parallèlement, la préproduction s'enrichit avec l'arrivée des maquettes physiques (notamment chez Kubrick), puis des scénarimages numériques et des outils de prévisualisation dans les années 2000. L'usage de la composition numérique se généralise avec Jurassic Park (1993), et la capture de mouvement devient un pilier avec Le seigneur des anneaux (2001–2003). Toutes ces techniques sont perfectionnées, jusqu'au StageCraft illustré en figure 1.8, un type d'OSVP (Production Virtuelle sur Plateau), utilisé dans *The Mandalorian* (2019) par Lucasfilm, combinant les concepts de décor virtuel en temps réel avec les murs d'écrans LED que l'on peut souvent voir dans les emissions de journal TV , ancrant la production dans une logique hybride mêlant préproduction, tournage et postproduction dans un même espace. Les acteurs qui jouaient devant des écrans verts d'incrustation peuvent désormais voir le décor qui va les entourer dans le film dès le tournage. Trois OSVP existaient dans le

monde en 2019 contre plus de 300 en 2022, après la popularisation de la technique par le StageCraft.



FIGURE 1.8 – **Le StageCraft, la salle de production virtuelle sur plateau qui a popularisé la technique.** Les décors ne sont plus incrustés en postproduction sur un fond vert , mais sont générés en temps réel pendant le tournage, selon la perspective et la parallaxe de la caméra. Les acteurs peuvent jouer tout en voyant le décor. (Crédit : [M3])

1.3 Le cinéma aujourd'hui : La réinvention de l'expérience face à la vidéo

1.3.1 L'apparition de la chronologie des médias en France

Ce qui s'appelle la chronologie des médias est un concept spécifique à la France. Ces législations définissent les délais d'attente à respecter avant qu'un film sorti au cinéma puisse être disponible en d'autres formats (c'est-à-dire DVD, Blu-ray, télévision, plateformes de *streaming*). Cette chronologie a pour but de protéger l'industrie du cinéma français, via la protection en amont de l'attractivité des salles de cinéma physique. C'est aussi dans la chronologie des médias que sont stipulés les quotas minimum d'œuvres européennes et françaises que les chaines de télévision françaises doivent diffuser, ainsi que le reversement de chiffre d'affaires à la création audiovisuelle française.

Avant même la chronologie des médias, son prédécesseur s'appelait l'exclusivité des cinémas, qui devient la norme dans les années 1920. Dû au prix élevé des copies de films (tirées autour de 40 par film), les plus grands cinémas de France, et les plus luxueux étaient ceux qui avaient l'exclusivité des copies à la sortie du film. Le billet pour les spectateurs étaient aussi le plus élevé. Une fois que ces cinémas d'exclusivité délaissent cette copie, cette dernière revient à la deuxième catégorie de cinémas ; les cinémas de continuation. Puis, à leur tour, les cinémas de troisième catégorie c'est-à-dire les cinémas de quartier récupèrent les droits de diffusion. Avec ce système pyramidale en trois étages, l'ORTF a par la suite naturellement permis aux films d'être affichés à la télévision cinq ans après leurs sorties en salle quand il en était question, mais ce n'était pas un terme

légal, seulement la pratique usuelle. Tant que le support cassette n'existait pas, le besoin de législation officielle ne s'est pas fait ressentir [14]. L'apogée de la TV ainsi que des deux chaines françaises publiques dans les ménages français apparait dans les années 1960 [15] Le nombre de TV en France passe d'un million en 1958 à dix millions en 1968, ce qui engendre la chute de fréquentation des salles de cinéma. Pour y répondre, les premières législations, avec des principes similaires à ceux de la chronologie d'aujourd'hui, tels que des règles de quota de diffusion et de reversement apparaissent enfin en 1972. L'arrivée du support cassette à partir de 1972, (VCR par Philips en 1972, Betamax par Sony en 1975 et enfin VHS en 1976) ainsi que l'annonce d'une troisième chaine de télévision française pour 1974 continuent d'alimenter la peur de l'industrie du cinéma pour leurs salles obscures. C'est ainsi qu'en 1982, le délai pour la sortie en vidéo physique est de 6 à 18 mois, selon les entrées. En 2000, il est fixé à 6 mois, s'alignant ainsi au délai pratiqué dans les pays européens [15]. En 2009, la législation se penche sur les services de vidéo à la demande par abonnement (SVOD), qui obtiendra une fenêtre de 36 mois d'attente, dans les toutes dernières places de la chronologie des médias.

Les hostilités entre le cinéma et la vidéo continuent. En 1993, le PDG de Canal+ André Rousselet dit qu' « il veut voir la vidéo crever » [15]. Ce genre d'hostilité continuera même lorsque la vidéo sur support physique évoluera en vidéo numérique à la demande, via les plateformes de SVOD et internet.

1.3.2 Le cinéma se réinvente face à la SVOD

Bien que la France reste le pays dont la fréquentation en salle a le moins souffert de la télévision [15], la télévision fait chuter la fréquentation en salle. Malgré cela, la peur de cette dernière a eu pour conséquence la chronologie des médias française, un cercle vertueux de l'industrie du cinéma qui s'autofinance par le biais de la télévision. L'industrie s'attend à un phénomène aussi dévastateur avec internet et le piratage qui rend le film quasi instantanément disponible à un cout inexistant, contrastant de manière ahurissante avec le deuxième moyen d'obtenir des films sur internet, c'est-à-dire les moyens légaux, de payer un abonnement de SVOD, après avoir attendu 36 mois après la sortie en cinéma. Pourtant, il n'en est rien, la fréquentation ne diminue pas, même après la démocratisation d'internet dans les foyers français. [15] Elle n'a d'ailleurs jamais été aussi haute qu'en 2017. En 2014, la société Orange refuse d'inclure un abonnement Netflix dans leurs offres, affirmant ne pas vouloir devenir un « Cheval de Troie » malgré eux au profit de Netflix [16].

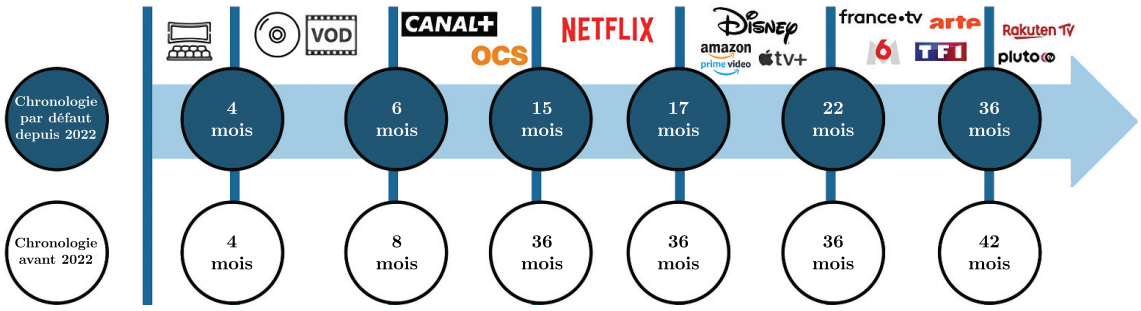


FIGURE 1.9 – **Illustration de la chronologie des médias.**[16] Les films sont d'abord exclusivement au cinéma, puis ils sont disponibles en support physique puis en plateforme à abonnement. Avant 2022, Netflix et Disney+ doivent attendre 36 mois après la sortie d'un film pour pouvoir les diffuser à leurs clients.

Jusqu'en 2022, les plateformes de SVOD dont Netflix, Prime Video et Disney+ seront toujours contraintes à ce délai de 36 mois comme il est illustré sur la figure 1.9, ce qui est une éternité par rapport aux autres pays et ce qui est même plus que les chaines de télévision françaises, quelles qu'elles soient (fixées à 30 mois maximum). En réponse, Netflix produit des films et des séries exclusives, sans les sortir au cinéma en France. Certains de ces films sont présents au festival de Cannes de 2017 ce qui engendre l'indignation [16]. Spielberg affirmera à ce sujet : « Une fois que vous vous engagez dans un format télévisuel, vous devenez un film de télévision... Vous méritez certainement, si c'est une bonne émission, un Emmy, mais pas un Oscar. Je ne crois pas que les films bénéficiant seulement d'une sortie symbolique dans quelques salles [dans le monde] pendant moins d'une semaine devraient être éligibles à une nomination aux Oscars. »[17]. Le festival demande alors à ce que les films Netflix présentés à Cannes soient disponibles au cinéma français, et c'est ainsi que Netflix a pris la décision d'arrêter de présenter ses films aux festivals de Cannes suivants [16].

En 2022, Des négociations sont faites, Disney et Prime parviennent à obtenir un délai de 17 mois contre 15 mois pour Netflix qui doit respecter sa condition de consacrer 20 % de son CA français à des créations audiovisuelles françaises, dont 4 % particulièrement destinés au cinéma français. En 2025, Disney négocie un délai de fenêtre de 9 mois en consacrant 25 % de son CA français en création audiovisuelle française dont 14 % pour le cinéma[18]. La réponse ne se fait pas attendre des autres acteurs : Canal+ menace alors de réduire son engagement (ce qu'il réussira à faire)[19] et Netflix saisit aussitôt le conseil d'état pour déloyauté [20].

Face à cette contraction temporelle et à la concurrence accrue des visionnages à domicile, les salles de cinéma ont été contraintes de proposer une expérience unique, plus engageante pour le client, pour se différencier de celle des écrans domestiques. Le passage de la pandémie, qui a augmenté les utilisations des plateformes de SVOD a également amené à cette direction [21], et on assiste aujourd'hui à la montée des formats prémium [21]. Pour autant, si l'engagement du spectateur peut être amélioré par un important investissement technique de la part des grands circuits de cinéma, les plus petits cinémas ont tout de même trouvé une alternative. On parle alors d'augmenter l'interaction avec le public [17], que ce soit avec des ciné-débats qui sont monnaies courantes chez nos cinémas français certifiés « art et essai », des émissions en direct, des quiz sur le film avec lots à gagner, de diner à thème spécial « Wicked »[22], d'autres utilisations de la salle de cinéma comme par des séminaires d'entreprises, des prises de paroles, des diffusions en direct , des matchs de sports ou d'*e-sports* [23], des activités dans le hall [17] telles que la présence des personnages de *Vaiana 2* joués par des comédiens, une démonstration du club local de taikwendo pour la sortie de *Karate Kid : Legends*. Le CNC a également mis en place un prix de l'innovation pour les cinémas [17], encourageant cette direction.

« Nous découvrons toutes sortes de nouvelles manières de transformer les films en évènements, afin qu'ils deviennent une expérience plus immersive », a déclaré la cheffe Marketing du circuit américain Alamo Drafthouse Cinema, « allant des expériences éphémères aux fêtes pour enfants, jusqu'aux projections déguisées : il y aura des perruques et des lunettes façon Bob Dylan pour Un Parfait Inconnu »[21]. Certains cinémas américains tentent même un *speed dating* autour du thème du cinéma [21], d'autres comme Cinepolis Junior vont jusqu'à créer un terrain de jeux pour enfant dans leurs salles de cinéma [24].

Grâce à tout cela, d'après le patron de la Fédération Nationale des Cinémas Français Richard Patry, le PAF (Paysage Audiovisuel Français) offre plus de diversité que n'importe quelle plateforme numérique internationale [17].

En ce qui concerne les innovations technologiques immersives, la stéréoscopie (cinéma 3D), expérimentée dès les années 1950, revient régulièrement comme tentative de spatialisation de l'image. Ses premières déclinaisons reposaient sur des lunettes anaglyphes à filtres rouge et bleu, qui permettaient de séparer les deux images projetées (chacune destinée à un œil) pour créer l'illusion du relief restituée par le cerveau. Les odeurs synchronisées (Smell-O-Vision, concept inventé en 1960, mais réitéré en 2024 avec le film *Heretic* [21]), ou plus récemment les sièges haptiques, la « cadence élevée » ou technologie HFR (pour *High Frame Rate*) permettant la projection de film en 48 i/s au lieu de 24 i/s ou encore les dispositifs immersifs comme le 4DX témoignent de cette volonté persistante d'englober le spectateur dans un environnement sensoriel total [25][23]. D'autres dispositifs immersifs sortent du cadre de la salle de cinéma tout en restant dans le monde de l'audiovisuel ou du jeu vidéo. Les vestes haptiques ou les casques VR sont des exemples marquants.

Avec la prise de risques d'autant de formats novateurs, certains formats expérimentaux n'ont inévitablement pas survécu : leur échec peut être dû à des limitations techniques, un catalogue trop restreint, un cout d'installation de salle important, un cout de production par film élevé ou encore un manque de confiance initial dans la viabilité du concept. C'est le cas notamment de ScreenX (qui produit une quantité restreinte de films par an) et de Barco Escape et Cinerama (projets éteints). À l'inverse, des formats comme IMAX, bien que très couteux à mettre en place et à produire, ont prouvé leur rendement économique et leur valeur ajoutée reconnue, ce qui a créé une confiance intarisable autant chez les spectateurs que chez les ayant-droits, ainsi qu'une inertie, ce qui explique leur implantation durable aujourd'hui.

L'ensemble de ces dynamiques illustre pourquoi et comment les cinémas ont dû se réinventer pour rester pertinents dans un paysage où la commodité du visionnage chez soi devient la norme.

Notre thèse s'inscrit dans la convergence de deux dynamiques : la course technologique impulsée par l'intelligence artificielle, et la quête d'immersion portée par le concept ICE . Ce dernier permet de contourner certains obstacles majeurs rencontrés par ScreenX, IMAX ou Barco Escape, tels que la lourde main-d'œuvre artistique nécessaire ou le cout prohibitif du matériel, deux facteurs qui freinent considérablement la production annuelle de contenus adaptés, ralentissant ainsi l'adoption à large échelle de ces formats.

Ce manuscrit traite ainsi de la place de l'apprentissage profond au sein d'acteurs de l'industrie cinématographique, avec comme acteur principal, la salle ICE.

1.3.3 Description du format ICE comme innovation immersive

Le format ICE repose sur un concept singulier : prolonger visuellement le film au-delà de l'écran principal en projetant des extensions dynamiques de l'image sur les côtés de la salle à l'aide de panneaux LED dédiés. Si d'autres modalités techniques complètent l'expérience, telles que l'éclairage additionnel par des lumières « PAR » et l'éclairage arrière (*backlights*), les projecteurs à tête motorisée , le confort des sièges, ou encore le son Dolby Atmos (une technologie permettant une spatialisation du son jusqu'à 128 canaux sonores différents), elles ne sont pas retenues ici pour l'analyse. L'équipement lié à l'éclairage de la salle ICE est visible en figure 1.10.

Le traitement d'image mobilisé pour l'extension du film dans les panneaux ICE reste simple d'un point de vue mathématiques : des duplications d'images, des anamorphoses

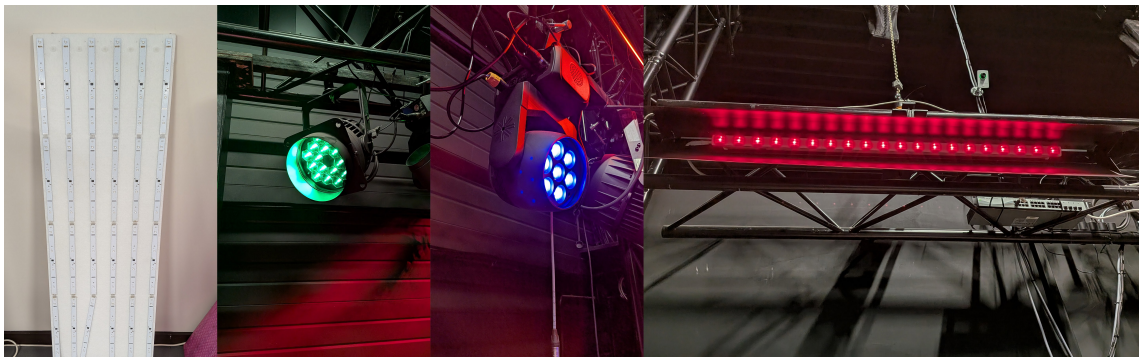


FIGURE 1.10 – **Photographies des équipements ICE.** a) Panneau ICE sans sa toile diffusante. Six centimètres séparent chaque LED. b) Lumière PAR. c) Projecteur à tête motorisée. d) Éclairage arrière.

2D (déformation de la hauteur et largeur de l'image), changement d'espaces de couleur, un suréchantillonnage des trames pour allonger une séquence vidéo, etc. En revanche, la réflexion derrière l'adoption ou non de telle ou telle technique, pour un plan donné, est complexe et difficile à formaliser. Cette réflexion doit être réalisée individuellement pour l'ensemble des plans du film, et ce, sur des outils de travail particulièrement lourds et exigeants en ressources, mémoire et calculatoire.

Le format ICE présente une adoption significative, comme en témoignent son expansion internationale et l'intérêt soutenu des studios et du public pour ce concept. Aujourd'hui, le réseau ICE compte 57 salles dans le monde, dont 11 à l'international (Équateur, Estonie, Inde, États-Unis, Arabie Saoudite, Espagne et Allemagne). Depuis sa création en 2016, ICE Theaters collabore activement avec des studios de production et a développé une compréhension approfondie de la vision des réalisateurs, s'efforçant de la restituer à travers sa technologie. *Lightstorm Entertainment*, la société de production à l'origine d'Avatar, s'est confiée officieusement à l'équipe ICE en expliquant que « ICE est le seul format de salle qui fait du cinéma », en ce sens qu'il intègre véritablement la réalisation de film dans son travail de postproduction.

Le deploiement du format ICE soulève plusieurs enjeux complementaires. Il constitue tout d'abord un enjeu **économique** : chez CGR, environ 33 % des bénéfices liés à la billetterie proviennent des projections en salle ICE, soulignant l'intérêt indiscutable des spectateurs pour ce format, malgré un prix supérieur au billet standard. Cet aspect à forte incidence financière s'accompagne évidemment d'un **enjeu de temps de production** : un film adapté au format ICE mobilise une équipe de postproduction de cinq experts pendant trois semaines. Enfin, ces defis techniques et temporels sont étroitement liés à un **enjeu de ressources humaines et de formation** : chaque nouvel expert ICE doit suivre un mois complet d'exercices pratiques avant d'être autonome et operationnel sur les premières tâches liées à la salle immersive. Ensemble, ces trois dimensions illustrent la complexité de la mise en œuvre de ICE et la nécessité de coordonner économie, temps et expertise pour que la salle immersive soit durable.

On peut voir en figure 1.11 le document interne permettant d'expliquer le travail de l'équipe aux visiteurs du siège. Les étapes 4 et 5 sont celles qui nous intéressent puisque ce sont ces étapes que nous devons modéliser. Ces étapes sont faites sur les films, mais elles sont aussi réalisées en amont sur les bandes-annonces. À titre de comparaison d'expertise, le montage d'une bande-annonce ou d'une bobine de film nécessite 1 mois de formation. Les

calibrations d'une bande-annonce et d'une bobine de film nécessite respectivement 6 mois et 1,5 an. Cette dernière étape qui consiste à calibrer des couleurs d'un film (c'est-à-dire faire correspondre les couleurs de la toile avec celles des panneaux ICE), bien que rébarbatif et laissant peu de place à la créativité, requiert une grande expérience, tout en prenant un temps considérable. En effet, ce travail de colorimétrie fait à vue a un poids très élevé pour que l'expérience au global (et son immersion) puisse fonctionner.

L'immersion dont il est question est le point central de ICE, en stimulant la vision périphérique. Il est donc important de définir ce qu'est cette immersion, et comment il est possible de l'améliorer.

1.4 Immersion

1.4.1 Immersion : définition et amélioration

Le concept d'immersion dans les expériences virtuelles peut être défini et quantifié de multiples façons. Cette section examinera certains des principaux concepts impliqués dans les aspects immersifs de ce projet. L'idée d'être entouré et de vivre des stimuli virtuels multisensoriels a été explorée dans de nombreuses publications, bien que les termes utilisés ne soient pas toujours identiques. Agrawal [26] unifie les termes des auteurs sous le concept de « immersion perceptive ». Cette dernière peut être vécue lors d'une expérience virtuelle de différentes manières combinables. La première consiste à inhiber les stimuli du monde réel [27], l'autre manière est de remplir le système sensoriel humain, de la manière la plus complète possible, par des stimuli virtuels [27]. Dans le même esprit, Lombard et Ditton [28] affirment que l'immersion perceptive peut être mesurée en comptant le nombre de sens fournis à l'utilisateur et en observant dans quelle mesure les entrées de l'environnement physique sont désactivées. Allant plus loin dans cette direction, Slater et Wilbur [29] abordent la question de la multimodalité des stimuli en déclarant : « Plus le système fournit des affichages dans toutes les modalités sensorielles, tout en préservant la fidélité par rapport à leurs équivalents sensoriels du monde réel, plus l'expérience est immersive ». Ermi et Mäyrä [30] affirment que les jeux peuvent offrir une immersion perceptive, en submergeant les informations sensorielles de l'environnement réel par de grands écrans et des sons puissants pour concentrer entièrement l'utilisateur sur le stimulus.

Tous ces auteurs nous aident à comprendre la définition de l'immersion applicable à notre application cinématographique. L'extinction des lumières dans une salle de cinéma, associée à une bonne isolation sonore, permet d'inhiber les stimuli du monde réel. Un écran plus large, un système audio multicanal comme le Dolby Atmos, des écrans latéraux pour fournir un contexte visuel périphérique sont des solutions pour entourer les sens du public de manière optimale. Les sièges à mouvement haptique, la 3D stéréoscopique, les flux d'air ou les projections d'eau comme dans le 4DX sont des exemples de stimuli diégétiques multimodaux.

1.4.2 L'immersion appliquée à ICE : facteurs disruptifs et actions permises

Lorsqu'il s'agit de « briser l'immersion », cela fait référence à l'immersion psychologique. Cet état psychologique se produit lorsqu'un utilisateur est engagé, impliqué, absorbé ou captivé [26]. L'immersion psychologique est donc liée au récit et se situe au niveau diégétique. Bien que l'accent ne soit pas mis ici sur le récit des films, il est toujours possible

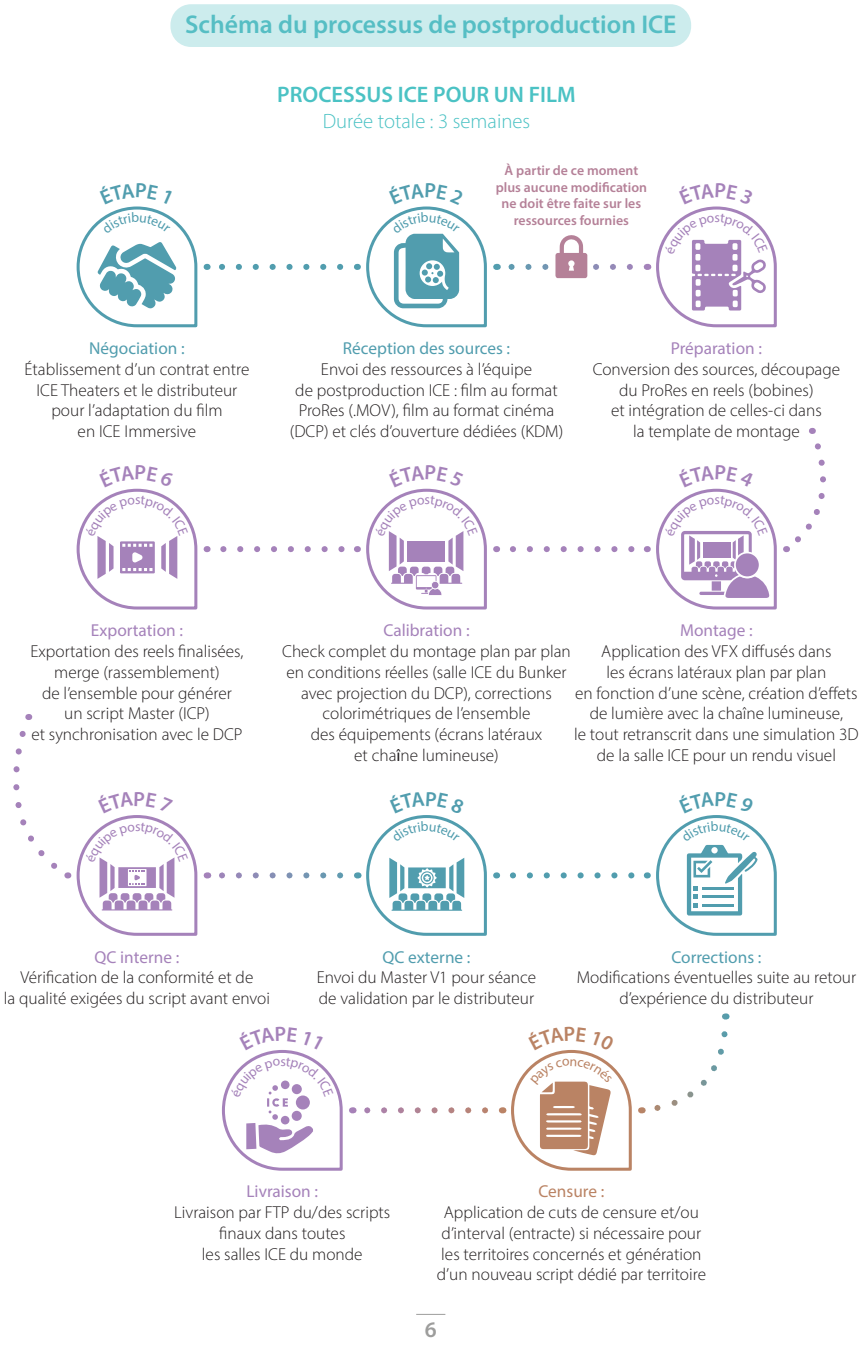


FIGURE 1.11 – **Processus de postproduction ICE pour un film**, montrant les différentes rôles des distributeurs des films et de l'équipe de postproduction ICE. (document interne).

d'améliorer la qualité de l'immersion psychologique des utilisateurs par des moyens techniques. Comme le souligne Thon [31], l'immersion perceptive (mentionnée dans la section précédente), par une couverture efficace des stimuli du monde réel, peut empêcher l'attention de l'utilisateur d'être détournée de l'expérience virtuelle (c'est-à-dire, des décalages

exogènes de l'attention), pour maintenir ainsi cette immersion psychologique. L'immersion dépend donc en partie des entrées sensorielles (l'autre partie étant la narration). Dans notre contexte, cela concerne principalement la vision et l'étendue de sa couverture spatiale. Le champ visuel humain n'est pas uniforme et peut être divisé en plusieurs zones distinctes selon leur fonction perceptive. Ces zones sont la zone centrale, paracentrale, maculaire, puis les zones périphériques proche, moyenne et lointaine, comme illustré en figure 1.12.

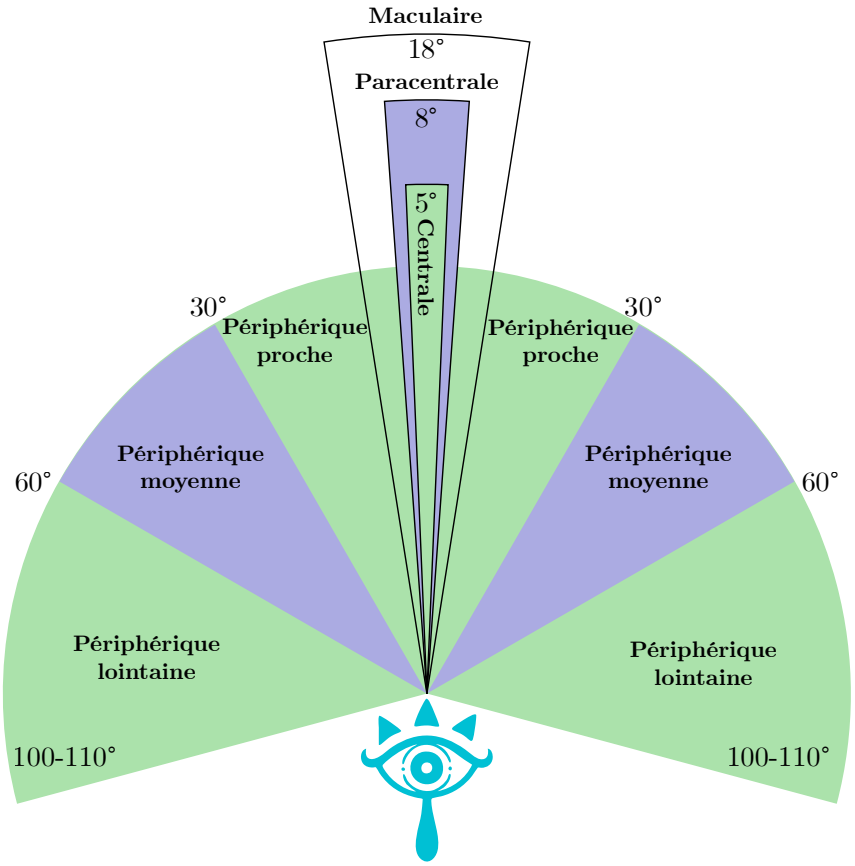


FIGURE 1.12 – **Division du champ de vision humain**[32], de la zone centrale, à la zone périphérique lointaine.

L'acuité visuelle, qu'il s'agisse de la reconnaissance de texte [2] comme sur la figure 1.13 ou de la perception des couleurs [33], diminue progressivement avec l'excentricité par rapport à la fovéa.

Cette variation s'explique par la distribution inégale des photorécepteurs sur la rétine, qui est représentée en figure 1.14. La fovéa est riche en cônes, spécialisés dans la vision fine et la perception des couleurs, tandis que la périphérie est dominée par les bâtonnets, plus sensibles au mouvement et aux faibles niveaux de luminosité.

Ces limitations sont compensées par des capacités améliorées de détection de mouvement [34], de différenciation de la vitesse de mouvement [34], et d'apparition d'un nouveau stimulus. Le scintillement et sa perception par les humains ont été étudiés pour définir la Fréquence Critique de Scintillement (FCS), au-delà de laquelle un stimulus visuel scintillant est perçu comme continu en raison de la persistance rétinienne [35]. Les sources indiquent que la FCS en vision périphérique est d'environ 90 Hz [36], tandis qu'en vision fovéale, elle peut être perçue jusqu'à environ 76 Hz [37]. Cela suggère que la perception du scintillement

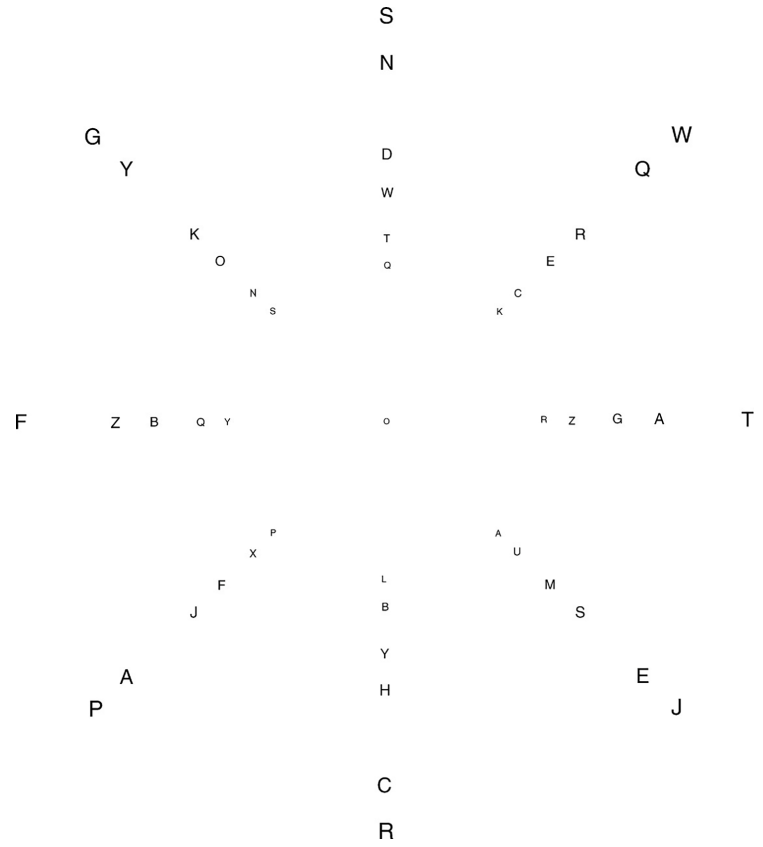


FIGURE 1.13 – **Démonstration de l'acuité périphérique des lettres par l'œil humain** [2]. Les tailles de lettres sont ajustées pour correspondre au seuil de taille lors d'une fixation centrale. Chaque lettre, bien que placée à une excentricité croissante, reste lisible car son agrandissement compense la diminution de résolution visuelle périphérique.

en vision périphérique est au moins aussi efficace, sinon supérieure, à celle en vision fovéale. À l'inverse, laisser un espace entre l'écran principal et les écrans périphériques maintient l'immersion, tant que la cohérence de continuité est préservée (c'est-à-dire, la loi de clôture de la théorie Gestaltique [38] qui est illustrée en figure 1.15).

D'après les résultats présentés dans cette section, il est essentiel d'éviter le scintillement sur les écrans périphériques ainsi que les différences de vitesse de mouvement entre les panneaux et l'écran principal. Néanmoins, il est possible d'avoir un contenu périphérique flou et des espaces entre les écrans sans perturber l'immersion.

1.5 État de l'art par rapport à ICE

1.5.1 Les systèmes « Focus+Contexte »

Le concept de Focus+Contexte a été introduit pour la première fois dans [39]. Ce terme désigne des dispositifs d'affichage combinant un écran central haute définition (le focus) avec un affichage secondaire périphérique en plus basse définition (le contexte). Cette configuration s'inspire du système visuel humain, qui combine vision fovéale précise et vision périphérique apportant des informations complémentaires sans nécessiter une haute définition.

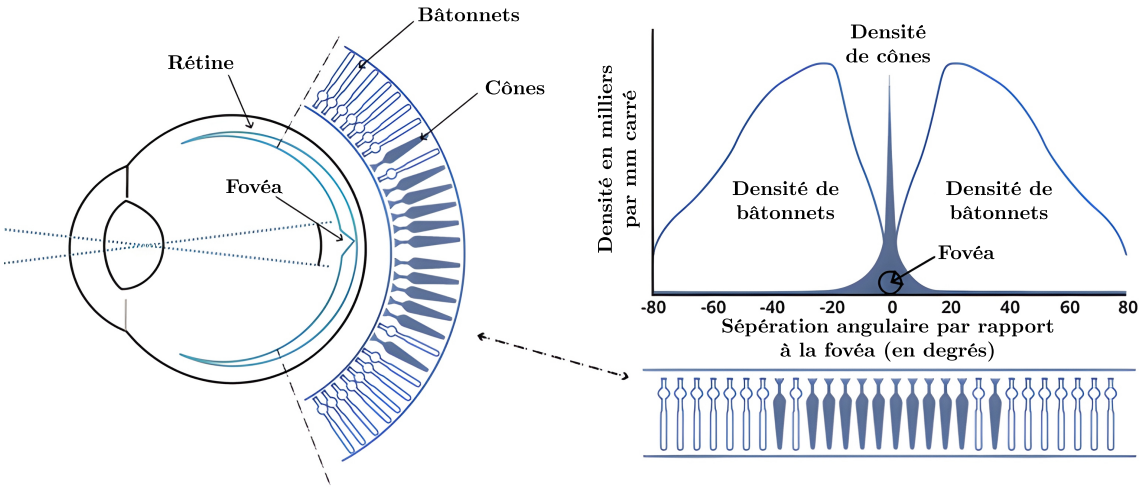


FIGURE 1.14 – **Schéma de la répartition des photorécepteurs dans la rétine humaine.** La fovéa, au centre du champ visuel, présente une forte densité de cônes responsables de la vision fine et colorée, tandis que la densité des bâtonnets augmente en périphérie, assurant la sensibilité à la lumière et la vision nocturne. Le graphique montre la variation de la densité des cônes et des bâtonnets en fonction de l'excentricité angulaire par rapport à la fovéa.(Crédit : [M7])

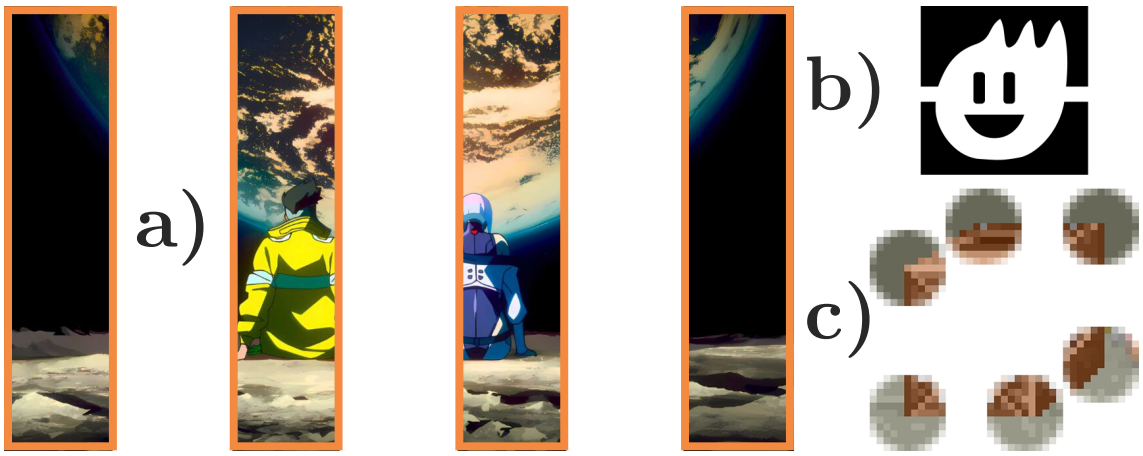


FIGURE 1.15 – **Illustration de la loi de clôture de la théorie Gestaltique par l'exemple.** Les formes pourtant incomplètes (a. la Terre circulaire, le sol rectiligne, b. le visage souriant pourtant ouvert, c. la caisse en bois) sont interprétées et complétées par le cerveau pour former un tout cohérent. (Crédit : auteur [M8])

Ce dispositif est fréquemment mis en œuvre par le biais d'un écran d'ordinateur ou de télévision couplé à un vidéoprojecteur qui prolonge la zone de focus. Philips Ambilight [40] est l'exemple le plus connu : des LEDs RVB disposées derrière l'écran créent une ambiance colorimétrique cohérente autour de l'image. Une étude utilisateur menée par Philips a montré un engagement accru lors du visionnage sur leur téléviseur Ambilight. Ambilux [41] (que l'on peut voir sur la figure 1.16), aujourd'hui abandonné, proposait une version améliorée du système en remplaçant les LEDs par neuf pico-projecteurs projetant une image basse résolution derrière l'écran.

ExtVision [42] combine apprentissage profond et extrapolation spatiale vidéo (communément appelé *video outpainting*) pour former un dispositif Focus+Contexte. L'*outpainting*



FIGURE 1.16 – **Réalisations commerciales ayant des points communs avec ICE, à savoir un dispositif d'affichage à but immersif mettant à profit la vision périphérique.** a) Philips Ambilight b) Philips Ambilux c) ScreenX d) Barco Escape e) Razer Ariana Project.

est un domaine de recherche visant à étendre une image au delà de ses limites. Ce domaine trouve ses origines dans l'*inpainting*, c'est-à-dire l'interpolation spatiale d'image visant à combler des trous dans l'image, ou encore à gommer un élément en se basant sur ce qu'il se trouve autour. Contrairement à Ambilight et Ambilux, ExtVision ne fonctionne pas en temps réel et partage les limitations des méthodes d'*outpainting*, c'est-à-dire confabulations [2] de contenus.

D'autres travaux ont tenté d'intégrer plus finement l'environnement physique du spectateur, de manière à projeter le contenu contextuel sur des surfaces non idéales tout en appliquant une contre-transformation afin d'éviter de fortes distorsions. C'est le cas de Razer Ariana [43], Microsoft IllumiRoom [44], et BBC SurroundVideo [45], qui ciblent les jeux vidéo et les contenus immersifs. Microsoft RoomAlive [46] introduit une interaction physique entre l'utilisateur et le contexte projeté, avec un fort accent sur le jeu vidéo.

La limitation principale de ces approches Focus+Contexte est leur confinement à un même plan visuel. Dans cette section, nous nous inspirons de ces travaux pour concevoir un dispositif plus immersif, en tirant pleinement parti du potentiel perceptif humain, tel que présenté en Section 1.4. Plus précisément, en nous appuyant sur les bases de l'immersion exposées en Section 1.4, nous jugeons essentiel d'exploiter pleinement le potentiel du champ visuel humain. Pour cela, il devient nécessaire, à un certain stade, de s'affranchir de la contrainte d'un affichage confiné à un seul plan.

1.5.2 Les systèmes « CAVE » et casques VR

Les dispositifs immersifs tels que les systèmes de type CAVE [47] (visibles sur la figure1.17) et les casques de réalité virtuelle, augmentée ou mixte (HMD pour *Head-Mounted Display*, traduit par visiocasque ou casque-écran) constituent des environnements tridimensionnels entourant le champ visuel de l'utilisateur. Les premiers, par leur configuration physique enveloppante, se rapprochent le plus de notre environnement cinématographique, tandis que les seconds représentent aujourd'hui la forme la plus répandue de dispositif immersif, soutenue par leur essor commercial et leur accessibilité. Ils sont particulièrement adaptés à la couverture du champ visuel humain, car la portion du champ visuel remplie par

2. Plus communément appelé hallucination de contenu par abus de langage. L'hallucination est une perception sensorielle erronée : cela signifierait qu'une IA de LLM telle que ChatGPT verrait votre *prompt* de manière déformée, ce qui n'est pas le phénomène observé. La confabulation est le comblage d'un vide par des fictions inconscientes, ce qui est neurologiquement plus correct.

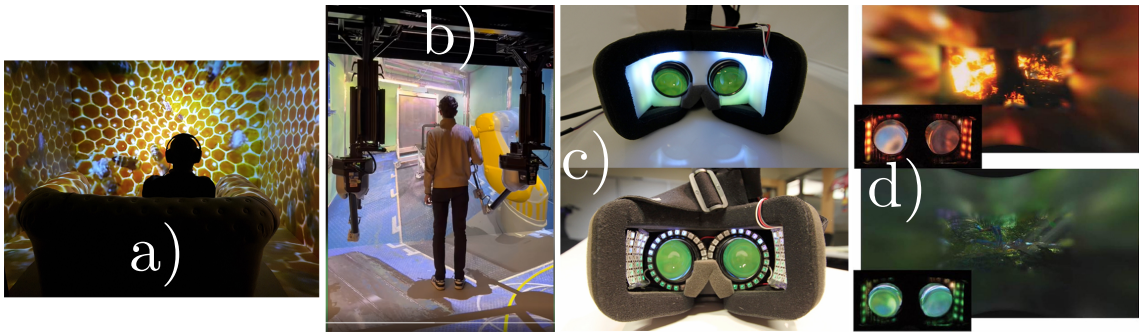


FIGURE 1.17 – **Réalisations scientifiques ayant des points communs avec ICE, à savoir un dispositif d'affichage à but immersif mettant à profit la vision périphérique.** a) Système CAVE sur un documentaire b) Système CAVE de simulateur industriel du CEA c) SparseLightVR d) DeepDive.

l'image, via un écran rapproché de la rétine, requiert moins de surface matérielle qu'un écran conventionnel. Par exemple, un écran de 23 pouces placé à 50 cm offre un champ horizontal de 54°, tandis que les casques VR standards atteignent environ 110°. Des casques professionnels permettent même d'atteindre des FoV (champ de vision, ou *Field of View*) plus larges (150° à 210°), au prix de la complexité de fabrication des lentilles de Fresnel [48].

Certaines modifications matérielles des HMD ont été proposées dans la littérature afin d'en renforcer l'immersion. Ambiculus [49], SparseLightVR v2 [50] et Deep Dive [51] intègrent des LEDs à proximité des lentilles pour prolonger le contenu dans la vision périphérique. Ces trois projets poursuivent un objectif similaire à celui des salles ICE : renforcer l'immersion en maximisant l'occupation du champ visuel humain. Ils s'appuient pour cela sur une stratégie de type Focus+Context, combinant une zone centrale riche en détails avec des extensions visuelles périphériques, moins détaillées, projetées dans la vision périphérique proche et éloignée.

1.6 L'apprentissage profond et sa place au sein de l'IA

1.6.1 Définitions

Bien que certains chercheurs ne trouvent aucun intérêt à redéfinir l'intelligence artificielle dans une thèse dont c'est le sujet principal, il est du devoir éthique et intègre du scientifique de redéfinir un concept lorsque ce dernier est à ce point répandu et employé à tort dans les médias populaires. Les définitions associées à l'intelligence artificielle présentent une complexité souvent sous-estimée, y compris par certains chercheurs qui se disent du domaine, et les frontières entre ses sous-domaines restent parfois floues. À des fins pédagogiques, nous commencerons par une première approximation, que nous nuancerons par la suite.

Le schéma le plus communément admis représente l'intelligence artificielle (ou IA) comme un ensemble englobant l'apprentissage automatique, lui-même contenant l'apprentissage profond comme visible dans la figure 1.18, avec les définitions suivantes. **L'IA** étant définie comme la capacité des machines à effectuer des tâches typiquement associées à l'intelligence humaine. **L'apprentissage automatique** est défini comme toute IA capable de s'améliorer grâce à un entrainement. Cet entrainement peut être plus ou moins encadré et guidé par l'humain. Du fait de cet entrainement, l'IA améliore ses performances.

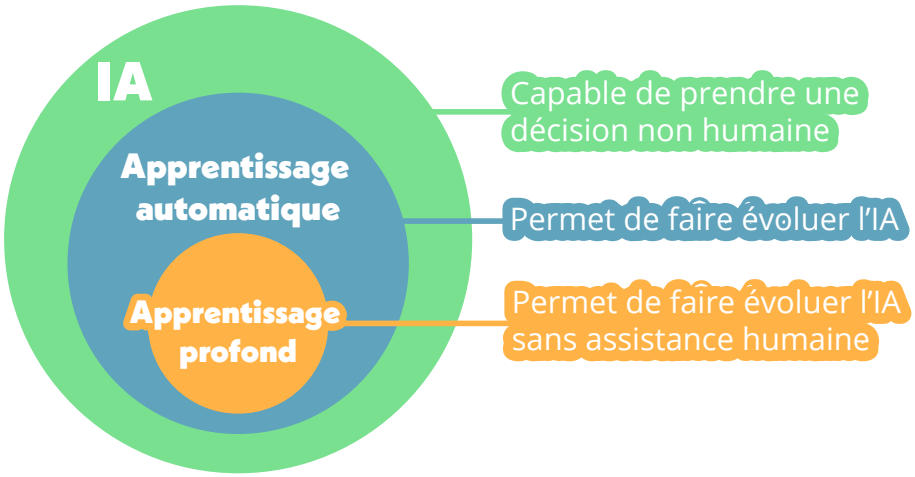


FIGURE 1.18 – **Schéma de la catégorisation de l'IA la plus commune, jugée incomplète dans ce manuscrit, mais permettant une approche simplifiée pédagogique.** L'intelligence artificielle englobe l'apprentissage automatique qui englobe l'apprentissage profond. Une catégorisation plus adéquate est proposée plus loin. (Crédit : auteur)

L'apprentissage profond est défini comme tout apprentissage automatique à base de « réseaux neuronaux artificiels ». Ces réseaux neuronaux, encore aujourd'hui, sont tous des améliorations itératives basées sur le neurone artificiel aussi appelé perceptron (1943), puis du MLP (Multilayered perceptron, 1958).

Chacune de ces catégories (IA, apprentissage automatique, apprentissage profond), permet une abstraction des données et des consignes à suivre pour le modèle d'IA en question. L'IA simple permet d'en garder en contrôle total, l'apprentissage profond quand à lui peut résoudre un problème que même l'humain ne saurait formaliser. L'apprentissage automatique comme l'apprentissage par renforcement et le *Q-Learning*, permet de garder un contrôle, de donner des règles, mais d'avoir une IA qui s'entraine tout de même selon ces règles, comme on peut le voir sur la figure 1.19.

Ceci est un modèle de définition très pratique et assez élastique pour être incontestable, et cette définition de l'IA est celle donnée par Le Monde, le Larousse, le Petit Robert et l'ISO. En revanche, dû à sa flexibilité, définir l'apparition de l'IA ou l'inventeur de l'IA s'avère mission impossible.

1.6.2 Le paradoxe de « l'effet IA » et redéfinitions

La question de savoir si une tâche est « typiquement associée à l'intelligence humaine » soulève des interrogations complexes, notamment en ce qui concerne la communication. Il est essentiel de noter que de nombreuses espèces animales possèdent des capacités de communication avancées, souvent comparables à celles des humains. Par exemple, le perroquet gris d'Afrique « Alex » a démontré une compréhension contextuelle des mots, répondant correctement à des questions portant sur la couleur, la forme, la taille et le nombre d'objets avec une précision supérieure à 80 % [52]. Lors d'une étude, des chimpanzés ont appris à utiliser des signes adaptés de l'American Sign Language, démontrant une capacité à communiquer de manière symbolique et à enseigner ces signes à d'autres individus [53]. De plus, les abeilles mellifères possèdent un système de communication sophistiqué, le *« waggle*

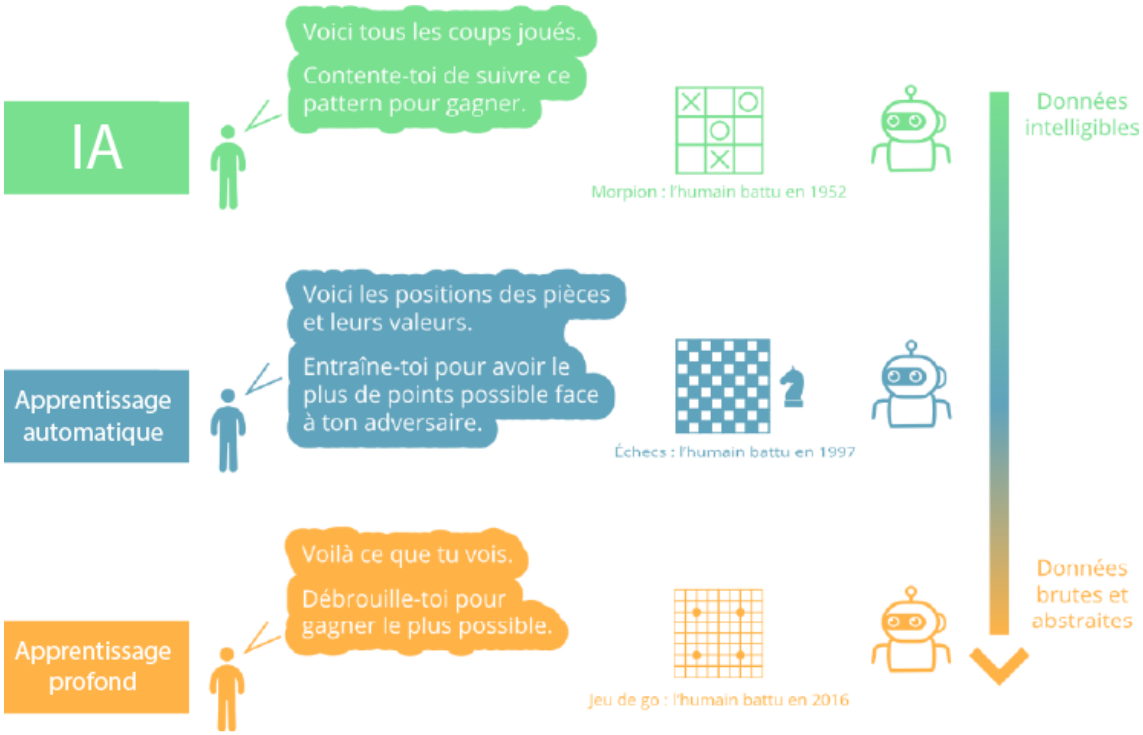


FIGURE 1.19 – **L'abstraction des données va de paire avec la spécialisation de l'IA vers l'apprentissage automatique puis l'apprentissage profond.** (Crédit : auteur [M9])

dance », qui permet à une abeille butineuse d'informer ses congénères de la direction et de la distance d'une source de nourriture [54].

Ces exemples illustrent que la communication n'est pas exclusivement humaine et que de nombreuses espèces animales possèdent des formes de communication complexes, remettant en question la notion de tâches « typiquement associées à l'intelligence humaine ». Il en va de même pour d'autres domaines que la communication. Un robot capable de se déplacer en prenant en compte son environnement est usuellement associé à de l'intelligence artificielle. Or, là encore, quasiment la totalité des animaux en sont capables.

Si ces domaines d'application de l'intelligence (la communication et le déplacement) sont problématiques pour la définition, qu'en est-il alors des perroquets qui peuvent établir une stratégie pour ouvrir une suite de verrous interdépendants ? À partir de quel critère la cognition, ou réflexion, est-elle réservée à l'humain ?

Puisque cette approche semble être une impasse, revenons alors à la définition. Doit-on plutôt comprendre « intelligence humaine », dans un sens plus large comme intelligence poussée, qu'elle soit humaine ou animale ? Dans ce cas, la définition d'intelligence artificielle se définit elle-même sur ce qui est vivant ou artificiel, ce qui en fait une définition récursive, autoréférentielle. Ceci a une conséquence intéressante : tant que la tâche « trouver le meilleur coup aux échecs » n'avait jamais été résolue sans humain, elle relevait de l'intelligence humaine. Depuis qu'elle a été résolue par machine, elle n'est plus réservée à l'intelligence humaine, ce qui fait que les « tâches typiquement associées à l'intelligence humaines » évoluent en fonction du temps. Ceci est appelé « L'effet IA ». À partir du

moment où quelque chose qui était qualifié à l'origine d'intelligence devient commun, celui-ci est relégué au rang d'algorithme. On peut citer de nombreux auteurs :

- John McCarthy, créateur en 1956 du terme « intelligence artificielle » : « dès que ça fonctionne, plus personne n'appelle ça de l'IA. » (2011, [56])
- Rodney Brooks, ancien directeur du MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) : « chaque fois que nous en découvrons un morceau, ça cesse d'être magique ; on dit : « Oh, c'est juste un calcul. » »(2002, [57])
- Larry Tesler, chercheur au Xerox PARC puis chez Apple, a formulé vers 1970 ce que l'on appelle depuis l'*AI effect*, puis a complété ses propos bien plus tard : « « L'intelligence, c'est ce que les machines n'ont pas encore accompli ». De nombreuses personnes définissent l'humanité en partie par notre intelligence prétendument unique. Tout ce qu'une machine — ou un animal — peut faire doit (disent ces gens-là) être autre chose que de l'intelligence. » [58]
- Nick Bostrom, philosophe à l'Université d'Oxford et fondateur de l'Institut pour l'Avenir de l'Humanité (FHI), observe que « les chercheurs en IA se plaignent parfois que dès qu'une technique fonctionne réellement, elle cesse d'être qualifiée d'« IA ». Certaines méthodes aujourd'hui courantes dans les logiciels et la robotique constituaient autrefois des avancées pionnières passionnantes de la recherche en intelligence artificielle. »[59]

De la même manière que l'espèce humaine s'est sentie spéciale et inimitable par une machine, elle s'est également sentie comme la seule forme de vie intelligente, spéciale par rapport aux autres animaux. Pourtant, il n'existe pas une frontière aussi franche que l'on voulait le croire. Comme on l'a vu plus haut dans cette section, la communication avec d'autres individus ainsi que l'analyse de leurs environnements sont courantes dans les espèces animales.

La mémoire et l'apprentissage des animaux ne sont pas en reste. Même les animaux raillés comme les poissons rouges peuvent être entrainés à suivre un labyrinthe avec des marqueurs visuels pour obtenir de la nourriture même après plus d'un mois (émission scientifique MythBusters, saison 2024, épisode 4), ou à se souvenir d'un son, ce qu'ils retiendront ensuite jusqu'à 55 jours [60].

C'est dans cette même logique que l'humanité l'a montré avec le géocentrisme et la condamnation de Galilée, l'humain a trop d'ego et se trouve supérieur au reste de la nature qui l'entoure : Descartes affirme par exemple que seuls les humains ont une âme pensante et que les animaux sont des machines sans pensée rationnelle. Pour lui, cette séparation entre l'humain et l'animal est la capacité de raisonner, de réfléchir et le langage [61], ce qui semble aujourd'hui absurde d'après les études montrées plus haut sur les capacités robotiques et les aptitudes animales. Kant adopte des idées similaires, que les animaux ne sont pas des agents rationnels et qu'ils n'ont qu'une valeur instrumentale [62].

En poussant la réflexion plus loin, nous pourrions alors proposer le schéma suivant visible en figure 1.20 : L'outil englobe l'automatisme, qui englobe lui-même l'IA. Enfin, l'IA englobe l'apprentissage automatique, qui englobe l'apprentissage profond. Tous ces concepts ont des barrières floues entre eux.

L'outil est un objet fabriqué, utilisé manuellement ou non pour réaliser une opération déterminée. L'automatisme réalise une suite d'actions et ce, sans intervention de l'humain entre chaque action, (ou presque, l'automatisation peut être partielle). Un automate est l'acteur matériel qui va réaliser ces opérations prédéterminées. Que l'automate soit artistique ou

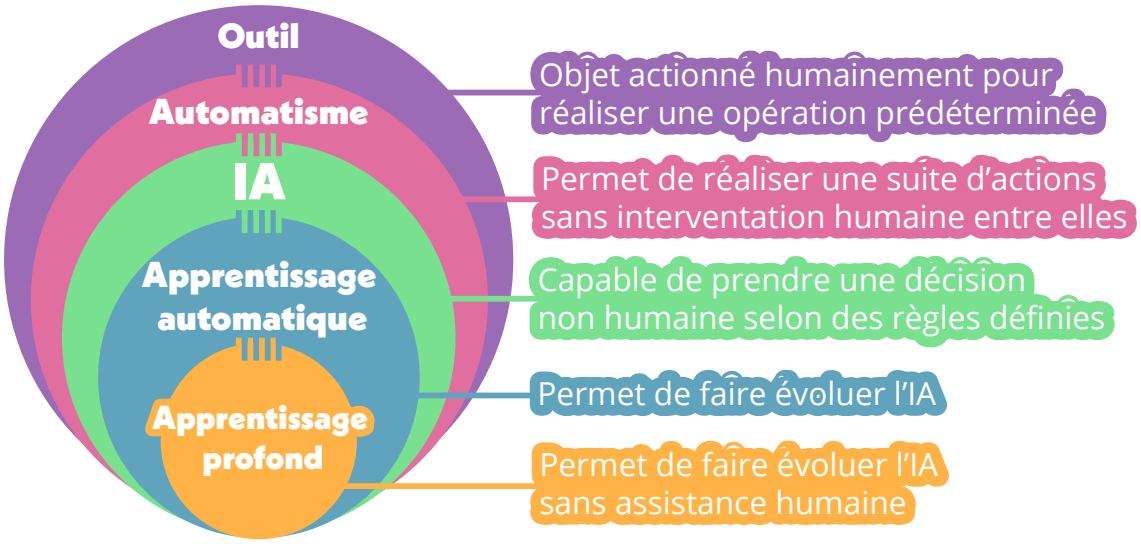


FIGURE 1.20 – **Catégorisations de l'IA, corrigée en accord avec la proposition de définition de ce manuscrit.** Comme des poupées russes, l'outil contient l'automatisme, qui contient l'IA, l'apprentissage automatique, puis l'apprentissage profond. Les frontières entre chaque catégorie sont floues, et dépendent parfois du point de vue du locuteur. (Crédit : auteur)

utile, il est un outil plus spécialisé qui a été fabriqué par l'homme avec un but à accomplir, qui n'a plus besoin d'être activé manuellement une fois actionné.

Si la tâche réalisée par l'automate semble assez complexe aux yeux de l'humain, l'automate peut être appelé une intelligence artificielle. Dans ce manuscrit, là où l'automate est considéré comme la version physique de l'automatisme ou de l'automatisation, le robot est également la version physique de l'IA, en ce sens qu'ils ont un corps dans le monde réel en plus de leur module d'action ou de décision. Les outils quant à eux, peuvent être virtuels comme matériels. Les logiciels, comme ceux de traitement de texte par exemple, sont des outils virtuels.

1.6.3 L'IA, de la fiction à aujourd'hui

Ces automates, androïdes ou autres robots ont toujours fasciné, et ont donné lieu à une multitude d'artéfacts, que ca soit dans la fiction ou la création physique et ce, depuis l'antiquité. **Dans la mythologie grecque** [63] (transmise oralement depuis environ −1800, et écrite à partir de −850 par différents auteurs comme Homère, avec l'Iliade), plusieurs exemples d'automates sont décrits, faisant d'eux les plus anciens automates de fiction connus : Héphaïstos aurait construit des automates pour son atelier. Talos, un homme artificiel en bronze forgé par Hephaistos, les chiens d'or et d'argent du roi Alkinoos. Également, selon Aristote, la statue en bois d'Aphrodite par Dédale, animée grâce à du mercure.

Homère lui-même décrit dans ses textes l'ouverture automatique de portes, introduisant le terme *automaton*[63] (concept qui nous parait aujourd'hui très éloigné de l'intelligence artificielle). **Le pigeon volant d'Archytas** (−400) est l'un des premiers exemples mécaniques de reproduction de comportements biologiques. **La machine d'Anticythère** (avant -87) est un calculateur analogique des positions astronomiques ; elle permet de prédire les éclipses en plaçant des aiguilles et en tournant une manivelle, interaction assimilable à une entrée utilisateur. **Le joueur de flûte** (1737) puis **le canard de Vaucanson**

(1738) sont des automates capables de reproduire des actions humaines ou animales de manière mécanique. **La machine à marées** (1872–1873), développée par Lord Kelvin, Edward Roberts et Alexander Légé, applique l'analyse harmonique formalisée par Kelvin pour prédire les niveaux de marées. **La Bombe de Turing** (1939) est une amélioration anglaise d'une machine polonaise destinée à décrypter Enigma, posant les bases d'un calcul algorithmique automatisé. **OXO** (1952), de A.S. Douglas, est un jeu de morpion contre un ordinateur, avec une IA effectuant des parties parfaites, parfois considéré comme le premier jeu vidéo. **Le programme de dames d'Arthur Samuel** (1959) constitue un jalon majeur de l'apprentissage automatique : Samuel était convaincu « qu'apprendre aux ordinateurs de jouer pouvait être très fructueux pour développer des stratégies appropriées pour des problèmes généralisés ». Cette intuition s'est vérifiée dans la suite des recherches, notamment avec les IA consacrées aux échecs (Deep Blue, IBM, 1997), à l'adaptation généralisée sur les jeux d'Atari 2600 (Google DeepMind, 2013), au jeu de go (AlphaGo, Google DeepMind, 2014), à Starcraft 2 (AlphaStar, Google DeepMind, 2019), à Dota 2 (OpenAI Five, 2018) etc. Toutes ces tâches et ces jeux sont plus complexes les uns que les autres, avec un nombre de parties possibles qui semble infinies et une stratégie à adopter impossible à formaliser en amont. **MENACE** (1961), une machine analogique conçue par Donald Michie, apprend progressivement à gagner au morpion. La particularité de cette machine est que, n'ayant pas le materiel mécanique, elle est faite de boites d'allumettes et de billes, l'ingéniosité résidant dans l'algorithme conçu et dans le fait de pouvoir se détacher du support physique, démontrant ainsi les premiers mécanismes d'apprentissage. **Shakey the robot** (1966) est le premier robot mobile intégrant plusieurs modules : il peut comprendre des instructions en langage naturel, telles que « va dans l'autre pièce et pousse le cube », tout en analysant visuellement son environnement et planifiant ses déplacements pour éviter les obstacles. **ELIZA** (1966) simule une conversation avec un psychothérapeute via une interface textuelle ; elle a donné lieu à l'*effet ELIZA*, où l'humain projette des caractéristiques humaines (compréhension, empathie) sur une machine à logique simple. **ALVINN** (1988) est l'un des premiers systèmes de conduite autonome sur base d'image, utilisant un réseau de neurones pour contrôler un véhicule ; cependant, sa généralisation était limitée par la puissance de calcul, la taille des jeux de données et la faible robustesse aux environnements variés. **Deep Blue** (1997) bat le champion du monde Garry Kasparov aux échecs, relançant les débats sur l'intelligence artificielle, la cognition humaine et les limites entre calcul et stratégie.

Les critères d'évaluation de l'intelligence artificielle, tels que ceux inspirés du test de Turing, évoluent au fil du temps. Par exemple, l'adaptation récente du test pour identifier ce qui « semble être produit par ChatGPT », que ce soit du code informatique ou du texte, reflète une redéfinition des critères d'humanisation visant notamment à détecter la triche. Un texte généré par un modèle d'aujourd'hui, aurait systématiquement été identifié comme écrit par un humain il y a quinze ans. Pourtant, plusieurs travaux montrent que les grands modèles de langage ne passent pas ce test [65][64], illustrant l'évolution dans le temps du test de Turing et des critères d'évaluation de l'intelligence artificielle.

1.6.4 La place de l'IA dans la société : Hivers de l'IA et luttes sociales

L'IA vue par les financements et le capital

L'utilité de l'intelligence artificielle moderne et des réseaux de neurones qui la composent est aujourd'hui indéniable, pourtant leur émergence n'a pas été sans heurts. Dès l'invention du perceptron, théorisé en 1943 par McCulloch et Pitts puis mis en application en 1957

par Rosenblatt, les espoirs sont immenses. Le New York Times annonce alors que le perceptron est « l'embryon de l'ordinateur électronique que la US Navy s'attend à voir un jour marcher, parler, voir, écrire, se reproduire et être conscient de son existence ». On mesure aujourd'hui à quel point cette déclaration était absurdement prématurée. Alors qu'elle engendre une euphorie colossale à court terme, faisant affluer les financements et une attention médiatique, elle contribue également à la désillusion, qui sera à la hauteur des espoirs placés. En effet, ces architectures et toute l'IA par extension, tombent en disgrâce après la mise en évidence des limites fondamentales du perceptron individuel par Minsky et Papert [67] en 1969, provoquant un abandon massif de cette approche. Cette période sera nommée l'hiver de l'intelligence artificielle, qui désigne des périodes prolongées (années 1970–1990) où l'engouement initial pour les systèmes intelligents retombe, faute de résultats concrets, de puissance de calcul suffisante, et d'approches théoriques robustes.

Dans les années 1990-2000, des chercheurs comme Yann Le Cun, Geoffrey Hinton ou Yoshua Bengio continuent de travailler sur les réseaux de neurones malgré un rejet académique quasi-total. Le terme « réseau de neurones » est alors mal vu dans les comités de financement, les revues, et les conférences : utiliser ce mot-clé suffisait souvent à faire rejeter un article ou une proposition. Le Cun lui-même raconte que les financements étaient presque impossibles à obtenir, et qu'il a dû détourner les intitulés de ses projets pour éviter de mentionner explicitement les réseaux neuronaux.

Les choses changent brutalement avec AlexNet (2012), lorsque Hinton et ses étudiants remportent le challenge ImageNet, un challenge de reconnaissance et de classification d'images avec la plus grande base de données à cette époque, avec une large avance, en utilisant un CNN entrainé sur GPU. Ce succès déclenche une renaissance de l'apprentissage profond, et avec elle la sortie fulgurante de l'hiver de l'IA.

Depuis cette époque, ces pionniers sont des acteurs importants des sociétés de recherche en IA les plus actives et les plus performantes d'aujourd'hui, à savoir Google DeepMind, Facebook FAIR et OpenAI.

En plus de cet engouement instable concernant principalement le doute sur l'efficacité de l'IA, menant à des coupes budgetaires d'investissement, l'exact opposé, la crainte sur l'efficacité de l'IA, ou plutôt de l'automatisme en général existe aussi, de manière tout aussi présente. Les CNN ont donc joué un rôle central dans ce basculement : ils ont montré que l'apprentissage profond, longtemps marginalisé, pouvait surpasser toutes les méthodes classiques et ont ouvert la voie à la révolution de l'IA moderne.

L'IA vue par le peuple

Il est prévisible et intemporel qu'une entreprise cherche à simplifier, standardiser, mécaniser et automatiser les tâches qui lui incombent : c'est sa nature même, orientée vers la rentabilité. L'ubérisation illustre d'ailleurs une forme particulière d'automatisation partielle d'un métier : la mise en relation automatique entre un client et un travailleur, la gestion d'une plateforme mutualisée à l'échelle mondiale, et l'utilisation d'outils numériques pour la paie, le GPS ou la notation mutuelle entre clients et travailleurs en sont des indices concrets. La règle des 80/20 est particulièrement pertinente ici : 20 % des tâches d'un corps de métier concentrent 80 % du temps de production, ce qui signifie qu'automatiser seulement ce premier cinquième du procédé peut diviser par cinq le temps de travail. Ces premières tâches sont généralement les plus simples à automatiser ; en revanche, les

derniers pourcents d'un métier, ceux qui requièrent jugement, adaptation et expérience, sont extrêmement difficiles à déléguer à une machine.

Il est donc assez naturel de penser que l'automatisation des tâches métiers favorisée par le capitalisme (basé sur la recherche de rentabilité par les entreprises) est tout aussi ancienne que les métiers eux-mêmes. Cependant, cette automatisation vient avec des tensions entre progrès technique et préservation des tâches humaines, et ce, avec des preuves remontant à l'Antiquité.

En 69 de notre ère, l'empereur romain Vespasien se voit proposer une machine capable de hisser des colonnes de marbre jusqu'au Capitole sans effort humain. Les ouvriers romains transportaient alors ces blocs à la seule force des bras et des cordes, un travail éreintant mais qui faisait vivre des milliers de familles. Vespasien récompense l'inventeur, mais refuse d'utiliser la machine : « il me faut nourrir mon petit peuple », rapporte Suétone. C'est l'une des plus anciennes confrontations connues entre automatisation et emploi.[68]

Au début du XIXe siècle, ce sont les ouvriers eux-mêmes qui passent à l'action. Les *Luddites*, tisserands et bonnetiers anglais, voient arriver les métiers mécaniques qui permettent à un seul opérateur non qualifié de produire autant que dix artisans formés pendant des années. Entre 1811 et 1816, ils s'organisent la nuit sous le nom de « Général Ludd » et détruisent systématiquement les machines dans les usines du Nottinghamshire et du Yorkshire. Le Parlement britannique réagit en votant la peine de mort pour destruction de machines.[69][70]

En 1886, la machine Linotype révolutionne les imprimeries : là où un typographe qualifié assemblait laborieusement les caractères en plomb un à un à la main, la Linotype compose une ligne entière au rythme d'une frappe de clavier, remplaçant l'équivalent de plusieurs ouvriers. Des années de formation devenaient obsolètes en quelques semaines. Aux États-Unis, plutôt que de détruire les machines, les syndicats de typographes négocient âprement et obtiennent que les Linotypes soient confiées en priorité aux anciens compositeurs à la main, préservant ainsi les emplois au prix d'une reconversion. Ce compromis est souvent cité comme l'un des premiers exemples modernes de négociation collective face à l'automatisation.[71]

Dans les années 1960, le conteneur maritime transforme radicalement les ports du monde entier. Avant son invention, décharger un navire était un travail de plusieurs jours : des centaines de débardeurs chargeaient et déchargeaient manuellement chaque caisse, chaque sac, chaque ballot. Le conteneur standardisé permet probablement à une grue et deux hommes de faire le travail d'une centaine. À San Francisco, le syndicat ILWU, l'un des plus puissants des États-Unis, engage une bataille historique : plutôt que d'interdire les conteneurs, il arrache en 1960 le *Mechanization and Modernization Agreement*, un accord garantissant aux dockers en poste leur emploi à vie et des retraites anticipées financées par les gains de productivité. C'est l'un des rares cas dans l'histoire où les travailleurs ont réussi à négocier directement une part des bénéfices de l'automatisation.[72]

Au milieu des années 1970, l'arrivée de la photocomposition et de l'informatique menace le métier des ouvriers typographes français, dont les compétences manuelles deviennent peu à peu inutiles. En France, ce métier est verrouillé par un syndicat exceptionnellement puissant, le Syndicat du Livre CGT, qui détient depuis 1944 un monopole d'embauche dans les imprimeries de presse parisiennes : aucun journal ne peut être imprimé sans sa main-d'œuvre. En février 1975, le patron du *Parisien libéré*, Émilien Amaury, décide d'imprimer ses éditions en dehors de Paris pour échapper à ce monopole et réduire ses coûts. La

riposte du syndicat est d'une violence rare : vingt-neuf mois de grève, de mars 1975 à août 1977, avec blocages, occupations d'imprimeries et destruction de tirages entiers. Le syndicat l'emporte, les ouvriers licenciés sont reclassés, mais le journal sort exsangue du conflit, passant de 750 000 à 350 000 exemplaires vendus. Une victoire syndicale qui n'aura fait que retarder l'inévitable modernisation de la presse.[73][74]

En 2023, c'est Hollywood qui se soulève contre l'intelligence artificielle. Scénaristes et acteurs craignent deux usages : que les studios utilisent l'IA pour écrire des scénarios à la place des auteurs, et qu'ils scannent le visage et la voix des comédiens pour en faire des « répliques numériques » réutilisables à l'infini, sans nouveau cachet. Les figurants étaient particulièrement visés : payés une demi-journée pour être numérisés, puis remplaçables par leur double virtuel. Les deux syndicats, celui des scénaristes puis celui des acteurs, déclenchent des grèves simultanées qui paralysent l'industrie du divertissement pendant une grande partie de l'année, parmi les plus longues de l'histoire d'Hollywood. Les accords finalement signés fin 2023 encadrent strictement l'usage de l'IA : le consentement explicite de l'acteur devient obligatoire pour créer ou utiliser sa réplique numérique, et il est posé qu'il faut être humain pour être reconnu comme auteur ou interprète.[75][76]

Plus récemment, la même inquiétude gagne la France avec le doublage. Lorsque la voix du comédien Alain Dorval, voix française de Sylvester Stallone, est recréée par IA peu après sa mort pour faire parler l'acteur dans un film, la profession prend peur. Des outils capables de cloner une voix et de l'adapter à n'importe quelle langue menacent directement les comédiens de doublage, un savoir-faire dont la France est particulièrement fière. En janvier 2024, l'association Les Voix et le syndicat SFA-CGT lancent la pétition « Touche pas à ma VF », qui dépassera les 200 000 signatures, soutenue par des voix connues comme celle de Brigitte Lecordier (Sangoku dans *Dragon Ball*) et de Pierre-Alain de Garrigues (Heimerdinger dans *League of Legends* et la voix d'origine de L'aubergiste dans *Heartstone*). Le mouvement organise des manifestations à Paris en 2024 et 2025, et alerte sur les 12 500 emplois de la filière potentiellement menacés. L'affaire remonte jusqu'à l'Assemblée nationale, où plusieurs députés interpellent la ministre de la Culture sur la nécessité de réguler l'IA pour protéger la profession.[77][78]

Pour le dernier exemple, revenons en Angleterre dans les années 1950. L'industrie du charbon y est considérable et plus d'une centaine de mines existent un peu partout sur le territoire. Depuis 1974, le programme MINOS est déployé par morceaux, un système informatique pilotant depuis une salle de contrôle en surface l'ensemble de l'activité du fond : convoyeurs, ventilation, et même le guidage automatique des foreuses [79]. Couplé à un équipement lourd, ce système permet d'exploiter un front entier avec deux mineurs, là où il en fallait au moins seize auparavant, transférant progressivement les compétences du mineur vers le logiciel. Le métier de mineur requiert ainsi moins de qualifications théoriques et scientifiques pour être exercé. Dans le même temps, la Colombie, l'Australie et les États-Unis développent des mines de charbon à ciel ouvert, dont le coût d'extraction est une fraction de celui des mines profondes britanniques : le charbon inonde le marché mondial à des prix que l'Angleterre ne peut concurrencer. Les mines manuelles britanniques deviennent ainsi invendables, non pas parce que les mines automatisées produisent beaucoup plus, comme l'avance Winterton [79], mais parce que le charbon mondial est tout simplement devenu moins cher qu'elles ne peuvent produire. En 1981, face à un premier projet de fermeture de vingt-trois puits, la seule menace de grève suffit à faire reculer le gouvernement. En effet, les mineurs avaient déjà provoqué en 1972 et en 1974 des coupures d'électricité jusqu'à neuf heures par jour, par la mise en place de piquets

volants : des mineurs mobiles forçant leur collègues à cesser le travail dans les villes voisines, ce qui contraint le gouvernement à déclarer l'état d'urgence et à rationner l'énergie [80][81]. Thatcher tire la leçon de cet épisode : entre 1981 et 1984, elle constitue méthodiquement un stock de six mois de charbon dans les centrales électriques et met en place des unités de police mobiles coordonnées à l'échelle nationale. En 1984, le gouvernement remet sur la table le projet de fermeture, et le bras de fer entre grêvistes et le gouvernement éclate, donnant lieu à la plus longue grève de masse de l'histoire britannique, qui tourne à l'affrontement physique, où des milliers de policiers chargent les grévistes. Pour ajouter au conflit entre population et l'automatisation imposée, le syndicat mandate un groupe de recherche qui calcule que cette automatisation ferait disparaitre 70k emplois. Ce chiffre est présenté comme alarmiste et complètement faux par le gouvernement. La défaite syndicale a lieu en 1985, et dans les deux années suivantes, plus de 74k emplois disparurent.

Cet épisode soulève une vérité inconfortable : l'ennemi des travailleurs n'est pas l'automatisation ou l'IA en elle-même, mais les commanditaires de celle-ci qui choisissent de la déployer, comment le faire et comment en présenter les conséquences. Surtout, freiner l'automatisation localement n'a de sens que si l'ensemble du marché accessible (aujourd'hui souvent mondial dû au libre-échange) adoptait la même politique. Si ce n'est pas l'Angleterre qui améliore sa productivité, c'est l'Australie ou la Colombie qui le fait à sa place, et c'est elle qui coule. La vraie question n'est donc pas « automatiser ou non ? » puisque l'automatisation et tout le progrès technique est hélas inexorable et arrivera tôt ou tard, par des portes dérobées s'il le faut. Elle serait plutôt « Y a-t-il réellement quelque chose de possible pour protéger et conserver la culture des métiers menacés dans un système capitaliste de rentabilité tel qu'en place aujourd'hui ? ». De plus, ironie du sort, après avoir numérisé une grande partie des domaines l'un après l'autre pendant des années afin de pouvoir simplifier les procédés (l'enregistrement sonore, la photographie, la vidéo, l'écriture, l'imprimerie, la comptabilité, la téléphonie, le réseau informatique, la cartographie, la radiodiffusion et la télévision, l'imagerie médicale, le dessin technique, l'algorithmie, la gestion, etc.), l'apprentissage profond peut entrainer un modèle d'IA sur n'importe quelles données numériques et en prédire d'autres, ce qui fait sa puissance aujourd'hui.

1.6.5 Les rouages de l'apprentissage profond

Sortons maintenant de la vision globale de l'IA et des applications, et penchons-nous plutôt sur le fonctionnement et les mathématiques de l'apprentissage profond dont nous aurons besoin pour comprendre cette thèse.

Le perceptron de Rosenblatt

L'histoire de l'apprentissage profond débute formellement avec l'introduction du *perceptron* par Frank Rosenblatt en 1958 [66]. Ce modèle constitue l'élément de base des réseaux de neurones artificiels et repose sur une analogie du fonctionnement des neurones biologiques, dans une forme algorithmique. Il vise à modéliser une fonction de décision binaire linéairement séparable.

Le perceptron, visible en figure 1.21, prend en entrée un vecteur $\mathbf{x} = [x_1, x_2, \ldots, x_n]$, auquel est associé un vecteur de poids $\mathbf{w} = [w_1, w_2, \ldots, w_n]$ et un biais b. Le neurone calcule une somme pondérée des entrées :

$$z = \sum_{i=1}^{n} w_i x_i + b \tag{1.1}$$

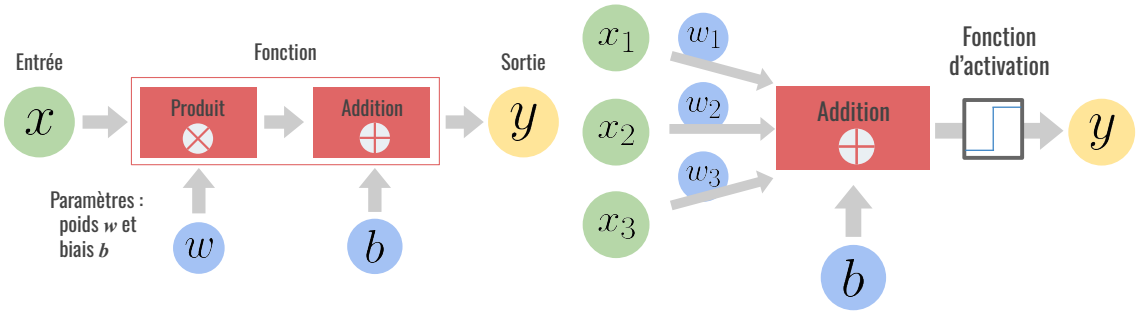


FIGURE 1.21 – **Similarité entre la fonction paramétrique (à gauche) et le perceptron (à droite).** Tous deux calculent une somme pondérée $\sum x_i w_i + b$, avec $i = 0$ à gauche (une seule valeur, soit une entrée dite « scalaire », et un calcul qui devient $x_1 \times w_1 + b$) et i allant de 0 à N à droite (plusieurs valeurs, une entrée dite « vectorielle », soit un calcul $x_1 \times w_1 + x_2 \times w_2 + \cdots + b$). Le perceptron applique également une fonction d'activation.

Cette somme est ensuite passée dans une fonction d'activation, qui décide de l'activation ou non du neurone. Celle de Rosenblatt qui n'est plus utilisée aujourd'hui est la fonction « marche d'escalier » :

$$f(z) = \begin{cases} 1 & \text{si } z \geq 0 \\ 0 & \text{sinon} \end{cases} \tag{1.2}$$

Ainsi, le perceptron réalise une classification linéaire dans l'espace des entrées. Toutefois, il présente une limite fondamentale : il est incapable de modéliser des fonctions *non linéairement séparables*, comme l'opérateur logique OU exclusif [67]. Cette faiblesse a freiné les recherches dans ce domaine jusqu'à l'introduction des modèles multicouches.

Fonctions d'activation modernes

Pour dépasser les limites du perceptron classique, les chercheurs ont introduit des fonctions d'activation *continues et différentiables* [82], indispensables à l'apprentissage par rétropropagation du gradient. Ces fonctions permettent d'introduire de la non-linéarité tout en autorisant le calcul analytique des gradients. Il existe parmi les fonctions les plus utilisées, qui sont illustrées en figure 1.22 :

— **Fonction sigmoïde** :

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{1.3}$$

Elle écrase toute valeur dans l'intervalle $[0; 1]$, mais souffre de saturation et phénomène de disparition du gradient (*vanishing gradient*) pour les valeurs extrêmes.

— **Fonction tangente hyperbolique** :

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{1.4}$$

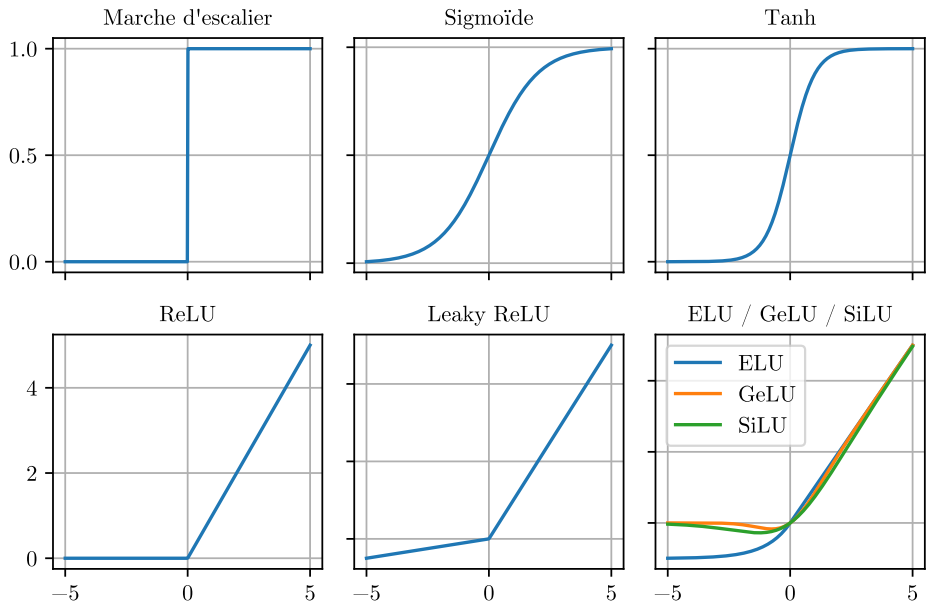


FIGURE 1.22 – **Les différentes fonctions d'activation utilisées depuis leur apparition :** La fonction Marche d'escalier est la première à servir de fonction d'activation. La fonction sigmoïde, tangeante hyperbolique, ReLU [3] et Leaky ReLU [4] sont très communes. Les fonctions ELU [5], GeLU [6], et SiLU [7] sont des propositions d'améliorations des fonctions ReLU et Leaky ReLU.

Elle recentre la sortie autour de zéro dans l'intervalle $[-1; 1]$, ce qui facilite l'apprentissage, mais reste sujette aux mêmes limitations que la sigmoïde.

— **Fonction ReLU** *(Rectified Linear Unit)* :

$$\text{ReLU}(x) = \max(0, x) \tag{1.5}$$

Elle est largement adoptée pour sa simplicité, sa non-linéarité effective, et son absence de saturation pour $x > 0$. Elle peut cependant entrainer l'extinction de certains neurones si ceux-ci reçoivent trop souvent des entrées négatives.

D'autres variantes comme *Leaky ReLU*, *ELU*, *GELU* ou *SiLU* ont été proposées pour surmonter certaines limites spécifiques. Néanmoins, ReLU reste à ce jour la fonction la plus couramment utilisée dans les architectures modernes d'apprentissage profond.

Le perceptron multicouche et la descente de gradient

Le perceptron multicouche ou MLP est une évolution directe du perceptron [83]. Contrairement au perceptron simple, qui ne permet de résoudre que des problèmes linéairement séparables, le MLP connecte plusieurs couches de neurones comme on peut le voir sur la figure 1.23, permettant ainsi la modélisation de fonctions non-linéaires complexes comme la fonction OU exclusif.

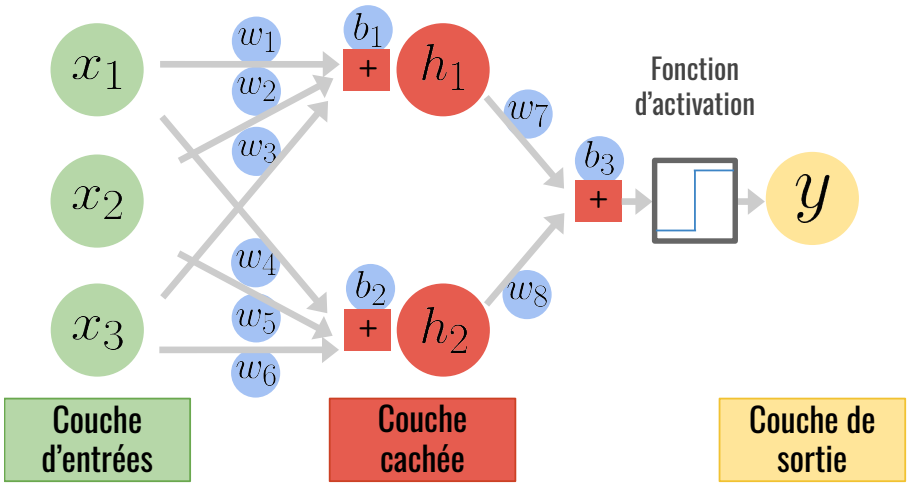


FIGURE 1.23 – **Perceptron multicouche**, ou MLP. Les fonctions d'activation ne sont pas représentées ici.

Au sein d'une architecture MLP, chaque neurone applique une fonction d'activation à une combinaison linéaire de ses entrées :

$$y = \phi\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{1.6}$$

où ϕ est la fonction d'activation, w_i les poids, x_i les entrées, et b le biais.

L'apprentissage du réseau repose sur la **descente de gradient** [84], qui consiste à minimiser une fonction de cout L (pour *loss function*). Cette fonction de cout doit être choisie correctement pour représenter le taux d'erreur d'une inférence (une prédiction) faite par le réseau de neurone, par rapport au label qu'on essaie de faire apprendre au réseau (la vérité terrain). Des fonctions de cout communes sont l'erreur quadratiques ou l'entropie croisée, et la minimisation se calcule comme suit, en ajustant les poids dans la direction opposée au gradient :

$$w \leftarrow w - \eta \frac{\partial L}{\partial w} \tag{1.7}$$

où η est le taux d'apprentissage.

Avant l'invention de la rétropropagation, l'apprentissage dans les réseaux de neurones multicouches était entravé par l'absence de méthode efficace pour ajuster les poids internes. Les approches comme le perceptron de Rosenblatt se limitaient aux architectures à une seule couche, tandis que les tentatives de réseaux profonds restaient anecdotiques et inefficaces, faute d'une règle de mise à jour généralisable. L'introduction de l'algorithme de rétropropagation de l'erreur [85], fondé sur la différentiation automatique et la règle de la chaine, a marqué une rupture : elle a permis de propager le gradient de l'erreur depuis la sortie jusqu'aux couches internes, rendant possible l'apprentissage supervisé à grande échelle. Ce mécanisme a ouvert la voie à l'explosion des réseaux profonds, devenus centraux dans l'apprentissage automatique contemporain. La rétropropagation permet de calculer efficacement les gradients des poids de toutes les couches via la règle de la chaine, rendant possible l'entrainement profond. Cette avancée a permis de redonner de l'intérêt

à l'apprentissage neuronal à une époque où l'IA symbolique [3] dominait encore largement.

Couches entièrement connectées

Les **couches entièrement connectées** (ou **FC**, *Fully Connected*) réalisent une **transformation affine** :

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b} \tag{1.8}$$

C'est l'opération de base dans les premiers réseaux de type MLP (*Multilayer Perceptron*). En anglais, on parle de *linear layer*, bien que mathématiquement il s'agisse d'une fonction affine (due au biais **b**). En français, le terme **couche affine** est plus exact, même s'il reste peu employé dans la littérature actuelle [86].

1.6.6 Réseaux de neurones convolutifs

Travailler sur de l'imagerie implique presque inévitablement l'usage de **réseaux neuronaux convolutifs** [4] [5]. Les **CNN** restent aujourd'hui une référence incontournable, car ils exploitent finement la structure spatiale des images.

Un **réseau convolutif** est essentiellement constitué de **couches de convolution** successives, généralement accompagnées de fonctions d'activation, de sous-échantillonnages et de couches entièrement connectées en fin de réseau. Les **couches de convolution** appliquent un ensemble de filtres appris qui extraient des motifs locaux à différentes échelles, tels que les bords, textures ou contours, et dont la combinaison progressive permet de construire des représentations de plus en plus abstraites des structures présentes dans l'image. Les **couches de sous-échantillonnage** quant à elles visent à réduire la redondance spatiale et la dimensionnalité des cartes de caractéristiques, tout en préservant les informations discriminantes essentielles à la tâche de reconnaissance.

Convolution 1D

L'opération de **convolution** 1D sur un signal discret $x[n]$ par un noyau $h[k]$ consiste à faire glisser le noyau sur le signal et à effectuer un produit scalaire local à chaque position :

$$y[n] = \sum_k x[n-k] \cdot h[k] \tag{1.9}$$

Cela permet d'extraire des motifs temporels ou spatiaux locaux, en respectant leur **ordre** et leur **proximité**.

3. L'IA symbolique, en opposition à l'IA connexionniste, représente les systèmes d'IA traditionnels privilégiant une formalisation explicite de règles, guidant ainsi le système. Nous y reviendrons en section 2.3.3

4. Les réseaux convolutifs sont aussi appelés ConvNet ou CNN ou *Convolutional Neural Networks*

5. Bien que les architectures à base de *transformers* soient désormais courantes en vision par ordinateur, elles sont apparues plus tard et reposent sur des fondations différentes. Elles seront vues en chapitre 3.7

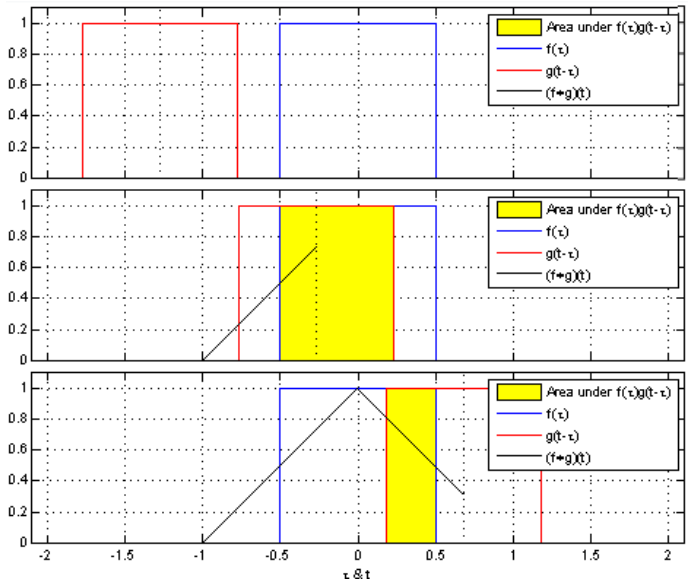


FIGURE 1.24 – **Explication visuelle d'une convolution 1D. Le fonctionnement peut être vu comme un glissement d'une fonction sur l'autre. Pour chaque position, la valeur de sortie est l'aire sous la courbe du produit des deux fonctions.** Sur la 1ère image, la fonction porte rouge nommée g est centrée sur $-1,25$, le produit des deux fonctions est alors nul partout, son aire est donc nulle. Sur la 2ème image, g est centrée sur $-0,25$, le produit des fonctions est égale à 1 sur la longueur de l'intersection de g et f, 0 ailleurs. L'aire sous la courbe de ce produit est représentée en jaune, et est calculée par « largeur × hauteur », soit visuellement $0,75 \times 1 = 0,75$. La fonction noire associe donc $y = 0,75$ lorsque $t = -1,25$. Après la 2ème trame, l'aire augmente jusqu'à ce que f et g soient confondues, avant de diminuer, créant ainsi ce résultat en fonction triangle. (Crédit : [M10])

Convolution 2D

Pour une image, la **convolution 2D**, dont une représentation est visible en figure 1.25, applique un noyau K sur une fenêtre locale de pixels :

$$Y[i,j] = \sum_m \sum_n X[i+m, j+n] \cdot K[m,n] \tag{1.10}$$

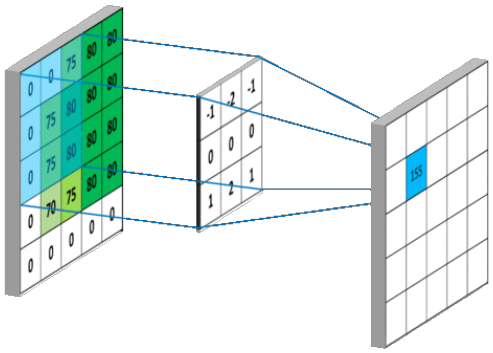


FIGURE 1.25 – **La convolution 2D**, dont le filtre va parcourir les positions de la matrice d'entrée (à gauche) afin de calculer toutes les valeurs des positions de la matrice de sortie (à droite). (Crédit : [M11])

Ce mécanisme préserve à la fois la **structure locale** (proximité des pixels) et leur **organisation spatiale** (positions relatives en x et y), ce qui est **primordial** pour l'analyse d'images. Contrairement aux MLPs classiques, les CNN capturent donc la **richesse contextuelle locale** dès les premières couches [87, 86].

Sous-échantillonnage ou *pooling*

Le sous-échantillonnage est utilisé dans les réseaux de neurones convolutifs (CNN) afin de réduire la dimension spatiale des représentations tout en conservant les informations essentielles. Cette réduction permet de diminuer le nombre de paramètres, d'accélérer les calculs et de limiter le surapprentissage. Parmi les différentes méthodes, le sous-échantillonnage par maximum (*max pooling*) est la plus courante : il sélectionne la valeur maximale dans une région locale (souvent 2×2), ce qui permet de conserver les activations les plus élevées. À l'inverse, le sous-échantillonnage par moyenne (*average pooling*) calcule la moyenne, mais tend à lisser les caractéristiques, ce qui le rend moins populaire dans les architectures modernes.

Architecture globale des CNN

Les CNN sont un arrangement de ces couches. Comme mentionné plus tôt, la théorie veut que les premières couches du CNN vont chercher à capturer des caractéristiques globales, la taille des filtres de convolutions sera alors la plus grande possible. Au fur et à mesure de l'avancée dans le réseau, ces couches de convolution contiendront des filtres de plus en plus petits et de plus en plus nombreux. Les CNN les plus connus sont l'AlexNet [88] (2012), le VGG16 [89] (2014) et le ResNet50 [90] (2015). Ce schéma de construction se visualise très bien avec les architectures des modèles, comme celui de VGG16 en figure 1.26.

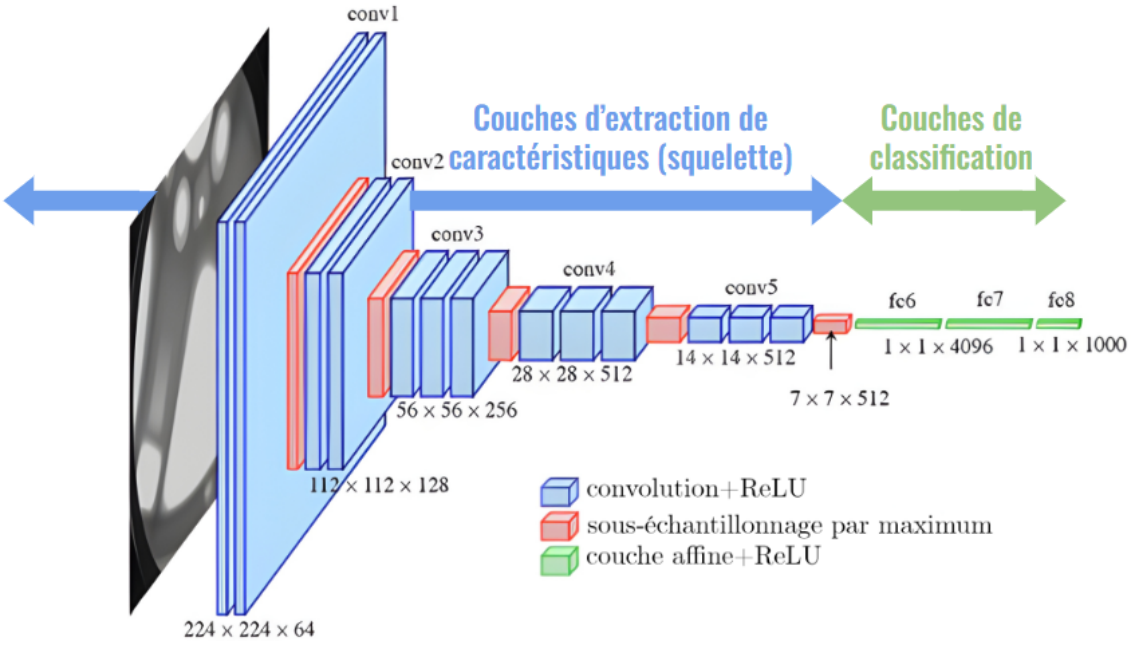


FIGURE 1.26 – **Schéma de l'architecture du VGG16.** Le modèle est composé d'alternances entre des couches de convolution, de fonction d'activations ReLU et de sous-échantillonnage par maximum. L'architecture est constituée de deux parties, une partie d'extraction de caractéristiques et une partie classifieur.

Les CNN ont une place importante car ce sont par eux que le grand public a pu réellement voir les prouesses dont sont capables les IA modernes et ainsi, recréer un engouement général pour les réseaux de neurones. Ceci n'était pourtant pas garanti, après les deux périodes appelées « hivers de l'IA » qui ont eu lieus précédemment.

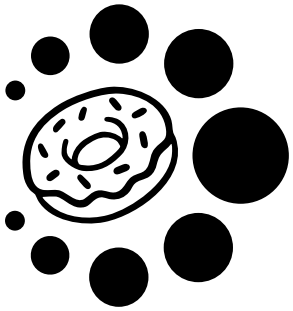
1.7 Conclusion

Dans ce chapitre, il a été montré que la recherche d'une expérience immersive, mêlant différentes modalités sensorielles telles que la vision et l'audition afin de contribuer à l'engagement du spectateur, est intemporelle et remonte à -37 000 ans. Le cinéma est une manifestation de cette recherche, parmi tant d'autres. Elle a été l'aboutissement de plusieurs itérations d'inventions, qui ont constitué le précinéma. D'abord les jouets optiques avec une animation dessinée de deux secondes, puis le kinétoscope avec une animation photographique individuelle, et enfin, le cinéma projeté en salle collective, avec les dessins animés projetés d'Émile Reynaud puis avec les films photographiques des frères Lumière. Cependant, l'évolution du cinéma ne peut pas s'arrêter à son invention, dû aux autres modes de consommation des films qui apparaissent. La télévision, puis le support cassette et enfin les plateformes de streaming bousculent les règles et les législations qui tentent plus ou moins adroitement de protéger l'industrie du cinéma, dont dépendent les autres moyens de transmission du film.

Afin de protéger l'intérêt qu'a le public pour le cinéma autrement qu'en imposant des contraintes aux autres médias, le cinéma profite de ses installations physiques pour innover. Les ciné-débats, les évènements thématiques, mais aussi des formats plus couteux, basée sur des nouvelles technologies, tels que ICE font irruption. Des recherches dans la littérature scientifique existantes concernant l'immersion nous on permis de valider le potentiel immersif de ICE, ainsi que de savoir la direction dans laquelle le format devrait ou ne devrait pas évoluer pour améliorer encore l'immersion du spectateur. Un état de l'art sur les dispositifs d'affichage immersif a été réalisé, mettant en avant les inconvénients et avantages des technologies similaires afin de mieux cerner la particularité de notre application par rapport à celles des autres dispositifs. Enfin, une redéfinition de l'intelligence artificielle a été mise en place et l'apprentissage profond, la base de nos travaux, a été présenté en dernière partie.

CHAPITRE 2

ICE-HOLSS : UNE ONTOLOGIE DE LA SÉMANTIQUE LATENTE DES SCÈNES POUR ICE



En raison de difficultés techniques, votre environnement de test ne fera pas l'objet d'une supervision.

Pierre-Alain de Garrigues, pour Aperture Laboratories

Table des matières

2.1 Introduction 69
2.2 Postproduction et décision ICE 70
2.2.1 Environnement de travail ICE 70
2.2.2 Décisions et génération des contenus ICE 70
WALL 72
WALL EXPAND 72
SEAM 72
WALL EXPAND FH 72
WALL COLOR 72
2.2.3 Contraintes technologiques et de la technique actuelle 72
Validation par les studios et obligation de cohérence 73
Déphasage temporel 74
Amplification de la vitesse 74
2.2.4 Décisions des graphistes ICE 75
Principe de base 75

Cas remarquables 76
Techniques avancées 77
2.3 Formalisation de la décision ICE à l'aide d'ontologie 78
2.3.1 Problématique 78
2.3.2 État de l'art des ontologies du cinéma 78
2.3.3 Qu'est-ce que l'ontologie, les graphes sémantiques et l'IA symbolique 80
2.3.4 Analyse filmique 83
Échelle de plan 84
Le cadrage 86
Mouvement dans l'image et mouvement de caméra 87
Image 89
Son 90
Profondeur 91
Composition 92
Lumière 93
Montage 94
2.4 Retours d'expérience sur la création d'une ontologie dans le cadre de ICE . 99
2.5 Conclusion 101

2.1 Introduction

Ce chapitre s'inscrit dans le cadre d'une revue complète autour de la technologie ICE [1], ce dispositif de projection de cinéma qui étend latéralement un film au-delà de l'écran principal grâce à deux panneaux LED installés de part et d'autre de la salle. L'objectif de cette technologie est de renforcer l'immersion du spectateur en prolongeant visuellement l'univers du film dans l'espace physique du cinéma.

Cependant, la création des contenus ICE repose sur un travail artistique complexe, où les graphistes conçoivent manuellement les extensions latérales selon des critères à la fois esthétiques, narratifs et perceptifs. Ces choix ne suivent aucune règle fixe : ils dépendent du contexte filmique, du rythme des plans, et de l'intention du réalisateur, ce qui confère à la production ICE un caractère profondément subjectif et empirique.

Pour cerner les problèmes qui seront à adresser tout au long de ce manuscrit, il est nécessaire de comprendre le fonctionnement de travail de l'équipe de postproduction ICE. Les contenus ICE sont générés par les graphistes à l'aide de logiciel **Adobe Premiere Pro**, et plus précisément, via sa surcouche **ICE Immersive Creator** visible sur la figure 2.1, qui facilite le démarrage de tout nouveau projet ICE.

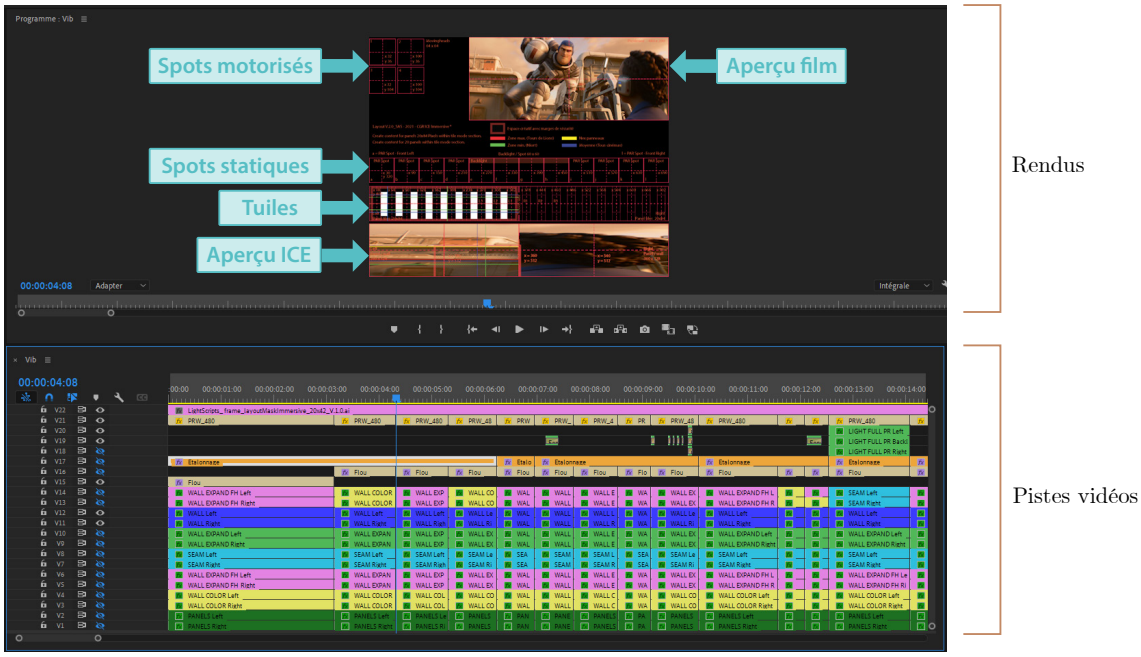


FIGURE 2.1 – **Capture d'écran du logiciel de travail de l'équipe de postproduction ICE, lors de la création ICE d'une bande-annonce de Buzz L'éclair.** Le logiciel est une préconfiguration d'Adobe Premiere Pro. Les pistes vidéos (en bas) permettent au graphiste de choisir quels seront les traitements d'image appliqués aux bords gauche et droit du film afin d'être placés dans les panneaux ICE et les rendus (en haut) montre le visuel des équipements ICE. Ce procédé doit être recommencé autant de fois que nécéssaire en parcourant tout le film.

Dans ce chapitre, nous proposons une analyse globale du dispositif ICE, en mettant en évidence ses caractéristiques techniques ainsi que son mode de fonctionnement en salle. Nous étudierons ensuite les méthodes de travail des graphistes impliqués dans la

1. Immersive Cinema Experience

production ICE, afin de comprendre les logiques visuelles et narratives qui guident leurs décisions. Enfin, une analyse filmique et l'ontologie ICE-HOLSS (pour *Holistic Ontology of Latent Scene Semantics* ou l'Ontologie Holistique de la Sémantique Latente des Scènes) du dispositif seront présentées pour formaliser les relations entre l'espace filmique, la perception du spectateur et la mise en scène immersive propre à ICE. Cette étude constituera la base conceptuelle et méthodologique du chapitre suivant, consacré à la modélisation automatique de ces choix visuels.

2.2 Postproduction et décision ICE

2.2.1 Environnement de travail ICE

Dans cette section, nous détaillons le fonctionnement du logiciel de travail dédié à la création ICE illustré dans la figure 2.1. La moitié du haut correspond au rendu du film en ICE et la moitié du bas représente les couches pistes vidéos. La zone `Aperçu film` affiche le film, et la zone `Aperçu ICE` montre ce qui sera appliqué aux panneaux ICE latéraux.La zone `Spots statiques` concerne les projecteurs non motorisés, notamment les PAR situés à gauche et à droite de la salle, ainsi que les lumières arrières positionnés entre les deux. Ces équipements sont visibles dans le chapitre précédent sur la figure 1.10. Ces lumières, incluant les spots motorisées sont pilotables en RVB pour s'adapter à la couleur de la scène à l'écran. Dans une salle de cinéma, les PAR sont les lumières situés sur les côtés et les lumières arrières sont à l'arrière de la salle, sous la fenêtre de cabine de projection. Les spots motorisés sont calibrés pour se répartir un éclairage vers le bas, vers les fauteuils de la salle.

Afin d'étendre le film vers les écrans latéraux, la technologie ICE se base sur les 360 pixels de la partie gauche et de la partie droite du film. Sur un format CinemaScope (858×2048) qui est la définition standard actuelle du cinéma, cela signifie que deux rectangles 858×360 [2] sont récupérés (environ 35 % de la largeur complète). C'est ce qu'on peut apercevoir dans la partie 2 de la figure 2.1.

En ce qui concerne la moitié du bas de l'écran sur la figure 2.1, on peut voir les différentes pistes vidéos du logiciel de montage. Ces dernières réagissent de la même manière que les calques des logiciels d'éditions d'image : chaque piste vidéo vient recouvrir celle d'en dessous comme si elles étaient empilées physiquement les unes sur les autres. Pour choisir un effet plutôt qu'un autre, l'effet est remonté au dessus de la pile pour devenir visible. Un descriptif de chacune de ces pistes est donné dans la figure 2.2.

2.2.2 Décisions et génération des contenus ICE

Après avoir présenté l'environnement de travail des graphistes ICE, nous nous intéressons à présent au processus de création associé à chaque production ICE, ainsi qu'aux différentes décisions conceptuelles et techniques qui le guident.

Les choix créatifs à la disposition des graphistes ICE constituent la base de travail des graphistes. Ce sont ceux que l'on peut voir sous la forme de différents calques de la figure 2.2 et leurs rendus sont montrés en figure 2.3. Leurs descriptions techniques et leurs utilités sont décrites ci-dessous :

2. Dans ce manuscrit, toutes les dimensions sont écrites dans l'ordre hauteur × largeur, standard mathématique et fréquemment informatique

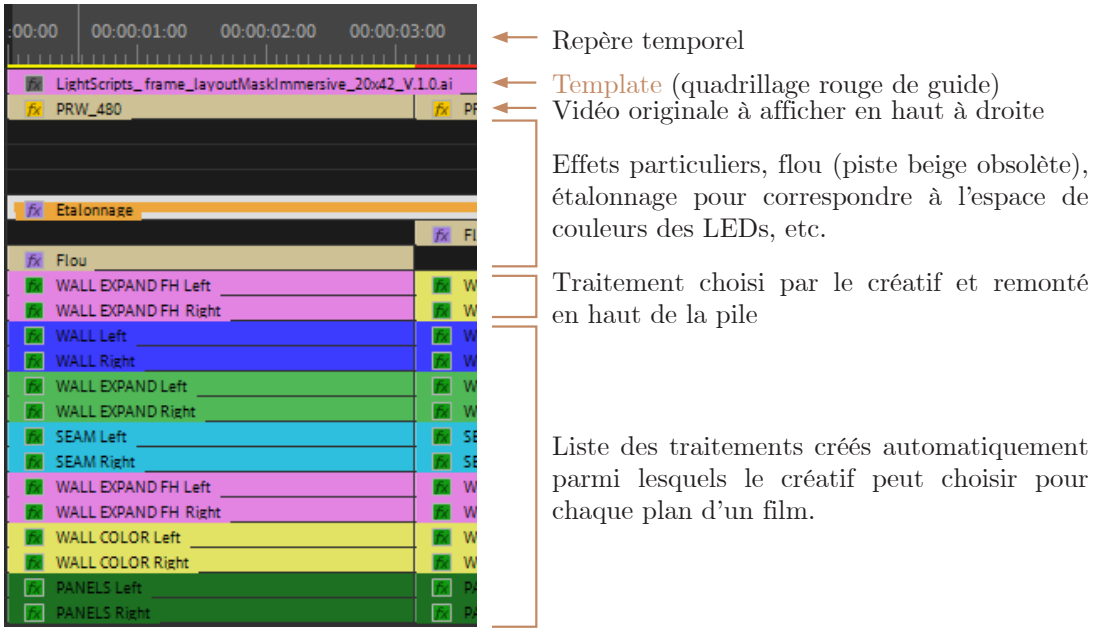


FIGURE 2.2 – **Zoom sur les pistes vidéos de l'outil Première Pro de l'équipe ICE.** Description des différentes pistes vidéos empilées lors d'une adaptation de film en ICE. Les pistes peuvent être cliquées-glissées par le graphiste pour être déplacés afin d'être choisies pour un certain plan du film.

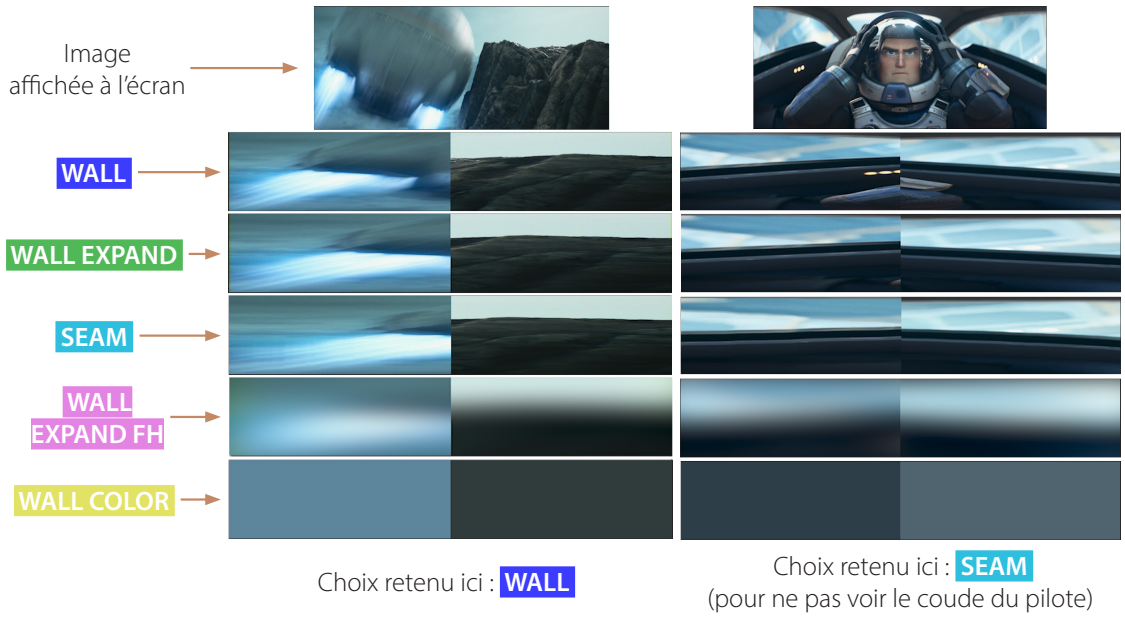


FIGURE 2.3 – **Illustration des différents prétraitements d'image disponibles à l'usage pour l'équipe ICE.** Ces derniers varient selon la frame d'origine affichée (en haut) et sont visibles en dessous, pour le panneau ICE de gauche et pour celui de droite. On peut remarquer que ces traitements sont triés du plus détaillé au moins détaillé. Le choix retenu pour chacun de ces choix est visible en bas de la figure.

WALL

Duplication des [858×360] pixels latéraux dans les deux espaces `WALL` [128×360]. Ce traitement est celui qui donne le plus de détails.

WALL EXPAND

Duplication des [858×240] pixels latéraux vers [128×360]. Ce traitement permet de donner du détail, en esquivant un élément perturbant qui se trouverait trop proche des bords de l'écran.

SEAM

Traitement développé le plus récemment (juin 2022), qui applique un `WALL` après avoir appliqué un traitement d'image de « seam carving » pour élargir le film de départ avec peu de perte de qualité. Le `SEAM` permet d'esquiver un élément perturbant, d'une manière similaire au `WALL EXPAND`, mais en zoomant moins, et en déformant moins le rendu de l'image.

WALL EXPAND FH

Applique un `WALL EXPAND` puis ajoute sur ce dernier un flou vertical très important fourni par Premiere Pro (valeur 50,0 de Premiere). C'est le choix le plus utilisé, permettant d'écarter la majorité des perturbations sans pour autant perdre entièrement le dynamisme du plan.

WALL COLOR

Aplat de couleur unie (autrefois par défaut une moyenne du bord de référence, aujourd'hui une couleur prise d'un point fixe dans ce bord). Elle est très souvent réajustée (teinte/luminance/saturation) par les graphistes ICE sur une autre couleur. C'est souvent le choix de référence lors d'un dialogue, surtout lors d'un plan fixe, et d'autant plus dans une scène en intérieur ou avec une faible luminosité. Il permet de laisser le spectateur se concentrer sur la narration.

Les choix disponibles aujourd'hui aux graphistes qui donnent le plus de détails lors du film pourrait être ordonnés de cette manière :
WALL > SEAM > WALL EXPAND > WALL EXPAND FH > WALL COLOR
Même si un plus grand niveau de détail ne garantit pas une meilleure solution, il est intéressant de garder cet ordre en tête. Nous verrons dans les décisions du graphiste les raisons qui peuvent amener à un choix créatif gardant moins de détails.

2.2.3 Contraintes technologiques et de la technique actuelle

Les contraintes propres à ICE se manifestent à la fois sur le plan technologique et sur le plan technique. Par « technologie », nous entendons ici un dispositif matériel à visée commerciale, en l'occurrence, des panneaux LED latéraux. Ils sont destinés à diffuser des contenus complémentaires au film, dans le but d'amplifier l'immersion du spectateur et de renforcer l'intégration spatiale de l'œuvre dans la salle. La technologie englobe donc le matériel éléctronique physique (*hardware*) lui-même, comme la faible résolution des panneaux LED, leur position orthogonale à la toile principale, ou encore les différences d'espace

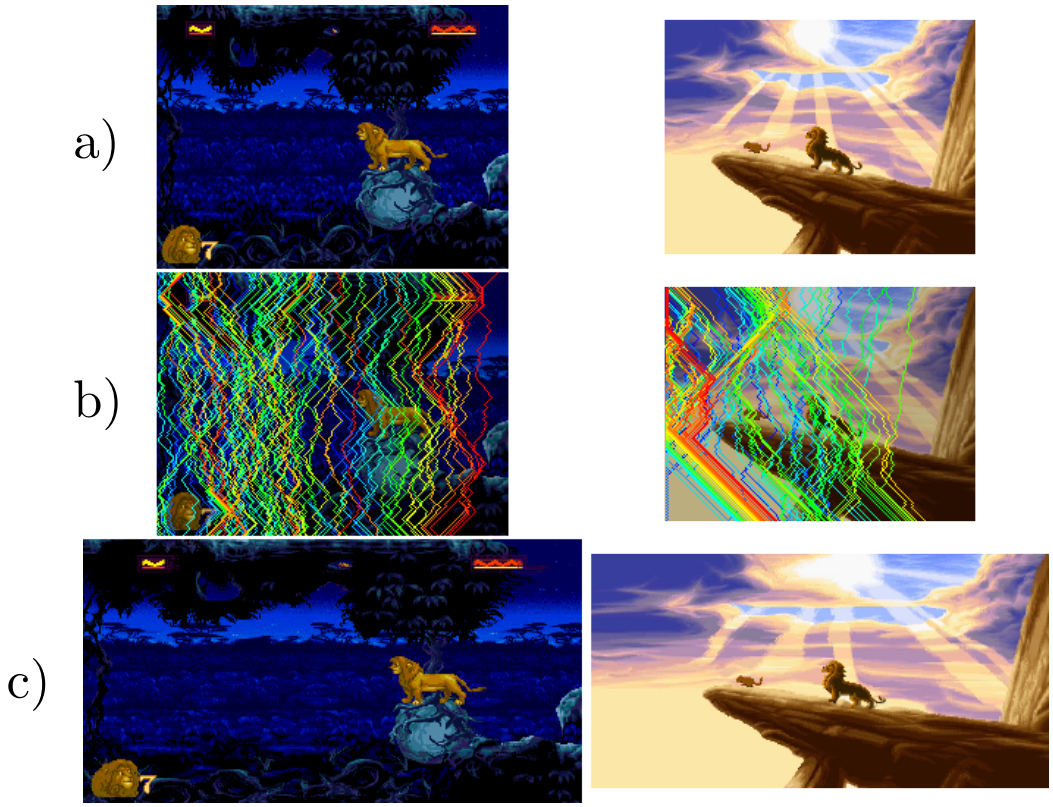


FIGURE 2.4 – **Illustration de l'algorithme de *Seam Carving* pour l'agrandissement d'image.** Cet algorithme permet d'étirer ce qui attire le moins l'œil et de ne pas déformer ce qui contient le plus d'informations, avec a) représentant les images d'origine, b) les 150 sillons de plus faible énergie et donc les moins chargés en détail, qui vont être agrandis artificiellement et c) les images agrandies résultantes. (Crédit : auteur [M12])

colorimétrique par rapport au projecteur central. À cela s'ajoute la « technique », c'est-à-dire l'ensemble des méthodes, choix de production et contraintes (souvent implicites) que les créateurs et studios adoptent pour tirer parti du dispositif. Par exemple, le fait que les contenus diffusés soient issus d'une duplication latérale des bords du film, ou encore la validation obligatoire de chaque version ICE par les studios, relèvent de considérations techniques. Ces pratiques définissent la manière dont la technologie ICE est concrètement exploitée dans le cadre de la production cinématographique immersive.

Validation par les studios et obligation de cohérence

Le processus de validation des contenus ICE constitue une étape cruciale pour garantir leur qualité. En particulier, ces contenus doivent être approuvés par les distributeurs des films. Cette étape de QC [3], consiste à présenter le film accompagné de ses contenus ICE en salle aux distributeurs, généralement aux responsables VFX [4]. Ces derniers se distinguent par un niveau d'exigence élevé et une grande précision dans leurs observations.

Deux contraintes importantes liées aux spectateurs doivent également être respectées :

3. pour *Quality Check*, ou contrôle qualité en français
4. VFX : effets spéciaux vidéos

1. Les panneaux ne doivent pas détourner l'attention du spectateur du film. Plus précisément, ils ne doivent jamais inciter le spectateur à les regarder directement, ni provoquer une quelconque perturbation de son expérience. Pour éviter toute perturbation, ce qui est affiché dans les panneaux doit donc être cohérent avec le film.
2. Si un spectateur lambda parvient à identifier que les éléments des panneaux correspondent à des images dupliquées des bords du film, cela signifie que la magie immersive n'a pas fonctionné et que la création ICE doit probablement être revue afin de produire des contenus plus cohérents et harmonieux entre eux.

Déphasage temporel

Par la technique actuelle de duplication des bords du film pour la génération des contenus étendus, ce qui est vu par le spectateur lors d'un mouvement s'apparente à un déphasage temporel (le contenu étendu semble être légèrement en avance par rapport au film). Une discontinuité spatiotemporelle est ainsi visible. Pour une question d'explicabilité nous allons transcrire ce phénomène en 2D, c'est-à-dire avec des écrans sur un même plan :

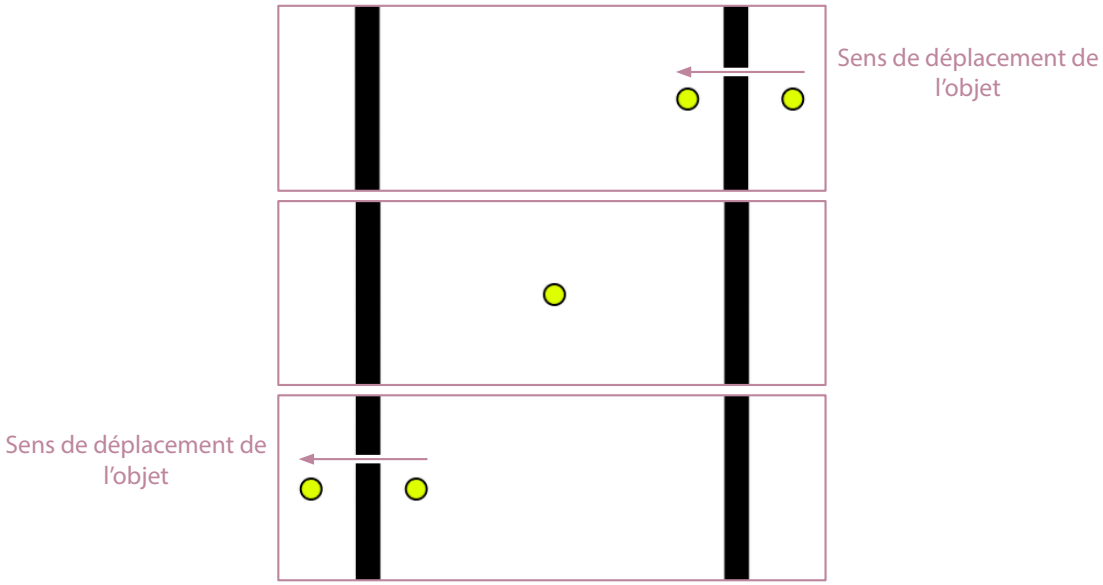


FIGURE 2.5 – **Illustration de la contrainte du déphasage temporel en ICE.** Reproduction d'une toile et des panneaux ICE, ici mis à plat, lorsqu'un objet traverse horizontalement l'écran. Dû à la technique ICE basée sur la réplication des bords de l'écran dans les panneaux ICE, on peut voir certaines trames où la balle est dupliquée.

On peut voir sur figure 2.5 qu'à certains instants, le spectateur peut voir la balle dupliquée. Elle peut se trouver à la fois sur la toile et sur les panneaux ICE.

Amplification de la vitesse

Les déformations des zones [858×360] dans deux espaces [128×360] engendrent, pour la même hauteur, un étirement horizontal d'environ ×6,70, ce qui est visible en figure 2.6. Pour un choix de `WALL EXPAND`, on passe de [858×240] à [128×360] soit un étirement horizontal d'environ ×10,05. Bien évidemment, ce calcul très simple ne prend pas en compte l'espace 3D d'une salle et la disposition latérale des panneaux, qui se rapprochent du spectateur.

L'occupation du champ de vision de ces panneaux latéraux est beaucoup plus importante lorsqu'ils sont placés autour du spectateur, plutôt que sur le même plan que la toile. Sur l'exemple de la dernière salle ICE montée en France (Carcassonne), la projection spatiale des panneaux latéraux sur le même plan que la toile nous montre que ces derniers couvrent le même champ visuel que si un écran de 100 mètres de long se trouvait de chaque côté de la toile de 10 mètre de long.

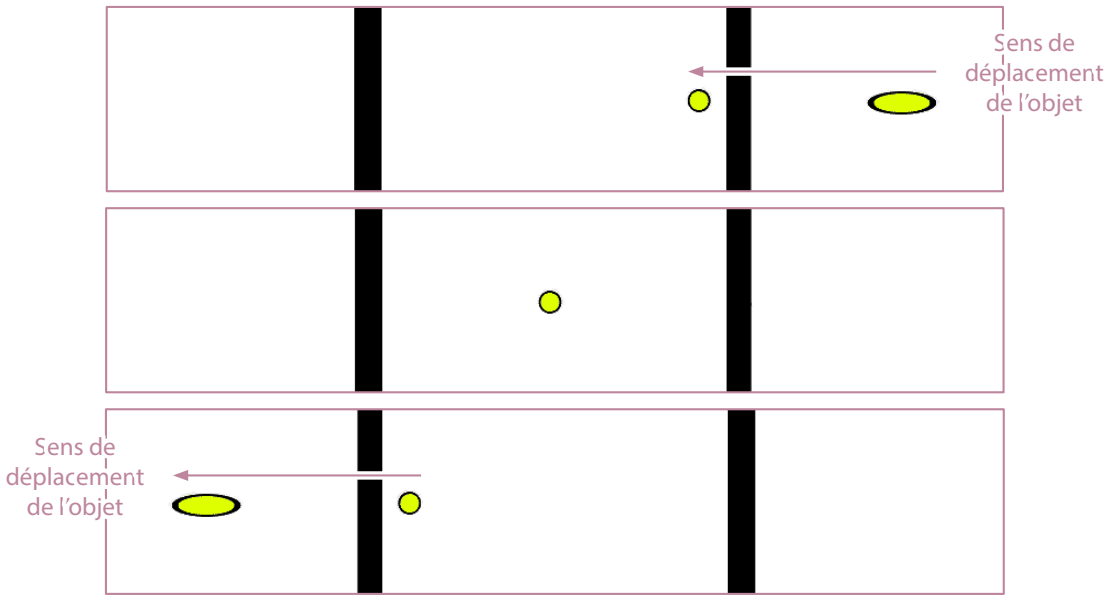


FIGURE 2.6 – **Contrainte de l'amplification de la vitesse.** Dû aux panneaux ICE qui couvrent une plus grande zone physique que l'image qu'ils accueillent, les images sont étirés et la vitesse horizontale est grandement amplifiée.

On voit que la vitesse horizontale de l'objet est bien plus importante dans les panneaux que dans la séquence centrale. Dans la pratique, le phénomène est tolérable pour un choix créatif de `WALL` lorsque l'objet est rapide. En revanche, il peut vite être gênant pour un `WALL EXPAND`, pour un `SEAM`, ou pour un `WALL` si le mouvement est léger, car la différence de vitesse devient trop perceptible.

2.2.4 Décisions des graphistes ICE

Bien que certaines règles générales guident la création des contenus pour ICE, de nombreux cas particuliers peuvent se présenter. Il ne s'agit donc pas ici de proposer un guide exhaustif et rigide : un tel guide n'existe pas, tant la décision finale dépend du rendu dans une salle réelle, et des impressions procurées dans ce contexte. Nous présentons plutôt quelques situations concrètes permettant d'illustrer les principes directeurs qui orientent les choix graphiques.

Principe de base

De manière générale, le choix du flou se prête bien à la création d'une ambiance immersive, tandis que le détail visuel est préférable pour les séquences dynamiques. Enfin, le respect de l'intention du réalisateur reste fondamental : une scène d'action rapide, contrastée, avec un rythme percutant, peut s'accommoder de panneaux latéraux dynamiques, tandis

qu'une scène calme ou intimiste appelle une continuité visuelle plus douce et discrète. Le déphasage temporel, évoqué comme une contrainte inhérente à la technique actuelle, peut rester tolérable, y compris lors de conservation de détails dans les panneaux (via traitements `WALL`, `SEAM`, `WALL EXPAND`) dans les cas suivants :

— si aucun élément saillant, visuellement marquant ou susceptible d'attirer fortement l'attention ne se situe dans les bords latéraux de 360 px ;
— si un objet en mouvement traverse rapidement l'écran avec une vitesse horizontale élevée, sans marquer d'arrêt dans les zones de duplication.

Cas remarquables

Description du cas	**Décisions à prendre**
Plan large, pas d'élément perturbant dans les 360px des extrémités	Choix gardant du détail possible et appréciable (WALL) quand le plan est fixe.
Image avec répétition de *patterns*, de motifs ou de formes (vagues, dunes de sables, effets spéciaux)	Choix gardant du détail possible et appréciable (WALL)
Plan large, un élément saillant ou perturbant dans les 360px des extrémités mais pas collé aux bords	Essayer SEAM ou WALL EXPAND pour voir si l'objet gênant est toujours présent dans les contenus générés.
Lors d'un plan moyen, mouvement de sortie (ou entrée dans de plus rares cas) de champ rapide d'un élément	Choix gardant du détail possible et appréciable (WALL) du côté du mouvement (Si un élément perturbe de l'autre côté, travailler en asymétrie avec un choix moins détaillé du côté de l'élément perturbant). Le mouvement n'est pas obligé d'appartenir à un élément particulier, des gerbes de feu qui proviennent d'une explosion ou d'un choc peuvent suffire.
Intermittence de couleurs très contrastées	Si on veut garder un niveau de détails supérieur ou égal à WALL EXPAND FH, pour diminuer la grande différence d'intensité ou de teinte entre plusieurs couleurs consécutives, une solution est la superposition de la méthode par un WALL COLOR qui viendra atténuer cette différence.

Champ/Contre-champ	S'il y a plus de 2 plans champ/contre-champ, garder un niveau de détail très bas (quasi-systématiquement WALL COLOR) pour éviter au maximum un effet de pompage ou de changements d'ambiance trop brusques et trop fréquents de la salle de cinéma, ce qui est perturbant pour le spectateur. Il est possible d'appliquer le même effet sur tous les champ/contre-champ de la même scène pour qu'il n'y ait pas de changement. Il est souvent préférable garder une ambiance sombre selon l'ambiance du plan pour recréer l'atmosphère intimiste et calme que voulait sans doute le réalisateur. Si trop de mouvements parasites sont visibles avec un WALL EXPAND FH, utiliser WALL COLOR.
Gros plan	Souvent préférable de garder une ambiance sombre selon l'ambiance du plan pour recréer l'atmosphère intimiste et calme que voulait sans doute le réalisateur. Si trop de mouvements parasites sont visibles avec un WALL EXPAND FH, utiliser WALL COLOR.
Fondu enchainé entre deux plans	Oblige à mettre un fondu enchainé aussi entre les différents choix créatifs s'ils sont différents pour les deux plans.

Techniques avancées

Lors de cas particuliers, les techniques avancées permettent d'affiner le rendu souhaité d'un cas remarquable. Elles viennent en complément des précédentes règles. Ces techniques avancées sont combinables entre elles. **Asymétrie :** Il est possible de fonctionner de manière asymétrique, donc d'avoir un certain choix créatif à gauche et un choix différent à droite. **Superposition :** Il est possible de créer une superposition d'un `WALL COLOR` en opacité diminuée par dessus un autre choix créatif (`WALL EXPAND FH` par exemple). Cela a pour but de masquer des artéfacts visibles dans les panneaux ou bien pour atténuer des scènes d'action avec des couleurs ou des mouvements trop vifs. **Traitement additionnel :** Il est possible d'ajouter un flou supplémentaire (ou autre effet) par dessus un choix créatif. **Ralentissement :** Lors d'un choix de `WALL EXPAND` lorsqu'il y a du mouvement à l'écran, ou lors d'un choix de `WALL` lorsque le mouvement est trop faible, la différence de vitesse causée par l'amplification de vitesse dans les panneaux peut être perturbante, comme mentionné dans la section 2.2.3. Il est possible de ralentir le temps d'un contenu généré dans les panneaux ICE afin d'y remédier.

Si un léger mouvement de panoramique est présent, l'amplification de la vitesse dans les panneaux due à un `WALL` sera perturbante. Dans ce cas, s'il est souhaité de garder du détail, on peut ralentir le temps du plan dans les panneaux, afin de baisser l'amplification de vitesse. Il est possible de ralentir le temps d'un contenu généré, par exemple pour les `WALL EXPAND` qui amplifie beaucoup le mouvement comme mentionné dans la section 2.2.3. **Subdivision fondue :** Lorsque le début et la fin d'un même plan présentent des caractéristiques visuelles différentes (par exemple, un mouvement rapide suivi d'une phase plus calme), il est possible de diviser temporellement le plan en deux segments. Chaque segment reçoit un choix créatif distinct, avec une transition en fondu enchainé entre les deux. **Ajustement colorimétrique :** Si les extrémités d'un plan (d'où proviennent les zones dupliquées vers les panneaux ICE) n'ont pas des couleurs représentatives de

l'ambiance colorimétrique générale de la scène, il est possible d'ajuster la teinte d'un `WALL EXPAND FH`, ou de modifier la couleur unie d'un `WALL COLOR` via un filtre. Cet effet peut aussi se faire par superposition.

Cette section a permis de mettre en lumière la richesse et la subtilité des décisions à effectuer par les graphistes en vue d'une postproduction d'un film en ICE. Chaque décision repose sur une compréhension des intentions esthétiques et narratives, difficile à formaliser de manière fixe et incontestable. Au contraire, bien que les graphistes ICE arrivent à un consensus commun, ces décisions reposent sur des ressentis perceptifs, directement éprouvés dans une salle de cinéma, et sur une subjectivité artistique assumée. Afin d'explorer des approches capables d'assister ce processus dans des contextes complexes ou à grande échelle, il est pertinent de s'appuyer sur les avancées en compréhension de la vidéo. La section suivante propose ainsi un état de l'art structuré des méthodes actuelles d'analyses et de compréhension de vidéo [5], qui servira de socle à la construction du modèle présenté par la suite.

2.3 Formalisation de la décision ICE à l'aide d'ontologie

2.3.1 Problématique

Là où le premier chapitre a expliqué dans quel contexte et pourquoi ICE a été inventé, il a aussi montré les principes fondamentaux d'immersion que le concept respecte, ainsi que les autres travaux, scientifiques et commerciaux, tirant profit de la vision périphérique lointaine pour améliorer l'immersion et l'engagement de l'utilisateur. Notre objectif global étant l'assistance à ICE par une IA, il nous faut d'abord comprendre le domaine précis de l'intention filmique.

Dès le début de cette thèse, et dans une démarche de CGR, nous avons commencé par la conceptualisation du domaine filmique afin d'élaborer une ontologie de domaine. Cette ontologie vise à faciliter la compréhension du domaine et, surtout, à renforcer l'explicabilité des modèles d'IA. La décision de créer cette ontologie a été motivée par l'expertise des graphistes de contenus ICE.

Sous la demande non négotiable d'Arthur Loosveldt pour CGR, ayant démissionné quelques mois plus tard (septembre 2022), nous allons effectuer des recherches d'analyse filmique. L'analyse filmique est le fait de transmettre des émotions ou des sens cachés via le cadre, les formes, les couleurs ou tout autre indice non-explicite, à la limite du subliminal ou du subconscient. En conclusion après ces recherches en analyse filmique, nous allons créer un diagramme des classes puis une ontologie sur les analyses filmiques dans le cinéma.

2.3.2 État de l'art des ontologies du cinéma

Alors qu'il existe peu d'ontologies exploitant le contenu visuel du film et ses effets perceptifs, de nombreux travaux portent sur les métadonnées, la réception du public et les analyses de sentiment. Ainsi, [91] met en place un système d'analyse linguistique fondé sur le graphe sémantique SentiWordNet, afin d'identifier dans les critiques en ligne les phrases

5. communément : *video understanding*

subjectives et de leur associer un jugement positif ou négatif selon le vocabulaire employé pour décrire différents aspects du film tels que le scénario, la musique ou le jeu des acteurs. De leur côté, [92] et [93] combinent analyse de sentiment et analyse de réseaux sociaux pour prédire le succès d'un film. Le premier exploite les données issues de Tencent Weibo (un réseau social chinois comparable à Twitter) afin d'extraire à la fois des indicateurs quantitatifs comme le volume de publications et l'activité des comptes, et des indicateurs sémantiques liés au contenu émotionnel des messages, dans le but de prédire les recettes au box-office. Le second s'appuie sur les discussions des utilisateurs d'IMDb [6], la base de données cinématographique en ligne, pondérées selon leur centralité dans le réseau, c'est-à-dire selon leur degré d'influence et leurs connexions avec d'autres contributeurs, pour montrer que la structure de ces échanges peut anticiper les résultats commerciaux et les nominations aux Oscars. [94] propose quant à lui un système de recommandation combinant le filtrage collaboratif, qui repose sur les similarités entre utilisateurs partageant des goûts proches, et le filtrage par contenu, qui s'appuie sur les caractéristiques internes des films, tout en intégrant l'analyse des émotions exprimées sur Twitter pour adapter les suggestions aux réactions réelles du public. Enfin, [95] intègre dans un modèle de graphe reliant films, utilisateurs et ressentis des représentations de sentiment extraites à l'aide du modèle de langage BERT, démontrant que l'inclusion de ces signaux émotionnels améliore la pertinence des recommandations. **Ces approches s'intéressent toutes à la dimension émotionnelle du spectateur, mais à travers des traces textuelles et sociales recueillies après visionnage. Elles diffèrent de notre démarche, qui vise à formaliser les liens causaux entre les éléments visuels concrets du film et leurs effets perceptifs immédiats, c'est-à-dire à relier ce que l'on voit à ce que l'on ressent, plutôt que d'inférer l'émotion à partir des discours du spectateur.**

D'autres travaux comme ceux de Li et Al. [96], proposent une ontologie sémantique visant à structurer les concepts visuels d'un film (scènes, objets, activités, personnages, évènements) afin de faciliter l'indexation et la recherche de séquences par contenu. Sa force réside dans la construction d'un vaste lexique hiérarchisé et interconnecté, adapté au domaine cinématographique, qui pose les bases d'une annotation automatique. Cependant, cette approche reste strictement descriptive et centrée sur la catégorisation d'éléments observables, sans modéliser leurs effets perceptifs ou émotionnels sur le spectateur. **Elle ne considère donc pas les relations de causalité ou de ressenti, ce qui la rend inadaptée à une ontologie d'analyse filmique fondée sur les effets visuels et leurs impacts psychologiques.**

D'une autre manière, Cannavò avec ses travaux [97] s'intéresse comme nous à la réponse émotionnelle induite par le contenu visuel d'un film. En l'occurence, l'objet d'étude est le point de vue du spectateur, première personne et personne externe. Sa force est d'établir empiriquement à posteriori des effets du point de vue sur l'immersion narrative, l'attention au récit, les mouvements oculaires, la compréhension et l'implication émotionnelle. **Ces résultats offrent des indicateurs validés, de signaux d'évaluation et de validation, mais sans formaliser des liens de causalité fine entre gestes de mise en scène spécifiques et ressentis. Notre ontologie vise précisément cette cartographie causale explicite entre faits visuels et effets perçus.**

Ces différents travaux, bien qu'ils abordent chacun à leur manière la relation entre le film, son contenu et la réaction du spectateur, révèlent une lacune commune : l'absence d'un cadre formel capable de décrire, relier et raisonner explicitement sur les éléments

6. Internet Movie Database

constitutifs du langage filmique et leurs effets perceptifs. Pour dépasser cette limite et offrir un outil d'analyse à la fois structuré et interprétable par des systèmes intelligents, il devient nécessaire d'introduire une représentation symbolique du savoir filmique. C'est précisément ce que permet l'ontologie, en s'appuyant sur les graphes sémantiques et les principes de l'IA symbolique, que nous présentons dans la section suivante.

2.3.3 Qu'est-ce que l'ontologie, les graphes sémantiques et l'IA symbolique

Une ontologie est une formalisation explicite et partagée d'un domaine de connaissances, composée de concepts, de relations et d'axiomes logiques, destinée à permettre la représentation de ces connaissances [98]. Elle peut être comprise comme une modélisation formelle des connaissances d'un domaine, mais elle ne se limite pas à une structure de données : elle regroupe des concepts, leurs relations et des axiomes logiques permettant un raisonnement automatisé.

L'ontologie est une brique fondamentale créée pour contribuer à l'écosystème du web sémantique [99]. Ce dernier symbolise toutes les motivations de passer du « web », qui relie l'utilisateur à des pages internet toutes indépendantes destinées uniquement à l'humain, à un « web » lisible par une machine, et de permettre aux machines d'établir des relations d'une ressource web à une autre. Tout cela se fait par l'ajout de métadonnées et de relations clairement définies et normalisées, d'abord explicitées par un concepteur humain pour que les machines puissent les interpréter de manière cohérente. En fait, le web sémantique est né de l'idée de Tim Berners-Lee, qui est aussi l'inventeur du WWW, des adresses URL, du HTML et le fondateur du W3C [7], ces quatres sigles représentant chacun un pilier de fondation sur lequel repose tout le web d'aujourd'hui. Il avait alors déjà permis de passer de l'âge des pages statiques locales aux pages web en tant que ressources accessibles par tous. Avec le web sémantique, il continue ce principe de partage global du web, en le rendant accessible par le biais du plus grande nombre d'interconnexions à faire entre les ressources du web [100]. L'objectif est que les ordinateurs puissent non seulement lire des pages destinées aux humains, mais aussi exploiter directement des métadonnées explicites, uniformes et adéquatement formatées, afin de permettre une interopérabilité d'ordinateur à ordinateur et une indexation plus riche que le simple référencement textuel. Par exemple, un site web pourrait indiquer à un moteur de recherche que « Paris » est une ville et que cette ville est la capitale de la France, permettant à la machine de comprendre et relier cette information automatiquement à d'autres données sur les pays ou les capitales.

Pour cela, la première pierre a été posée avec le modèle de structure de données RDF [8] [101]. RDF introduit un modèle simple mais universel basé sur des triplets (sujet - prédicat - objet), où chaque élément est identifié par un URI [9]. On peut d'abord le voir de façon intuitive avec un exemple concret :

$$\text{Alice} : \text{aPourEnfant} : \text{Bob} \tag{2.1}$$

Ici, « Alice » et « Bob » représentent des ressources identifiées par un URI, et « aPourEnfant » la relation entre elles. Un URI est un identifiant unique et global qui permet de

7. WWW : *World Wide Web*, URL : *Unifier Resource Locator*, HTML :*HyperText Markup Language*, W3C : *World Wide Web Consortium*

8. RDF : *Resource Description Framework*

9. URI : *Uniform Resource Identifier*

nommer une ressource de manière univoque sur le web ou dans un espace d'information numérique, indépendamment de sa localisation ou de son mode d'accès. Il peut désigner toute ressource, que ce soit un document, un concept, une personne ou un objet numérique. LURI' est la généralisation des URL bien connues : ces derniers identifient non seulement une ressource mais fournissent également son adresse et protocole pour y accéder, par exemple `https://site.com/page`, ce qui n'est pas toujours disponible pour toutes les ressources du web.

Pour formaliser ce concept de triplet, chaque triplet appartient à un ensemble typé :

$$(s, p, o) \in U \times U \times (U \cup L) \tag{2.2}$$

où s (sujet) et p (prédicat) sont des URI et o (objet) peut être un URI ou un littéral. U représente l'ensemble des URI et L l'ensemble des littéraux (texte, nombres, etc.). L'ensemble des triplets forme un graphe RDF :

$$G \subseteq \{(s, p, o) \mid s \in U, p \in U, o \in U \cup L\} \tag{2.3}$$

où G correspond au graphe RDF et (s, p, o) un triplet valide appartenant au graphe.

Cette structure peut être vue comme un multigraphe orienté étiqueté : les ressources sont les nœuds et les relations sont les arcs. RDF se distingue du XML [10], qui est un langage de structuration syntaxique : XML peut dire « comment organiser des données », mais pas « ce que ces données signifient ». RDF ajoute justement cette sémantique partagée. Pour interroger ces graphes, un langage spécifique, SPARQL [11], permet des requêtes de sélection, de construction, de filtrage ou de parcours de chemins de propriétés.

Cependant, RDF seul ne permettait pas de définir des hiérarchies ou des contraintes sur les données. C'est là qu'intervient RDFS [12] [102], qui ajoute des briques de vocabulaire telles que les classes, les sous-classes, les propriétés et leurs domaines et portées. Par exemple, on peut formaliser qu'« un étudiant est une sous-classe de personne » par :

$$\text{Étudiant} \sqsubseteq \text{Personne} \Longrightarrow \forall x, x \in \text{Étudiant} \ \Rightarrow \ x \in \text{Personne}$$

De même, on peut préciser qu'une propriété comme aPourEnfant relie des individus de type *Personne* à d'autres *Personnes*, ce qui s'écrit :

$$\text{domaine}(p) = C_s, \quad \text{portée}(p) = C_o \Longrightarrow \forall (s, p, o), s \in C_s \ \wedge \ o \in C_o$$

Avec RDFS, il devient donc possible d'obtenir un premier niveau d'inférence logique : si Alice est une Étudiante, on peut automatiquement en déduire qu'elle est une Personne. En revanche, RDFS reste limité : il ne peut pas exprimer des contraintes cardinalitaires, des relations complexes ou des disjonctions.

C'est cette limite qui a conduit à l'introduction des ontologies au sens strict. Une ontologie, en informatique, vise à formaliser la connaissance d'un domaine en combinant des classes, des propriétés, des individus et des axiomes. Ce n'est pas seulement un « format de

10. *Extensible Markup Language*
11. *SPARQL Protocol and RDF Query Language*
12. RDF Schema

données », mais un cadre logique permettant à des raisonneurs automatiques d'inférer de nouvelles connaissances, de vérifier la cohérence et de classifier les concepts. Formellement, une ontologie peut être vue comme un quadruplet :

$$O = (C, P, I, A)$$

où C, P, I, et A désignent respectivement l'ensemble des classes (concepts), l'ensemble des propriétés, l'ensemble des individus et l'ensemble des axiomes logiques.

Pour standardiser ces besoins, le OWL [13] [103] a été développé. Basé sur les logiques descriptives, reprend les fondations / mais les enrichit considérablement, permettant de déclarer des classes équivalentes ou disjointes, et d'exprimer des restrictions. Par exemple la restriction « un parent est une personne qui a au moins un enfant » peut être formalisée comme suit :

$$\text{Parent} \equiv \text{Personne} \sqcap \exists \text{aPourEnfant.Personne}$$

on peut aussi spécifier des cardinalités minimales ou maximales, par exemple « une personne avec au moins deux enfants » :

$$\text{PersonneAvecDeuxEnfants} \equiv \text{Personne} \sqcap (\geq 2\ \text{aPourEnfant.Personne})$$

et définir des propriétés avec des caractéristiques logiques, telles que **transitives** (si A est lié à B et B est lié à C, alors A est lié à C), **fonctionnelles** (un sujet ne peut avoir qu'un seul objet pour cette propriété), **symétriques** (si A est lié à B, alors B est lié à A), **inverses** (si A est lié à B par cette propriété, alors B est lié à A par une propriété correspondante), etc :

$$\text{Transitif(ancestreDe)} \Longrightarrow (x\ \text{ancestreDe}\ y\ \wedge\ y\ \text{ancestreDe}\ z)\ \Rightarrow\ x\ \text{ancestreDe}\ z$$

Grâce à cela, OWL permet un raisonnement beaucoup plus riche, notamment à travers des raisonneurs automatiques. La contrepartie est que, plus l'ontologie est expressive, plus le raisonnement devient couteux, voire indécidable dans certains cas. Le logiciel libre utilisé ici pour simplifier l'écriture du XML avec une syntaxe OWL ainsi que pour pouvoir inférer le graphe est l'éditeur d'ontologies Protégé. On peut voir l'interface du logiciel ainsi que quelques statistiques de l'ontologie crée sur la figure 2.7.

Ces technologies du web sémantique constituent ainsi un pilier de l'IA symbolique. Elles incarnent l'approche classique de l'IA : découper le monde en entités atomiques et en relations formelles, imposer des structures et des règles claires, puis laisser un moteur logique inférer les conclusions qui en découlent. Formellement, ce principe de raisonnement peut s'écrire :

$$O\ \vDash\ \phi \Longleftrightarrow \text{l'axiome } \phi \text{ est déductible de } O$$

13. Web Ontology Langage

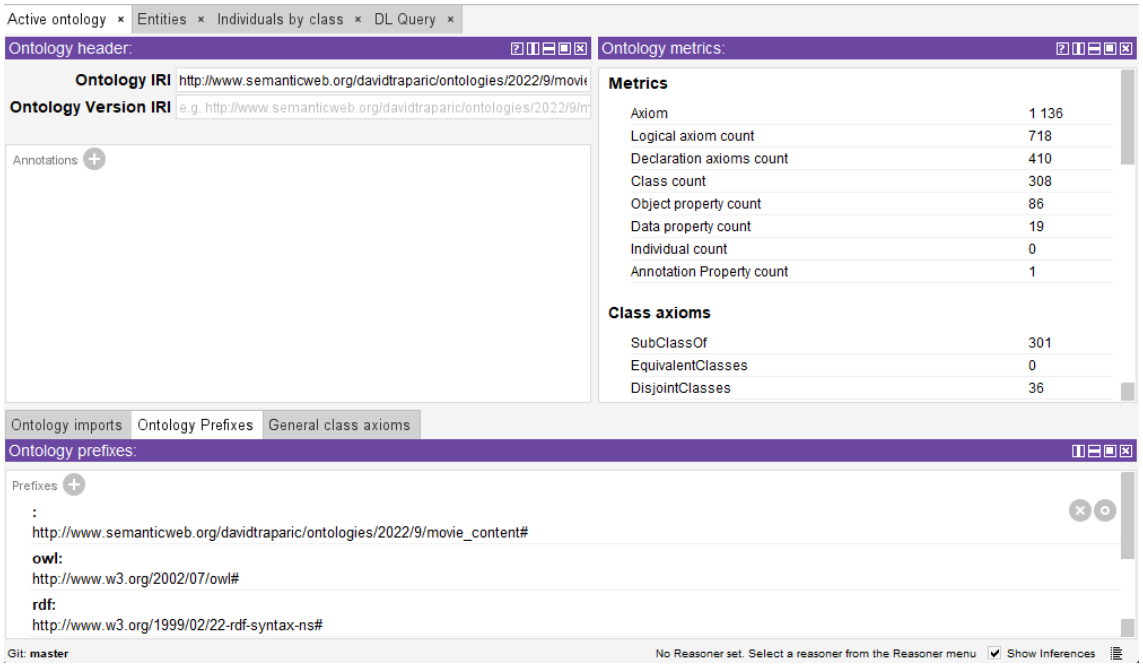


FIGURE 2.7 – **Interface graphique de Protégé**, le logiciel libre d'édition d'ontologie.

À l'opposé, les IA connexionnistes comme l'apprentissage profond ne partent pas d'une structuration imposée : elles ingèrent directement les données brutes (images, textes, vidéos) et optimisent des représentations internes de manière statistique. Leur fonctionnement repose sur la minimisation d'une fonction de perte :

$$\theta^* = \arg\min_{\theta} \mathcal{L}\big(f_\theta(X), Y\big)$$

avec θ les paramètres du modèle, θ^* les paramètres optimaux appris, $\mathcal{L}$ la fonction de cout ou fonction de perte, soit l'erreur à minimiser, $f_\theta(X)$ la sortie du modèle pour des données d'entrée X et Y la vérité terrain.

L'IA symbolique cherche à construire une chaine de raisonnement explicite et interprétable ; l'IA connexionniste trouve de manière implicite des paramètres atomiques qui, agrégés, produisent la vision globale. Ces deux approches ne s'excluent pas, mais traduisent deux visions complémentaires : la première fondée sur la logique et la connaissance explicite, la seconde sur l'apprentissage statistique et la généralisation.

2.3.4 Analyse filmique

L'analyse filmique est la capacité à interpréter l'implicite provoqué chez le spectateur. Comme on peut s'en douter, cette affirmation est quelque peu problématique en elle-même puisque la manière de ressentir tel ou tel effet à l'écran dépend fortement de la personne spectatrice, sa personnalité et son vécu. Pour autant, les réalisateurs de films se basent tout de même sur des consensus, certains bien mieux établis et universels (comme le plan débullé, visible sur la figure 2.8) que d'autres.

Bien qu'il existe des consensus, l'analyse filmique, faisant partie de l'interprétation **artistique**, constitue un domaine très subjectif et dépend de l'interprétation de chacun.



FIGURE 2.8 – **Exemples de plan débullé**, aussi appelé plan cassé, ou *dutch angle* en anglais. Le rôle du plan débullé fait largement consensus parmi les réalisateurs. Le spectateur se retrouvant en déséquilibre, il pourra percevoir dans la narration de la scène un sentiment de tension, d'angoisse ou d'instabilité. (Crédit : auteur [M13])

Il n'est pas rare de trouver des ressources cinématographiques différentes proposant des significations opposées pour un même évènement à l'écran, révélant ainsi qu'il existe certains consensus moins universels que d'autres. La majorité des concepts connus viennent du milieu de la photographie, qui viennent eux-mêmes assez communément de la peinture. Également, l'analyse filmique ne reposant pas sur des ressentis vérifiables ou étudiés, la recherche bibliographique peut se faire uniquement en recoupant plusieurs sources, majoritairement des sites internet pédagogiques ou de formations universitaires pour réaliser un film.

Échelle de plan

L'échelle de plan correspond au cadre plus ou moins resserré sur un personnage filmé par la caméra. On peut voir les différentes échelles de plan usuelles, employés par les réalisateurs sur la figure 2.9.

Le **plan général** montre l'entièreté du décor, dans le but de créer un contexte autour de l'action qui s'y déroule. Il vise à décrire un lieu, un paysage, une ville, ou tout autre décor. Des personnages peuvent y être intégrés mais ne sont pas obligatoires. Un décor est un ensemble d'éléments intégré dans un environnement. Le décor ne bouge pas lorsque le plan est statique. Si un élément du décor se met à bouger, l'élément peut alors être considéré comme un objet. Le décor n'est pas le sujet du plan et il n'attire pas le regard. Le décor n'est pas forcément à l'arrière-plan (c'est-à-dire qu'il peut être à la même distance de la caméra que le sujet, voire plus proche dans certains cas particuliers). Le décor est entier si aucun élément ne semble manquer ou être coupé du cadre. L'environnement, lui, est continu et infini et est forcément coupé à un endroit par le cadre. Si un personnage est intégré, il est très petit et seule sa silhouette est discernable.

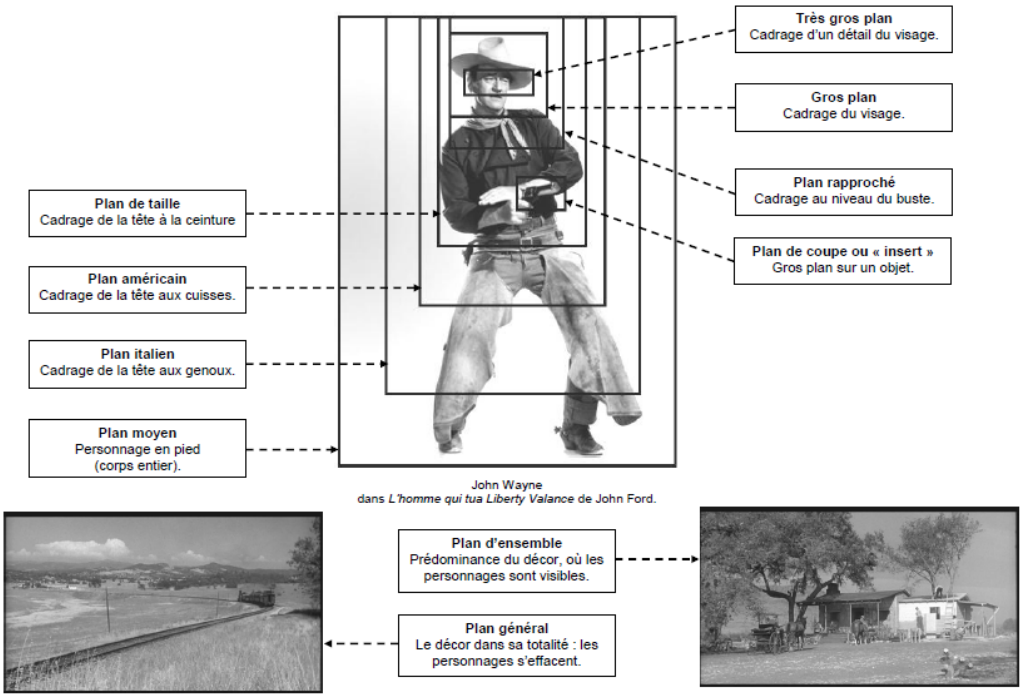


FIGURE 2.9 – **Schéma des différentes échelles de plan employés dans le cinéma**, de l'échelle la plus large, le plan général, à la plus étroite, le très gros plan.[M14]

Le **plan d'ensemble** se focalise sur un lieu tel qu'une place ou une rue en ville. Les personnages sont suffisamment visibles pour que le spectateur comprenne leur action, on décrit alors le contexte de l'action à échelle humaine. Des éléments du décor sont coupés, mais assez d'éléments sont montrés pour pouvoir contextualiser une scène, avec un (ou plusieurs) sujet. L'environnement du décor est visible.

Le **plan moyen**, ou **plan pied** cadre un ou plusieurs personnages en entier, et concentre l'attention sur eux. Le plan moyen sert à positionner l'action et les différents personnages de façon plus nette et significative que les plans larges. On peut aussi découvrir certains éléments du décor. Le spectateur se focalise ainsi sur les personnages et leurs actions. Les pieds du personnage sont visibles en bas du cadre, et la tête du plus haut personnage se trouve dans la partie supérieure de l'écran.

Le **plan américain** cadre les personnages à mi-cuisses. Il permet par exemple de montrer plusieurs personnages lors d'un dialogue et de montrer la posture des personnages : le jeu du personnage est clairement mis en avant. Le réalisateur n'a pas obligation de combler l'espace avec le décor ou un objet. La tête du plus haut personnage se trouve dans la partie supérieure de l'écran. Cette largeur de plan permet tout de même d'apercevoir l'environnement du décor, même si aucun élément distinct ou discernable ne s'y trouve.

Le **plan rapproché** cadre les personnages au niveau de la ceinture. Le personnage est en confidence, le spectateur profite des détails de son visage. On met l'accent sur ce qu'ils disent ou font, sans oublier leur corps et leur gestuelle. Seul le haut du corps d'un personnage est visible et le bas du cadre doit à minima inclure la poitrine (plan rapproché poitrine) et au plus, le haut de la taille, tout en laissant le haut de la tête visible.

Le **gros plan** isole le visage d'un personnage, le coupant généralement juste au-dessus

des épaules. Le regard du personnage est mis en avant et les spectateurs sont amenés à s'intéresser à ses pensées. Il permet de lire directement les émotions d'un personnage, ses réactions les plus intimes. Le fond est partiel et minoritaire, il ne permet plus de discerner l'environnement ou un élément du décor.

Le **très gros plan** montre un détail du visage (comme les yeux, la bouche) ou du corps (une cicatrice, un doigt). Il focalise l'attention du spectateur sur un détail particulier et a un but narratif. L'insert est un très gros plan particulier, qui se focalise sur un objet. Le fond est partiel et minoritaire voire inexistant, en particulier si le sujet est un détail du visage. Cette échelle de plan ne permet plus de discerner l'environnement ou un élément du décor.

Le cadrage

Le **cadrage** au cinéma désigne ce que le cinéaste capture durant la prise de vue. Il correspond au choix des limites de l'image, incluant les angles de prise de vue, l'échelle des plans ainsi que l'organisation des objets et des personnages dans le champ. Le cadrage regroupe ainsi le choix du format du cadre, les valeurs de plans et les angles de prise de vue.

Le **format**, défini par le ratio d'image, constitue un paramètre essentiel du cadrage. Si le format 2,35 :1 reste le plus couramment utilisé au cinéma, son choix influe directement sur la perception de l'espace et des émotions. Dans *Mommy* de Xavier Dolan (2014), par exemple, le format carré 1 :1 accentue l'isolement du personnage face à son environnement, tandis que l'élargissement ponctuel du cadre en 1,85 :1 accompagne visuellement son épanouissement, lui laissant plus d'espace et traduisant une harmonie retrouvée avec le monde extérieur.

La **règle des trois tiers** divise l'image en zones de tension visuelle formant **des lignes et des points de force**. Placer des éléments importants sur ces zones confère au cadrage une composition équilibrée et harmonieuse. Un sujet positionné sur un point de force produit généralement une image stable et agréable, alors qu'un cadrage centré ou décentré modifie la perception du spectateur.

La position du sujet est également importante dans le cadrage. Un sujet centré, bien qu'équilibré géométriquement, peut générer un sentiment de tension ou de malaise par l'effet de symétrie trop rigide qu'il impose. À l'inverse, un sujet **décadré**, placé sur le bord du cadre comme visible sur la figure 2.10, traduit souvent l'isolement, l'exclusion ou l'angoisse. Le rapport entre le sujet et son orientation dans l'espace est également crucial : il est généralement plus harmonieux de laisser de l'espace dans la direction de son regard ou de son mouvement, sous peine de provoquer une impression d'oppression.

En ce qui concerne l'**orientation du sujet**, un personnage **de dos** introduit une part de mystère en dissimulant ses émotions de la même manière que le profil caché vu sur la figure 2.10, créant une distance interprétative avec le spectateur. Cette position peut également servir à présenter un paysage à travers le regard du personnage, unissant ainsi la découverte visuelle du héros et celle du spectateur. Un cadrage **de trois quarts** permet de maintenir une expressivité lisible tout en conservant un léger retrait, souvent utilisé pour équilibrer présence et pudeur. Le **regard caméra**, quant à lui, rompt la continuité narrative : en fixant directement l'objectif, le comédien interpelle silencieusement le spectateur, créant un lien de complicité ou une mise en abîme du regard.



FIGURE 2.10 – **Illustration du décadrage.** Le personnage est fortement décadré, proche du bord de l'écran, faisant écho à l'angoisse et l'incertitude vécue par le personnage dans la narration. (Crédit : [M15])

Les **angles de vue** sont clairement porteur de sens, le **plan incliné** (ou *Dutch Angle*) consiste à pencher la caméra sur le côté, produisant un effet d'instabilité visuelle. Même en **plongée totale**, un sujet incliné peut renforcer cette sensation de déséquilibre. La **plongée**, où la caméra se situe au-dessus du sujet, place celui-ci en position de faiblesse ou d'enfermement, parfois jusqu'à la déshumanisation. Elle peut aussi être utilisée pour des plans d'ensemble ou pour dévoiler un décor, souvent accompagné d'un personnage vu **de dos**, découvrant la scène en même temps que le spectateur. La **plongée totale**, presque verticale, accentue encore cet écrasement visuel, évoquant tour à tour la punition ou la délivrance selon le contexte. À l'opposé, la **contre-plongée**, plaçant la caméra en dessous du sujet, confère à ce dernier une impression de domination, de puissance ou de grandeur. Elle peut également souligner un malaise ou une disproportion entre le personnage et son environnement. Enfin, un cadrage **plat**, c'est-à-dire à hauteur du regard, reste la norme par défaut, traduisant une neutralité de point de vue et une absence d'effet dramatique marqué.

Mouvement dans l'image et mouvement de caméra

Le **mouvement dans l'image**, par opposition au **mouvement de caméra**, désigne tout déplacement interne au cadre alors que la caméra reste statique. Il peut s'agir du mouvement d'un personnage ou d'un objet dans la scène. Lorsque la caméra est en mouvement, c'est le fond qui se déplace visuellement, à l'inverse d'un plan fixe où seul le sujet est net et où un éventuel flou affecte le fond. Il est possible de combiner mouvement dans l'image et mouvement de caméra : dans un plan de **suivi**, par exemple, le fond bouge avec un flou cinétique tandis que le sujet demeure stable. Dans d'autres cas, le fond et le sujet peuvent bouger de manière indépendante, avec des vitesses ou directions différentes, créant des effets dynamiques ou de tension visuelle. Le **travelling** correspond à un déplacement physique de la caméra pendant la prise de vue, que ce soit pour suivre un sujet, s'en rapprocher, s'en éloigner ou le contourner afin d'en révéler de nouveaux aspects. Il modifie le point de vue spatial et contribue à rendre la narration plus fluide et immersive. Sans travelling, l'enchainement de plans fixes peut paraître artificiel, le rythme monotone et l'évolution

narrative figée, car le spectateur, à travers l'œil de la caméra, n'a pas l'habitude d'une vision fragmentée par des coupes successives. Le travelling favorise donc une immersion naturelle.

Le **travelling horizontal** suit généralement un personnage ou un véhicule, le mouvement de la caméra épousant la vitesse et la direction du sujet. Lorsqu'il est perpendiculaire à l'axe de la caméra, tous les points du cadre se déplacent de manière uniforme de droite à gauche ou inversement. Un déplacement **gauche-droite** suggère souvent une évolution positive, tandis qu'un mouvement **droite-gauche** traduit symboliquement un recul, une résistance ou la présence d'un antagoniste. Le **travelling vertical** est souvent utilisé pour introduire une séquence en dévoilant progressivement un lieu ou une information essentielle, comme une inscription sur une pierre tombale. Le **travelling avant** permet d'adopter le point de vue du personnage ; en s'approchant de lui, la caméra révèle progressivement ses émotions, créant une montée en tension. À l'inverse, un **travelling arrière** peut signifier une prise de distance, un retrait émotionnel ou une révélation de contexte. Le **travelling circulaire**, mouvement autour d'un sujet, possède une forte valeur dramatique : il amplifie la tension ou symbolise la confusion mentale du personnage. Enfin, le **travelling compensé** — combinaison d'un zoom avant avec un travelling arrière, ou inversement — produit une distorsion visuelle qui évoque la perte d'équilibre, la panique ou la désorientation, comme dans *Vertigo* d'Hitchcock ou *La Haine* de Kassovitz, où le point de stabilité varie au fil du plan.

Le **panoramique** est un mouvement de rotation de la caméra autour de son axe, tandis que son pied reste fixe. Il permet de suivre un personnage (**panoramique d'accompagnement**) ou de révéler un décor (**panoramique descriptif**). Le mouvement peut être horizontal, vertical ou combiné, et les points de départ et d'arrivée du panoramique possèdent souvent une valeur narrative ou esthétique particulière.

Une **caméra à l'épaule** est tenue directement par le cadreur, produisant un mouvement organique et légèrement instable, proche de la perception humaine. Ce procédé renforce l'immersion du spectateur et permet un **point de vue subjectif**, aligné sur celui d'un personnage. Sans stabilisation type **steadicam**, les tremblements visibles peuvent traduire une situation d'action, de désordre émotionnel ou un réalisme brut. Avec steadicam, la stabilité est améliorée tout en conservant une liberté de mouvement humaine et spontanée. La caméra épaule, par sa hauteur et ses micro-mouvements, évoque directement la présence d'un corps filmant, renforçant ainsi l'authenticité de la scène.

Montée sur une grue, la **caméra-grue** peut s'élever ou descendre continuellement, permettant des plans en hauteur ou des **plongées** sans support fixe. Ce type de mouvement modifie la perception spatiale et offre des transitions spectaculaires entre différents niveaux du décor. Un plan de **travelling aérien** obtenu par grue permet d'embrasser la scène dans son ensemble tout en conservant la fluidité du déplacement.

Le **zoom optique** (ou travelling optique), résulte d'une variation continue de la distance focale de l'objectif. Il modifie la taille apparente du sujet sans altérer la perspective. Bien qu'il permette d'ajuster rapidement le cadrage, il est souvent critiqué au cinéma pour son effet artificiel. À l'inverse, le **zoom numérique**, simple recadrage d'une image enregistrée, dégrade la qualité et ne constitue pas un véritable mouvement cinématographique.

Image

Le **style chromatique** d'une image regroupe l'ensemble des caractéristiques visuelles liées à la couleur, à la lumière et à leur perception. La **saturation** distingue une teinte d'un gris de même luminosité. Une faible saturation tend vers le gris et rend l'image fade, tandis qu'une saturation élevée donne un rendu vif et éclatant. Elle influence directement la sensation de vitalité et d'intensité perçue. Le **contraste** mesure la différence de luminosité entre les zones claires et sombres, caractérisant la répartition lumineuse globale. Un contraste marqué accentue les formes et les volumes, alors qu'un contraste faible adoucit l'image. L'**exposition** désigne la quantité totale de lumière reçue par la surface sensible (pellicule ou capteur). Une image **surexposée** perd des détails dans les zones claires (« zones brûlées »), tandis qu'une **sous-exposée** les perd dans les ombres (« zones bouchées »). L'analyse de l'histogramme permet de détecter ces déséquilibres : un pic dans les hautes lumières signale une surexposition, et un pic dans les basses lumières une sous-exposition. L'exposition influence la luminosité globale de l'image. Certaines approches visuelles jouent volontairement sur ce paramètre : **l'éclairage franc** [14], qui supprime les ombres par un éclairage frontal, produit une atmosphère éthérée, souvent associée à des rêves ou à des états de transition, comme on peut le voir en figure 2.11. À l'inverse, **l'éclairage discret** aussi appelé **clair-obscur** [15], reposant sur la sous-exposition, immerge le spectateur dans une obscurité partielle où seuls certains éléments sont révélés. Ce procédé intensifie la tension, notamment dans les films d'horreur, en plaçant le spectateur dans la même incertitude perceptive que le personnage.

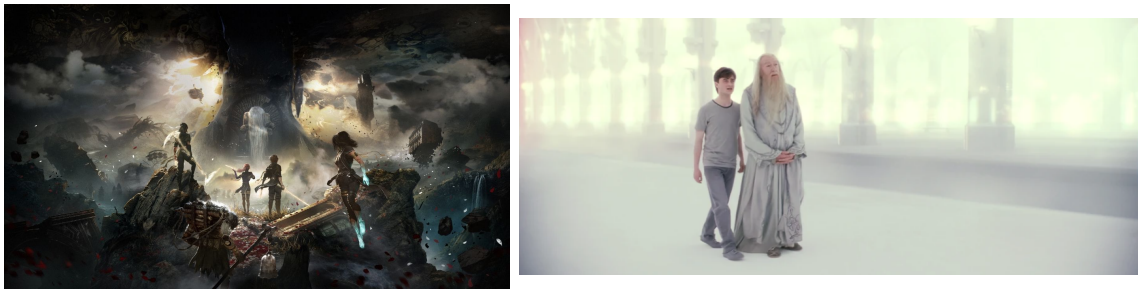


FIGURE 2.11 – **Exemples du clair-obscur et de l'éclairage discret.** La technique du clair-obscur à gauche est une forme d'éclairage discret en ce sens que les tons sombres sont dominants. Le clair-obscur ajoute à cela une mise en avant marquée du contraste entre les tons sombres et les tons clairs. Une scène en éclairage franc à droite, pour faire comprendre un espace entre la vie et la mort dans laquelle se trouve le personnage. Le blanc de la scène symbolise un espace immaculé de pureté et de neutralité absolue, ni angoissante ni accueillante. (Crédit : [M16])

Le **ton** ou la **teinte** désigne la couleur pure. Sur le cercle chromatique, les **tons froids** s'étendent du violet au vert clair, tandis que les **tons chauds** vont du jaune au rouge violacé. La **luminosité** correspond à la quantité de lumière émise ou perçue. Elle dépend de la **luminance**, mesurable sur les écrans selon divers standards où la luminance relative varie de 0 (pixel éteint) à 1 (pixel allumé). Bien que proportionnelle à cette mesure, la luminosité perçue reste subjective, influencée par la saturation et le contexte visuel environnant. Les grandeurs physiques associées incluent l'**intensité lumineuse** (candela), le **flux lumineux** (lumen) et l'**éclairement** (lux). L'intensité lumineuse caractérise la distribution directionnelle de la lumière émise, le flux lumineux mesure la quantité totale

14. ou *high-key lighting*
15. ou *low-key lighting*

émise, et l'éclairement quantifie la lumière reçue sur une surface. Ces paramètres, bien que mesurables, participent à la perception d'éclat ressentie par le spectateur.

Le **flou**, défini comme une perte de netteté, peut être volontairement introduit pour des raisons esthétiques ou narratives. Il permet de détacher un sujet, de traduire un mouvement ou d'exprimer un état psychologique. Sa mesure repose sur des critères mathématiques tels que la dérivée discrète, la variance d'intensité ou l'entropie, mais sa perception demeure largement qualitative. Le **flou cinétique**, causé par le mouvement du sujet pendant l'exposition, est directionnel et localisé. Il renforce l'impression de dynamisme : dans un plan de suivi, le fond devient flou tandis que le sujet reste net, alors que dans un plan fixe, c'est le sujet en mouvement qui se floute. Le **flou de bougé** résulte, quant à lui, du déplacement de la caméra pendant l'exposition. Si ce mouvement est perpendiculaire à l'axe optique, l'ensemble de l'image est flou ; s'il est aligné, les traînées de flou convergent vers le centre. En cas d'inclinaison, un **flou radial** apparait, créant des arcs circulaires autour du centre. Enfin, le **bokeh**, flou situé en dehors de la zone de netteté, découle des propriétés optiques de l'objectif. Il correspond à un filtrage de la lumière suivant une fonction d'étalement du point, très proche d'un **flou gaussien**, et confère à l'image une profondeur visuelle douce et esthétique.

Son

Le **son direct** correspond à un enregistrement effectué simultanément à la prise de vue, capturant la réalité sonore du tournage. À l'inverse, le **son postsynchronisé** est ajouté ou refait en studio après coup, afin d'améliorer la clarté des dialogues ou de maîtriser la qualité sonore. Ces deux approches traduisent des choix esthétiques et techniques qui influencent la perception de la scène.

Dans le **film muet**, seule l'image est enregistrée. Des bruitages ou musiques étaient alors produits en direct dans la salle de projection. Certains codes hérités de cette époque, comme l'intertitre, ont été repris dans des œuvres contemporaines telles que *The Artist*, pour évoquer la nostalgie ou la pureté du cinéma primitif.

Le **son *in*** (ou diégétique) désigne un son dont la source est visible à l'écran. Il regroupe la musique d'écran jouée dans la scène (par un personnage, une radio ou un instrument) ainsi que les bruits et les paroles produits par les éléments visibles. Ces sons renforcent la cohérence de l'espace diégétique et l'immersion du spectateur.

Le **son hors-champ** provient d'une source non visible mais appartenant au même espace-temps que la scène filmée. Il suggère la présence d'éléments extérieurs au cadre et participe à la continuité sonore, enrichissant la perception du hors-champ narratif.

Le **son *off***, quant à lui, émane d'un autre lieu ou d'un autre temps que celui de l'action. Il englobe la musique d'illustration, qui accompagne et colore l'ambiance, la musique de mimétisme, qui souligne les gestes et transitions, ainsi que la musique de contrepoint, qui crée un décalage expressif entre le son et l'image (par exemple une musique douce sur une scène violente). On y trouve aussi la musique dite « effet clip », utilisant des rythmes commerciaux ou des ralentis pour styliser la séquence. Le rythme et la tension de ces musiques influencent directement l'émotion et la dynamique du montage.

La **voix *off*** correspond à celle d'un narrateur ou d'un personnage absent du champ visuel. Elle guide la compréhension, commente ou ajoute une dimension introspective. La

voix intérieure, elle, exprime les pensées d'un personnage sans qu'il parle à voix haute, permettant d'accéder à sa subjectivité et à son état psychologique.

Le **point d'écoute** désigne la position d'où le spectateur perçoit le son. Il peut être réaliste, en phase avec la caméra, ou subjectif, lorsque le son reflète ce qu'entend un personnage. Ce **son subjectif** transporte ainsi le spectateur dans la perception auditive du protagoniste. Une variation du point d'écoute spatial permet parfois de dissocier l'audition de la vision, renforçant la tension ou l'immersion.

Les **paramètres acoustiques**, que sont le volume, le timbre et la hauteur, contribuent à l'identité sonore. Le volume traduit l'intensité, la hauteur dépend de la fréquence des vibrations, et le timbre définit la qualité propre d'un son, sa « couleur » perceptive (chaud, métallique, rauque, etc.). Le **décalage sonore** entre une source visuelle et le son perçu peut créer des effets de distanciation, d'ironie ou de malaise. Enfin, le **silence** agit comme un outil dramatique puissant : il peut susciter le suspense, l'attente ou l'émotion, tout comme le noir visuel suspend la narration par l'image.

Profondeur

La **profondeur** désigne la perception de la distance dans l'image, entre les éléments les plus proches et les plus éloignés de la caméra. D'un point de vue perceptif, elle est généralement plus marquée dans les grands plans que dans les gros plans ou très gros plans, car la distance entre les points extrêmes du champ visuel est plus importante dans un plan d'ensemble. Cette sensation de profondeur varie également en fonction de la **focale**, qui influence la manière dont l'espace est restitué à l'image.

La **perspective** constitue une transformation géométrique permettant de représenter un espace tridimensionnel sur une surface plane. Elle crée donc une illusion de profondeur, renforcée par le cadrage et la mise en scène. Les lignes de fuite, par exemple, sont les projections de droites parallèles de l'espace 3D convergeant vers un même point de fuite, ce qui donne au regard une direction et une sensation d'éloignement. Les droites perpendiculaires à l'axe optique de la caméra constituent une exception, car elles demeurent parallèles à l'image. Selon la complexité du point de vue, on distingue un point de fuite unique dans une vision frontale, deux points de fuite latéraux dans une vision oblique, et jusqu'à trois (deux latéraux et un vertical) dans une vision aérienne.

La **focale** détermine la largeur du champ couvert par la caméra : un objectif à courte focale offre un champ large, tandis qu'un objectif à longue focale restreint le cadre. Elle influe également sur la perception des distances : les espaces entre des objets situés à différentes profondeurs semblent plus grands avec une courte focale et plus comprimés avec une longue focale. Lors d'un zoom, la focale varie et s'allonge, tandis qu'elle reste constante lors d'un travelling.

La **profondeur de champ** correspond à l'étendue de la zone de netteté dans la profondeur du champ visuel. Elle dépend de plusieurs paramètres : plus la focale est courte, plus la profondeur de champ est grande ; plus le diaphragme est fermé, plus la zone nette s'étend ; enfin, bien que dans une moindre mesure, plus la distance de mise au point est importante, plus la profondeur de champ augmente. La transition entre zone nette et zone floue n'est pas brutale, mais progressive, formant un gradient de netteté. En jouant sur la mise au point, généralement effectuée sur le sujet principal, le réalisateur peut guider le regard du spectateur sans modifier la composition du plan, simplement en floutant certaines zones pour en mettre d'autres en valeur.

Composition

La **composition** désigne la manière dont les éléments présents dans le cadre sont agencés et mis en relation. Elle repose sur la disposition des éléments situés au premier plan, au second plan et à l'arrière-plan, et contribue à la lecture visuelle et symbolique d'une image. Elle établit un équilibre entre ces différentes zones, tout en guidant le regard du spectateur.

Les **lignes directrices** sont des éléments graphiques qui structurent le cadre et orientent le regard à travers le plan. Les **lignes horizontales** transmettent une impression de stabilité, de calme et de sérénité. Les **lignes verticales**, quant à elles, apportent de la force et peuvent, selon le contexte, accentuer un sentiment de tension ou d'agressivité. Les **lignes diagonales** introduisent un fort dynamisme, guidant le regard dans la profondeur ou le mouvement de l'image. Enfin, les **lignes courbes** évoquent la douceur, la fluidité et la rondeur. Lorsque ces lignes, qu'elles soient horizontales ou verticales, apparaissent devant un personnage, elles peuvent suggérer une sensation d'enfermement, comme un cadre dans le cadre évoquant les barreaux d'une prison.

Les **formes géométriques** présentes ou suggérées dans la composition renforcent la signification visuelle d'un plan. Le **cercle** évoque la douceur, la tendresse et la bienveillance. Le **triangle** suggère le danger, la rapidité ou le tranchant, tandis que le **carré** symbolise la stabilité, la solidité et la robustesse. Ces formes participent à la construction émotionnelle du cadre.

Les **symétries** jouent un rôle clé dans la perception d'une image. Elles peuvent instaurer une impression d'harmonie et de stabilité, ordonner une image visuellement encombrée en la rendant plus lisible, ou encore conférer à un lieu ou un personnage une dimension imposante et intimidante. La symétrie est ainsi un outil puissant pour équilibrer ou amplifier l'effet dramatique d'une scène.

Les **techniques de cadres saturés** consistent à remplir le cadre d'un grand nombre d'éléments visuels pour créer un effet d'intensité ou de surcharge sensorielle. Cela peut se faire par la quantité d'objets visibles, comme dans un paysage dense, ou par des procédés techniques tels que le **split screen**, qui juxtapose plusieurs images pour créer un effet de tension ou de décalage entre les actions. La **transparence**, héritée du collage et aujourd'hui assimilée au fond vert, permet d'ajouter un décor ou un arrière-plan derrière les personnages, comme dans *Vertigo* de Hitchcock. La **surimpression**, quant à elle, consiste à superposer plusieurs images pour renforcer l'intensité ou la symbolique d'une scène.

À l'inverse, les **techniques de cadres raréfiés** reposent sur l'économie visuelle : l'attention est concentrée sur un seul élément, un espace peut être laissé vide pour créer une respiration, ou un **hors-champ** important peut être suggéré, donnant au spectateur la responsabilité d'imaginer les éléments absents mais liés à la scène.

Le **cadre dans le cadre**, ou surcadrage, consiste à inclure des bordures internes dans l'image, comme une porte, une fenêtre ou un miroir. Ce procédé peut isoler un personnage en renforçant son enfermement ou sa solitude, orienter la lecture du spectateur vers une zone spécifique du plan, ou encore créer une immersion émotionnelle plus forte en intégrant le spectateur dans la situation du cadreur (comme le montre la figure 2.12).

La **scène** est composé de plusieurs **plans cinématographiques**. Elle définit le **lieu**, intérieur ou extérieur, les **décors** mis en place, que cela implique des objets, des meubles ou autres, le **moment de la journée**, comme le jour ou la nuit, et le nombre de personnage



FIGURE 2.12 – **Exemple de cadre dans le cadre** où l'attention du spectateur est portée sur le personnage, suivi par la fenêtre. Ce plan surprend le spectateur qui se retrouve pris en joue puis blessé directement par le personnage, ce qui a pour effet d'impliquer le spectateur lui-même dans la diégèse, et d'augmenter son immersion.

dans la scène, et pas seulement dans le cadre, mais tout ceux vu dans toute la longueur de la scène, et ceux qui sont hors-champ.

Lumière

Dans la **lumière au cinéma**, la direction et la qualité de l'éclairage guident le regard du spectateur en hiérarchisant les éléments de l'image. Le cinéma classique hollywoodien a établi le système d'**éclairage à trois points**, composé d'une lumière principale, d'une lumière diffuse et d'une lumière de contour.

La **lumière dure** provient d'une source ponctuelle, comparable à un projecteur Fresnel. Elle produit une image très contrastée, aux contours nets et aux ombres denses. Cette lumière directionnelle crée peu de zones intermédiaires, renforçant la netteté et la tension visuelle.

À l'inverse, la **lumière diffuse** émane d'une source plus grande que le sujet. Elle enveloppe l'ensemble de la scène, réduisant les contrastes et adoucissant les formes. Utilisée souvent comme lumière secondaire, elle sert à déboucher les ombres créées par la source principale, apportant une sensation de douceur et de naturel.

La **lumière par le haut** éclaire le sujet depuis une position supérieure, générant des ombres courtes mais denses. Elle fait briller les yeux d'un éclat caractéristique tout en gommant les détails du visage, ce qui confère une dimension dramatique ou mystique au personnage.

La **lumière de face**, placée proche de l'objectif, efface les reliefs et écrase les perspectives. Elle crée une apparence uniforme et un teint laiteux, souvent employée pour adoucir les traits d'un personnage.

La **lumière à 45°** correspond à l'angle d'incidence du soleil et permet de révéler la structure du visage avec précision. Ce type d'éclairage, à la fois naturel et valorisant, offre un rendu équilibré entre contraste et lisibilité.

La **lumière latérale**, venant du côté, accentue le contraste et dramatise la scène. En l'absence d'une source complémentaire, elle plonge une moitié du visage dans l'ombre, évoquant la dualité, le secret ou le conflit intérieur.

Le **contrejour** place la source lumineuse derrière le sujet. Il crée une séparation entre le sujet et le fond, parfois jusqu'à transformer le personnage en silhouette. Selon le traitement, on peut adoucir cet effet avec une lumière d'appoint ou accentuer le mystère en conservant un visage presque noir.

Enfin, la **lumière par dessous** inverse les ombres naturelles, donnant au sujet une apparence étrange ou inquiétante, souvent utilisée pour exprimer le surnaturel ou la folie.

La **lumière de contour**, petite source placée en léger contrejour, dessine les bords du sujet pour le détacher du décor. Sans elle, le sujet se fond dans l'arrière-plan, perdant en relief et en importance. Cela peut être volontairement exploité pour signifier la discrétion ou la dissolution du personnage dans son environnement.

Montage

Le **montage**, englobe le séquencement des plans, l'unité de temps d'un film et sa dissection, le rythme, les raccords, les transitions ainsi que la variation de la temporalité.

L'art du **séquencement des plans** consiste à assembler les prises les unes à la suite des autres afin de construire le rythme et la narration du film. Le **montage bout-à-bout** représente souvent la première version du film, sans retouches ni raccords, suivant strictement l'ordre du scénario. Le **montage chronologique** respecte la temporalité de l'histoire, tandis que le **montage alterné** juxtapose plusieurs actions simultanées ou consécutives, créant du dynamisme (comme dans une poursuite entre poursuivant et poursuivi). Le **montage parallèle**, lui, alterne des actions sans lien temporel, mais reliées par leur thème. Certains procédés particuliers, comme le **montage par adjonction d'images**, visent à provoquer des associations d'idées pour traduire des émotions. Le **montage par leitmotiv**, inspiré de la musique, repose sur la récurrence d'un thème visuel ou sonore, parfois très rapide, voire subliminal.

L'**unité de temps d'un film** s'articule autour de plusieurs niveaux : le **plan** est une prise continue, la **scène** regroupe les actions se déroulant dans un même lieu et une même temporalité, et la **séquence** rassemble plusieurs scènes formant un évènement narratif cohérent. Une ellipse ou une transition (comme un fondu) signale généralement un changement de séquence.

Le **rythme** du film nait du montage. Il dépend de la durée des plans, des mouvements de caméra et parfois du tempo musical. Des plans rapides et une musique en accord avec l'action accélèrent la perception, tandis qu'un montage lent ou une transition prolongée ralentissent le récit.

Les **raccords** assurent la continuité entre deux plans. Le **raccord cut** juxtapose deux prises successives d'un même plan sous un angle différent (au moins 30° d'écart pour éviter la saute visuelle). Le raccord **dans l'axe** conserve la même orientation mais modifie la valeur du plan : resserrer renforce l'intensité, élargir contextualise. Le **raccord d'entrée/sortie de champ** relie deux plans par le mouvement d'un personnage quittant ou entrant dans le cadre. Le **raccord dans le mouvement** généralise ce principe à tout geste ou déplacement continu. Le **raccord de regard** établit un lien logique entre le regard

d'un personnage et l'objet ou la personne observée. Le **champ/contre-champ**, utilisé notamment dans les dialogues, respecte la règle des 180°, sous peine de faux raccord. D'autres formes incluent le **raccord panoramique**, qui enchaine deux scènes par un mouvement continu, ou le **raccord par analogie**, basé sur une ressemblance visuelle ou symbolique entre deux images. Un **faux raccord** est un type de raccord assez particulier, car il rompt la continuité. Il peut s'agir d'une règle volontairement transgressée (faux raccord de mouvement, de regard, ou de 180°), ou d'une incohérence matérielle entre deux plans. Le **jump cut**, quant à lui est un faux raccord particulier. Il juxtapose deux plans presque identiques pour produire un effet de discontinuité et de saut brusque entre deux images, il souvent utilisé pour accélérer une action ou créer une tension visuelle, mais il a aussi déjà utilisé à des effets comiques.

Les **transitions** marquent un changement de séquence ou de temporalité. Le **cut** assure un enchainement direct, tandis que le **fondu enchainé** fait se chevaucher deux images. La **fermeture/ouverture au noir** ou à l'**iris** crée une pause narrative ou attire l'attention sur un élément. Le **volet** balaie l'image d'une direction à une autre, marquant une rupture nette. Ces procédés contrôlent le rythme global du film, le cut l'accélérant, les autres le ralentissant.

La **variation de temporalité** regroupe les effets d'ellipse, de *flashback* et de *flashforward*. L'**ellipse** omet une action jugée inutile pour la narration, accélérant le récit. Le ***flashback*** montre une action passée, tandis que le ***flashforward*** anticipe un évènement futur, souvent pour créer du suspense. Les effets spéciaux tels que l'**accéléré** (ou *time-lapse*) et le **ralenti** (ou *slow motion*) modifient la perception du temps : le premier condense l'action, le second l'étire pour souligner un moment poétique, dramatique ou marquant. Enfin, l'**arrêt sur image** fige le temps, soulignant un instant clé, une fin de récit, ou la prise d'une photographie dans la diégèse.

Nous avons tout d'abord défini le concept d'ontologie ainsi que les différents prédicats nécessaires à son élaboration. Nous avons ensuite présenté de manière détaillée les fondements de l'analyse filmique, en exposant l'ensemble des règles et conventions qui la structurent. À partir de ces éléments, nous pouvons désormais formaliser l'analyse filmique réalisée sous la forme d'un diagramme de classes, afin de rendre visibles les interactions entre les différents concepts et critères identifiés. La notation visuelle, ou syntaxe graphique, que nous proposons s'illustre de la manière suivante sur la figure 2.13. En analyse filmique, dans le domaine du **cadrage**, le critère *« Espace devant le personnage »* a une influence certaine sur le critère *« Effet à l'écran »* du domaine de la **réponse émotionnelle**. Selon la présence ou non de cet espace, cela peut faire transparaitre soit une harmonie, soit une oppression à l'image.

Le code d'ontologie OWL qui traduit ce graphe sémantique est présenté ci-dessous :

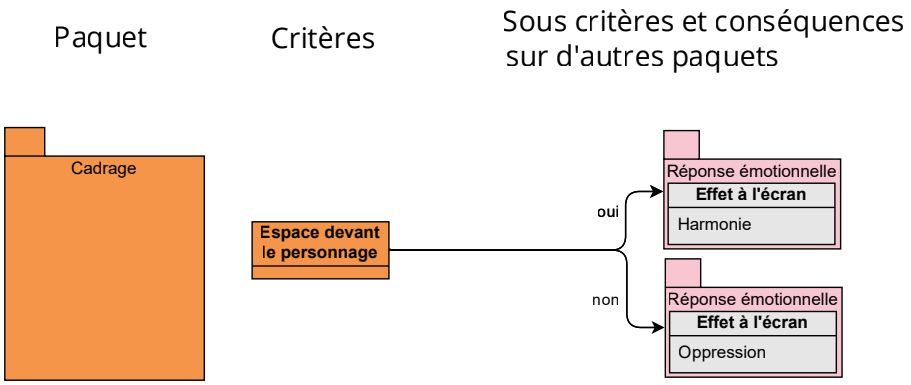


FIGURE 2.13 – **Fonctionnement de base de notre ontologie représentée sous la forme d'un diagramme de classe.** Le domaine du cadrage a plusieurs critères dont un montré ici : l'espace devant le personnage. Ce critère peut avoir soit pour conséquence l'harmonie de la scène et du personnage à l'écran, soit son oppression, qui sont deux critères du domaine de la réponse émotionnelle. (Crédit : [M17])

```
<!-- http://...#signifie_espaceDevantPersonnage -->
<owl:ObjectProperty rdf:about="http://...#signifie_espaceDevantPersonnage">
    <rdfs:subPropertyOf rdf:resource="http://...#signifie"/>
    <rdfs:domain rdf:resource="http://...#EspaceDevantLePersonnage"/>
    <rdfs:range rdf:resource="http://...#Harmonie"/>
    <rdfs:range rdf:resource="http://...#Oppression"/>
</owl:ObjectProperty>

<!-- http://...#EspaceDevantLePersonnage -->
<owl:Class rdf:about="http://...#EspaceDevantLePersonnage">
    <rdfs:subClassOf rdf:resource="http://...#CaractéristiqueDeCadrage"/>
</owl:Class>
```

LISTING 1 – Extrait de code XML RDF/OWL correspondant à la formalisation du graphe de la figure 2.13. Les trois points de suspension correspondent au chemin URI de l'ontologie.

Certains critères n'impliquent pas forcément une influence sur un autre critère, comme sur la figure 2.14 où ce type de conséquence est mise en valeur par une flèche en pointillés. Ainsi, ce critère est un indice de plus, mais il est être préférable de le recouper avec d'autres critères allant dans le même sens afin d'en conclure des effets résultants.

Le code d'ontologie OWL qui traduit le graphe sémantique de la figure 2 est présenté ci-dessous :

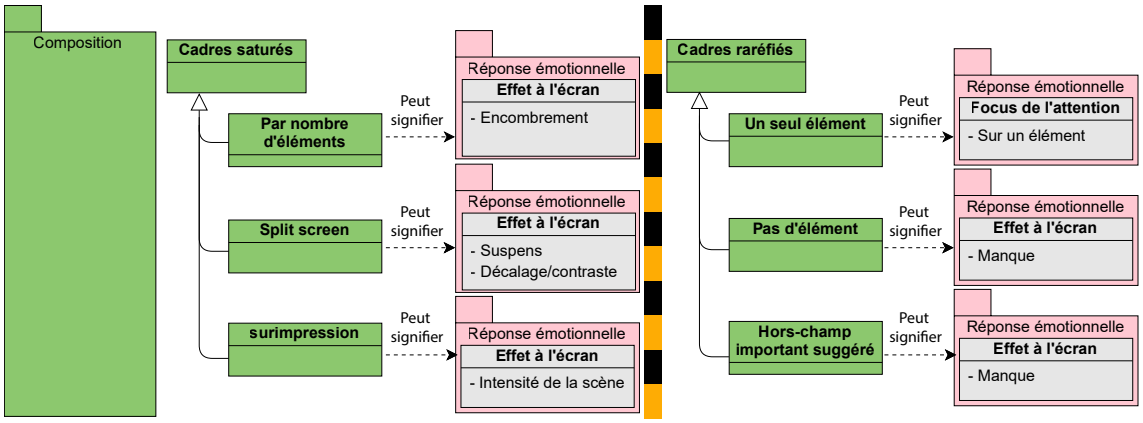


FIGURE 2.14 – **Illustration des significations potentielles mais non certaines dans le diagramme de classe.** Les différentes natures de cadres saturés et de cadres raréfiés peuvent avoir chacune des significations différentes sur le domaine de la réponse émotionnelle, mais il est plus avisé de les recouper avec d'autres critères allant dans le même sens.

```
<owl:Class rdf:about="http://...#CadreSaturé">
    <rdfs:subClassOf rdf:resource="http://...#CaractéristiqueDeComposition"/>
</owl:Class>
<owl:Class rdf:about="http://...#Splitscreen">
    <rdfs:subClassOf rdf:resource="http://...#CadreSaturé"/>
</owl:Class>
<owl:Class rdf:about="http://...#Surimpression">
    <rdfs:subClassOf rdf:resource="http://...#CadreSaturé"/>
</owl:Class>
<owl:Class rdf:about="http://...#ÉlémentsNombreux">
    <rdfs:subClassOf rdf:resource="http://...#CadreSaturé"/>
</owl:Class>
<owl:ObjectProperty rdf:about="http://...#peutSignifier_nbelmt">
    <rdfs:subPropertyOf rdf:resource="http://...#peutSignifier"/>
    <rdfs:domain rdf:resource="http://...#ÉlémentsNombreux"/>
    <rdfs:range rdf:resource="http://...#Encombrement"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:about="http://...#peutSignifier_split">
    <rdfs:subPropertyOf rdf:resource="http://...#peutSignifier"/>
    <rdfs:domain rdf:resource="http://...#Splitscreen"/>
    <rdfs:range rdf:resource="http://...#ContrasteDécalage"/>
    <rdfs:range rdf:resource="http://...#Suspense"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:about="http://...#peutSignifier_surimpr">
    <rdfs:subPropertyOf rdf:resource="http://...#peutSignifier"/>
    <rdfs:domain rdf:resource="http://...#Surimpression"/>
    <rdfs:range rdf:resource="http://...#IntensitéDeLaScène"/>
</owl:ObjectProperty>
```

LISTING 2 – Extrait de code XML RDF/OWL correspondant à la formalisation du graphe des cadres saturés. Les trois points de suspension correspondent au chemin URI de l'ontologie.

L'héritage d'un critère à l'autre est également représenté. Sur la figure 2.15, on peut voir le critère *Cercle*, *Triangle* et *Carré*, qui héritent tous du critère *Forme Géométrique*. Ainsi, toute forme géométrique a une incidence sur le critère *Cadre dans le cadre* du même domaine **Composition**, et chaque forme distincte a aussi son propre sens sur des critères du domaine de la **Réponse Émotionnelle**. On peut aussi remarquer que ces relations et

incidences peuvent concerner différents critères d'un autre domaine, comme ici le critère *Effet à l'écran* ainsi que le critère *Sensation*.

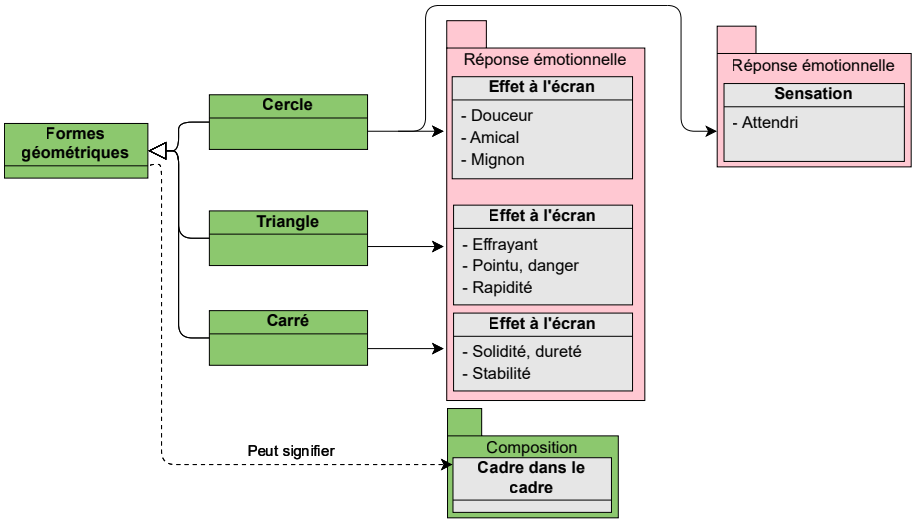


FIGURE 2.15 – **Illustrations de l'héritage en diagramme de classe.** La présence de cercle, de triangle ou de carré, sont toutes des classes héritantes de la classe *forme géométrique*. La présence d'une forme géométrique quelconque a une incidence, celle d'un cadre dans le cadre potentiel, et la présence d'une forme géométrique particulière définit d'autres implications.

Toutefois, on peut aussi tomber sur des critères qui doivent avoir lieu en même temps que certains autres critères précis pour avoir une signification sur un autre domaine. Par exemple, le critère du *cadre dans le cadre* en **composition** amène *de facto* l'attention du spectateur à l'intérieur du surcadrage et le sensation d'implication dans la scène, mais s'il est combiné à un personnage, cela peut créer l'effet à l'écran de l'isolement d'un personnage et d'un enfermement. En revanche, comme on peut le voir sur la figure 2.16 si le personnage est de dos, cela peut amener le spectateur à s'identifier à ce personnage et à découvrir en même temps que lui un paysage nouveau qui se dévoile.

Même la combinaison de plusieurs effets peut avoir des subtilités à prendre en compte pour considérer une incidence. Par exemple, sur la figure 2.17 on peut voir que là où un plan statique avec un sujet en mouvement va impliquer un flou cinétique sur le sujet, une caméra en mouvement combiné à un sujet en mouvement *lorsque celui-ci est suivi par la caméra* et uniquement dans ce cas-là, la situation va impliquer un flou de bougé sur le fond mais pas de flou cinétique sur le sujet.

Les nœuds de graphe sémantique montrés dans cette section sont une fine partie des graphes complets qui ont été faits à partir de toute l'analyse filmique présentée plus haut. Ces derniers sont représentés en annexe.

Ces graphes sémantiques et l'ontologie ont été faits en étroite collaboration avec les collaborateurs de CGR commanditaires de cette ontologie, qui ont approuvé les deux produits finaux, puis ont démissionné dans le mois. Un aperçu via le logiciel Protégé est disponible sur la figure 2.18.

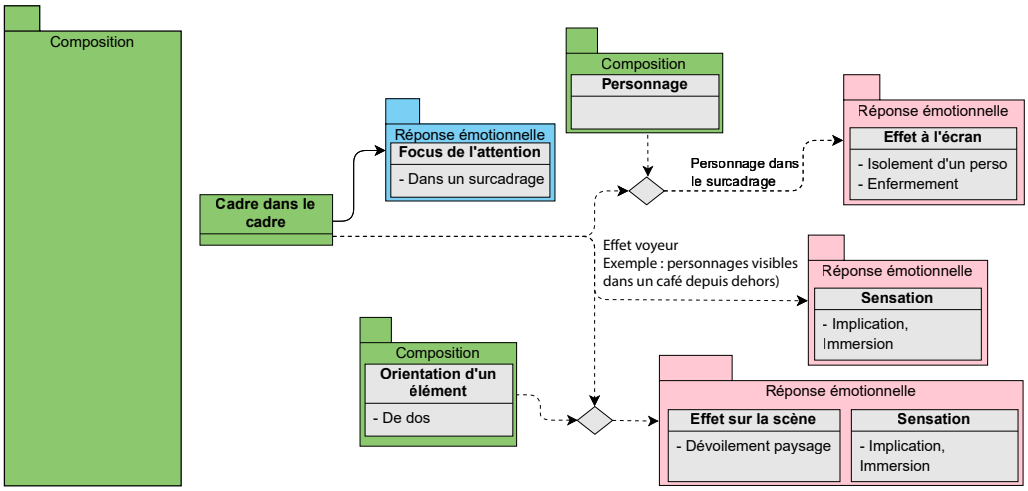


FIGURE 2.16 – **Illustration de la combinaison de critères en diagramme de classe.** Le cadre dans le cadre, seulement s'il est combiné avec un personnage de dos par le symbole losange gris, peut créer un effet sur la scène de dévoilement de paysage et de découverte, créant ainsi une association entre le personnage et le spectateur qui découvrent en même temps. À l'inverse, un personnage dans ce surcadrage peut signifier l'isolement et l'enfermement du personnage.

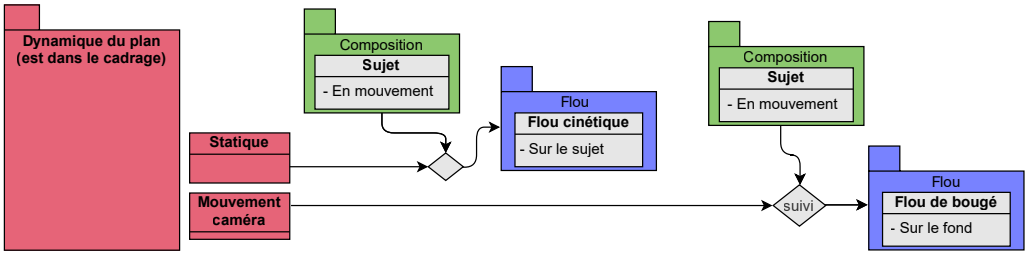


FIGURE 2.17 – **Illustration d'une association de critères sous conditions en diagramme de classe.** Un mot peut être inscrit à l'intérieur d'une combinaison de plusieurs critères afin d'ajouter une condition à cette association, comme ici ou le mouvement de caméra et un sujet en mouvement, s'il est suivi par la caméra, va engendrer un flou de bougé sur le fond.

2.4 Retours d'expérience sur la création d'une ontologie dans le cadre de ICE

Dans cette section, nous souhaitons partager notre retour d'expérience sur la construction d'une ontologie de domaine dans le cadre des contenus ICE et son exploitation. Afin de mener cette discussion, rappelons les objectifs principaux assignés à une ontologie, qui peuvent varier selon les contextes d'application. Dans un cadre industriel, les systèmes basés sur des ontologies peuvent, en théorie, remplir plusieurs fonctions :

1. Automatiser le raisonnement en substituant le jugement humain ;
2. Guider l'apprentissage machine en imposant un cadre sémantique ;
3. Rendre les modèles explicables grâce à la transparence de leurs règles symboliques ;
4. Aider à la recherche en structurant les connaissances.

Certes, la présence d'une ontologie pour les contenus ICE constitue un apport non négligeable pour comprendre le domaine filmique ainsi que le vocabulaire employé par les

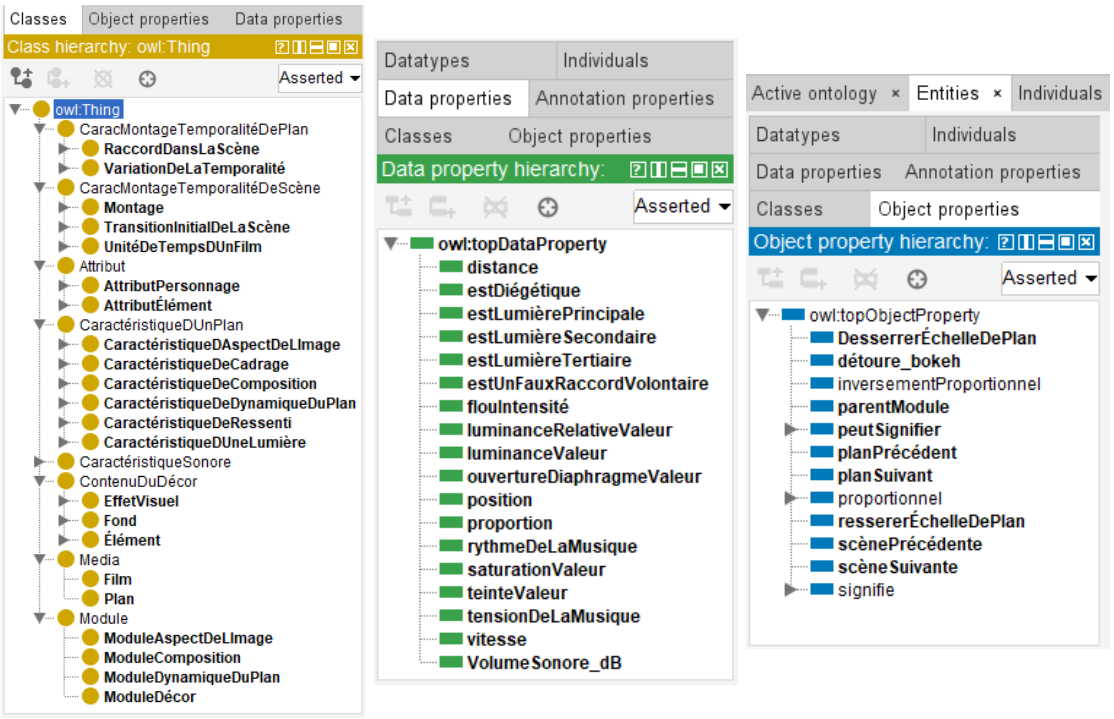


FIGURE 2.18 – **Aperçu des premiers niveaux des différentes entités de l'ontologie sur Protégé**, c'est-à-dire les classes, les propriétés de données et les propriétés d'objets

graphistes. Cependant, certains objectifs se heurtent à des contraintes importantes.

— Tout d'abord, l'automatisation complète du raisonnement apparait comme irréaliste. Pour inférer une décision à partir de l'ontologie, il est nécessaire d'instancier, pour chaque plan du film, les 307 concepts ontologiques, chacun devant être renseigné manuellement par un opérateur. Une telle opération, à répéter pour plusieurs centaines de plans, rend le processus extrêmement lent, même comparé à un graphiste ICE débutant. L'ontologie ne constitue donc pas une automatisation effective : elle déplace simplement la charge cognitive vers une étape de saisie manuelle inenvisageable à grande échelle.

  À l'inverse, les IA fondées sur le mécanisme d'Attention (comme les LLM [16] telles que ChatGPT) offrent une approche bien plus adaptée à ce cas, car ces modèles sont capables de hiérarchiser l'importance des informations selon le contexte, sans éliminer systématiquement les données secondaires. Ainsi, certaines caractéristiques peuvent être prises en compte uniquement dans des cas particuliers. Cette flexibilité de l'Attention contraste avec la rigidité des systèmes symboliques, pour lesquels chaque donnée supplémentaire introduit une instanciation supplémentaire systématique à réaliser par un opérateur.

— L'ontologie pourrait également jouer un rôle de guide pour l'apprentissage profond, en apportant une structuration conceptuelle ou en imposant des contraintes sémantiques aux modèles. Toutefois, cette idée suppose l'existence d'un jeu de données annoté selon l'ontologie. Cela supposerait dans le cadre d'un film que le réalisateur fournisse pour chaque plan les intentions et émotions visées, ce qui est, là encore,

16. Large Langage Model

irréaliste. Même une annotation approximative demanderait un effort considérable pour un bénéfice très limité, compte tenu du nombre de plans et de la complexité des critères impliqués.
— Les ontologies présentent également un intérêt du point de vue de l'explicabilité : les relations qu'elles décrivent étant entièrement définies par l'humain, elles permettent une interprétation directe des raisonnements produits. Cependant, la difficulté à formaliser la procédure cognitive des graphistes ICE réduit cet avantage. Il apparait plus pertinent d'utiliser des modèles de réseaux de neurones explicables, tels que Grad-CAM, capables d'indiquer les régions de l'image ou les caractéristiques ayant le plus influencé une décision donnée.
— Si l'ontologie pourrait effectivement servir de support à de futures recherches, son instanciation demeure incertaine en raison de la subjectivité des concepts. Les critères esthétiques, perceptifs ou émotionnels mobilisés par un réalisateur relèvent d'une logique floue [104] : leurs valeurs sont continues, contextuelles et difficilement délimitées. Il devient alors impossible d'obtenir une vérité terrain fiable permettant de valider les inférences produites.

En résumé, si les ontologies constituent un cadre rigoureux pour modéliser et organiser la connaissance, leurs contraintes pratiques et théoriques en limitent l'usage dans le contexte ICE. Ces constats motivent le passage vers des approches d'apprentissage profond, plus souples face à la subjectivité et mieux adaptées à la complexité perceptive du raisonnement artistique.

Compte tenu de ces difficultés, l'entreprise a écarté la piste de développement et d'exploitation d'une telle ontologie.

2.5 Conclusion

Ce chapitre a d'abord présenté la procédure de travail des graphistes ICE et leur méthode de création d'un film en format immersif latéral. Nous avons détaillé leurs outils, leurs marges d'action, ainsi que la logique sous-jacente à leurs choix artistiques. Il a aussi exposé l'analyse filmique : nous avons fait état de tous les éléments consensuels connus chez les réalisateurs concernant les procédés filmiques et les sensations visées par le spectateur. Dans un troisème temps, nous avons introduit la notion d'ontologie et des graphes de raisonnement associés. Enfin, un diagramme de classe complet modélisant l'analyse filmique précédente a été réalisé, dont une partie a été expliquée dans ce chapitre.

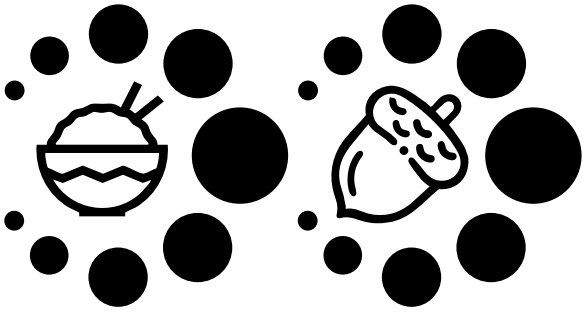
L'ontologie offre une structuration rigoureuse des connaissances en reliant des entités hétérogènes au moyen de relations formalisées. Elle a permis des avancées majeures dans l'indexation, le partage et la mise en cohérence des données, posant ainsi les fondations des IA symboliques qui ont été très utiles en tâches de référencements et de recommandations avant l'apparition des IA connexionnistes de l'apprentissage automatique. Cependant, son efficacité dépend fortement du formalisme imposé par l'humain : la qualité du raisonnement inférentiel repose sur la précision des concepts initiaux ainsi que de leurs relations. Cette contrainte rend l'approche inadaptée à des domaines où la subjectivité, la continuité des valeurs ainsi que l'ambiguïté des critères jouent un rôle central, comme dans le raisonnement ICE. Dans notre contexte, l'instanciation manuelle des 307 concepts identifiés dans le diagramme de classes, pour un seul plan cinématographique (sur les quelques 8100 plans d'un film) rend tout processus d'inférence très peu réaliste. Par ailleurs l'ontologie ne peut pas modéliser un opérateur de « et logique », afin d'identifier plusieurs concepts

présents simultanément actifs. C'est pourquoi les décisions ICE reposant souvent sur des critères flous et contextuels, parfois issus de l'occurence conjointe de plusieurs facteurs perceptifs, ne peuvent pas être correctement modélisées et interprétées par l'ontologie, limitant ainsi sa capacité à représenter la complexité réelle du raisonnement ICE et la richesse des interactions perceptives impliquées.

Ces limites conduisent naturellement vers l'apprentissage profond, qui marque une rupture avec l'approche symbolique. Contrairement aux ontologies, les modèles neuronaux s'affranchissent du formalisme explicite en apprenant directement les régularités à partir des données, même lorsqu'elles sont subjectives ou difficilement formalisables. Les modèles convolutifs présentés au Chapitre 3 et ceux à base d'attention au Chapitre 3.7 permettent de dépasser les contraintes liées à la mise en place et à l'instanciation d'une ontologie, tout en conservant des possibilités d'explicabilité. En somme, l'apprentissage profond s'impose comme la suite naturelle de cette étude : il offre la possibilité de modéliser les choix artistiques et perceptifs des graphistes ICE sans exiger de formalisation exhaustive, tout en maintenant une certaine interprétabilité des décisions du modèle. Par ailleurs, l'analyse approfondie de la réflexion ICE et de l'analyse filmique menées dans ce chapitre fournissent désormais une base solide pour identifier les données les plus pertinentes à intégrer en entrée de ces prochains modèles d'apprentissage profond.

CHAPITRE 3

R-ICE ET ICE-AGE : PRÉDICTION DU NIVEAU DE DÉTAIL ADÉQUAT EN POSTPRODUCTION



Ce ne sont pas les personnes bien payées qui sont réellement importantes à la société.

Kento Nanami

Table des matières

3.1 Introduction 105
3.2 État de l'art sur l'analyse de vidéo 105
La classification d'action 106
Le résumé de vidéos 106
Le Sous-titrage vidéo 107
La détection d'actions 108
La génération de propositions temporelles d'action (TAPG) 108
Localisation d'action temporelle (TAL) 108
Localisation spatiotemporelle d'action (STAL) 109
« Compréhension de tâches humaines » et autres sous-genres 109
3.3 Problématique et objectifs du projet 110
3.4 Gestion des données 110
3.4.1 Constitution de la base de données 113
3.4.2 Méthodologie de stockage des données 115
3.5 Méthodologie et appropriation d'architectures existantes 116

3.5.1 Prétraitement des données . . . 117
3.5.2 Descripteurs issus de l'apprentissage profond . . . 117
3.5.3 Descripteurs sur-mesure . . . 119
3.5.4 Classification . . . 120
3.6 Expérimentations . . . 121
3.6.1 Jeu de données . . . 121
3.6.2 Détails d'implémentation . . . 123
3.6.3 Résultats et discussions . . . 124
3.6.4 Étude d'ablation . . . 127
3.7 ICE-AGE : Extension Graphique basée sur l'Attention . . . 129
3.7.1 Motivations . . . 129
3.7.2 Mécanisme d'attention . . . 130
3.7.3 Méthodologie et architecture . . . 136
3.8 Conclusion . . . 140

3.1 Introduction

Nous abordons dans ce chapitre la mise en œuvre d'une approche basée apprentissage profond nommée R-ICE [1], visant à prédire automatiquement le niveau de détail à appliquer aux panneaux latéraux d'une scène donnée.

Cette prédiction constitue une étape essentielle du processus de création ICE car elle détermine l'aspect visuel du contenu à générer pour chaque plan et conditionne les traitements graphiques ultérieurs. Notre objectif est donc de concevoir un modèle capable de capter les régularités esthétiques et contextuelles identifiées dans les pratiques des experts.

Pour ce faire, nous présenterons un état de l'art ciblé sur les approches d'apprentissage profond appliquées à la vidéo, en soulignant les limites et potentiels des architectures existantes. Nous montrerons notamment que les réseaux convolutif 3D, par leur capacité à modéliser conjointement l'espace et le temps, offrent une base solide pour cette tâche prédictive. Enfin, nous détaillerons notre méthodologie d'adaptation, nos choix techniques, la récupération et la gestion de nos données, ainsi que les stratégies d'entrainement mises en œuvre pour maximiser les performances du système R-ICE. Enfin, nous verrons la possibilité d'intégration de mécanisme d'« attention » avec la proposition de ICE-AGE , ce mécanisme omniprésent dans les modèles à base de *transformers* et dans les IA conversationnelles actuelles.

3.2 État de l'art sur l'analyse de vidéo

Les tâches concernant les vidéos retrouvent très souvent les mêmes problématiques que l'image, avec tout de même une certaine adaptation à considérer et une difficulté accrue due à la dimension temporelle supplémentaire. Un parallèle avec les tâches équivalentes en image permet de mieux les comprendre ainsi que leurs défis.

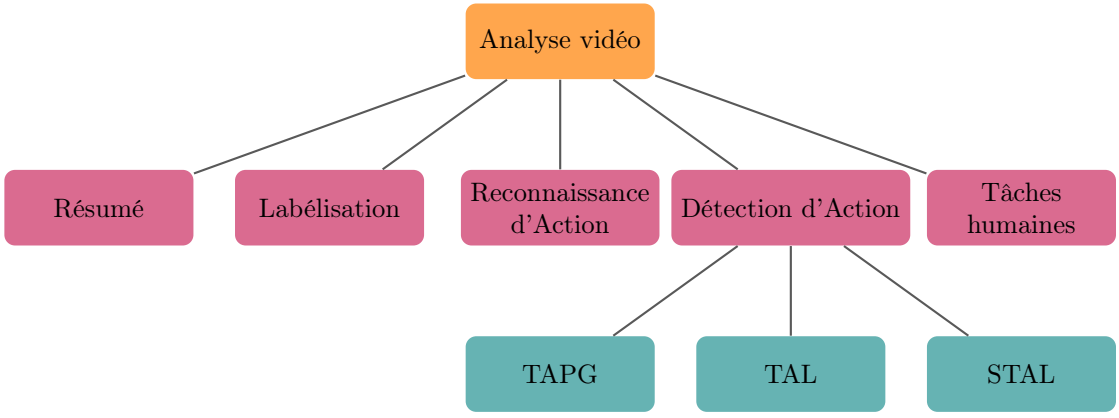


FIGURE 3.1 – **Arborescence des différentes sous-tâches qui composent la compréhension de vidéo**, toutes ces tâches sont des défis à résoudre par l'intelligence artificielle et notamment l'apprentissage profond, avec des objectifs et des applications différentes.

Le domaine de la compréhension de vidéo [105] est en plein essor et regroupe l'ensemble des tâches d'apprentissage profond visant à extraire automatiquement du sens à partir des contenus sémantiques à travers l'analyse ou l'interprétation, à partir de séquences vidéo.

1. R-ICE pour Robotic-ICE

Comme on peut le voir sur la figure 3.1, ces tâches s'articulent autour de différents objectifs, comme le résumé de vidéo, le sous-titrage vidéo, la classification d'actions et la localisation de tâches [2].

La classification d'action



FIGURE 3.2 – **Exemple visuelle d'une action nécessitant la dimension temporelle pour être comprise, en l'occurence, contenu audiovisuel interactif où un personnage doit se projeter à l'aide de son marteau.** Est-ce que l'homme en mouvement vient de sauter ou au contraire de chuter ? Impossible de le déterminer sans connaitre les trames précédentes ou suivantes. (Crédit : [M18])

Alors que la classification d'images consiste à identifier le contenu visuel d'une image parmi un ensemble de classes connues du modèle, la classification d'actions consiste à reconnaitre l'action dominante dans une vidéo ou un clip donné, caractérisée à la fois par son aspect spatial (contenu visuel) et son déroulement temporel [3]. De la même manière que la position relative des pixels (les uns par rapport aux autres) est essentielle dans la compréhension spatiale d'une image, l'ordre chronologique des trames (les unes par rapport aux autres) est tout autant porteur d'information pour la compréhension temporelle. Une vidéo ne se résume pas à un ensemble désordonné d'images, mais constitue une séquence structurée. Par exemple, les actions « sauter » et « chuter » illustrées par la figure 3.2, peuvent reposer sur des trames identiques, mais leur ordre chronologique permet de distinguer l'une de l'autre.

Le résumé de vidéos

Le résumé de vidéo a pour objectif de condenser une vidéo en sélectionnant des images clés ou de courtes séquences de trames. L'enjeu consiste à supprimer les trames redondantes tout en conservant l'intégralité des segments ou des trames qui portent de l'information et qui sont pertinents pour la compréhension globale de la scène.

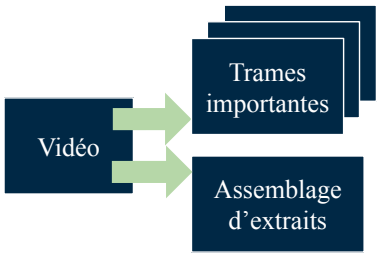


FIGURE 3.3 – **Principe du résumé de vidéo.** La sortie du modèle correspond soit à certaines trames importantes soit à un assemblage de petites séquences vidéos, qui permettent de résumer la vidéo d'entrée du modèle.



FIGURE 3.4 – **Illustration du sous-titrage vidéo.** Chaque évènement est localisé temporellement, avec ses propres instants de début et de fin, ce qui implique que certains peuvent se chevaucher ou se dérouler simultanément. (Crédit : auteur [M19])

Le Sous-titrage vidéo

Cette tâche consiste à générer automatiquement une description textuelle d'une vidéo, en langage naturel. Elle mobilise à la fois la compréhension visuelle (objets, actions, scènes) et l'analyse temporelle (enchainement des évènements), et est particulièrement utile pour l'accessibilité, les moteurs de recherche, la surveillance automatisée ou encore l'extraction ciblée d'informations dans des vidéos longues non-éditées issues de plateformes comme YouTube, telles que les tutoriels ou les démonstrations techniques. Contrairement à la classification ou la détection d'actions, le résumé de vidéo ne se limite pas à prédire une étiquette ou localiser un évènement, mais cherche à produire une narration cohérente et contextuelle.

2. traduits respectivement de leurs noms de tâche d'origine mieux référencés,*video summarizing*, *Video captioning*, *Action Classification*, *Task Localization*

3. Elle est aussi parfois appelé reconnaissance d'action ou action recognition plutôt que classification.

La détection d'actions

La détection d'actions va plus loin que leur classification. Il s'agit ici de les localiser précisément dans le temps (voire dans l'espace). Cette tâche se décline en plusieurs sous-parties, qui sont elles-mêmes parfois étudiées comme tâche à part entière :

La génération de propositions temporelles d'action (TAPG)

Là où la segmentation binaire d'une image serait de détecter de manière binaire, si tel ou tel pixel fait partie de l'avant-plan ou non, du personnage principal ou non, etc., la TAPG [4] [106] est une tâche de segmentation binaire temporelle où chaque portion de la vidéo est classée comme contenant une action ou non. Cette distinction peut être utile pour filtrer les parties pertinentes d'une longue vidéo [107], comme des enregistrements de vidéosurveillance, isoler des séquences utiles dans des vidéos sportives [108]. L'objectif est d'identifier les segments porteurs d'évènements dynamiques.

Comme représenté sur la figure 3.5, la proposition des segments temporels des modèles de TAPG peuvent être **basés sur des ancres**, en se basant sur des intervalles temporels prédéfinis, ou **sans ancre**, en détectant directement les bornes temporelles des actions sans contrainte initiale de durée ou de position [106]. Les méthodes basées ancre sont efficaces mais limitées par la rigidité des intervalles prédéfinis, tandis que les approches sans ancre offrent plus de flexibilité pour détecter des actions de durées variées, au prix d'une localisation souvent moins stable ou précise, et d'une complexité d'implémentation accrue.

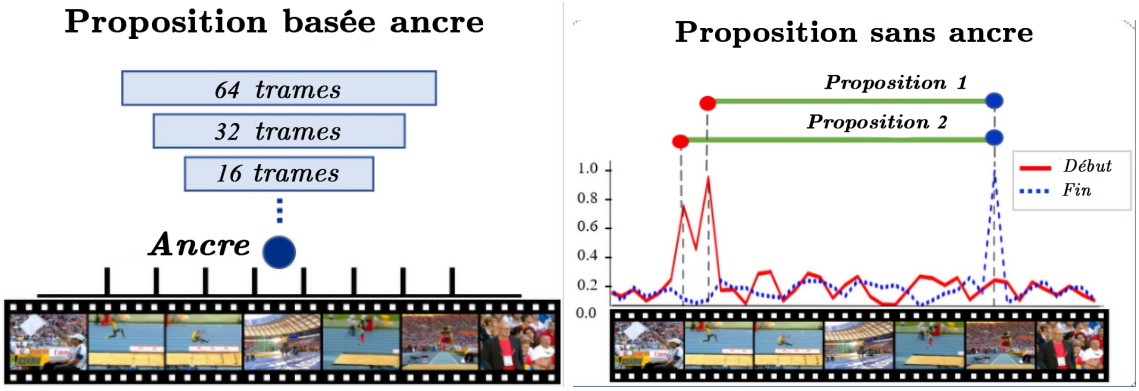


FIGURE 3.5 – **Différences entre les modèles TAPG d'approche basées ancre et d'approche sans ancre.** Les approches basées ancre ont un nombre fini d'intervalles possibles et sont plus simples à implémenter et plus stables, tandis que les approches sans ancre sont plus complexes mais plus flexibles. (Traduite depuis : [114])

Localisation d'action temporelle (TAL)

La localisation d'action temporelle [5] [109] reprend le principe de la TAPG mais ajoute une classification. Il ne s'agit plus seulement de détecter qu'une action se déroule, mais également de déterminer laquelle. On localise donc précisément dans le temps le début et la fin de chaque action, et on l'étiquette à l'aide du vocabulaire de classes préalablement appris par le modèle.

4. ou *Temporal Action Proposal Generation* en anglais
5. ou TAL pour *Action Temporal Localization*

Localisation spatiotemporelle d'action (STAL)



FIGURE 3.6 – **Utilisation d'un modèle de Localisation d'Action spatiotemporelle (STAL).** Le modèle de STAL permet de localiser l'endroit et la trame concernée par une action recherchée. Il est ici utilisé pour détecter les vols en temps réel à l'aide d'une caméra de surveillance. (Crédit : auteur)

La localisation spatiotemporelle d'action [6] [110] va encore plus loin : en plus de déterminer l'intervalle temporel de l'action, elle en identifie la localisation spatiale dans l'image, sous forme de boîtes englobantes ou de masques de segmentation comme illustré sur la figure. 3.6. Elle permet donc de suivre l'action dans le temps, tout en isolant visuellement l'élément concerné dans chaque trame.

« Compréhension de tâches humaines » et autres sous-genres

La compréhension de vidéo ne se limitent pas à des vidéos bien découpées et scénarisées. Les jeux de données récents cherchent à se rapprocher de la complexité du monde réel. On est progressivement passés de vidéos courtes, à des vidéos non tronquées [111] et à des vidéos non scriptées capturées en conditions réelles. Cette tendance se poursuit avec des vidéos en vue subjective [7], des scènes impliquant des actions longues à étapes multiples, du multi-tâches, comme c'est le cas dans les corpus récents tels qu'EPIC-KITCHENS[112]. Pour qualifier ces nouvelles actions, certains travaux utilisent le terme de **compréhension de tâche humaine** [113], certains de **localisation d'étapes** [114]. Ces nouveaux jeux de données donnent lieu à de nouveaux défis comme l'estimation du degré d'avancement d'une activité, la prédiction des étapes suivantes, ou l'identification de l'intention globale.

Cet état de l'art sur la compréhension de vidéos est important afin de pouvoir préparer l'architecture la plus adaptée à notre besoin. Par exemple, notre application pour prédire

6. ou STAL, pour *Spatio-Temporal Action Localization*
7. aussi respectivement plus connues sous les noms techniques d'origine de ce type de jeux de données, *untrimmed video*, *unscripted video* et *egocentric video*

les traitements à effectuer sur l'un de nos segments vidéo présente quelques difficultés. Les vidéos à parcourir sont très longues puisqu'il peut autant s'agir de bandes-annonces (trois minutes environ) que d'un film entier de deux heures à l'avenir. L'infrastructure logicielle [8] proposée ne doit pas être bloquant à ce niveau et peut donc s'inspirer des modèles de TAL qui s'appliquent à des jeux de données de vidéos non tronquées.

Les approches de TAPG sans ancre quant à elles peuvent également nous inspirer pour un cas particulier assez rare du travail ICE, celui de la subdivision temporelle. Cette resegmentation arbitraire d'un plan pourrait être proposée avec une approche sans ancre mais pas avec une approche basée ancre. Elle n'a cependant pas été mise en œuvre, en raison du manque de données d'entrainement à ce sujet. Ce cas reste très rare puisqu'il apparait une vingtaine de fois par film et moins d'une fois par bande-annonce. Néanmoins, les perspectives d'évolution demeurent ouvertes et déjà anticipées grâce au choix de ce type de modèles.

De plus, la capacité à localiser spatialement pourrait être intéressante sur la dimension d'explicabilité des modèles exploités dans ce travail. Cela permettrait de vérifier la ou les zones de la scène ayant conduit aux choix faits par l'IA.

3.3 Problématique et objectifs du projet

Une partie essentielle du travail ICE consiste à déterminer le niveau de détail à appliquer sur les panneaux ICE, qu'il s'agisse d'aplats de couleur, de flou ou de détails. Cette tâche est fondamentale, car l'ensemble du processus ICE en est dépendant. En pratique, cette problématique est complexe et est abordée en profondeur dans le Chapitre 1 dédié au concept ICE. Notre système R-ICE vise à proposer une première prédiction de cette étape en classifiant le niveau de détail d'un plan donné, avant la vérification et l'intervention des concepteurs ICE. Il s'agit donc d'une tâche de classification temporelle, basée sur des critères inédits, partiellement subjectifs, difficiles à formaliser et de nature non absolue (voir section 2.2.4 pour plus de détails.).

3.4 Gestion des données

La part de travail reliée aux données, comme souvent dans un travail en apprentissage profond, est très importante. Cette thèse et ce projet R-ICE ne font pas exception. Les données dont CGR Cinémas dispose ne sont pas directement accessibles. Une part considérable de cette thèse consiste en la gestion et le traitement des données pour qu'elles soient intelligibles pour la machine. Les seules données fournies sont les projets d'Adobe Premiere Pro, le logiciel de montage utilisé par l'équipe de postproduction ICE. Le logiciel étant propriétaire, il ne permet pas l'accès à l'ensemble des données en utilisant un programme informatique. On peut consulter, avec une licence Adobe et le logiciel Premiere Pro, le contenu vidéo manipulé ainsi que les annotations [9] associées à chacun des instants du contenu vidéo. Cependant, seuls sont disponibles les données liées aux bandes-annonces.

8. ou plus précisément, le *framework*, aussi parfois traduit par cadriciel

9. En apprentissage profond, on appelle les annotations, ou labels, ou vérités terrains, les classes associées aux données sur lesquelles l'IA est entrainée. Dans notre cas, le label « Plan de détail » peut être associée à une donnée (16 trames RVB, soit une matrice informatiques de taille hauteur $\times$ largeur $\times$ 16 $\times$ 3)

En ce qui concerne les films, les annotations faites par ICE sur le film sont disponibles, mais le contenu vidéo ne l'est pas pour plusieurs raisons.

D'une part, les contenus fournis par les studios des films sont fortement floutés en leurs centres avec un filigrane visible pour des raisons de confidentialité, comme illustré en figure 3.7).

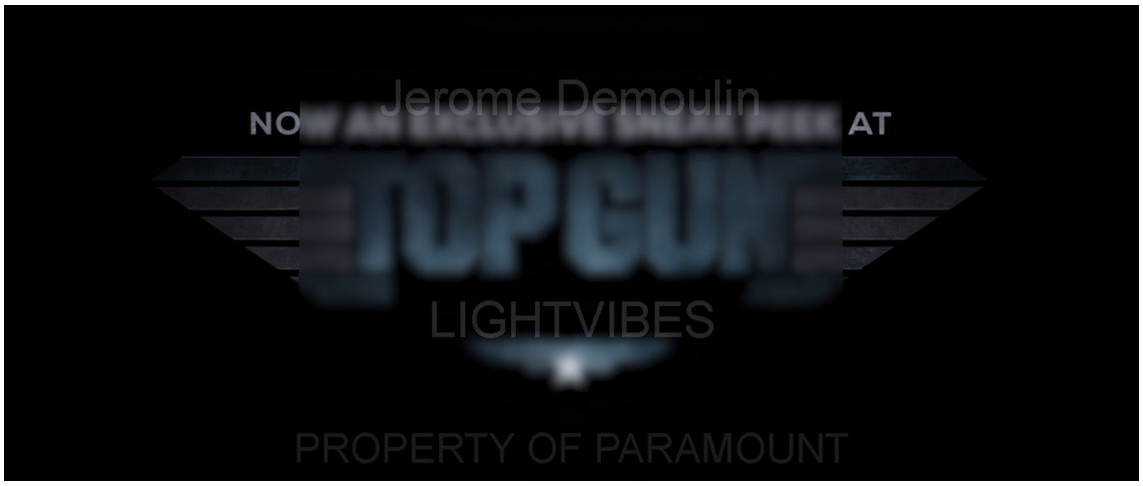


FIGURE 3.7 – **Aspect d'un contenu vidéo de travail fourni par les studios de film.** Trame d'un extrait long de Top Gun : Maverick (5 min), avec un flou central important et un filigrane. Cet extrait long a été dé-sécurisé après la sortie du film.

D'autre part, la salle de travail ICE est soumise de manière contractuelle à une sécurité intransigeante, exigée par les distributeurs afin d'éviter toute fuite avant la sortie du film. De plus, contourner le problème en utilisant des DVD ou Blu-Ray nécessiterait d'arriver à casser le système anti-copie. Enfin, même de cette manière, rien ne garantit d'obtenir la même version du film : les versions des films sont très nombreuses car les censures de certains pays peuvent parfois couper un plan du film. Les génériques et les « cartons » d'avant-film et d'après-film peuvent aussi varier comme illustré sur la figure 3.8. De cette manière, recalquer les annotations de l'équipe ICE sur des films devient un problème difficile à résoudre.

Le seul moyen trouvé pour exploiter ces projets Première Pro est d'exporter le projet dans un format compatible pour Adobe Final Cut Pro. De cette manière, ces projets sont transcrits au format XML, qu'il devient possible de parcourir avec un langage informatique de script, tel que Python.

Même dans ces cas-là, la récupération des données tels que le début d'un plan annoté, sa fin, et le label associé n'est pas garanti. Une transition fondue appliquée par l'équipe ICE fait disparaître les deux bornes qui se trouvent dans la transition fondue lors de la récupération XML.

Un changement de teinte appliqué au contenu ICE d'un plan n'est pas non plus visible via l'un des exports possibles. L'analyse des différents fichiers XML générés après l'export de dizaines de projets a permis de détecter que ce fichier d'annotations peut aussi arborer deux différents décalages cachés (quatre pour les films séparés en bobine). Ces derniers décalent tous les horodatages de début et de fin de plan du fichier, visibles sur la figure 3.9. Ainsi, le numéro des trames dans le fichier vidéo F_v est donné par l'équation suivante :

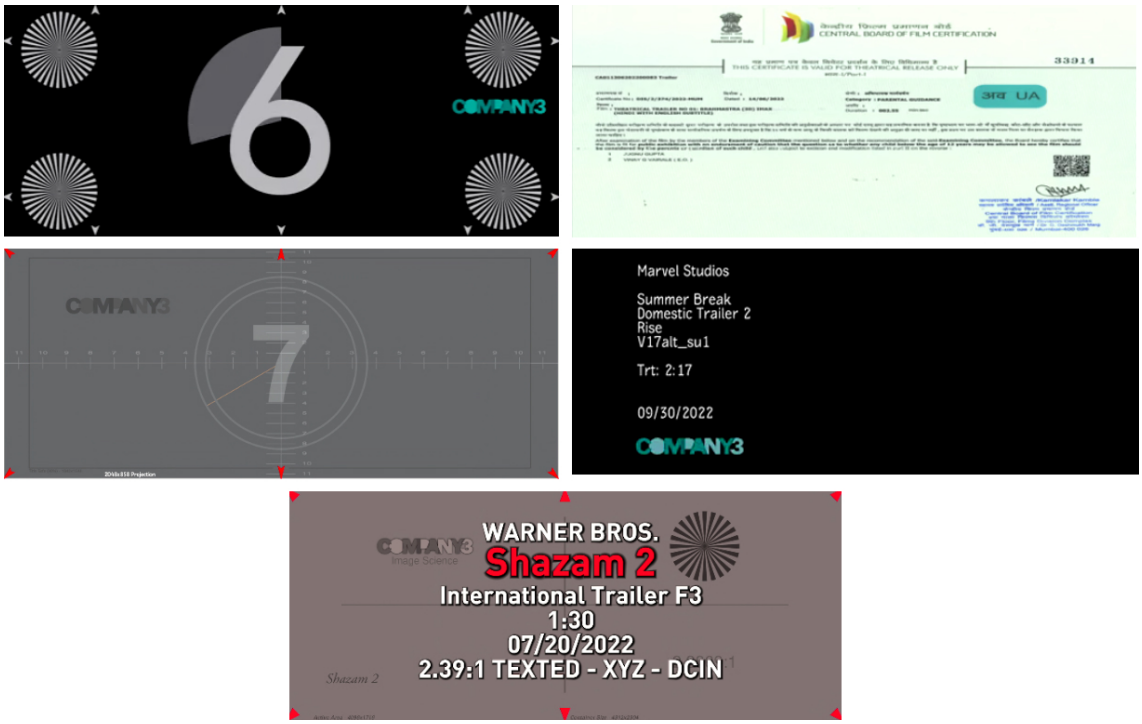
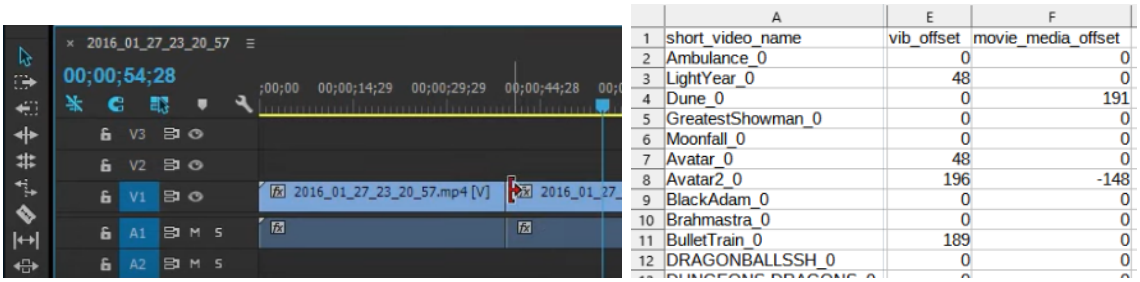


FIGURE 3.8 – **Illustrations de cartons divers** : mires, décomptes, visa d'exploitations. Chacun de ces cartons peuvent varier en longueur, selon la région de sortie et la version du film, ce qui modifie le contenu vidéo.

$$F_v = F_p + O_m + O_v \tag{3.1}$$

Où F_p représente le numéro des trames indiquées dans Premiere Pro, O_m les décalages de media, O_v les décalages de vidéo.

À l'issue de cette analyse de données, une solution a été développée permettant de récupérer les informations pertinentes disponibles, puis les placer dans un fichier CSV. Une arborescence de ces fichiers est créée comme système de gestion de nos données.



	A	E	F
1	short_video_name	vib_offset	movie_media_offset
2	Ambulance_0	0	0
3	LightYear_0	48	0
4	Dune_0	0	191
5	GreatestShowman_0	0	0
6	Moonfall_0	0	0
7	Avatar_0	48	0
8	Avatar2_0	196	-148
9	BlackAdam_0	0	0
10	Brahmastra_0	0	0
11	BulletTrain_0	189	0
12	DRAGONBALLSSH_0	0	0

FIGURE 3.9 – **Origine et prise en compte des décalages temporels entre les fichiers vidéos des données de vérité terrain et les labels associés à ces données.** À gauche, la possibilité de cacher une partie du média ou d'ajouter un espace vide dans Premiere Pro. Ces opérations parfois faites par l'équipe de postproduction ICE créent un décalage à prendre en compte pour faire concorder les numéros de trame exportés et le fichier vidéo de source. À droite, les décalages récupérés des projets de travail des bandes-annonces.

3.4.1 Constitution de la base de données

Le but recherché est de constituer une base de données intelligible de tous les films ou bandes-annonces incluant ces plans, leurs débuts, leurs fins et leurs labels associés. Plusieurs obstacles ont été rencontrés lors de cette phase. Les seules informations dont nous disposions dans le fichier structurés résident dans l'objet informatique d'Adobe Premiere Pro de *« clipitem »*, son nom, et sa piste vidéo. Les pistes gauche et droite sont dissociées, et l'on peut voir que ces dernières sont alternées en règle générale. Ces *clipitems* sont analogues à un traitement d'image particulier et correspondent souvent à un plan cinématographique. Comme certains *clipitems* s'appliquent à droite alors qu'ils sont labellisés gauche (lorsqu'un effet miroir est appliqué), il est indispensable d'inspecter le fichier structuré (XML) et ne pas se fier au premier *clipitem*. Cela permet de savoir de manière appropriée si une piste s'applique à gauche ou à droite. Cette approche fonctionne pour la majorité des pistes sauf dans les cas où une seule piste de surimpression existe sur les deux. Cette dernière peut avoir très peu de *clipitems* en plus d'avoir une défaillance sur le nommage (gauche et droite). Dans ce cas précis, le projet Premiere Pro doit être modifié pour lever cette ambiguité. La procédure suivie pour la préparation des données du fichier XML se déroule comme suit :

Algorithme 1 : Procédure de préparation et de filtrage des pistes

Entrée : Fichier XML du projet Premiere Pro
Sortie : Liste des pistes principales

1. **Recherche de la composition `Vib` :** en Python, recherche récursive de la composition nommée `Vib`
2. **Attribution des noms de pistes :** récupération du nom du premier `clipitem` de chaque piste pour attribuer un nom aux pistes
3. **Sélection par liste de traitements connus :** conservation uniquement des pistes dont le nom correspond à un traitement existant (`WALL`, `WALL COLOR`, `WALL EXPAND FH`, etc.)
4. **Filtrage des pistes masquées :** exclusion des pistes désactivées par l'équipe de postproduction
5. **Détermination des pistes principales :** regroupement des pistes portant le même nom et sélection de celles situées au-dessus de la source des choix mais en dessous des surimpressions optionnelles

Cette dernière étape (regroupement des pistes par leur nom et détermination des pistes principales par leur position) est faite parce que les pistes de surimpression peuvent avoir le même nom que les pistes principales ainsi que celle de source des choix créatifs. Il peut y avoir jusqu'à six instances de pistes avec le même nom. En revanche, les pistes principales sont par essence celles qui sont placées au dessus de la source des choix, mais en dessous des pistes optionnaux de surimpression, comme on a pu l'observer précédemment sur la figure 2.2, et c'est sur cette information que cet algorithme se base.

Une fois qu'on a pu extraire toutes les informations du fichier structuré XML, il est nécessaire de les rendre cohérentes et de les nettoyer.

La première étape pour pouvoir fusionner les labels de gauche et de droite sur les mêmes plans est d'essayer de fusionner les informations redondantes entre les pistes gauche et droite. Quand bien même les contenus vidéos ont des séparations nettes entre leurs plans, les séparations des plans sur le projet de travail ICE ne sont pas les mêmes pour la piste

gauche et la piste droite. Un exemple de ces cas est illustré sur la figure 3.10, montrant une subdivision et une fusion arbitraire opérées par l'équipe ICE, ce qui se répercute dans nos données. Nous corrigeons les plans de manière à ce que les coupes (ou séparations de plans) corrigées correspondent à :

$$C_{\text{gauche}}^{\text{corrigé}} = C_{\text{gauche}} \cup C_{\text{droit}} \tag{3.2}$$

$$C_{\text{droit}}^{\text{corrigé}} = C_{\text{gauche}} \cup C_{\text{droit}} \tag{3.3}$$

Où C correspond à l'ensemble des coupes identifiées par leurs horodatages. On a ainsi deux ensembles suréchantillonnées sur les plus petits pas de coupes possibles.

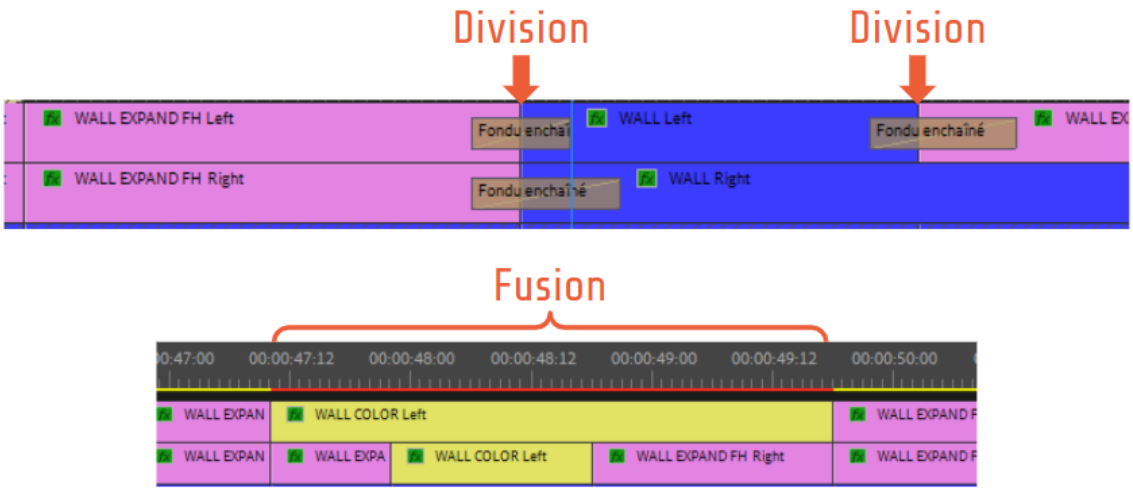


FIGURE 3.10 – **Fondu enchainé et fusion entre les plans d'une piste vidéo.** Les graphistes ICE peuvent être amenés à réaliser ces opérations. Ces dernières dérèglent la piste gauche et la piste droite qui sont censées être équivalentes pour pouvoir les fusionner et associer un label par plan, ce qui est nécessaire du point de vue des données d'entrainement. Il faut sous-échantillonner ces plans selon leurs nouvelles définitions données en équation 3.2.

Puisque nous avons des informations reliées aux vidéos à des échelles différentes, un seul fichier CSV ne permet pas de tout renseigner. Nous allons avoir cinq échelles de fichier CSV différentes, de la plus grande à la plus fine comme schématisé sur la figure 3.11.

Un fichier donne des informations à l'échelle des fichiers vidéo. Elles permettent de relier un fichier vidéo à son film et retrouver quels sont les bandes-annonces différents d'un même film, et ainsi quelles sont les données corrélées entre elles. Nous verrons dans la prochaine section 3.4.2 que ceci a son importance pour obtenir un entrainement viable de modèle d'apprentissage profond. Le fichier CSV concernant les plans est le fichier principal puisqu'il contient la vérité terrain, c'est-à-dire le choix fait par l'équipe ICE, comme illustré sur la figure 3.12. Il contient les informations de la bobine ou la bande-annonce utilisé (champ video_uid), le début ainsi que la fin du plan, les labels mais aussi le fait que le bord droit du film soit utilisé pour les panneaux ICE de gauche, et inversement.

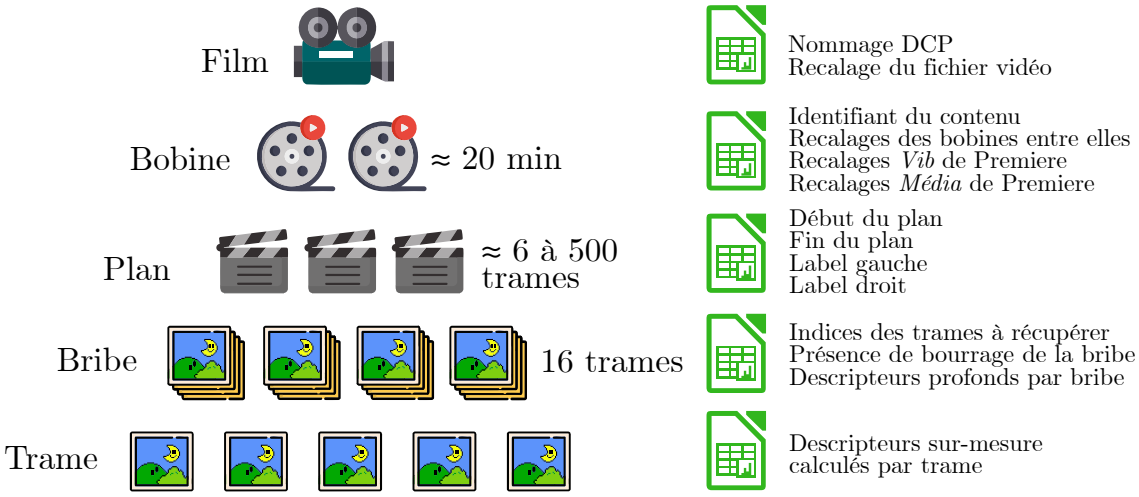


FIGURE 3.11 – **Hiérarchie de la structure de données mise en place.** (Crédit : auteur [M20])

	A	B	C	D	E	F	G	H	I	J	K	L
1	video_uid	f_start	f_stop	left_choice _from_right	choice_type_left	choice _index_left	right_choice _from_left	choice_type_right	choice Index_right	len	fps	video -frames
2	ANT_MAN_3_REEL_1	29	598	0	WALL	0	0	WALL	0	569	24.0	569
3	ANT_MAN_3_REEL_1	598	629	0	WALL EXPAND FH	3	0	WALL EXPAND FH	3	31	24.0	31
4	ANT_MAN_3_REEL_1	629	732	0	WALL EXPAND FH	3	0	WALL EXPAND FH	3	103	24.0	103
5	ANT_MAN_3_REEL_1	732	911	0	WALL COLOR	4	0	WALL COLOR	4	179	24.0	179
6	ANT_MAN_3_REEL_1	911	979	0	WALL COLOR	4	0	WALL COLOR	4	68	24.0	68
7	ANT_MAN_3_REEL_1	979	1190	0	WALL EXPAND FH	3	0	WALL EXPAND FH	3	211	24.0	211
8	ANT_MAN_3_REEL_1	1190	1343	0	WALL EXPAND FH	3	0	WALL EXPAND FH	3	153	24.0	153
9	ANT_MAN_3_REEL_1	1343	1534	0	WALL EXPAND FH	3	0	WALL COLOR	4	191	24.0	191
10	ANT_MAN_3_REEL_1	1534	1605	1	WALL COLOR	4	0	WALL COLOR	4	71	24.0	71
11	ANT_MAN_3_REEL_1	1605	1672	1	WALL COLOR	4	0	WALL COLOR	4	67	24.0	67
12	ANT_MAN_3_REEL_1	1672	1815	0	WALL EXPAND FH	3	0	WALL EXPAND FH	3	143	24.0	143
13	ANT_MAN_3_REEL_1	1815	1976	1	WALL COLOR	4	0	WALL COLOR	4	161	24.0	161
14	ANT_MAN_3_REEL_1	1976	1999	0	WALL EXPAND FH	3	0	WALL COLOR	4	23	24.0	23
15	ANT_MAN_3_REEL_1	1999	2066	1	WALL COLOR	4	0	WALL COLOR	4	67	24.0	67
16	ANT_MAN_3_REEL_1	2066	2105	0	WALL EXPAND FH	3	0	WALL COLOR	4	39	24.0	39
17	ANT_MAN_3_REEL_1	2105	2239	0	WALL EXPAND FH	3	0	WALL EXPAND FH	3	134	24.0	134
18	ANT_MAN_3_REEL_1	2239	2364	0	WALL COLOR	4	1	WALL COLOR	4	125	24.0	125
19	ANT_MAN_3_REEL_1	2364	2526	0	WALL COLOR	4	1	WALL COLOR	4	162	24.0	162
20	ANT_MAN_3_REEL_1	2526	2624	0	WALL	0	0	WALL	0	98	24.0	98
21	ANT_MAN_3_REEL_1	2624	2722	0	WALL COLOR	4	1	WALL COLOR	4	98	24.0	98
22	ANT_MAN_3_REEL_1	2722	2841	0	WALL COLOR	4	1	WALL COLOR	4	119	24.0	119

FIGURE 3.12 – **Aperçu des données de film par plan cinématographique.** 46 251 plans pour les 22 films sont ainsi répertoriés. Ce fichier contient respectivement les champs de la vidéo concernée, le début et la fin du plan, l'inversion gauche-droite des contenus des panneaux ainsi que le choix (détail, flou, aplat) correspondant au plan.

3.4.2 Méthodologie de stockage des données

Plusieurs systèmes de stockage des données sont aussi étudiés. Afin d'entrainer des modèles d'apprentissage profond, il sera nécessaire d'accéder aux données par trame et non par vidéo, et ce, de manière ininterrompue.

La plupart des bibliothèques de codes disponibles pour les modèles d'apprentissage profond destinés à la vidéo adoptent un mode de gestion des données calqué sur celui utilisé pour les modèles d'images. Dans ce cadre, chaque trame d'une vidéo est stockée indépendamment sur le disque. Ce système de fichiers est le plus rapide mais demande des espaces mémoires titanesques. En effet, là où la bande-annonce compressée sans perte requiert 2,19 Go, le stockage de la somme de toutes les trames requiert 5 Go sur disque.

Étant donné que nous allons travailler avec une quarantaine de bandes-annonces, en stockant plusieurs versions altérées des bandes-annonces, ce gain d'espace n'est pas à négliger. Dans ce premier projet R-ICE, nous allons privilégier l'optimisation en stockant le plus souvent possible les données sous forme de vidéo. Cela implique de décoder la vidéo à chaque instant où notre modèle devra accéder à une trame, ce qui ajoute fatalement un temps d'attente d'accès mémoire. Afin de ne pas altérer la qualité de la vidéo suite au processus d'encodage et décodage succesifs de la vidéo, nous privilégions un codec sans perte, à savoir « H.264 Profile Haut 4 :4 :4 Predictif en mode RVB ».

En apprentissage profond, les données disponibles doivent être partitionnées en trois ensembles : l'ensemble d'entrainement, l'ensemble de validation et l'ensemble de test [10]. Les deux derniers termes sont parfois interchangés à tort dans les articles de la communauté scientifique. Les données d'entrainements sont celles présentées au modèle d'apprentissage profond pendant l'entrainement pour que le modèle évalue les performances en utilisant une fonction de perte et permettent de mettre à jour ses paramètres (poids et biais).

Lors de cette étape, le modèle infère sur des données qu'il n'a jamais vu durant son entrainement : les données de validation. Cela permet de quantifier sa capacité à généraliser, depuis les données d'entrainement, vers des données inconnues. C'est cette capacité-là que nous recherchons à maximiser lors de l'entrainement de tout réseau de neurones.

Les données de tests quant à elles, permettent une évaluation à la fin de l'entrainement et donc à la fin des validations itératives du modèle. Ces données viennent résoudre un problème existant lorsque seuls deux jeux de données sont utilisés. Comme les données de validation orientent le modèle qui sera sélectionné, le modèle est indirectement biaisé par ces données. Les données de tests viennent alors confirmer une bonne généralisation du modèle entrainé.

Par ailleurs, dans l'entrainement d'un modèle d'apprentissage profond, les instances d'un même contenu ne doivent pas apparaitre dans les différents ensembles [11]. Dans notre contexte de travail, les bandes-annonces sont des instances différentes mais qui peuvent parfois venir du même film. Cela fait partie des informations qui seront retenues dans notre structure hiérarchique de données évoquée plus tôt pour pouvoir garder ces instances dans le même ensemble.

3.5 Méthodologie et appropriation d'architectures existantes

L'architecture logicielle incluant les différentes chaines de traitement de la solution proposée est présentée en figure 3.13. Les données en entrée, à savoir un film, sont segmentées en plans cinématographiques, chacun étant associé à une annotation de vérité terrain extraite manuellement des projets Adobe Première Pro provenant des graphistes ICE. L'objectif est de prédire, pour tout plan donné, le traitement le plus approprié afin d'offrir à l'utilisateur une expérience immersive et confortable.

10. On peut aussi croiser le terme « jeu de données » d'entrainement, de validation et de test, calqué sur l'anglais « training, validation and test data set »

11. Le problème qui aurait lieu serait de faire entrainer et valider le modèle sur des données quasi-identiques, et ce qui biaiserait nos indices de performances et de bonne généralisation du modèle, qui seraient alors plus élevés que ce qu'ils devraient.

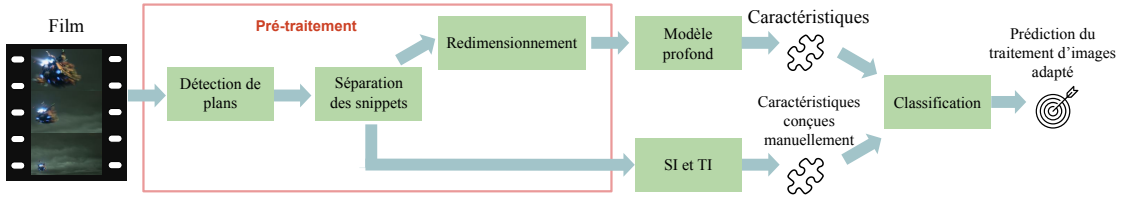


FIGURE 3.13 – Architecture de notre modèle R-ICE.

Pour ce faire, il s'agit d'extraire des descripteurs représentatifs permettant de caractériser chaque plan. Des descripteurs spécifiques à la vidéo sont extraits à l'aide de modèles d'apprentissage profond issus de la littérature. Étant donné que notre problématique concerne à la fois des objectifs de classification subjective, mais aussi les notions d'immersion et de vision extrafovéale, des descripteurs conçus manuellement sont également calculés afin de fournir des informations spécifiques adaptées à notre modèle, dans le but d'optimiser le choix du contenu immersif périphérique.

3.5.1 Prétraitement des données

Pour prétraiter convenablement un film donné, celui-ci est d'abord segmenté en plans à l'aide de l'outil Adobe Scene Detection, puis chaque plan est sous-divisé en extraits de longueur fixe appelés bribe [12] ou clip (16 images). Si la dernière bribe d'un plan contient moins de 16 images, elle sera ignorée. Cette taille de 16 trames successives correspond à la condition d'entrées d'origine des bribes pour deux des trois réseaux convolutifs 3D évalués. Elle correspond aussi à la taille médiane d'un plan entre deux coupes dans nos données d'entrainement. C'est pour ces deux raisons que cette taille d'entrée est choisie.

3.5.2 Descripteurs issus de l'apprentissage profond

Dans le cadre de notre projet R-ICE, il est nécessaire d'utiliser des modèles d'apprentissage profond capables de traiter conjointement les informations spatiales et temporelles. Cela permet notamment de résoudre des cas où des objets saillants n'apparaissent que dans les premières images du plan. Par ailleurs, comme certains plans peuvent durer plusieurs centaines d'images, il est important de considérer et de s'inspirer des modèles à mémoire temporelle long-terme. Certains modèles, par la nature des tâches pour lesquelles ils ont été conçus, requièrent une telle mémoire temporelle étendue. Par exemple, l'interpolation temporelle de vidéo qui est l'action de générer des trames interstitielles pour élever la fluidité du mouvement ne nécessite que quelques trames successives. En revanche, la compréhension de vidéo dont il a été question plus haut demande une large temporalité pour comprendre certaines actions [109].

Les modèles de traitement d'image sont déjà connus pour leur forte consommation en ressources, bien plus que les modèles de traitement du langage naturel. Les modèles vidéo, quant à eux, repoussent davantage ces limites en raison de la dimension temporelle qu'ils intègrent. D'ailleurs, Yann Le Cun souligne que les modèles larges d'apprentissage de langage (LLM) reposent sur la prédiction de jetons rendant la gestion de l'incertitude relativement aisée en raison de l'ensemble fini des jetons possibles. [13] En revanche, prédire

12. calqué sur le terme anglais *snippet*
13. Les jetons sont une unité de délimitations inférieur ou égal à un mot, pour le représenter dans un espace mathématique latent. Le nombre de mots d'une langue étant un ensemble fini, le nombre de jetons différents possibles est également fini.

des images ou des séquences vidéo implique de modéliser une distribution sur un espace de données **continu** et potentiellement de très haute dimension, ce qui rend la tâche nettement plus ardue [115] [14]. Ici, dans le discours de Yann Le Cun, il est question de prédiction temporelle, mais la compréhension des données d'entrées suit les mêmes lois. L'entrainement par le biais d'une rétropropagation (et plus la simple inférence) de tels modèles vidéo représente une charge computationnelle supplémentaire considérable, à la fois en termes de puissance de calcul que de mémoire vive. Pour cette raison, ces entrainements restent quasi exclusivement l'apanage des grandes entreprises technologiques disposant d'infrastructures massivement parallèles. Les structures de recherche plus modestes, quant à elles, se reposent sur les modèles préentrainés rendus disponibles par ces grands acteurs. Elles utilisent ces architectures comme squelettes qu'elles affinent par spécialisation *(fine tuning)*, en les réentrainant sur des domaines d'application plus ciblés et compatibles avec leurs ressources limitées. Les modèles C3D [116], I3D [117], TSP [118] [15] et SlowFast [119] illustrent des approches conçues pour extraire des caractéristiques sensibles à la dynamique temporelle d'une vidéo.

Notre problématique partage plusieurs points communs avec les tâches ciblées par ces modèles. Cela nous a conduit à sélectionner trois modèles largement utilisés et reconnus dans la littérature : C3D, I3D et TSP, en tant qu'extracteurs de caractéristiques spatiotemporelles profondes.

Ces trois modèles représentent trois étapes distinctes dans l'évolution des modèles convolutifs 3D. C3D [116] est le premier modèle à exploiter des convolutions 3D à tous les niveaux pour analyser des clips vidéo courts (16 images). Il encode à la fois les informations spatiales et temporelles, mais sans bénéficier de préentrainement : tous les poids doivent être appris depuis zéro, ce qui rend l'apprentissage plus couteux et moins généralisable. I3D [117], développé par DeepMind (Google), reprend la structure d'un réseau 2D (Inception-V1) et la « gonfle » (*inflate*) en 3D, en dupliquant les noyaux 2D préentrainés sur ImageNet le long de l'axe temporel. Ce transfert permet d'obtenir de meilleures performances avec moins de données annotées. En contrepartie, le modèle est lourd et nécessite des vidéos d'entrée qui prennent une place en mémoire colossale ainsi qu'une forte puissance de calcul pour être entrainé efficacement. TSP [119] adopte quant à lui une approche plus légère et ciblée. Il réentraine un modèle comme I3D avec une tâche autosupervisée, conçue pour renforcer la sensibilité à la structure temporelle (par exemple, reconnaitre si l'ordre des images a été perturbé). TSP utilise des clips plus petits, en nombre de pixels et en nombre de trames et parvient néanmoins à surpasser I3D en précision, tout en réduisant la taille des entrées par un facteur quatre sur l'axe pixel et un facteur huit en nombre de trames. Cela en fait une alternative nettement plus accessible à réentrainer.

Ces modèles, souvent utilisés comme dorsale de modèle dans des architectures plus complexes, diffèrent sur plusieurs aspects liés à leurs procédures d'entrainement. C3D est entrainé sur des extraits de 16 images RVB à 25 i/s. I3D utilise deux flux : l'un basé sur 64 images RVB, l'autre sur les flots optiques correspondants, également à 25 i/s ; les vidéos trop courtes sont bouclées pour satisfaire cette exigence. TSP, de son côté, travaille sur 16 images RVB extraites par sous-échantillonnage temporel, ramenant les vidéos de 30 i/s à

14. Puisqu'une vidéo est aujourd'hui surtout numérique, l'espace de données n'est pas à proprement parler continue, puisque tout est constitué nécessairement d'une suite binaire de longueur finie, d'un nombre de caractères fini (0 ou 1). Cependant l'image et la vidéo sont effectivement des espaces de données à des ordres de grandeur de dimensions bien supérieurs que les espaces de données textuels.

15. Noms complets des modèles : *Convolutional 3D* (C3D), *Inflated 3D ConvNet* (I3D), *Temporary-Sensitive Pretraining* (TSP)

TABLE 3.1 – **Comparaison des conditions d'origine des trames de chaque modèle.** Comparaison entre les trois dorsales de modèle C3D, I3D et TSP, les différents choix sur le redimensionnement de la trame d'entrée, sur le recadrage pendant l'entrainement, pendant le test et comparaison de la taille du réseau de neurones en nombre de paramètres (M pour millions).

Modèle	Redimensionnement de la taille de la trame d'entrée	Pendant l'entrainement	Pendant le test	Nombre de paramètres
C3D	côté le plus petit à 128, en conservant le ratio	recadrage aléatoire 112×112	recadrage centré 112×112	78,0 M
I3D	côté le plus petit à 256, en conservant le ratio	recadrage aléatoire 224×224	recadrage centré 224×224	12,3 M
TSP	côté le plus petit à 128, en conservant le ratio	recadrage aléatoire 112×112	recadrage centré 112×112	31,3 M

15 i/s. Ainsi, chaque bribe vidéo correspond à une « empreinte temporelle » c'est-à-dire la largeur de bande temporelle d'où proviennent les informations des cartes de caracteristiques avant leur classification. Cette empreinte temporelle correspond à la borne maximale de mémoire temporelle que peut considérer le modèle pour sa décision, en supposant que ce dernier soit correctement entrainé et que son architecture permette une bonne propagation de ces informations tout au long du réseau. L'empreinte temporelle est ainsi l'analogue 1D du champs receptif 2D dans les modèles d'images [16]. Elle est différente selon le modèle : 0,64 seconde pour C3D, 2,56 secondes pour I3D et 1,875 seconde pour TSP.

Alors que les dorsales de modèle C3D, I3D et TSP ont été entrainés sur des images recadrées au centre et au format carré, les décisions de notre tâche reposent davantage sur les bords gauche et droit des images, là où se produit une discontinuité avec l'extension de la toile de cinéma (voir la section 2.2). L'hypothèse intuitive serait qu'un recadrage centré risque donc de nuire à l'apprentissage de notre modèle. Cette hypothèse sera étudiée plus en détails en section 3.6.4, afin de la confirmer ou l'invalider. Les trames sont alors redimensionnées à 112×267 (rapport d'aspect cinéma conservé) sans recadrage.

Concernant les dimensions spatiales, les trois modèles procèdent à un redimensionnement de la vidéo, suivi d'un recadrage aléatoire ou un recadrage centré selon les étapes d'entrainement ou de test, comme illustré dans le tableau 3.1. Dans notre cadre expérimental, une bribe vidéo de 16 images produit 4096 descripteurs pour C3D, 1024 pour I3D et 512 pour TSP.

3.5.3 Descripteurs sur-mesure

Les caractéristiques spatiales et temporelles d'un plan cinématographique donné sont des attributs essentiels lors de la phase de classification, afin de déterminer quel choix créatif appliquer selon la méthodologie de travail de l'équipe postproduction ICE. Pour prendre en compte ces caractéristiques et enrichir l'entrée du classifieur avec des descripteurs sur-mesure (*hand-crafted features* ou descripteurs faits-main) supplémentaires, nous avons

16. Le champs receptif ou *receptive field* est intéressant à examiner pour les modèles d'image ou de vidéo. Trop peu de couches de convolution dans un modèle d'image pourrait nous donner un champ réceptif trop étroit et serait dans l'incapacité d'avoir accès à toutes l'image dans son ensemble.

sélectionné le descripteur spatial SI et le descripteur temporel TI [17], respectivement liés au niveau de détail et à l'intensité des mouvements à l'échelle d'une trame.

SI et TI sont définis dans la recommandation [120]. Le SI est exprimé comme suit :

$$SI_n = \sigma^*(\text{Sobel}(F_n)) \tag{3.4}$$

basé sur les contours du contenu, où SI_n représente l'information spatiale d'une image à l'instant n, $\sigma^*(\cdot)$ l'écart-type sur les pixels d'une image, Sobel$(\cdot)$ le filtre de Sobel appliqué à une matrice, et F_n le plan de luminance de l'image à l'instant n, tel que défini dans la recommandation. Le TI est quant à lui calculé entre images successives, en mesurant la différence entre les valeurs de pixels de deux images consécutives :

$$TI_n = \sigma^*(M_n) \tag{3.5}$$

Où

$$M_n(i,j) = F_n(i,j) - F_{n-1}(i,j) \tag{3.6}$$

avec $F_n(i,j)$ la valeur du pixel situé à la i^{e} ligne et la j^{e} colonne de l'image F à l'instant n, et M_n la matrice des différences (pixel à pixel) entre deux images consécutives.

Les descripteurs sur-mesure se composent d'une valeur de SI et d'une valeur de TI pour chaque image de la bribe vidéo analysée, soit un vecteur de 32 scalaires fourni en entrée du classifieur pour une bribe de 16 images.

3.5.4 Classification

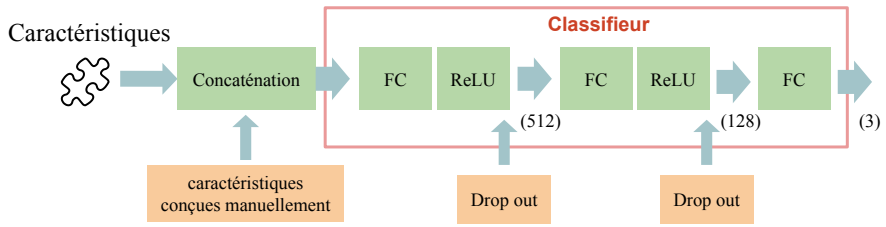


FIGURE 3.14 – Structure du classifieur placé après l'extraction des caractéristiques par les différents backbones temporels.

Le classifieur de base utilisé est un perceptron multicouche (MLP), dont l'architecture est représentée en figure 3.14, où FC désigne les couches entièrement connectées. Les descripteurs sur-mesure, représentés par les entrées *SITI* du classifieur, sont une option explorée afin de tenir compte de la variabilité du contenu spatiotemporel et d'améliorer les performances du modèle. Il en va de même pour les couches de *dropout*, utilisées afin de limiter le surapprentissage, en forçant le modèle à exploiter l'ensemble des descripteurs fournis, ce qui se révèle particulièrement utile pour l'apprentissage à partir de nos descripteurs sur-mesure.

D'autres variantes ont été testées pour améliorer les performances du modèle, comme l'ajout d'une couche de convolution avant le MLP ou de couches affines supplémentaires.

17. SI et TI pour respectivement *Spatial Perceptual Information* et *Temporal Perceptual Information*

Dans ce dernier cas, la couche affine [18] est suivie d'une fonction d'activation ReLU vue en sectionss :MLP ou d'une couche de *dropout*, pour rester cohérent avec le reste de l'architecture. Une autre alternative consiste à injecter les descripteurs manuels dans un flux de caractéristiques séparé ; l'ajout ou non de *dropout* sur ce second flux permet d'évaluer la sensibilité du modèle à cette régularisation.

Les prédictions de ce classifieur sont produites à l'échelle des bribes, les plans ayant des durées variables qui ne sont pas compatibles avec l'entrée fixe du modèle. Ainsi, pour obtenir une prédiction au niveau d'un plan, une majorité des votes issus des prédictions des segments correspondants est calculée.

3.6 Expérimentations

3.6.1 Jeu de données

Comme mentionné en section 3.4 l'accès aux films est extrêmement contraint en raison des restrictions imposées par les studios. C'est pourquoi les travaux de ce chapitre se sont concentrés sur un jeu de données de bandes-annonces, librement accessibles via CGR Cinémas. Ces dernières nous ont permis d'obtenir les informations de vérité terrain nécessaires concernant les traitements appliqués aux images pour générer les contenus gauche et droit destinés aux panneaux ICE.

Ce jeu de données est composé de 30 bandes-annonces en résolution 858×2048, avec des labels fournies par des experts ICE. Leurs durées varient entre 40 et 176 secondes. Le modèle C3D étant incompatible avec notre résolution de 112×267, le recadrage central initial a été conservé spécifiquement pour ce dernier.

Il est reconnu que l'activité temporelle des bandes-annonces diffère notablement de celle des films. Un ajustement ultérieur sera probablement requis lorsque le contenu complet sera disponible.

Comme mentionné dans le chapitre 2, les experts sélectionnent un traitement parmi cinq à appliquer au plan concerné.

Nous avons mesuré l'occurrence de chaque type de traitement possible à appliquer à un plan donné, aussi bien dans notre jeu de données construit que dans les films complets. Les résultats sont présentées en figure 3.15.

On observe que `Seam` et `WALL EXPAND` sont rarement, voire jamais utilisés, quel que soit le corpus. Afin d'éviter des déséquilibres à l'entrainement, nous avons fusionné `Seam` et `WALL EXPAND` avec `WALL`, en raison de la similarité des traitements appliqués. Dans un but de clarté, nous pouvons également appeler les trois labels restants `Détail`, `Flou` et `Aplat de couleur`.

Cette décision de fusion est faite selon deux hypothèses, validées par les experts. 1) Les rendus visuels entre ces trois choix créatifs sont proches et 2) les buts recherchés lors de l'application de l'un ces trois choix est le même : conserver un maximum de détails dans les panneaux ICE.

Étant donné la quantité limitée de données disponibles pour cette application (seulement 30 bandes-annonces), l'augmentation de données s'avère être une solution essentielle pour

18. aussi appelée entièrement connectée

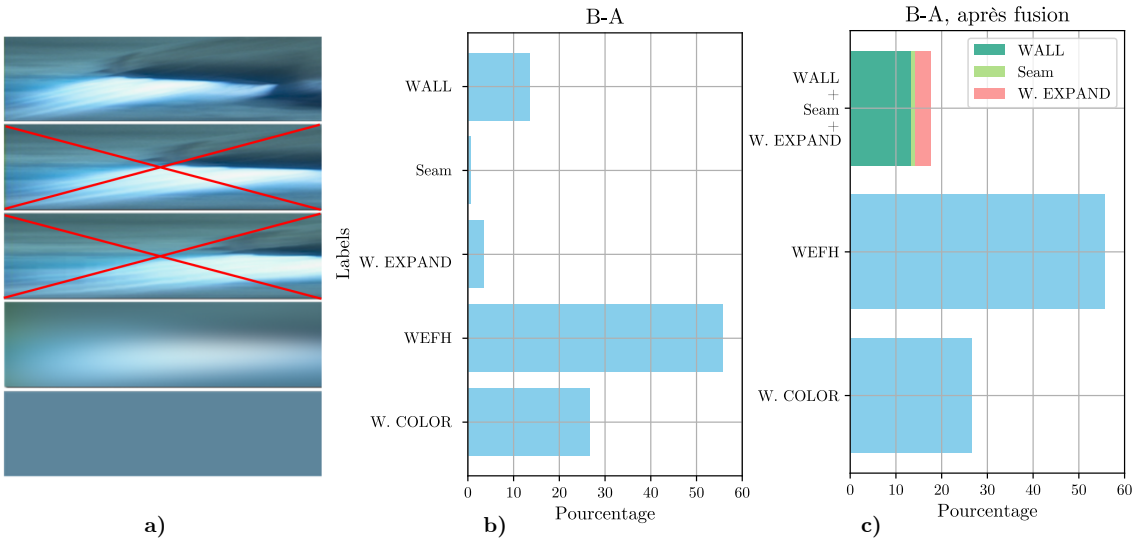


FIGURE 3.15 – **Distribution des labels de notre base de données**, c'est-à-dire des traitements choisis par les experts, et fusion des labels. a) Un exemple visuel des traitements d'images, b) la distribution des occurences dans les bandes-annonces. On peut voir une base de données très fortement déséquilibrée. c) La fusion des choix `WALL`, `SEAM` et `WALL EXPAND` comme un seul label « Détail » est la solution choisie pour permettre un entrainement correct du modèle.

compenser ce manque. En collaboration avec les experts, nous avons validé une série de transformations d'images pouvant être appliquées aux plans cinématographiques sans altérer la vérité terrain. Ces transformations sont visibles sur la figure 3.16 et incluent : modification de teinte, de saturation et de luminosité dans l'espace TSL (teinte, saturation, luminosité) [19], augmentation du contraste, rotation de l'image entre -5 ° et $+5$ °, et retournement vertical. Les bornes précises de ces opérations ont été établies empiriquement avec l'aide des experts. Le vide laissé par la rotation de l'image est comblé par le miroir de l'image à son extrémité.

Plusieurs travaux suggèrent qu'il est bénéfique d'introduire les transformations faibles en début d'entrainement, puis de renforcer progressivement leur intensité et leur variété. Cette approche s'inscrit dans le cadre du *curriculum learning*, une méthode d'apprentissage inspirée du développement humain, où un modèle est d'abord exposé à des données d'entrainement simples, avant d'être confronté progressivement à des cas plus complexes. Dans la littérature scientifique, il est explicitement recommandé d'adopter cette montée graduelle en difficulté des augmentations en tant que stratégie d'optimisation efficace pour la généralisation[121]. Ce principe est d'ailleurs commun dans les recherches et les entrainements de modèles d'IA non supervisés, qui structurent les représentations en apprenant d'abord sur les régularités simples avant de capturer des relations plus complexes, sans supervision explicite [20]. On applique ainsi une stratégie progressive d'augmentation où

19. le nom d'origine est HSL pour *Hue, Saturation, Lightness*

20. Les travaux de ce manuscrit concernent tous des IA supervisées. Le domaine de recherche des IA faiblement supervisées ou non supervisées s'intéresse à l'entrainement d'IA sans vérité terrain c.-à-d. sans avoir de label. L'IA doit comprendre à détecter des patterns sans avoir de réponse humaine pour le guider durant l'entrainement. Un exemple connu d'IA non supervisée peut être une galerie photos d'un smartphone. Bien que l'IA n'ait jamais eu accès aux noms des personnes, elle classe les photos dans des dossiers différents en regroupant les visages qu'elle identifie comme appartenant à une même personne.



FIGURE 3.16 – **Augmentation de données progressives au fil de l'entrainement.** Une même trame est altérée de plus en plus en fonction de l'avancée de l'entrainement de l'IA. Les changements de teintes, de rotations, de contraste, de luminance, et de saturation sont déterminés aléatoirement au sein d'intervalles de plus en plus importants. Les retournements verticaux sont aussi possibles avec une probabilité de 0,5. (Crédit : auteur [M21])

l'entrainement commence par des transformations légères (retournement vertical), puis l'intensité d'autres variations (rotations, fluctuations de couleur, saturation, contraste, luminosité) augmente au fil des *epochs* [21]. Cette montée en complexité est régulée par un paramètre de difficulté D, croissant avec l'avancée de l'entrainement, afin de guider le réseau vers une robustesse croissante.

3.6.2 Détails d'implémentation

Alors que les dorsales utilisés ont été entrainés sur des images carrées recadrées, notre tâche repose davantage sur l'analyse des bords gauche et droit de l'image, zones critiques où s'opèrent les discontinuités avec les extensions latérales. De ce fait, il est estimé que le recadrage central des images vidéo nuirait fortement à l'apprentissage du modèle. Lorsque cela est possible, le choix est donc entrepris de redimensionner les images vidéo à 112×267 (rapport cinéma conservé) sans recadrage. I3D et TSP tolèrent ce format d'entrée, mais ce n'est pas le cas de l'implémentation C3D utilisée ici, qui conserve donc un recadrage central en 112×112.

Les expériences ont été menées sur une GPU NVIDIA Quadro RTX 4000 disposant de 8 Go de VRAM [22]. À l'issue de tests empiriques, le taux d'apprentissage optimal pour le classifieur a été fixé à 1×10^{-3}, à l'exception de C3D pour lequel il a été abaissé à

21. Une *epoch*, terme d'origine anglaise calqué sur la même étymologie que notre mot français « époque », est une unité de progression d'un entrainement. Elle correspond au passage de toute la base de données une fois dans le modèle d'IA.

22. La VRAM est la RAM, ou mémoire vive, dédiée à la carte graphique (GPU).

1×10^{-4} afin de garantir une phase d'apprentissage stable. L'optimiseur utilisé est Adam, associé à un plan de décroissance du taux d'apprentissage à chaque *epoch* selon la formule $lr = lr \times 0{,}95$. La fonction de perte choisie est l'entropie croisée, largement répandue pour les tâches de classification.

L'entrainement s'effectue sur 50 *epochs*, avec 9 répartitions aléatoires fixées du jeu de données entre entrainement et validation. Afin d'assurer une couverture suffisante de validation, au moins 20 % des bandes-annonces sont réservées à la validation dans chaque répartition. Une séparation stricte est maintenue entre les bandes-annonces des ensembles d'entrainement et de validation afin d'éviter toute corrélation indésirable.

En dépit de la fusion de trois traitements en un seul label, le déséquilibre de classes demeure un enjeu important. L'augmentation de données pour y remédier n'est efficace que si elle s'appuie sur une stratégie de pondération des échantillons lors du tirage [122] [123], utilisée notamment pour pallier les déséquilibres de classes. Cette stratégie d'échantillonnage est mise en œuvre à chaque *epoch*. Elle consiste à supprimer aléatoirement des bribes vidéo associées aux labels surreprésentées dans les ensembles d'entrainement et de validation, jusqu'à atteindre un équilibre parfait entre les classes [122]. Cette technique permet de minimiser la perte de performance due à un jeu de données déséquilibré.

3.6.3 Résultats et discussions

TABLE 3.2 – Précision (taux de prédiction correcte) des différents modèles d'extracteurs de caractéristiques temporels (squelette, ou dorsale), selon les différentes configurations de classifieurs possibles.

Classifieur \ Dorsale	TSP	C3D	I3D
Par défaut	**68,02 %**	54,99 %	65,44 %
Augmentation de données	**68,85 %**	52,69 %	68,18 %
Convolution	**66,92 %**	54,24 %	66,91 %
Drop out	**70,32 %**	46,97 %	66,42 %
SITI	**69,95 %**	56,97 %	68,94 %
Double-flux	**69,53 %**	57,42 %	68,48 %
Drop Out sur le 2nd flux	**69,66 %**	55,14 %	67,58 %

Le tableau 3.2 présente les performances individuelles des modèles pour chaque configuration testée isolément, tandis que le tableau 3.3 illustre les performances obtenues lorsqu'on combine plusieurs configurations. Dans les deux tableaux, les cellules vertes indiquent une amélioration par rapport au classifieur de référence, tandis que les cellules rouges signalent une dégradation. Le meilleur extracteur de caractéristiques (dorsale, ou squelette) pour une configuration donnée est mis en gras.

Une observation notable issue de ces résultats est que C3D, bien qu'utilisé comme modèle temporel pour notre tâche spécifique, affiche systématiquement des performances inférieures aux deux autres modèles, indépendamment de la configuration du classifieur. À l'inverse, TSP s'impose comme le modèle le plus performant pour notre cas d'usage, suivi de près par I3D. Cette différence pourrait s'expliquer par le fait qu'I3D soit entrainé sur des segments de 64 images, ce qui ne correspond pas de manière optimale aux longueurs de plans présentes dans notre jeu de données (voir figure 3.20).

TABLE 3.3 – Précision (taux de prédiction correcte) des différents modèles d'extracteurs de caractéristiques temporels, selon les différentes combinaisons de configurations de classifieurs possibles.

Aug. Données	Drop out	SI TI	Double-flux	Drop out 2nd flux	TSP	C3D	I3D
✓	✓				**69,56 %**	47,78 %	67,93 %
	✓	✓			**71,72 %**	53,03 %	69,12 %
	✓	✓	✓		**71,21 %**	50,11 %	67,42 %
	✓	✓	✓	✓	**69,89 %**	48,99 %	67,16 %
✓		✓			**71,43 %**	56,40 %	68,69 %
✓		✓	✓		68,21 %	57,88 %	**68,85 %**
✓		✓	✓	✓	**68,28 %**	54,90 %	68,14 %
✓	✓	✓			**70,88 %**	54,48 %	68,87 %
✓	✓	✓	✓		69,17 %	51,52 %	**70,54 %**
✓	✓	✓	✓	✓	69,28 %	47,69 %	**69,70 %**

Lors des tests menés sur chaque configuration séparément (tableau 3.2), nous avons observé que l'augmentation de données, l'ajout des descripteurs SI et TI et l'utilisation du dropout améliorent significativement les performances des modèles utilisant TSP et I3D comme extracteurs de caractéristiques. Le dropout procure le gain le plus important avec les caractéristiques TSP, l'ajout séparé du flux SITI bénéficie le plus à C3D, tandis que la concaténation du flux SITI avec les caractéristiques I3D entraine l'amélioration la plus marquée dans ce cas. Ces résultats soulignent l'efficacité de ces techniques dans l'amélioration des performances en fonction des architectures utilisées.

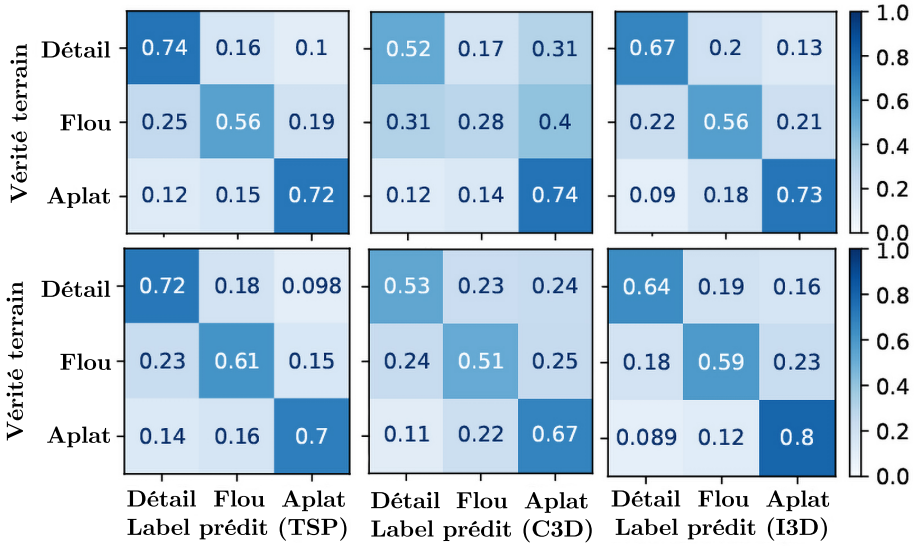


FIGURE 3.17 – **Matrices de confusion entre les étiquettes prédites et les annotations de référence** : configuration MLP par défaut (en haut), meilleure configuration MLP (en bas).

Les matrices de confusion du classifieur par défaut sont présentées en haut de la figure 3.17. Il apparait clairement que C3D ne fournit pas des caractéristiques suffisamment pertinentes

(au regard de notre tâche) pour permettre au classifieur de distinguer correctement les labels `flou`. Bien que ce label soit le plus difficile à prédire correctement pour tous les extracteurs de caractéristiques, il atteint néanmoins une précision de 56 % avec TSP et I3D. La principale différence entre les précisions d'I3D et de TSP réside dans leur capacité à prédire correctement les labels `detail`, avec une précision de 74 % pour TSP contre 67 % pour I3D.

Les meilleures configurations des modèles, listées dans le tableau 3.3, impliquent une combinaison de plusieurs techniques. Pour TSP, la configuration optimale est SITI+Dropout, atteignant une précision de 71,72 %. Pour I3D, la meilleure configuration inclut SITI+augmentation de données+dropout+2stream, avec une précision de 70,54 %. Les matrices de confusion correspondant à ces meilleures configurations sont visibles en bas de la figure 3.17. Il est important de noter que, dans ces cas, la prédiction du label `flou` est améliorée pour tous les extracteurs de caractéristiques.

Enfin, pour un contrôle qualité en situation réelle, une bande-annonce du film « Shazam ! La Rage des Dieux » a été entièrement montée en postproduction ICE en simulant le rendu des effets prédits par R-ICE sans intervention humaine.

Il a été remarqué que l'erreur commise sur un plan n'était pas toujours du même cout. Par moments, la prédiction faite était indéniablement incohérente avec la scène. Cependant, dans plus de la moitié des cas d'erreurs, les observateurs lambda ne percevaient pas clairement ce qui rendait le choix de R-ICE erroné. Pour eux, ce choix apparaissait simplement différent, mais tout aussi crédible que celui de l'équipe ICE. Deux de ces exemples sont visibles en figure 3.18. Seule l'équipe ICE elle-même pouvait identifier l'erreur et en expliquer la nature.

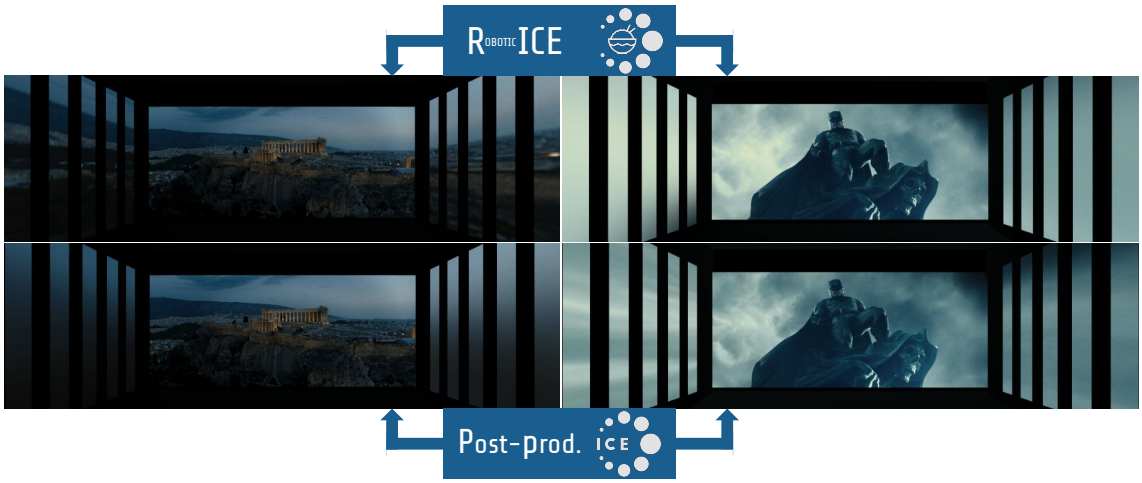


Figure 3.18 – **Illustration des mauvaises prédictions de R-ICE lorsque celles-ci semblent tout à fait plausibles pour un œil non expert.** Sur le plan à gauche, la prédiction de R-ICE (en haut) de mettre du détail sur un plan général concorde avec les règles générales, alors que l'équipe ICE préfère mettre du flou. Sur le plan de droite, R-ICE prédit du flou alors que l'équipe ICE laisse du détail. À droite, aucun de ces choix ne brise l'immersion.

3.6.4 Étude d'ablation

Perceptron multicouches plus profond

L'ajout de couches affines supplémentaires, accompagnées de couches ReLU et de dropout si nécessaire, n'a entrainé aucune amélioration des performances des modèles testés dans cette étude. Cela suggère que les trois couches affines présentes dans notre MLP sont suffisantes pour capturer toutes les associations de caractéristiques possibles, du moins dans la limite du nombre d'epochs utilisé pour l'entrainement.

Classifieurs de régression traditionnels

Plutôt qu'un MLP en tant que classifieur, des modèles de régression traditionnels ont été testés en utilisant les caractéristiques extraites par TSP, concaténées avec nos descripteurs sur-mesure. Les résultats ont montré une précision de 60,40 % pour le Gradient Boosting, 65,70 % pour la Forêt Aléatoire et 57,90 % pour le SVR. Ces performances se sont révélées inférieures à celles obtenues avec l'utilisation du MLP comme classifieur.

Le bourrage de bits à zéro

Le bourrage de bits à zéro ou *zero padding* est le fait de combler l'espace manquant par une valeur nulle. Dans notre contexte, puisque notre taille de bribe est fixé à 16 trames, une décision doit être prise sur le surplus de trames lorsqu'un plan n'est pas divisible par 16. Soit ce reste est complété par un bourrage à zéro, comme illustré sur la figure 3.19, soit ce reste est supprimé.

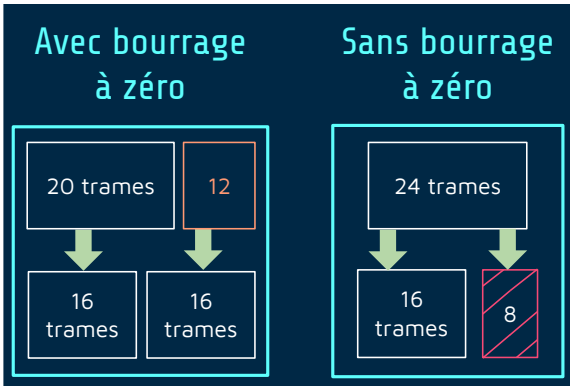


FIGURE 3.19 – Le **bourrage à zéro** permet de combler des données via l'ajout de données factices (*dummy data*) à zéro afin qu'elles aient la taille requise pour leur utilisation.

Comme l'on pouvait s'y attendre, les modèles sans rééchantillonnage mais avec bourrage obtiennent systématiquement de moins bonnes performances que ceux sans bourrage. Il n'est donc pas surprenant que la configuration avec la dorsale TSP, une entrée vidéo rééchantillonnée de 24 i/s à 15 i/s et avec bourrage, suive cette même tendance.

Ce remplissage a conduit à une dégradation des performances, quelle que soit la configuration du classifieur ou du modèle utilisé. À partir des caractéristiques extraites par TSP, la diminution de précision était de 0,93 % pour le MLP par défaut, 1,87 % pour le MLP avec augmentation de données, 2,59 % pour le MLP avec dropout, et 2,95 % pour le MLP avec SITI.

Rééchantillonnage

Nos données originales sont basées sur le standard cinématographique de 24 images par seconde.

Étant donné que TSP sous-échantillonne ses vidéos à 30 i/s pour un entrainement à 15 i/s, une approche possible consisterait à rééchantillonner nos propres vidéos de 24 i/s à 15 i/s.

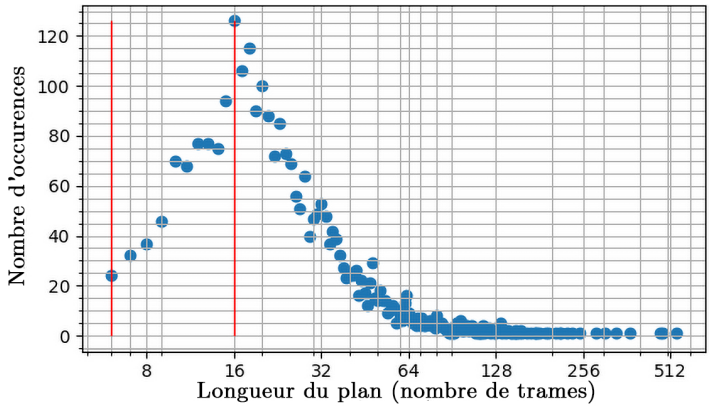


FIGURE 3.20 – **Longueur des plans de notre jeu de données de bandes-annonces.** La première ligne rouge définit les plan les plus courts (6 trames) et la deuxième ligne rouge correspond à la longueur de plan la plus fréquente (16 trames).

Toutefois, comme illustré en figure 3.20, la distribution des longueurs de plans dans notre jeu de données révèle que les séquences de 16 images sont les plus fréquentes.

Cette observation peut être attribuée au fait que les données disponibles sont des bandes-annonces de films plutôt que des films complets. Ainsi, lors du rééchantillonnage des données à 15 i/s, les extraits résultants doivent être complétés par du bourrage pour atteindre une longueur de 16 images. Sans ce bourrage, le jeu de données serait insuffisant. Afin d'évaluer l'impact de ce remplissage, un test complémentaire utilisant TSP sans rééchantillonnage mais avec bourrage a été inclus dans l'étude d'ablation mentionnée précédemment.

TSP à 15 i/s avec bourrage obtient la plupart du temps de meilleures performances que TSP à 24 i/s avec bourrage pour une configuration donnée : −0,04 % de précision pour le MLP par défaut mais +1,77 % pour le MLP avec augmentation de données, +0,20 % pour le MLP avec dropout, et +2,93 % pour le MLP avec SITI. Cette observation est probablement liée au fait que TSP soit entrainé sur des vidéos à 15 i/s dans l'article d'origine. On peut ainsi comprendre que TSP à 15 i/s sans bourrage constituerait une option viable pour améliorer les performances : le rééchantillonnage permettrait d'être conforme à la fréquence d'images utilisée lors de l'entrainement du modèle et donc des conditions optimales du modèle. Il nous est impossible de se remettre dans ces conditions dû à la nature de notre jeu de données. Une partie trop importante de notre jeu de données ne serait plus prise en compte si nous rééchantillonnions nos trames sur l'axe temporel. Elle passerait donc en dessous des 16 trames réglèmentaires après rééchantillonnage, ce qui signifierait que ces plans ne seraient pas pris en compte (voir figure 3.20).

Recadrage central

TABLE 3.4 – Comparaison des performances de notre classifieur selon les différentes conditions d'image proposées dans la littérature (TSP, I3D, C3D) et celles conçues pour cette thèse.

Origine du prétraitement	**Redimensionner le plus petit côté**	**Puis recadrer un carré central**	**TSP**	**I3D**	**C3D**
Nôtre	112 (×267)	✗	67,36 %	65,44 %	✗ (impossible)
TSP (original)	128 (×171)	112×112	63,89 %	60,94 %	54,17 %
C3D (original)	112 (×267)	112×112	63,64 %	60,69 %	54,99 %
I3D (original)	256 (×611)	224×224	63,26 %	62,81 %	✗ (impossible)
Nôtre	224 (×535)	✗	65,49 %	67,82 %	✗ (impossible)

Le tableau 3.6.4 compare les performances de TSP, I3D et C3D selon différentes stratégies de redimensionnement et de recadrage spatial des entrées vidéo. L'objectif est de tester l'hypothèse selon laquelle les bords de l'image contiennent des informations importantes pour la classification, et que leur suppression par un recadrage central systématique nuit aux performances.

Les résultats montrent que, pour TSP et I3D, éviter le recadrage central (✗) après un simple redimensionnement (ligne 1) conduit à de meilleurs résultats que les configurations classiques avec recadrage (lignes 2 à 4). Par exemple, TSP atteint 67,36 % sans recadrage, contre 63,89 % dans sa configuration d'origine. De même, I3D passe de 60,94 % à 65,44 %, confirmant que préserver les bords améliore la performance. I3D progresse davantage (jusqu'à 67,82 %) lorsqu'on l'alimente avec sa résolution native (224×224 sans recadrage), mais cela implique une entrée 32 fois plus volumineuse que celle de TSP (224×224×64 opposé à 112×112×16), alourdissant considérablement l'inférence. Enfin, C3D ne supporte que les entrées 112×112, ce qui empêche l'exploration d'autres tailles ou de versions sans recadrage. Ces observations soutiennent notre hypothèse de départ : les stratégies de prétraitement dominantes dans l'entrainement des modèles de la littérature, c'est-à-dire basées sur un recadrage central, éliminent des indices visuels importants à la décision. Leur remise en question améliore nettement les performances.

3.7 ICE-AGE : Extension Graphique basée sur l'Attention

3.7.1 Motivations

R-ICE a permis d'ouvrir la voie dans la détermination de contenus à afficher dans les panneaux ICE. Grâce à cela, nous avons pu examiner en détail la manière dont les décisions étaient prises pour chacun des plans. Cela nous a également permis d'identifier des cas non anticipés, comme la corrélation entre la décision gauche et celle de droite, qui ne coïncident pas toujours. D'autres situations complexes sont aussi apparues, telles que les divisions, les changements de teinte, les regroupements, ou encore la présence de plusieurs bandes-annonces pour un même film, dont il faut tenir compte lors de la séparation. Enfin, le système de traitement différent entre les bandes-annonces et les films nous a empêchés de progresser davantage vers l'analyse des films. On a pu également décider d'une manière

d'agencer nos données, par une structure multiéchelle, que l'on saura réutiliser à bon escient pour les autres travaux.

R-ICE nous a permis d'explorer les données, de comprendre les enjeux applicatifs et d'identifier les limites de notre première approche. Malgré des prédictions satisfaisantes de 72,71 %, l'expérience acquise avec ces travaux précédents a mis en lumière plusieurs pistes d'amélioration, tant sur le plan méthodologique que conceptuel. Ces nouveaux travaux ICE-AGE visent à approfondir l'explicabilité du modèle et à proposer une architecture repensée intégralement pour correspondre parfaitement au contexte et dépasser les performances précédentes.

Nous avons vu dans la section 2.2.4 que les informations à prendre en compte dans la vidéo pour une décision ICE sont très nombreuses. La profondeur dans l'image, la profondeur de champ ainsi que la présence d'un bokeh qui en résulte, l'échelle de plan, le nombre de personnages et d'objets, la présence de discussion ou non, la quantité de mouvement, qu'ils proviennent des personnages, d'objet, de la scène entière ou de la caméra sont tous des concepts qui peuvent intervenir dans une décision pour ICE. Un problème qui peut toucher les modèles de réseaux de neurones, et notamment les réseaux convolutifs comme celui mis en place **R-ICE** est le fait qu'avoir une grande quantité de données d'entrée pour un modèle peut légèrement noyer le modèle et rendre son entrainement plus compliqué pour démeler les informations pertinentes des autres, selon le contexte. Le modèle mettra ainsi plus de temps a converger, voire ne convergera même plus du tout. C'est pour cette raison (en plus des limites computationnelles et de ressources mémoires) qu'on a pu, avec le modèle R-ICE, ne donner que 16 trames d'entrées. Tout ce qui apparaissait après ces 16 trames, soit 1,875 seconde selon l'empreinte temporelle de R-ICE, ne pouvait pas être pris en compte par ce dernier pour sa décision, ce qui reste limité.

C'est là que le mécanisme d'attention, le paradygme apparu après la convolution peut nous aider. Les mathématiques qu'il cache sont plus complexes que la convolution, mais elles permettent de créer un système de relation entre les données sans que chacune intervienne nécessairement dans chaque inférence. Ce système est apparu pour la compréhension du langage, ou le NLP [23], et ce n'est pas anodin. Dans certaines phrases, les mots trop polyvalents n'apportent pas d'informations. Dans d'autres, ils sont déterminants pour savoir si on parle « de » quelqu'un ou « à » quelqu'un. Leurs sens et leurs importances sont déterminés par leurs contextes. Si l'on ajoute cet aspect au fait que trop de données d'entrées peut rendre moins efficace l'entrainement, traiter le langage écrit, qui peut contenir des phrases de longueur indéterminée, n'était pas fait efficacement avant l'apparition du mécanisme d'attention. Ce dernier permet justement de considérer des sens différents d'un mot et les relations différentes qu'il peut entretenir avec d'autres.

3.7.2 Mécanisme d'attention

Pour comprendre le mécanisme d'attention, il est plus pédagogique de passer par le domaine du NLP, qui sera plus naturel à la compréhension. Ce mécanisme fonctionne avec le plongementPlus communément appelé *Embedding* dans des espaces à hautes dimensions. Ces espaces sont aussi appelés espaces latents. Est dite « latente » une chose qui existe de manière diffuse, sans être apparent, mais qui peut à tout moment se manifester. Le plongement sémantique de différents mots consiste à leur attribuer des vecteurs en établissant une corrélation entre la proximité spatiale de ces mots dans l'espace et leur rapprochement

23. NLP : traitement du langage naturel, ou Natural Language Processing

sémantique. On peut s'attendre à ce que les mots « Pain » et « Brioche » soient représentés par des vecteurs relativement proches. Cette idée de plongement sémantique nous vient du linguiste Zellig Harris en 1954 [124], bien avant que son utilité soit mise à profit pour les IA génératives d'image ou de texte afin de faire le pont entre les paramètres de l'IA et la sémantique linguistique humaine. À cette époque, le linguiste établissait le plongement sémantique des mots via le contexte dans lesquels ils apparaissent, ce qui s'appelle l'« Hypothèse Distributionnelle » et les « Sémantiques Statistiques ».

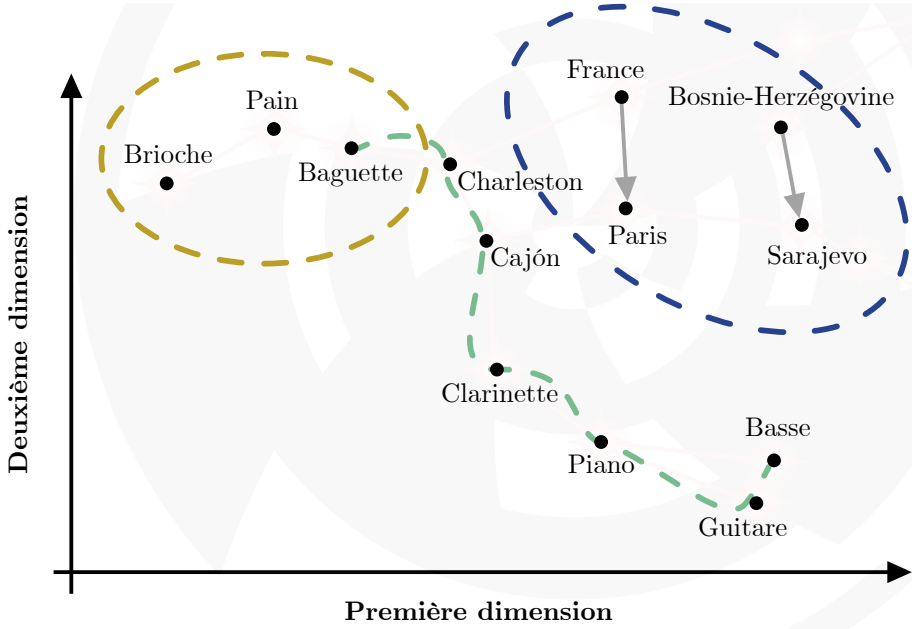


FIGURE 3.21 – **Plongement sémantique de mots.** On peut remarquer que l'addition d'un vecteur à un mot (France → Paris) correspond à une relation similaire lorsque l'addition de ce vecteur est effectuée sur un autre mot (Bosnie-Herzégovine → Sarajevo). Le pain, la brioche et la baguette forment une grappe (*cluster*). Du côté des instruments, les cordes pincées sont très proches (guitare et basse), vient ensuite les cordes frappées (piano), les vents (clarinettes) et enfin les percussions (Cajón, Charleston et Baguette). L'espace latent permet une représentation des double-sens, avec Baguette étant proche du Pain et du Charleston. (Crédit : auteur)

Dorénavant, le plongement lexical n'est plus basé sur des matrices de cooccurences, mais est appris automatiquement en apprentissage profond. Le plongement dépend d'une couche appelée couche de plongement qui se situe en amont du modèle.

Cette couche de plongement est mathématiquement équivalente à une matrice d'encodage *un* parmi n multipliée par une couche affine entrainable, en prenant moins de ressources mémoires et computationnelles[125].

Dans notre cas, les extracteurs de caractéristiques des modèles particuliers seront nos plongeurs dans des espaces latents entrainés chacun différemment. En clair, notre extracteur prévu pour la génération de carte de profondeur servira à un plongement particulier, notre extracteur prévu pour la génération de flot optique servira à un autre plongement, etc. C'est sur ces caractéristiques que nous allons calculer l'attention croisée.

Pour percevoir l'intérêt de l'attention, le cas de la traduction de texte est intéressant [126], dû aux différences de structures grammaticales entre les langues.

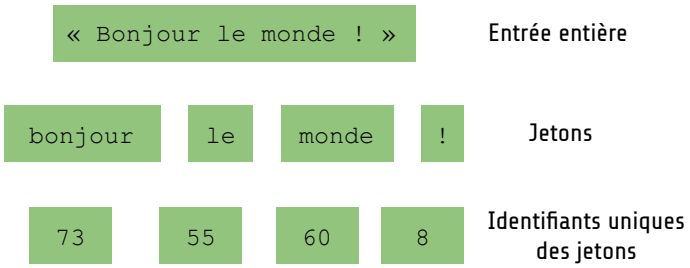


FIGURE 3.22 – **Exemple de jetonisation d'une phrase.** Toute entrée d'un modèle de NLP, c'est-à-dire tout mot (et caractère d'une phrase) connu et inconnu doit être manipulé de manière à être intelligible pour un système informatique comme notre réseau de neurones. La jetonisation transforme donc les mots dans les identifiants composés de chiffres qui les représentent dans un dictionnaire préalablement choisi. Le point d'exclamation, ayant un sens sémantique dans une phrase, a également droit à son identifiant. Un mot inconnu est le plus souvent sous-découpé en jetons de préfixes et suffixes connus. Anticonstitutionnellement pourrait ainsi être découpé en quatres jetons anti-constitution-nelle-ment

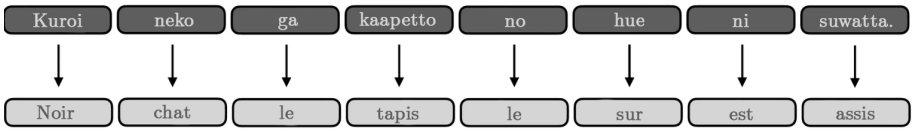


FIGURE 3.23 – Traduction mot à mot d'une phrase japonaise vers le français. Le texte traduit n'a plus de sens.

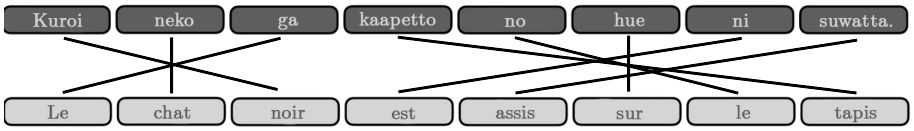


FIGURE 3.24 – Traduction efficace d'une phrase vers le français. Pour être traduite, la phrase doit prendre en considération tout son contexte, et les mots de la langue cible dépendent des mots qui apparaissent plus tôt ou plus tard dans la phrase de la langue d'origine.

L'attention [127][128] permet de capturer ces dépendances de mots qui ne sont pas à la même position, même avec un changement de position très variable, grâce à cette attention qui s'adapte.

Pour construire ce mécanisme, plusieurs étapes sont nécessaires. Le but est d'obtenir des poids d'attention, de chaque mot contextuel par rapport au mot d'intérêt. Sur la figure 3.25, les mots concernés par le module d'attention sont présents, les mots sont ensuite

Le plongement sémantique d'un mot donne du sens formel interprétable par la machine à ce mot, cependant ce dernier est alors plongé dans un espace très général de tout le vocabulaire possible. Le mécanisme d'attention vise à rendre plus spécifique ce plongement en associant un sens et des relations de ce mot par rapport au reste de son contexte, ce plongement plus spécifique s'appelle le vecteur de contexte. Ainsi le mot « Ton » suivi de « métier » aura une forte relation contextuel avec ce dernier, là où le plongement sémantique sans contexte de « Ton » ne permet pas d'obtenir beaucoup d'information. Il est la combinaison pondérée des vecteurs de plongement de tous les mots du contexte, par leurs poids d'attention respectifs selon le mot considéré.

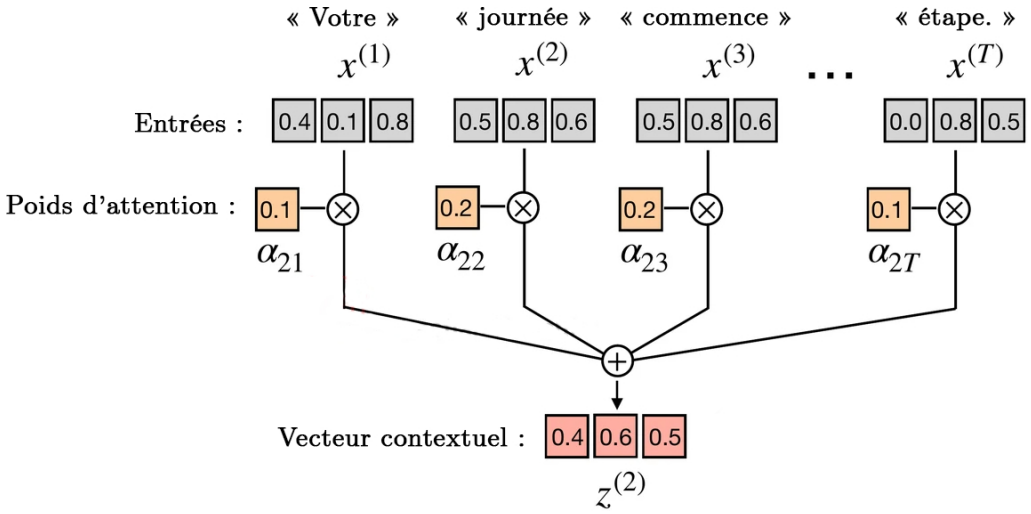


FIGURE 3.25 – **Construction des vecteurs contextuels comme première étape du mécanisme d'attention.** (Traduite depuis : [125])

Par notation mathématique, considérons avec $\tilde{s}_k^{(i)}$ le score d'attention simplifiée qui mesure le degré d'importance du token plongé $\mathbf{x}^{(k)}$ sur le token $\mathbf{x}^{(i)}$. On écrit :

$$\underbrace{\tilde{s}_k^{(i)}}_{1\times 1} = \underbrace{\mathbf{x}^{(i)}}_{1\times L} \underbrace{(\mathbf{x}^{(k)})^\top}_{L\times 1} \tag{3.7}$$

Les vecteurs $\mathbf{x}^{(i)}$ et $\mathbf{x}^{(k)}$ représentent respectivement les plongements sémantiques des i-ème et k-ième tokens, chacun sous forme d'un vecteur de taille $[1 \times L]$, où L correspond à la dimension de l'espace latent.

L'ensemble des scores de tous les tokens peut être calculé simultanément par produit matriciel :

$$\underbrace{\begin{bmatrix} \tilde{s}_1^{(1)} & \cdots & \tilde{s}_n^{(1)} \\ \vdots & \ddots & \vdots \\ \tilde{s}_1^{(n)} & \cdots & \tilde{s}_n^{(n)} \end{bmatrix}}_{\tilde{\mathbf{S}}} = \underbrace{\begin{bmatrix} \mathbf{x}^{(1)} \\ \vdots \\ \mathbf{x}^{(n)} \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} (\mathbf{x}^{(1)})^\top & \cdots & (\mathbf{x}^{(n)})^\top \end{bmatrix}}_{\mathbf{X}^\top} \tag{3.8}$$

Ainsi, $\tilde{\mathbf{S}}$ contient les scores d'attention simplifiée non normalisés entre tous les tokens de la séquence.

Le but est d'obtenir des scores d'attention, qui nous donnent l'importance d'un mot contextuel par rapport à un mot donné de la phrase. Ces scores, une fois normalisées donnent les poids d'attention. Ainsi a21 est l'importance du mot x1 sur le mot x2 pour comprendre x2. Une fois tous ces poids d'attention obtenus, un vecteur de contexte z2 est dans le même espace latent que x2 une fois plongé sémantiquement, mais il est maintenant remis en lien avec son contexte.

Pour obtenir des **poids d'attention normalisés**, on applique la fonction softmax sur chaque ligne i de $\mathbf{S}$:

$$\mathbf{a}^{(i)} = \text{softmax}(\mathbf{s}^{(i)}) \quad \text{avec} \quad a_k^{(i)} = \frac{\exp(s_k^{(i)})}{\sum_{j=1}^{n} \exp(s_j^{(i)})} \tag{3.9}$$

Ainsi, chaque vecteur ligne $\mathbf{a}^{(i)}$ contient les poids d'attention normalisés associés au token i et satisfait :

$$\sum_{k=1}^{n} a_k^{(i)} = 1 \tag{3.10}$$

Le vecteur de contexte associé au token i s'écrit alors :

$$\mathbf{z}^{(i)} = \sum_{k=1}^{n} a_k^{(i)} \mathbf{x}^{(k)} = \mathbf{a}^{(i)} \mathbf{X} \tag{3.11}$$

Le vecteur $\mathbf{z}^{(i)}$ est de même dimension que le token plongé $\mathbf{x}^{(i)}$, mais intègre désormais l'information contextuelle issue des autres tokens.

Pour l'instant, les poids d'attention sont largement simplifiés. En réalité, eux aussi sont issus d'un entrainement spécifique. On remarque que le token joue en effet trois rôles différents pour le calcul du vecteur contextuel. Il en a déjà deux pour le calcul de son coefficient d'attention : « quelle importance a le mot n°1 sur le mot n°2 ? ». Ces deux rôles distincts sont nommés respectivement la clé, qui fournit l'information contextuelle, et la requête, qui détermine quelles informations sont pertinentes. Les tokens plongés jouent un dernier rôle, celui d'être multiplié par le poids d'attention afin d'être sommé comme vecteur contextuel. Ce dernier rôle est la valeur. Ces trois rôles étant maintenant distincts comme sur la figure 3.26, on peut leur attribuer des matrices entrainables, initialisées aléatoirement, afin que les clés, requêtes et valeurs de chaque token plongés soient une altération entrainée spécifiquement de ce dernier.

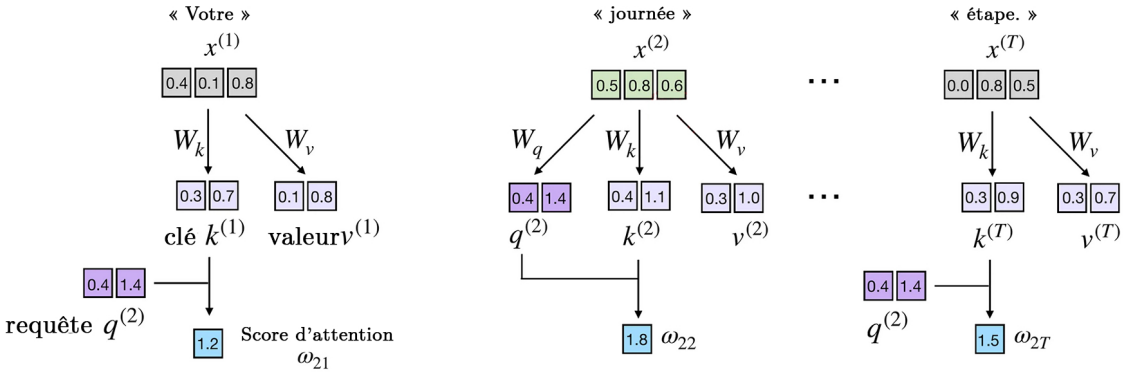


FIGURE 3.26 – **Mise en œuvre de matrices entrainables afin d'obtenir un poids d'attention efficace.** (Traduite depuis : [125])

Ces trois rôles sont obtenus via des projections linéaires entrainables :

$$\mathbf{q}^{(i)} = \mathbf{x}^{(i)} \mathbf{W}_q, \quad \mathbf{k}^{(i)} = \mathbf{x}^{(i)} \mathbf{W}_k, \quad \mathbf{v}^{(i)} = \mathbf{x}^{(i)} \mathbf{W}_v, \tag{3.12}$$

avec

$$\mathbf{W}_q, \mathbf{W}_k \in \mathbb{R}^{L\times d_k}, \quad \mathbf{W}_v \in \mathbb{R}^{L\times d_v}. \tag{3.13}$$

Les scores d'attention deviennent alors :

$$\mathbf{S} = \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}, \quad \text{où} \quad \mathbf{Q} = [\mathbf{q}^{(1)}; \ldots; \mathbf{q}^{(n)}], \quad \mathbf{K} = [\mathbf{k}^{(1)}; \ldots; \mathbf{k}^{(n)}], \quad \mathbf{V} = [\mathbf{v}^{(1)}; \ldots; \mathbf{v}^{(n)}]. \tag{3.14}$$

Après application de la softmax ligne par ligne :

$$\mathbf{A} = \text{softmax}(\mathbf{S}), \quad \text{avec chaque ligne} \quad \mathbf{a}^{(i)} = \text{softmax}(\mathbf{s}^{(i)}) \quad \text{et} \quad a_k^{(i)} = \frac{\exp(s_k^{(i)})}{\sum_{j=1}^{n} \exp(s_j^{(i)})}. \tag{3.15}$$

Le vecteur de contexte ajusté de chaque token est obtenu par :

$$\mathbf{Z} = \mathbf{A}\mathbf{V}, \tag{3.16}$$

où chaque ligne $\mathbf{z}^{(i)}$ combine les valeurs des autres tokens selon leur pertinence contextuelle apprise. Pour chacun des mots, on constitue une matrice de poids de requête, une matrice de poids de clé et une matrice de poids de valeur.

L'essentiel du mécanisme d'attention a été présenté, certaines améliorations restent importantes et sont applicables selon notre besoin de conception de l'IA. Le fait d'utiliser une attention à têtes multiples est incontournable : les matrices d'entrainements, comme $\mathbf{W}_q$, $\mathbf{W}_k$, $\mathbf{W}_v$ sont dupliquées afin de percevoir différentes relations distinctes. Cette amélioration est analogue au fait d'utiliser une multitude de filtres convolutifs par couche de convolution dans les CNN, et cela améliore grandement la performance du modèle. Les autres matrices du modèle sont adaptées en cohérence avec ce changement, comme pour Q, K, V, ainsi que Z comme le montre la figure 3.27.

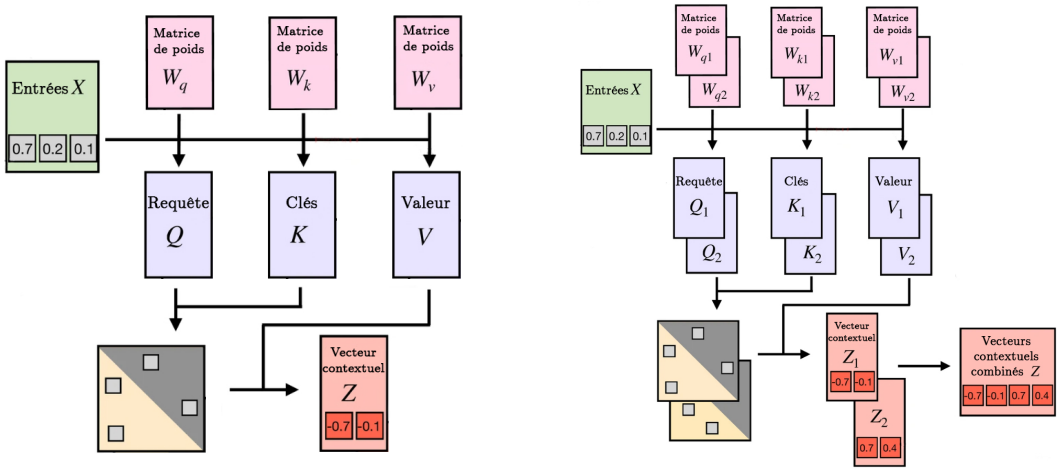


FIGURE 3.27 – **Mécanisme d'attention à tête multiple.** Il permet au modèle de se concentrer simultanément sur différentes parties de la séquence, en capturant plusieurs types de relations contextuelles entre les éléments. (Traduite depuis : [125])

3.7.3 Méthodologie et architecture

Grâce à l'étude de nos graphes sémantiques en Chapitre 2, il est évident que certaines données, que nous n'avions pas pu implémenter avec R-ICE, sont riches en information.

L'échelle de plan, sa carte de profondeur ainsi que son flot optique en font partie. Une échelle de plan large est plus propice à du détail. Une carte de profondeur, illustrée en figure 3.28 peut être porteuse de différentes informations utiles, de la présence d'éléments identifiés ou non à l'écran, mais aussi l'indication d'un cadre dans le cadre menant vers un paysage ouvert, ou bien de l'environnement fermé ou ouvert qui entoure la scène.



FIGURE 3.28 – **Exemple de carte de profondeur** obtenue sur un plan lors d'essais avec le modèle de profondeur P3Depth.

Bien que nous avions déjà une métrique corrélée à la quantité de mouvement avec la métrique TI dans R-ICE, elle ne nous permet pas d'avoir une information aussi détaillée qu'un flot optique, qui est un champ vectoriel de mouvement entre deux trames, donnant également la direction de ce mouvement. Un exemple de flot optique est représenté en figure 3.29

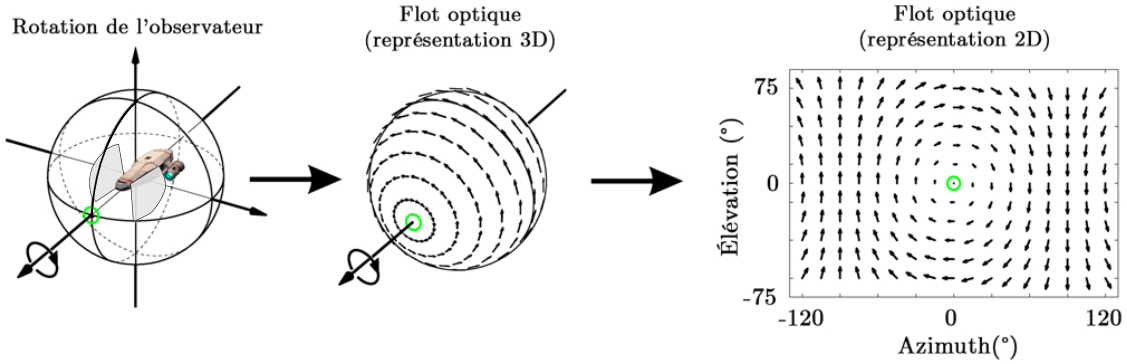


FIGURE 3.29 – **Illustration d'un flot optique**, en l'occurence qui correspond au champ de vision d'un vaisseau lorsque celui-ci effectue une bascule sur son axe de roulis. (Crédit : [M22])

Lorsqu'il est question d'un mouvement dans un espace continu, le champ vectoriel doit être discrétisé pour permettre cette représentation en vecteurs (direction et intensité), mais lorsqu'il est question d'une image comme pour nos données, cette approximation n'a pas lieu d'être, il est même plus lisible de convertir ces informations 2D en couleur de

l'espace TSV (teinte, saturation, valeur), avec la direction traduite en teinte et l'intensité en saturation (ainsi que la valeur fixée à son maximum), comme on peut le voir sur la figure 3.30.

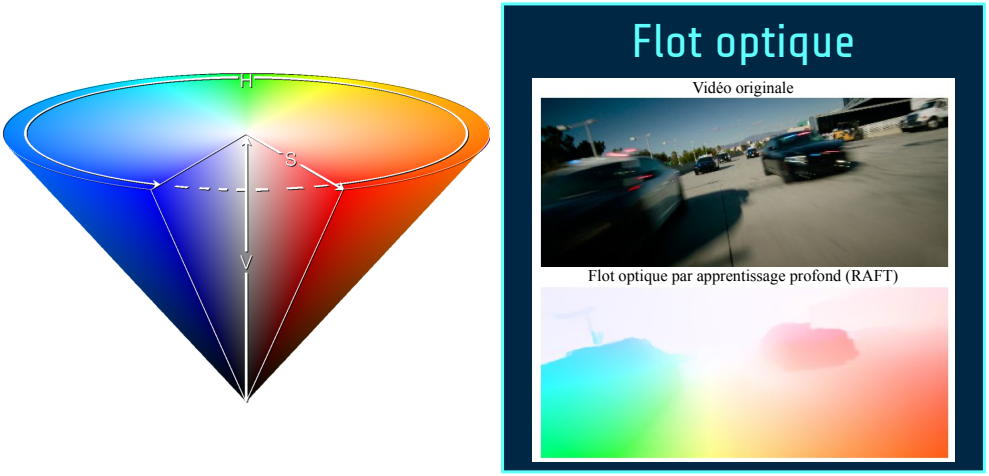


FIGURE 3.30 – **Représentation de l'espace de couleur TSV par un cône à gauche et son utilité pour représenter un champ vectoriel de flux optique à droite.** Le cyan représente un mouvement vers la gauche tandis que le rouge représente un mouvement vers la droite.

Le principe de base sur lequel notre architecture s'appuie est d'utiliser des caractéristiques issues de ces modalités, puis de les confronter par un système d'attention à des caractéristiques génériques, comme illustré en figure 3.31. Dans la littérature, lorsqu'il y a un besoin de caractéristiques d'images génériques, ces dernières sont souvent issues de modèles de classification. Un modèle de ResNet50, entrainé sur ImageNet est de mise pour cela. Ainsi, il est possible de voir ce mécanisme d'attention croisée comme la question : « Comment ces caractéristiques particulières nous aident-elles à prédire notre tâche de classification ICE ? ».

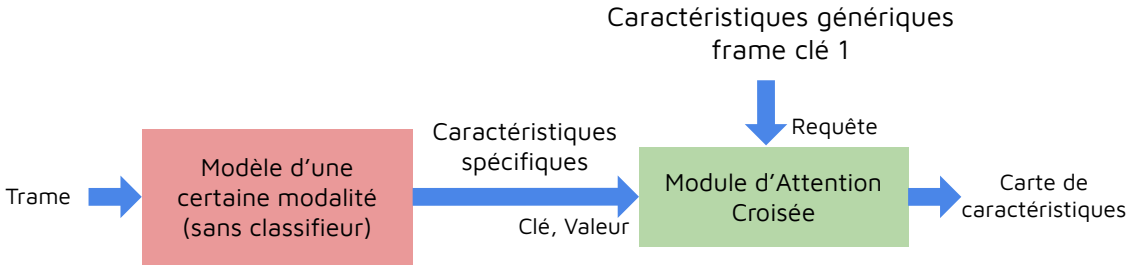


FIGURE 3.31 – Module d'attention croisée

L'étape suivante est de combiner le système unimodal précédent avec d'autres modalités intéressantes. Les premières modalités à intégrer vont être, comme conseillé plus haut, un modèle de détection d'objet sans classifieur, afin de maintenir des cartes de caractéristiques réexploitables plutôt que des tentatives de prédiction qui sont des données trop pauvres en information. Ce modèle sans classifieur sera un squelette (ou dorsale de modèle) préentrainé sur le jeu de données d'objet communs : COCO, en l'occurence. Un modèle de classification d'échelle de plan ainsi qu'un modèle de prédiction de la profondeur auront aussi chacun leur propre voie de traitement, comme illustré sur la figure 3.32.

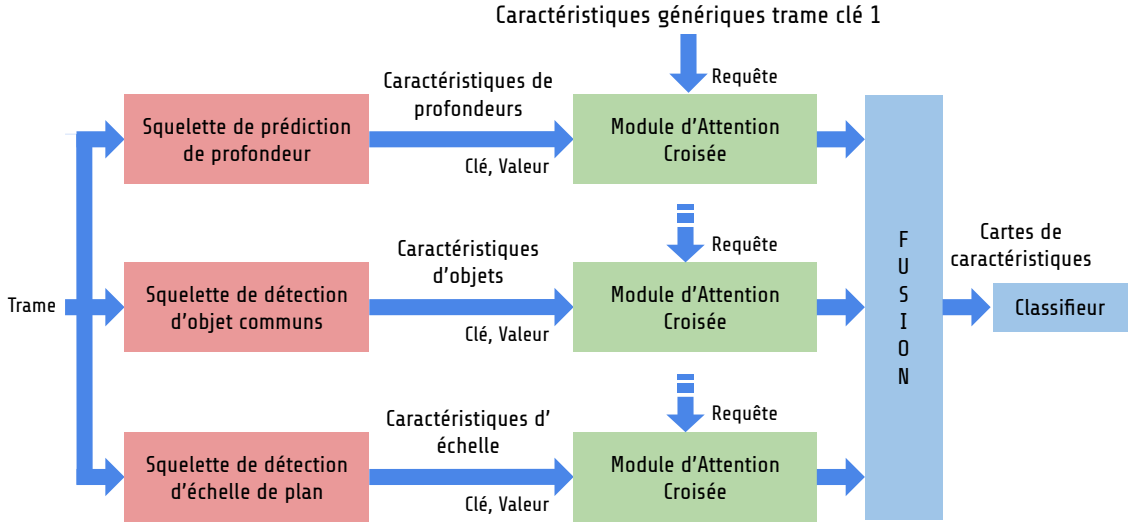


FIGURE 3.32 – Modules d'attention croisée venant de différentes modalités afin de classifier une trame.

Cependant, cette architecture ne concerne qu'une seule trame. Puisque nous travaillons en vidéo, il est nécesssaire de l'adapter avec un réseau de neurones récurrents comme classifieur sur ces cartes de caractéristiques, comme illustré sur la figure 3.33. Un aspect limitant des CNN était la fenêtre fixe de 16 trames, c'est pourquoi un LSTM, un réseau de neurones récurrents adapté aux entrées de grande taille, permet une classification efficace.

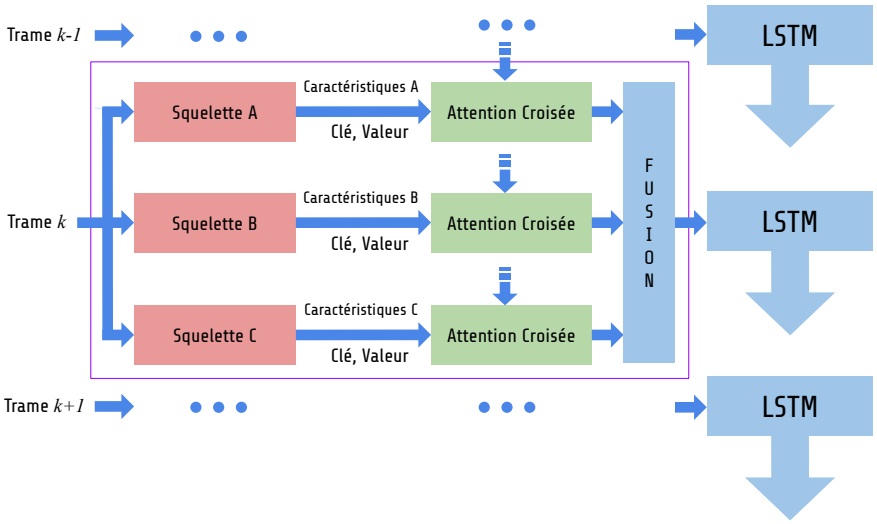


FIGURE 3.33 – Architecture globale de ICE-AGE, avec une attention croisée par modalité et par trame. Le classifieur est un LSTM qui considère plusieurs trames successives en entrée pour inférer sa sortie.

Avec une architecture comme ceci, le modèle devient très puissant, la modularité des différents éléments de l'architecture permet à la fois l'explicabilité mais aussi une modularité très importante. L'IA étant un domaine qui évolue extrêmement vite, ce serait dommage de ne pas en tirer parti en créant une architecture qui sera forcément obsolète d'ici quelques années. Avec notre architecture, les modèles intermédiaires peuvent être remplacés facilement par leurs homologues plus performants à l'avenir. La contrepartie

de ce modèle est sa complexité impraticable en l'état. Ces modalités sont évidemment importantes, mais inférer des données de haut niveau comme des cartes de profondeur (comme évoqué en section 3.7.3) sur chacune des trames n'est pas réaliste. En plus de cela, ces données sont redondantes d'une trame à l'autre. C'est pourquoi un sous-échantillonnage est nécessaire. En l'occurence, la solution optimale de sous-échantillonnage est d'obtenir les trames jugées les plus importantes, ce qui fait partie du résumé de vidéo, vu dans la Section 3.2. Notre solution est d'utiliser un algorithme de partitionnement adaptatif, la séparation de grappe par pic de densité (density-peak clustering) [129][130] sur des cartes de caractéristiques des trames, comme illustré sur la figure 3.34.

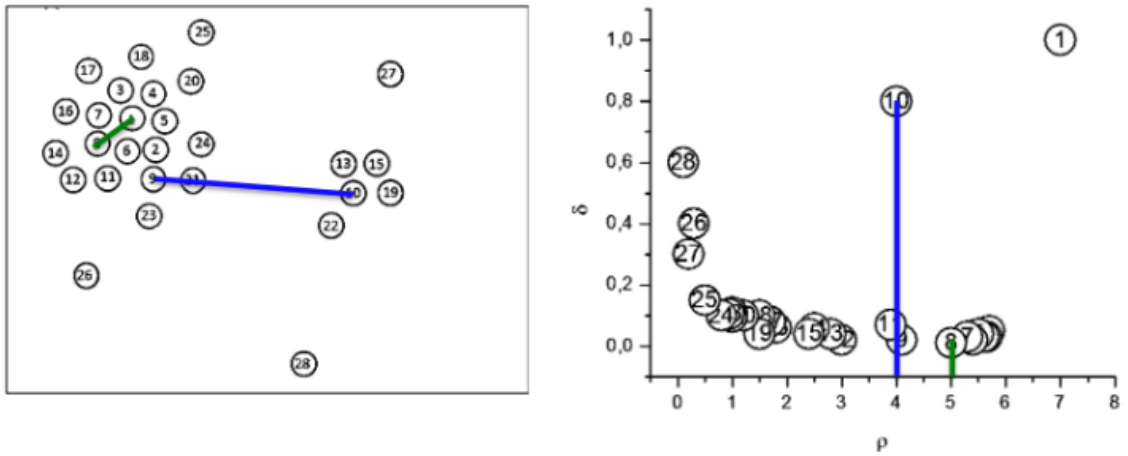


FIGURE 3.34 – **Graphiques de l'agorithme de séparation de grappe « Pic de densité ».** (Crédit : [129][130])

Afin d'être appliqué à plusieurs images pour s'en servir de résumé de vidéo, les images sont d'abord individuellement transformées dans un espace vectoriel réduit, au moyen d'un réseau convolutif de reconnaissance général d'image, le ResNet50. On peut ensuite appliquer l'algorithme de séparation de grappe par pic de densité dont le déroulé est précédemment détaillé. Un exemple de cette utilisation est illustré en figure 3.35.

En mettant ces cartes de caractéristiques sous la forme de vecteur, on peut placer ces positions sur un repère à N dimensions, et utiliser n'importe quel algorithme de séparation de grappe. Celui du « Pic de densité » est adaptatif en ce sens qu'il ne nous donnera pas nécessairement le même nombre de grappes pour des ensembles de vecteurs différents. Cet algorithme se déroule en trois étapes, le calcul de la densité locale de chaque point, le calcul de la distance minimale vers un point de densité plus élevée, ainsi que l'identification des différentes grappes.

Pour chaque point i, on définit une densité locale ρ_i comme la somme pondérée des distances vers tous les autres points, selon un noyau décroissant (souvent gaussien) :

$$\rho_i = \sum_{j \neq i} \chi(d_{ij} - d_c) \tag{3.17}$$

où :

$$\chi(x) = \begin{cases} 1, & \text{si } x < 0 \\ 0, & \text{sinon} \end{cases} \tag{3.18}$$

et d_c est un seuil de coupure. La **distance delta** δ_i mesure la distance minimale entre le point i et tout point ayant une densité supérieure :

$$\delta_i = \begin{cases} \min_{j:\rho_j>\rho_i} d_{ij}, & \text{si } \exists j \text{ tel que } \rho_j > \rho_i \\ \max_j d_{ij}, & \text{sinon} \end{cases} \tag{3.19}$$

Les centres de grappes sont sélectionnés comme les points présentant simultanément une densité élevée et une grande valeur de δ_i. On définit un **indice de décision** :

$$\gamma_i = \rho_i \cdot \delta_i \tag{3.20}$$

Les points maximisant γ_i sont choisis comme centres. Chaque point non-centre est ensuite assigné à la même grappe que son voisin le plus proche ayant une densité supérieure :

$$C(i) = C\left(\arg \min_{j:\rho_j>\rho_i} d_{ij}\right) \tag{3.21}$$



FIGURE 3.35 – **Test de « pic de densité » en guise de modèle de résumé de vidéo, afin d'extraire les trames principales des séquences.** Le mouvement de connexion des mains de Buzz l'Éclair et son acolyte sont résumés par les deux images qui apportent la majorité de l'information contenue dans la séquence.

Malgré ces optimisations et le précalcul de toutes les caractéristiques possibles, l'entrainement restant prend tout de même trois semaines à réaliser sur une GPU Tesla V100 à 32 Go de VRAM. Ceci crée fatalement une lenteur entravant l'amélioration de manière itérative et agile du modèle. C'est pour cette raison que les dernières performances obtenues par ce nouveau modèle d'apprentissage automatique n'est que de 68,42 % sur les mêmes tâches de prédictions que R-ICE, c'est-à-dire sur les trois labels à prédire à partir de plans.

3.8 Conclusion

Après avoir introduit en détail l'environnement de travail ICE et les spécificités techniques et technologiques du dispositif, ce chapitre a mis en lumière la complexité des choix graphiques à opérer dans ce cadre. Reposant sur l'exploitation des bords latéraux du film, le système ICE implique des contraintes fortes, telles que la faible résolution des panneaux LED, leur position perpendiculaire à l'écran principal, ou encore les différences d'espace colorimétrique. À cela s'ajoutent des considérations artistiques, souvent guidées par une

expérience sensorielle en salle : la décision d'appliquer un flou, de préserver le détail ou de renforcer une ambiance repose sur une subjectivité difficilement traduisible en règles explicites. Ces choix, bien que parfois récurrents, demeurent sensibles au rythme, à la dynamique du plan, et à l'intention du réalisateur.

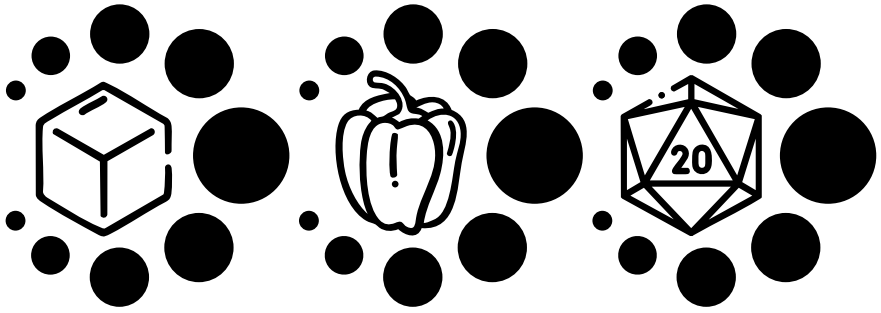
Pour faire émerger une solution capable d'assister ce processus, une étude approfondie de la littérature en compréhension de vidéo a été menée. Elle a permis d'identifier les approches les plus pertinentes pour la tâche de prédiction du traitement latéral, en particulier celles fondées sur des architectures temporelles telles que les 3D-CNN. Ces modèles ont la capacité de capturer conjointement les dimensions spatiale et temporelle, ce qui les rend particulièrement adaptés à la nature séquentielle et contextuelle des plans cinématographiques.

Sur cette base, une méthode de prédiction automatique du niveau de détail à appliquer aux panneaux latéraux a été développée. Elle constitue une étape fondamentale de la chaine de procédés ICE, conditionnant l'ensemble du traitement graphique en aval. En s'appuyant sur des modèles de classification temporelle, cette approche permet d'identifier, pour chaque plan, le traitement optimal à appliquer, et ce, avec une précision dépassant celle d'approches plus classiques comme les modèles génératifs ou les méthodes par correspondance d'imagette (ou *patchs* en anglais).

L'exploration de différentes configurations de classifieurs a mis en évidence des pistes d'optimisation supplémentaires, ouvrant la voie à une amélioration continue du système. Si les contraintes de confidentialité ont limité l'étude à un corpus de 30 bandes-annonces, les perspectives futures incluent un élargissement aux longs-métrages, afin d'accroitre la diversité et la robustesse des données d'entrainement. Par ailleurs, l'intégration de descripteurs perceptuels plus riches et de nouvelles composantes architecturales pourrait encore affiner la capacité du modèle à reproduire les subtilités des choix esthétiques propres à ICE.

CHAPITRE 4

ICE-CUBED : EXTENSION INTELLIGENTE DE CONTEXTES IMMERSIFS PAR INPAINTING



La science nous impose d'exploser le soleil.

Pye Nomai

Table des matières

4.1 Introduction 144
4.2 État de l'art 145
4.3 Méthodologie 150
4.3.1 Architecture du système 150
4.3.2 Segmentation d'objets 152
4.3.3 Création du logiciel libre d'annotation SSPICE et de la base de données 153
4.4 Résultats 155
4.4.1 Configuration expérimentale 155
4.4.2 Discussions sur la détection 155
4.5 Validation perceptuelle 157
4.5.1 Méthodologie de l'étude perceptuelle 157
4.5.2 Résultats de l'étude perceptuelle 159
4.6 Conclusion 161

4.1 Introduction

Bien que notre modèle R-ICE (cf. chapitre 3) peut prédire avec une précision significative le taux de détail qu'il convient d'appliquer sur un plan (72,71 %), il manque de sophistications dont est capable l'équipe de postproduction ICE. L'équipe doit alors tout de même vérifier ce qu'a prédit R-ICE, et le corriger. Cela implique, pour l'équipe, de parcourir le film, même avec la mise en production de R-ICE dans les processus de CGR Cinémas. De plus, une complication supplémentaire rencontrée par l'équipe de postproduction constitue un enjeu qu'il convient d'aborder dans ces nouveaux travaux. Les instants les plus impressionnants de ICE (et les plus valorisés commercialement par extension) sont les instants où les plans de cinéma se prêtent bien à afficher du détail sur les panneaux ICE. Cela peut correspondre à une scène d'action, une scène de paysage donnant une sensation de grandeur, ou une scène à forte dynamique de mouvement (voir la section 2.2.2 pour plus de détails). Toutefois, la technique ICE présente certaines limites à ces égards : les objets proches des bords de l'image du film interfèrent fréquemment avec ICE. Avec la présence de tels éléments, il devient impossible de répliquer des détails dans les panneaux ICE. Cela aurait pour effet de sortir le spectateur de la narration immersive du film, pour deux raisons possibles.

La première est liée au fait que cet élément pourrait attirer l'œil du spectateur, plus que l'écran central. Soit l'aspect visuel est trop détaillé par rapport au reste de l'image, soit ses caractéristiques temporelles (mouvement, scintillement, etc.) dépassent les seuils de perceptibilité. Il convient de rappeler que les mouvements dans les panneaux sont amplifiés dans les panneaux ICE (cf. section 2.2.3). La deuxième raison consiste en une incohérence spatiale, ayant lieu lorsqu'un élément, comme le visage d'un personnage, peut être visible simultanément face au spectateur (écran principal) et dans sa vision périphérique lointaine, à droite ou à gauche sur les panneaux ICE (voir figure 4.1). Cet effet double perturbe l'immersion et demeure fondamentalement incompatible avec les principes de la narration cinématographique selon ICE.

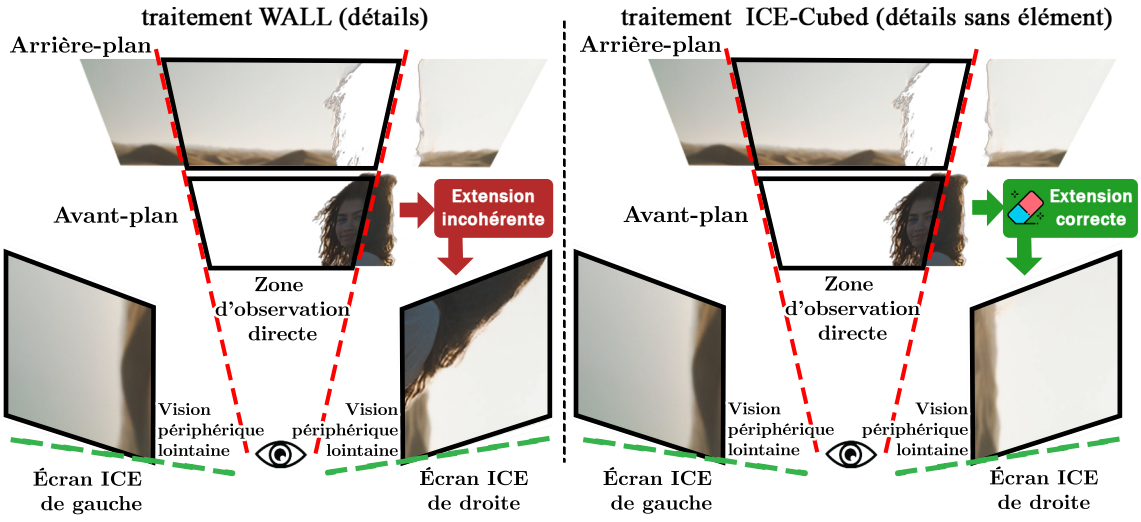


FIGURE 4.1 – **Illustration de la nécessité de cohérence spatiale pour ICE. L'expansion de l'image doit concerner uniquement l'arrière-plan en raison de la configuration tridimensionnelle des écrans.** Si le film était étendu dans le même plan que l'écran principal, les écrans sembleraient continus. Cependant, ce n'est pas le cas dans la disposition spatiale 3D de la salle, entrainant ainsi une incohérence spatiale. (Crédit : auteur)

Pour préserver la cohérence, en règle générale, seuls les éléments d'arrière-plan doivent s'étendre sur les écrans périphériques, comme le montre la figure 4.1. Les objets perturbants du premier plan doivent être traités différemement. Le cas traité dans la figure 4.1 est le plus simple car il représente un élément avec une forme bien définie, nette, sur un fond uni et flou. Les situations peuvent être plus compliquées, comme dans le cas où l'élément à segmenter peut être flou, en avant-plan très proche, sur un fond uni. L'élément peut également sur le même plan que le fond, en étant flou, ou même être net tout comme le fond.

Il peut ne pas être un objet, voire ne pas avoir de forme définie, ou être le simple contour d'un bâtiment. Enfin, dans de rares cas, l'élément peut être plus loin que le fond à étendre.

L'objectif de ce travail consiste en l'extension latérale de la vidéo afin de créer un continium n'attirant pas le regard tout en apportant une sensation d'immersion plus importante dans le film. Ainsi, il serait possible d'augmenter le faible nombre de choix créatifs détaillés appliqués sur les films, améliorant l'expérience.

En analysant la littérature qui se rapproche de ces besoins, il est possible de remarquer que les algorithmes d'*outpainting* poursuivent un objectif similaire puisqu'ils visent à étendre un contenu visuel existant. Toutefois, plusieurs limitations considérées ici comme rédibitoires par rapport à notre application ont été rélevées.

En cause mais pas seulement, sa propension à générer des confabulations [1] et des artéfacts notables, ce qui rend ces méthodes inappropriées dans un contexte cinématographique, où tout ajout visuel doit impérativement rester conforme à l'intention du réalisateur. Pour surmonter ces limitations, ce chapitre présente une solution ICE-Cubed qui consiste à détecter automatiquement les éléments qui empêchent de dupliquer le film dans la vision périphérique, puis de proposer une solution pour masquer ces zones. Bien que de nombreux modèles de détection soient disponibles dans la littérature, aucun n'a été conçu pour cette tâche spécifique aux éléments perturbant l'immersion lorsqu'ils sont situés en vision périphérique lointaine. Une étude comparative des différents modèles de segmentation les plus récents est donc menée. La constitution d'un jeu de données DICE adapté est une étape primordiale et pour laquelle nous avons conçu un logiciel d'annotation dédié SSPICE . Enfin, vient la phase de réentrainement du modèle jugé le plus adapté à notre cas d'usage. [2] Une chaine de traitement intégrant modèles d'apprentissage profond et techniques de traitement d'images permettant de garantir un niveau d'artéfact minimal est ensuite mise en place, afin de réduire la charge de postproduction et d'améliorer l'expérience perceptive des spectateurs.

4.2 État de l'art

L'inpainting et l'outpainting sont deux domaines qui vont nous intéresser dans ce travail, c'est donc naturellement vers ces domaines que va se tourner cet état de l'art. Ces deux domaines font partie de la plus grande catégorie des modèles génératifs. Les architectures destinées à générer du contenus étaient souvent à base de réseaux convolutifs avant l'arrivée

1. ou hallucination par abus de langage

2. SSPICE pour *Scene Segmentation Platform by ICE* ou plateforme de segmentation de scènes par ICE, DICE pour *Dataset of dIsturbing Cinematographic Elements* ou jeu de données d'élements cinématographiques perturbants, ICE-Cubed pour *Inpainting of Cinematographic Elements for Intelligent Context Expansion by ICE* ou inpainting d'éléments cinématographiques l'extension intelligente de contextes ICE.

des *transformers* et des modèles de diffusions, avec comme entrée une image, et comme sortie une image. Là où les modèles de compréhension d'images ont une architecture d'encodeur seul, les réseaux génératifs ont un encodeur et un décodeur (voir figure 4.2). Grâce à cela, le transfert de connaissances et les évolutions mutuelles des domaines de recherches comme la superrésolution, la restauration d'images, le transfert de style, etc., sont facilitées.

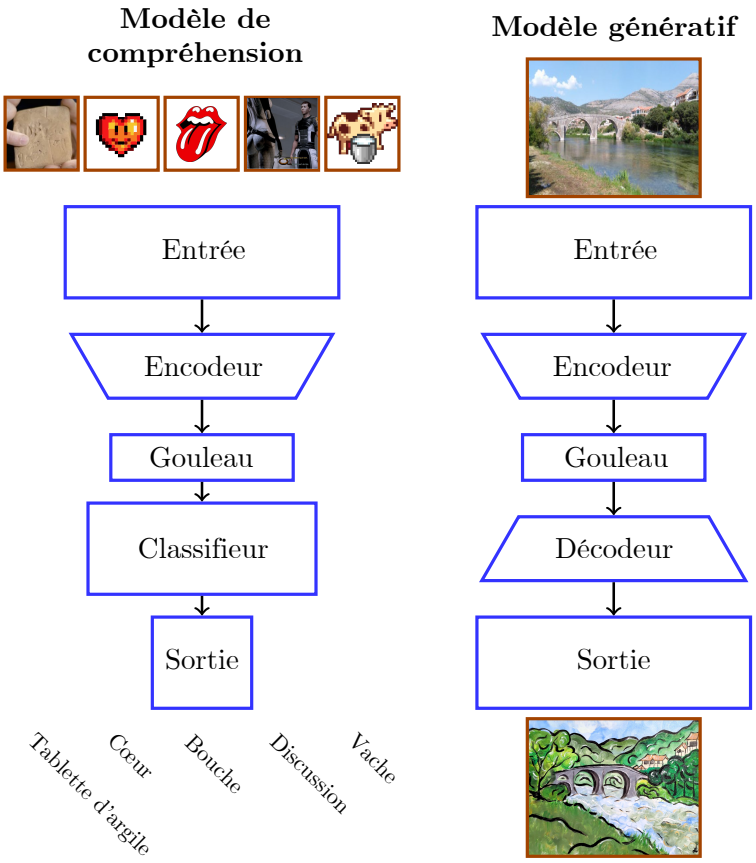


FIGURE 4.2 – **Similarités et comparaisons d'architectures entre les modèles profonds de compréhension et les modèles génératifs.** Encodeur à gauche pour les classifications et autres sorties du type « label ». Encodeur puis décodeur à droite pour la génération d'image, avec un modèle de transfert de style comme exemple. (Crédit : auteur [M23])

L'inpainting représente l'action de restaurer une image dont une partie des pixels est manquante pour diverses raisons. Une illustration de cette action, réalisée par des artistes humains, est visible en figure 4.3. La partie restaurée est alors reconstituée selon ce qui entoure cette zone.

L'inpainting peut être décrit comme la généralisation de l'interpolation d'un signal 1D à un espace image 2D. L'interpolation est l'estimation de données interstitielles entre les points de données existants, tandis que l'extrapolation est l'estimation de nouvelles données au-delà des données existantes. L'outpainting, qui consiste à étendre l'image au-delà de ses bordures, relève donc de l'extrapolation 2D (voir figure 4.4).

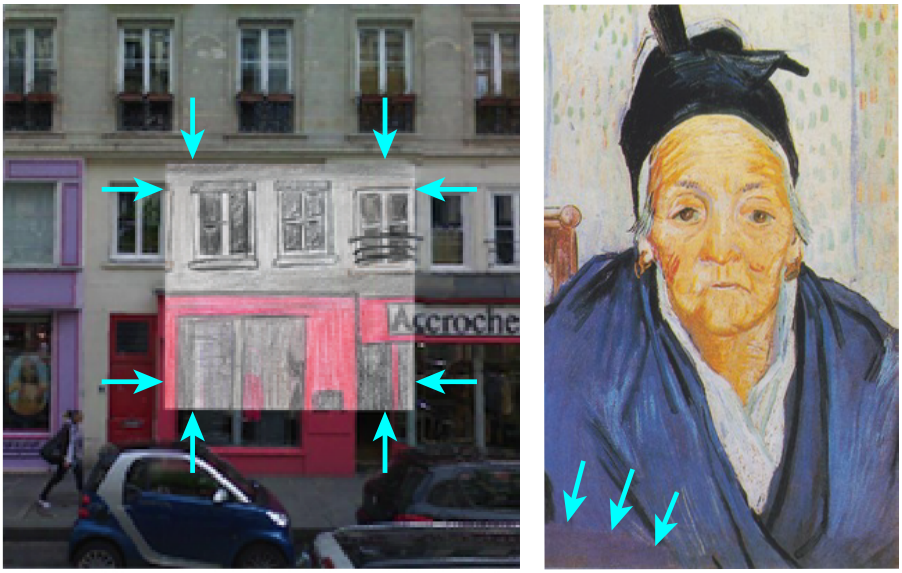


FIGURE 4.3 – **Exemples de tâches d'inpainting et d'outpainting réalisés par l'humain, tâches possibles bien avant l'apparition de l'intelligence artificielle.** À droite, la peinture « Une vieille femme d'Arles » par Vincent Van Gogh, dont le coin bas-gauche a été outpainté par une couleur neutre. (Crédits : modifiées depuis [133] et [134])

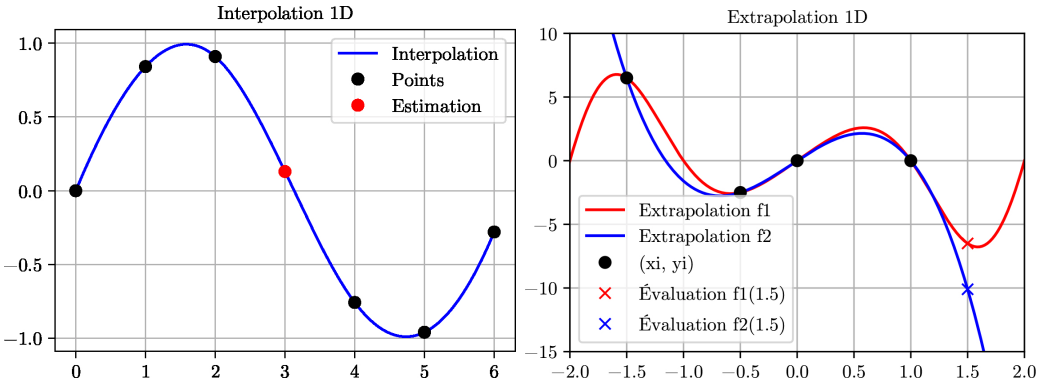


FIGURE 4.4 – **Interpolation et extrapolation d'un signal 1D.** Ces opérations permettent de générer une nouvelle donnée en cohérence avec celles d'origine. L'interpolation prédit une donnée au milieu d'autres données réelles, tandis que l'extrapolation prédit des données au-delà des données réelles. L'inpainting peut être vu comme une interpolation 2D d'une image et l'outpainting comme une extrapolation 2D d'une image. (Crédit : auteur)

Ce problème d'extrapolation est bien plus complexe que l'inpainting, puisqu'il existe une infinité de contextes environnants susceptibles de correspondre à divers critères prédéfinis [131], comme illustré en figure 4.5.

Puisque notre objectif est d'étendre une image, l'outpainting constitue l'approche la plus naturelle. Toutefois, il présente de fortes similitudes avec l'inpainting, tant du point de vue méthodologique que des objectifs poursuivis, à savoir la reconstruction de contenus manquants tout en préservant la cohérence globale.

Les algorithmes traditionnels cherchant à résoudre ces problématiques (qui n'étaient alors

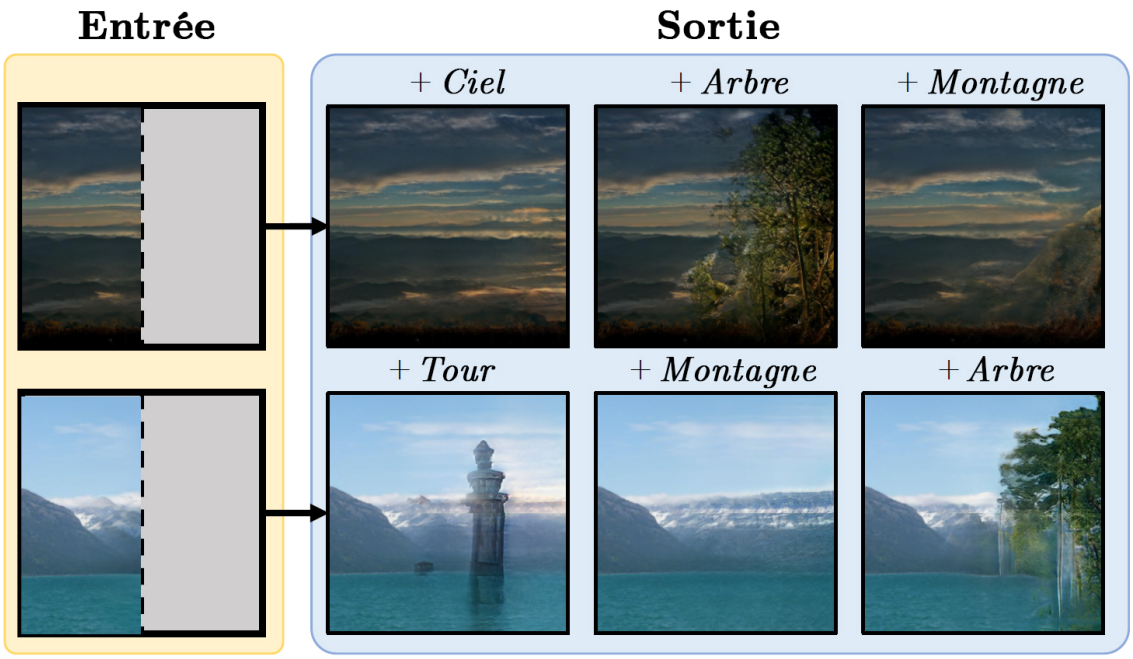


FIGURE 4.5 – **Exemples d'inpainting avec la méthode InOut (2022), tirant profit d'un espace latent** [8]. Cet espace latent permet d'orienter les possibilités de cohérence infinies qu'il est possible de créer avec le problème inverse mal posé qu'est l'outpainting. Ce principe d'espace latent est encore le plus répandu pour la génération d'image par instructions textuelles en 2026.

pas encore appelées inpainting ou outpainting) se basent sur une approche par « patch » , ou imagette[132] : le principe est, qu'à l'aide d'une fonction de cout conçue au préalable, on arrive à trouver un patch existant dans l'image qui correspondrait le mieux au patch que l'on cherche.

En raison des similitudes entre ces deux problèmes, les premiers modèles d'outpainting étaient directement adaptés des modèles d'inpainting, à la fois pour les approches par patch et pour les approches d'apprentissage profond. Ces derniers, par l'adaptabilité inhérente aux réseaux de neurones, peuvent même utiliser exactement le même modèle comme montré par Noord et al. [134] bien que cela peut être vu comme un objectif secondaire plutôt qu'une solution optimale. Ces premières méthodes d'outpainting par patchs assuraient une cohérence locale mais échouaient en général à maintenir une cohérence d'ensemble [135]. Des approches plus récentes s'appuient sur des architectures d'apprentissage profond. Dans l'ordre chronologique, les modèles génératifs ont été des autoencodeurs, des modèles GAN, des transformers et des modèles de diffusion [136, 137]. Les transformers améliorent la cohérence à longue distance mais demeurent fortement dépendants des jeux de données. Gao et al.[136] peinent à généraliser au-delà des paysages, tout comme Cheng et al. [8], tandis que les travaux de Wu et al.[137] se révèlent efficaces pour des environnements intérieurs structurés, mais manquent d'adaptabilité face à des compositions cinématographiques complexes.

Une application plus précise de l'outpainting d'images est la création de contextes immersifs pour la vidéo, en élargissant les images et par conséquent, le champ visuel, de la même manière que le but recherché par ICE. Même si le modèle d'outpainting d'images [131] utilise une dégradation fovéale de la qualité pour atténuer les artéfacts, le maintien de la

cohérence du contexte environnant entre les images successives d'une vidéo constitue un problème d'un ordre de grandeur supérieur. Dans cette reconstruction de contenus vidéo (tant en inpainting qu'en outpainting d'ailleurs), un enjeu majeur réside dans l'incohérence temporelle, qui se traduit par des effets de scintillement perturbant l'immersion. Ce problème est d'autant plus accentué en outpainting, où l'extrapolation non contrainte accroit l'instabilité et le risque d'hallucinations visuelles. Les travaux de Maggia [138] explorent l'outpainting appliqué au cinéma, dans le but d'adapter les résolutions des films 4/3 en 16/9. Toutefois, ils reposent sur des GANs, dont l'entrainement reste notoirement difficile en raison de la convergence mutuelle des réseaux adverses, et souffrent encore d'instabilités temporelles et de générations peu robustes, notamment lorsqu'un élément sort du cadre ou lorsque la caméra est en mouvement.

Pour pallier ces problèmes de mouvements, d'autres approches vidéo intègrent conjointement les informations de flot optique et la complétion de contours [139, 140, 141], améliorant ainsi l'alignement temporel en reconstituant des régions occultées. La complétion de contours quant à elle vise à rendre plus robuste la cohérence globale via ces contours qui agissent comme lignes directrices. Néanmoins, ces méthodes demeurent extrêmement sensibles aux mouvements rapides et aux occlusions, ce qui limite leur robustesse dans les scènes dynamiques.

Pour se rapprocher de notre contexte, l'application de ces domaines à la vidéo présentent des défis supplémentaires. Tant l'inpainting vidéo [141] que l'outpainting vidéo [140] basés sur des méthodes d'apprentissage profond ont donné des résultats prometteurs. Une autre méthode d'outpainting vidéo repose sur la déformation de maillage et l'assemblage d'images voisines pour générer une vidéo extrapolée [142].

L'outpainting préserve généralement le même niveau de détail que celui présent dans l'image originale à étendre. Cette approche engendre toutefois des contextes hallucinés hautement problématiques et de nombreux artéfacts, significativement plus marqués que dans le cas de l'inpainting. Cela découle du fait que l'outpainting constitue un problème inverse mal posé car une infinité d'images de plus grande taille peuvent correspondre à une seule image réduite [131]. Cette ambiguïté intrinsèque explique pourquoi l'outpainting est fondamentalement plus difficile à résoudre que l'inpainting.

Dans notre contexte d'application, trois limites de l'outpainting et de son usage actuel ne s'accordent pas parfaitement avec notre démarche. Considérant que les zones étendues sont situées dans la vision périphérique et alignées avec nos objectifs immersifs, il est bien préférable d'obtenir un contexte peu détaillé, qui ne capte pas l'attention, plutôt qu'un contexte extrêmement détaillé mais hallucinatoire, dépourvu de cohérence spatiale ou temporelle, ce qui casserait l'immersion. D'autre part, les modèles d'outpainting sont entrainés à extrapoler une extension à partir d'une image sur un même plan géométrique, ce qui n'est là encore pas adapté à notre besoin, puisque ces modèles résulteraient nécessairement en une incohérence spatiale, évoquée précédemment (voir figure précédente 4.1). Enfin, l'incohérence peut être spatiale, par la disposition perpendiculaire, par la présence d'artéfact et de contenu halluciné, mais l'incohérence peut également être temporelle avec un phénomène de scintillement, induit par les artéfacts et leur nonalignement spatial dans le temps. De plus, bien que la vision périphérique humaine soit moins sensible à la couleur et à la texture, elle reste particulièrement sensible au mouvement et au scintillement [143] [36].

Pour l'ensemble de ces trois points, les approches actuelles d'outpainting ne conviennent pas à la réalisation de nos objectifs. C'est pourquoi, dans la sous-section suivante, nous nous

concentrerons sur les travaux qui privilégient la cohérence maximale et la minimisation des artéfacts, quitte à réduire la résolution du contenu périphérique généré. Compte tenu de l'instabilité inhérente aux méthodes d'outpainting et de leur tendance à amplifier le scintillement temporel, nous privilégierons dans notre méthodologie les **techniques d'inpainting**, qui permettent une génération de contenus plus stable et garantissent une meilleure cohérence temporelle et structurelle. Nous pouvons observer ces techniques d'inpainting appliquées à nos données sur la figure 4.6.



FIGURE 4.6 – **Exemples d'inpainting appliqué à nos données.** En haut, les trames originales respectives du *Royaume des Abysses*, *Dune : Première Partie* et *Buzz L'Éclair*. En bas, les trames en sortie de notre modèle ICE-Cubed dont les objets perturbants ont été détectés automatiquement puis inpaintés.

4.3 Méthodologie

4.3.1 Architecture du système

Notre proposition consiste à détecter les éléments perturbant l'immersion dans le film, à les effacer par inpainting, puis à étendre le film ainsi modifié sur les panneaux ICE. La vue d'ensemble du l'architecture proposée pour ICE-Cubed est présentée dans la figure 4.7.

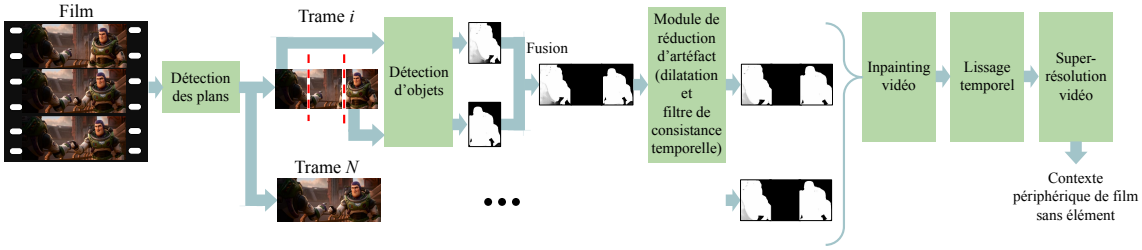


FIGURE 4.7 – **Schéma détaillé de l'architecture proposée pour générer des contextes périphériques sans élément, en tant qu'extension de trame.**

Tout d'abord, les plans constituant un film sont détectés et traités séparément. Chaque image d'un plan donné est divisée en sous-images « gauche » et « droite », afin d'orienter l'attention du modèle de segmentation sur une zone spécifique. Le modèle de segmentation utilisé est le BiRefNet [144], réentrainé sur des données que nous avons annotées. Après cette segmentation, les sous-images sont recombinées. On peut constater ce qu'apporte cette subdivision en figure 4.8.

Plusieurs opérations sont appliquées afin d'affiner les masques obtenus par le modèle de segmentation précédent. Chaque opération cible un type spécifique d'artéfact, comme détaillé ci-après.

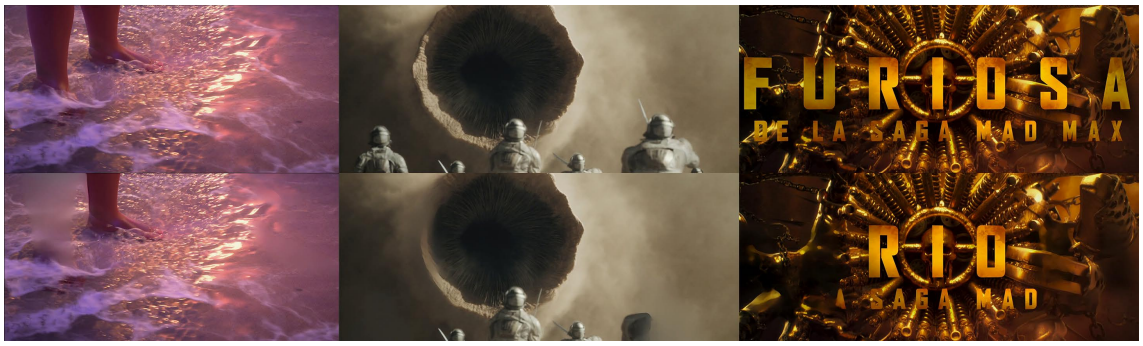


FIGURE 4.8 – **Illustration de l'effet de subdivision « gauche » et « droite » dans l'architecture ICE-Cubed pour l'inpainting de trames.** En haut, les trames originales respectives de *Vaiana 2*, de *Dune : Deuxième Partie* et de *Furiosa : Une saga Mad Max*. En bas, les trames en sortie de notre modèle ICE-Cubed dont les objets perturbants ont été détectés automatiquement puis inpaintés, lorsque l'attention du modèle de détection d'objet est volontairement dirigée vers les bords de l'image.

Une opération morphologique de dilatation à l'aide d'un noyau 7×7 de type disque est d'abord appliquée au masque, afin de produire un effet de croissance de ce dernier comme visible sur la figure 4.9.



a) Original **b) Dilatation** **c) Érosion**

FIGURE 4.9 – **Effets de l'opération morphologique de dilatation et d'érosion sur un logotype, avec silhouette et lettres stylisées.** Les filtres appliqués peuvent aussi être vus respectivement comme des filtres 2D maximum et minimum (en l'occurence, respectivement de de taille 5×5 et 3×3).

Comme il n'est pas nécessaire d'obtenir un masque d'une extrême précision et que l'objectif n'est pas un travail de composition photographique d'image ou de fond vert, mais bien de l'inpainting, nous privilégions un modèle qui sur-détecte de manière grossière, plutôt qu'un modèle très précis mais susceptible de sous-détecter. Ce dernier cas provoquerait des artéfacts spatiaux où la majeure partie de l'élément est inpaintée, mais dont les contours « débordent » dans le contenu régénéré. Il est donc important que l'élément perturbant soit entièrement inscrit à l'intérieur du masque qui sera inpainté.

De manière analogue, un filtre temporel maximum d'un noyau de taille 5 est appliqué sur le masque du plan afin de réduire les artéfacts temporels. Ces derniers apparaissent lorsqu'un élément perturbateur est correctement détecté dans toutes les trames, sauf dans une (ou quelques-unes) d'entres elles, entrainant la réapparition soudaine de l'élément inpainté durant une très courte durée, comme illustré en figure. 4.10). Cela se traduit alors par un effet de scintillement visuel, qui rompt inévitablement l'immersion et compromet entièrement le plan cinématographique.

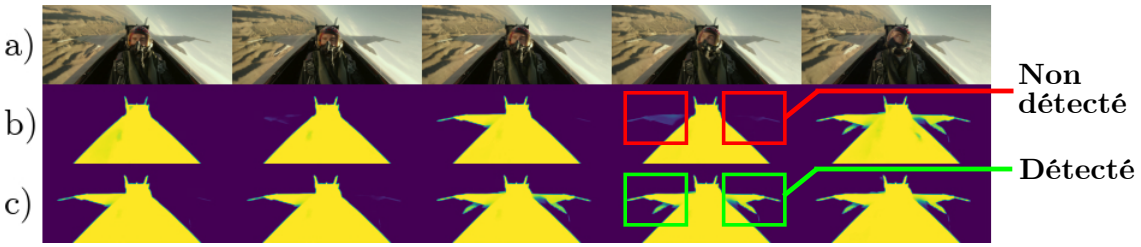


FIGURE 4.10 – **Démonstration de l'importance du module proposé de réduction d'artéfact.** a) Cinq images successives b) Masques avant le filtre temporel maximal c) Masques après le filtre temporel maximal. Les masques obtenus sont plus stables.

Le masque vidéo obtenu est ensuite inpainté à l'aide du modèle E2FGVI [3] [141]. Ce dernier procède à l'inpainting vidéo via trois modules entrainables : complétion de flot optique, propagation de caractéristiques et génération de contenu.Cela lui permet de générer un rendu visuellement cohérent et stable dans le temps, les régions complétées s'intégrant naturellement dans la dynamique de la scène. Les transitions spatiotemporelles sont fluides, avec peu d'artéfacts ou de scintillement. Néanmoins, le modèle tend à lisser les textures fines et à perdre certains détails dans les zones à forte dynamique visuelle. Il reste tout de même le meilleur modèle d'inpainting vidéo au moins jusqu'en 2022 sur deux jeux de données vidéo DAVIS [145] et Youtube-vos-2018[146], comme le classe le site *Open Research Knowledge Graph*, avec un PSNR à 33,71 dB et une SSIM de 0,9700 [4].

Le modèle réduit la résolution de l'image d'entrée à 214×512 afin de faciliter une reconstruction plus précise.

Pour atténuer les légers effets de tremblement pouvant apparaitre dans une zone inpaintée, un filtre de correction temporelle par médiane, avec un noyau de taille 3, est appliqué sur les séquences. La capacité du filtre médian à lisser tout en préservant les contours s'avère efficace dans le domaine temporel, en traitant toute variation locale soudaine de couleur comme des « contours temporels ». Ainsi, les coupes de plan resteront intactes.

La sortie inpaintée est ensuite transmise à RealBasicVSR [147], un modèle de superrésolution vidéo. La vidéo est alors suréchantillonnée par un facteur ×4, restituant une image en haute définition (858×2048, la résolution standard CinemaScope).

4.3.2 Segmentation d'objets

Il est important de comprendre la démarche d'une solution, ainsi que les différentes itérations qui justifient les éléments qui la constitue. Dans cette sous-section, nous allons chercher à optimiser notre solution selon les difficultés observées. L'approche générale repose sur la détection des éléments perturbants, leur inpainting, puis leur superrésolution. L'utilisation, pour chacune des phases, de modèles proposés pour l'image ou pour la vidéo est à étudier. Les modèles vidéo peuvent être extrêmement gourmands, tant en mémoire vive qu'en ressources calculatoires. Ces modèles seront plus longs en inférence et en entrainement. Il faut donc garder en tête ces problématiques, sachant que les itérations d'améliorations entre essais et erreurs sera beaucoup plus longues.

3. E2FGVI : *End-to-End framework for Flow-Guided Video Inpainting*
4. PSNR pour *Peak Signal-to-Noise Ratio*, le ratio du pic du signal à bruit et SSIM pour *Structural Similarity*, similarité de structure

Compte tenu de la nature spécifique et atypique de la tâche de détection imposée par ICE-Cubed, nous avons évalué de nombreux modèles de segmentation, couvrant à la fois des approches d'apprentissage profond et des méthodes traditionnelles. Les méthodes suivantes ont ainsi été testés : Itti saliency map [148], Itti+GrabCut [149], Mask-RCNN [150], TRACER [151], BRIAAI-RMBG-1.4, et BiRefNet [144].

L'algorithme de saillance d'Itti génère une carte de saillance basée sur la combinaison de contrastes de couleur, d'intensité et d'orientation. GrabCut est un algorithme interactif de segmentation par graphes, conçu pour un raffinage de contour.

Mask-RCNN est un modèle d'apprentissage profond destiné à la classification d'objets et à la segmentation d'instances. TRACER est quant à lui spécialisé dans la Détection d'Objet Saillant (SOD) [5]. BRIAAI-RMBG, basé sur une architecture de type transformeur, est un modèle commercial visant à supprimer les arrière-plans, relevant également de la SOD. Enfin, BiRefNet a été proposé avec le jeu de données DIS5K et une nouvelle tâche d'apprentissage profond à savoir la segmentation dichotomique d'images, qui généralise la SOD en supprimant les hypothèses sur le type ou le nombre d'objets à détecter. Dans son article de référence, BiRefNet a été testé et entrainé sur un ensemble varié de jeux de données fondamentalement différents, allant de la détection d'objets camouflés aux données liées à la détection d'objets saillants haute résolution [6]. Son architecture combine un module de localisation exploitant des informations sémantiques globales, avec un module de reconstruction assurant une segmentation fine et haute résolution.

4.3.3 Création du logiciel libre d'annotation SSPICE et de la base de données

La détection précise des éléments perturbateurs constitue une composante essentielle de ICE-Cubed. Ces éléments apparaissent parfois floutés, soit en raison d'effets de profondeur de champ affectant un personnage au premier plan, soit à cause d'un flou de mouvement. Or, les jeux de données de segmentation existants (DUTS [152], HRSOD [153], DIS5K [154]) contiennent très peu ou pas d'exemples de tels types de flous sur les objets segmentés.

Pour combler cette lacune, nous avons créé un jeu de données dédié et avons réentrainé le meilleur module de détection identifié pour notre tâche : BiRefNet. Le logiciel **SSPICE**, conçu pour ce travail, est publié en open source afin de permettre une annotation rapide des trames d'une vidéo (`github.com/dtraparic/SSPICE`).

Cette thèse étant portées sur des problématiques industriels, l'intégration continue de développement (DevOps) et l'intégration continue d'apprentissage (MLOps), sont des points qui ont été formalisés dans le sujet de thèse. SSPICE contribue intégralement à cette démarche. Certains principes, comme la traçabilité, la reproductibilité, l'automatisation de pipeline, et la collaboration ont été respectés. La traçabilité est assuré par des outils de versionning Git. La reproductibilité est assurée via les tests unitaires du code et le stockage des données de test. L'automatisation de la chaine de traitements, pour faciliter l'utilisation et la reprise du code à l'avenir par d'autres développeurs de l'entreprise. De même, la collaboration est assurée, comme le montre ce logiciel avec une interface homme-machine graphique et simple d'utilisation.

5. SOD pour *Salient Object Detection*
6. COD pour *Camouflaged Object Detection* et HRSOD pour *High-Resolution Salient Object Detection*

SSPICE a été développé pour répondre à nos besoins spécifiques, en tant que logiciel open-source destiné à faciliter l'annotation de segmentation sur des images issues de films. Son interface utilisateur est visible sur la figure 4.11. Imaginons une bande-annonce de 2 minutes 30 secondes, cette bande-annonce contient alors $24 \times (2 \times 60 + 30) = 5000$ trames. Détourer autant de trames sur un logiciel payant comme Adobe Photoshop, prendrait beaucoup trop de temps, car c'est un logiciel extrêmement lourd, qui peut prendre entre 10 et 15 secondes pour ouvrir ou enregistrer une image, clics compris. Une navigation fastidieuse à travers les menus [7] est également nécessaire, ce qui ralentit énormément le procédé lorsqu'une tâche est répétée autant de fois. SSPICE a donc été développé pour un objectif de rapidité, et pour correspondre au mieux aux données à traiter et aux besoins d'annotations.

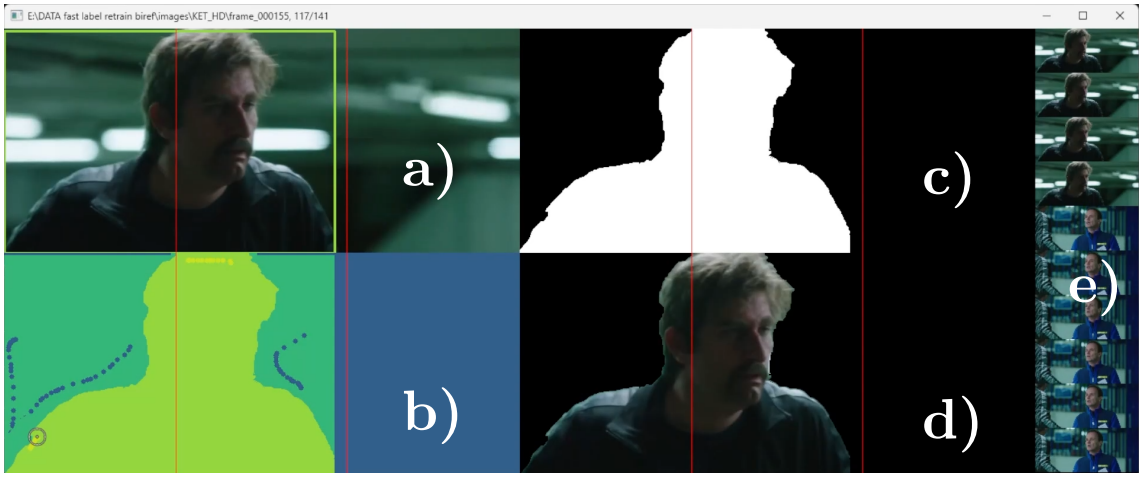


FIGURE 4.11 – **SSPICE, le logiciel open source de détourage de trames vidéo ultrarapide, créé pour cette thèse.** Chaque zone est modifiable afin d'ajouter ou de retirer des pixels du masque de segmentation. (a) Image complète. (b) Quatre valeurs possibles par pixel dessiné par l'utilisateur, chacune représentée par une couleur. Bleu : arrière-plan certain, vert foncé : arrière-plan probable, vert clair : premier plan probable, et jaune : premier plan certain. (c) Masque binaire résultant du raffinage de contours. (d) Avant-plan disposé sur fond noir pour une meilleure lisibilité. (e) Aperçu des dix images suivantes. (Crédit : auteur, avec démonstration vidéo [M24])

Le logiciel intègre l'algorithme GrabCut [149], permettant une sélection rapide et itérative à partir de clics successifs de l'utilisateur. Différents types de brosses sont disponibles : arrière-plan certain, premier plan certain, arrière-plan probable et premier plan probable, comme illustré dans le coin inférieur gauche de l'interface. Ces quatres types de brosse permettent ensuite à l'algorithme GrabCut de trouver la selection optimale, selon le contraste et les couleurs de l'image. Une possibilité ajoutée au logiciel est la modification de la taille de chacune de ces brosses. Cette fonctionnalité permet de couvrir une zone très large avec une certaine brosse, ou bien être très précis s'il le faut, comme sur les contours par exemple.

L'utilisateur peut, quand il le souhaite, appliquer une opération morphologique implémenté dans le logiciel, permettant d'agrandir ou de rétrécir le masque. Cette action, suivie de l'actionnement d'une itération de GrabCut, permet de suivre automatiquement le déplacement d'un objet entre plusieurs trames, le tout en deux touches de clavier.

Par défaut, SSPICE applique la détection qui a été faite par les modèles d'IA de détection choisis. Dans notre cas, le plus performant, c'est-à-dire BiRefNet. Le logiciel exploite par

7. Plus communément appelée « menuing » si on se réfère au lexique du *speedrun*

ailleurs la forte probabilité que deux images consécutives soient similaires. L'utilisateur peut appliquer un masque déjà réalisé sur la trame précédente, mais pas seulement. Il peut ainsi remplacer le masque de la trame qu'il manipule par un masque déjà corrigé de la trame précédente ou celle d'avant. Il a aussi la possibilité d'appliquer le masque de la prochaine trame ou celle d'après. En plus de remplacer le masque, il est aussi possible d'additionner l'un des masques disponibles au sien. Cela peut créer un effet de fusion entre deux trames successives, afin de faciliter le suivi d'un personnage. Enfin, l'utilisateur peut également ignorer la trame actuelle et passer à la suivante, le tout en prévisualisant les dix trames suivantes sur la droite de l'interface utilisateur afin d'anticiper au mieux ses prochaines actions.

Grâce à ce logiciel, nous proposons un jeu de données constitué de 2578 images issues de la bande-annonce de *Dune : Première Partie* et de 3106 images issues de la bande-annonce de *Top Gun : Maverick*. Ces images ont été segmentées en étroite collaboration avec le responsable de création de contenu ICE, afin d'identifier et de valider les objets susceptibles de devenir perturbateurs lorsqu'ils apparaissent en vision périphérique. L'ensemble constitue le jeu de données **DICE** (**D**ataset of d**i**sturbing **C**inematographic **E**lements).

4.4 Résultats

4.4.1 Configuration expérimentale

Données utilisées dans les expériences : La solution proposée a été testée sur plusieurs bandes-annonces de films, à savoir *Dune : Première Partie*, *Top Gun : Maverick*, *Buzz L'éclair*, *Ambulance*, *Dragon Ball Super : Super Hero*, *Le Royaume des Abysses*, *Vaiana 2*, *Furiosa : Une saga Mad Max*. Ces bandes-annonces partagent la même définition d'images que les films correspondants, à savoir (858×2048), (1000×2000) et (1080×1920). Toutes sont diffusées à une cadence de 24 images par seconde.

Configuration du réentrainement : Le modèle de détection BiRefNet a été réentrainé sur une GPU Tesla V100 doté de 32 Go de VRAM. L'entrainement a été mené avec un taux d'apprentissage de 10^{-3}, une taille de lot de 4, et sur 50 époques [8], pour une durée totale de 50 heures.

4.4.2 Discussions sur la détection

Les résultats visuels de ces détections sont présentés dans la figure 4.12. Les masques générés par l'algorithme Itti seuillé sont fortement corrélés avec les points caractéristiques de l'image, généralement plus détaillés que leurs zones environnantes. Cependant, ils échouent à fournir une estimation fiable des objets distracteurs, car les points moins détaillés situés à l'intérieur de régions très détaillées ne sont pas sélectionnés. Ce phénomène est observable à l'intérieur du personnage de *Dragon Ball*, sur le visage du personnage de *Dune*, ainsi que dans l'habitacle du cockpit.

Lorsque Itti est combiné avec GrabCut, les résultats sont sensiblement meilleurs, les masques correspondant davantage à des régions précises. Néanmoins, ces deux méthodes

8. Les termes référencés dans la littérature sont le *learning rate*, le *batch size* ou taille de batch, et les *epochs*

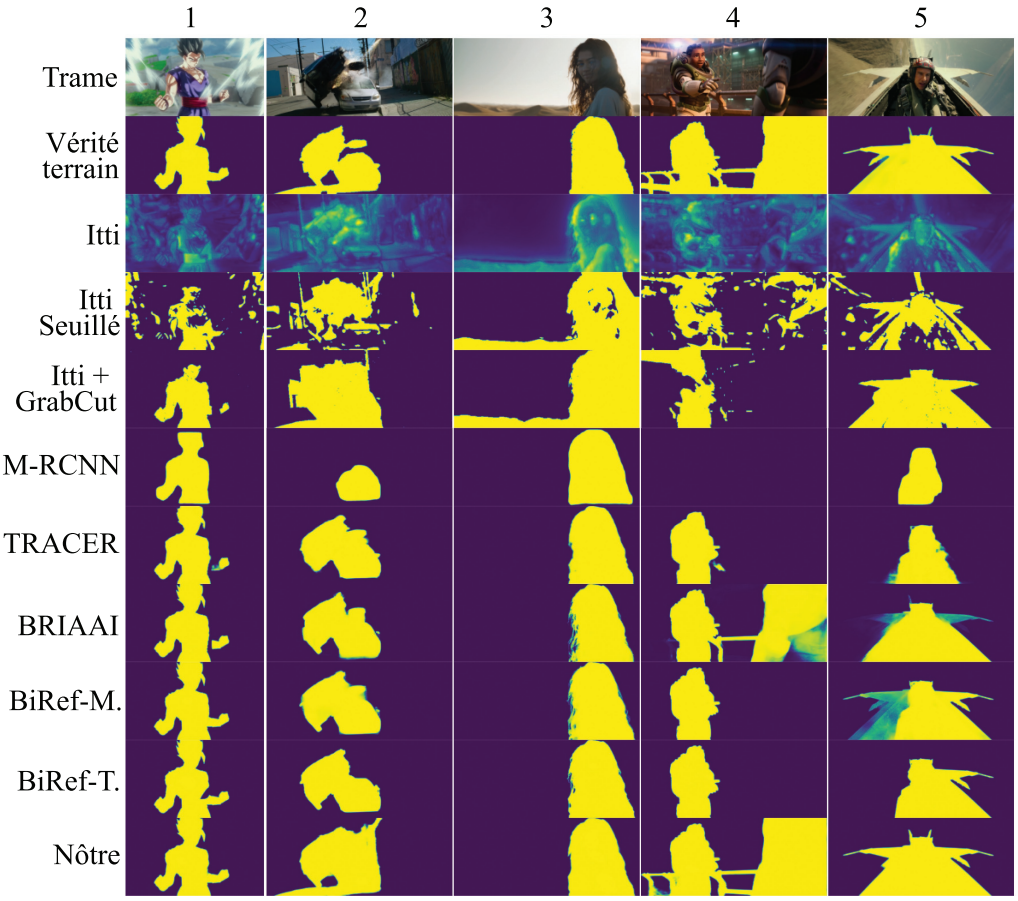


FIGURE 4.12 – **Comparaison visuelle et performances qualitatives de tous les modèles de segmentation utilisés**, respectivement sur les trames des films suivants : *Dragon Ball Super*, *Ambulance*, *Dune*, *Buzz L'éclair*, *Top Gun : Maverick*

restent très instables sur des images successives, entrainant un scintillement constant et important qui les rend inadaptées à un usage pratique.

Mask-RCNN obtient de meilleurs résultats, mais demeure limité par sa dépendance aux catégories d'objets prédéfinies lors de son entrainement. Ce modèle a en effet été initialement entrainé sur les classes du jeu de données COCO, qui inclut des personnages et des objets du quotidien. Or, le modèle montre ses limites face à la diversité artistique propre au cinéma, qui propose souvent des éléments inhabituels et très variés. Ces éléments non détectés peuvent être, comme illustrés sur la figure 4.12, des personnages non réalistes notamment l'acolyte de Buzz L'Éclair, une partie floue d'un personnage comme Buzz L'Éclair vu de dos, ou encore les ailes d'un avion vues sous un angle atypique comme sur l'image 5.

TRACER, le modèle DOS, montre de bonnes performances mais tend à ne considérer qu'un seul objet par image. Ce phénomène, visible dans l'image 4, limite son intérêt dans les situations où plusieurs objets apparaissent de part et d'autre de l'écran. Cette restriction est très probablement liée aux conditions d'entrainement et au jeu de données initialement utilisé pour ce modèle.

Enfin, BRIA et BiRefNet offrent des performances assez similaires, et la comparaison visuelle présentée ici ne permet pas de les distinguer clairement sur une seule image, d'autant que la stabilité temporelle entre images successives est cruciale pour notre application. De la

TABLE 4.1 – Comparaison des performances entre différents modèles d'apprentissage profond et méthodes traditionnelles sur notre jeu de données annoté. La MAE doit être faible, tandis que la MF et la MS doivent être élevées. Les meilleurs scores pour chaque métrique sont soulignés.

Méthode	MAE↓	MF↑	MS↑	Nombre de param.
Itti	0.370	0.296	0.461	/
Itti Seuillé	0.306	0.415	0.469	/
Itti+Grab.	0.266	0.460	0.496	/
M-RCNN	0.399	0.149	0.375	44.2 M
TRACER	0.257	0.443	0.490	66.2 M
BRIAAI	0.226	0.501	0.521	44.0 M
BiRef-M.	0.201	0.531	0.640	220.1 M
BiRef-T.	0.260	0.369	0.554	44.3 M
La nôtre	**0.147**	**0.651**	**0.692**	44.3 M

même manière que TRACER, BiRefNet ne prend pas suffisamment en compte la présence simultanée de plusieurs éléments distracteurs. Pour traiter ces cas particuliers, nous avons jugé plus adapté, dans l'architecture de ICE-Cubed, de diviser les images en deux régions distinctes (« gauche » et « droite »), afin de guider plus efficacement les modèles pour notre cas d'utilisation spécifique. Des améliorations peuvent ainsi être observées aux bords de l'image, tandis que la détection centrale (qui n'est pas pertinente pour notre application) demeure beaucoup moins précise.

Avec la création de notre jeu de données, nous disposons désormais d'images de référence issues du cinéma, permettant de quantifier la précision d'un masque prédit par rapport à son masque de référence. Parmi les métriques couramment utilisées pour évaluer la qualité des masques, trois options populaires sont l'erreur absolue moyenne (MAE), la mesure F pondérée [156] (MF), et la mesure de structure (MS) [157].

Ces métriques confirment dans le tableau 4.1 les performances observées des différents modèles testés, et valident l'amélioration de notre module de détection après son réentrainement afin de l'adapter à nos besoins spécifiques. La MAE est plus faible, tandis que la mesure F pondérée et la MS sont plus élevées, démontrant collectivement une amélioration nette de la précision des prédictions par rapport aux résultats précédents.

L'avis des experts créatifs d'ICE a été recueilli concernant ces différentes expérimentations visant à proposer une version enrichie d'ICE. Tous ont exprimé un vif intérêt pour le projet et l'ont approuvé en l'état. Ils ont souligné que son utilisation pouvait déjà faciliter leur travail et ont envisagé de nouvelles possibilités pour renforcer l'immersion dans des scènes cinématographiques complexes.

4.5 Validation perceptuelle

4.5.1 Méthodologie de l'étude perceptuelle

Afin d'évaluer la validité et la pertinence dans un contexte réel de ICE-Cubed et de la méthodologie proposée, nous avons mené une étude percetuelle avec des experts ICE. De plus, cette immersion s'inscrit dans un contexte particulier, puisque nos travaux portent sur des vidéos issues du monde cinématographique. Par conséquent, elle doit se rapprocher le

plus fidèlement possible de l'intention initiale du réalisateur pour les scènes qui composent le film. Ces experts avec qui a été menée l'étude sont les responsables des effets spéciaux vidéo et de la postproduction ICE. Ils sont au nombre de cinq et constituent les interlocuteurs directs des studios de production cinématographique. Ces derniers valident ou non les travaux réalisés par nos cinq experts pour les films diffusés en ICE. Depuis la création de ICE en 2017, ces cinq experts ont collaboré avec 30 studios de production différents ainsi que leurs chargés de VFX en chef respectifs. Ils disposent ainsi d'un savoir unique concernant les exigences des studios, tant sur le respect des scènes que sur la recherche d'une immersion au service du film. L'objectif de cette étude était d'analyser l'impact de notre approche fondée sur l'inpainting sur l'immersion visuelle, la cohérence spatiale et la perception des artéfacts. L'évaluation a porté sur 29 scènes issues de cinq bandes-annonces de films (*Top Gun : Maverick, Dune : Première Partie, Dune : Deuxième Partie, Vaiana 2, Furiosa : Une saga Mad Max*).

Le questionnaire d'évaluation a été conçu pour recueillir à la fois des retours qualitatifs et quantitatifs sur l'efficacité de notre méthode. Les experts devaient comparer les versions originales et inpaintées de chaque scène et formuler leurs évaluations. Pour chacun des critères suivants, la réponse donnée par l'expert était donnée sur une échelle discrète, soit de 3 niveaux comprenant les qualificatifs « Oui - Pas d'avis - Non », soit de 5 niveaux comprenant les qualificatifs « En désaccord total – Plutôt en désaccord – Neutre – Plutôt d'accord – Tout à fait d'accord ».
ICE-Cubed est préféré : les experts indiquaient s'ils préféraient la version inpaintée ou non. Réponses possibles : Oui - Pas d'avis - Non.
Tremblement : les experts évaluaient leur accord avec l'énoncé : « J'ai perçu une incohérence temporelle ou un scintillement dans les zones reconstruites. Ces artéfacts spatiaux étaient perceptibles et perturbants. ». Réponses possibles : En désaccord total – Plutôt en désaccord – Neutre – Plutôt d'accord – Tout à fait d'accord.
Effet de « Pop » : les experts évaluaient leur accord avec l'énoncé : « J'ai remarqué la réapparition soudaine d'éléments inpaintés. Cet effet de « pop » était visible, concernait une large portion de l'objet et était perturbant. ». Réponses possibles : En désaccord total – Plutôt en désaccord – Neutre – Plutôt d'accord – Tout à fait d'accord.
Continuité : les experts évaluaient leur accord avec l'énoncé : « L'écran principal et les panneaux latéraux semblaient connectés et la continuité visuelle était perceptuellement convaincante. ». Réponses possibles : En désaccord total – Plutôt en désaccord – Neutre – Plutôt d'accord – Tout à fait d'accord.
Ouverture : les experts indiquaient si la version inpaintée améliorait l'ouverture spatiale perçue de la scène. Réponses possibles : Oui - Pas d'avis - Non. Ils évaluaient ensuite leur accord avec l'énoncé : « La scène, accompagnée de son contexte inpainté, procurait une nette impression d'ouverture et d'expansion spatiale. ». Réponses possibles : En désaccord total – Plutôt en désaccord – Neutre – Plutôt d'accord – Tout à fait d'accord.
Immersion : les experts indiquaient si la version inpaintée renforçait l'immersion. Réponses possibles : Oui - Pas d'avis - Non. Ils évaluaient ensuite leur accord avec l'énoncé : « La scène, visionnée avec son contexte inpainté, offrait un très bon niveau global d'immersion. ». Réponses possibles : En désaccord total – Plutôt en désaccord – Neutre – Plutôt d'accord – Tout à fait d'accord.

Cette étude a fourni des enseignements essentiels sur l'efficacité de ICE-Cubed, en particulier pour le maintien de la cohérence spatiale, l'atténuation des artéfacts perturbateurs et le renforcement de l'immersion dans un environnement cinématographique ICE.

4.5.2 Résultats de l'étude perceptuelle

Les résultats de notre étude perceptuelle, présentés à travers deux visualisations distinctes, offrent une compréhension complète des préférences et perceptions des experts sur différents contenus cinématographiques. Cette étude a été conduite dans une réplique de salle ICE. Une illustration de ce que donne les deux types de contenus ICE (original et après ICE-Cubed) montrés lors d'un visionnage en salle est visible en figure 4.13.



FIGURE 4.13 – **Comparaison de rendus en salle ICE, entre le visionnage d'un contenu d'origine et sa transformation une fois ICE-Cubed appliqué.** La trame est issue de la bande-annonce du film *Moonfall*, et présente une écriture devant un paysage apocalyptique. Sans ICE-Cubed (en haut), on peut voir l'écriture être dupliquée dans les panneaux ICE et créer une incohérence de continuité. Après traitement ICE-Cubed (en bas), l'écriture est automatiquement détectée et inpaintée, recréant la continuité à travers les différents équipements.

Les 5 experts assis dans les sièges les plus hauts et les plus centrés de la salle, dans la même configuration que les contrôles qualités effectués par eux-mêmes ou par les distributeurs comme Disney. L'étude a été menée suivant un protocole comparatif explicite, permettant une analyse directe des différences perçues entre les conditions testées. Le plan d'origine est d'abord joué dans la salle, puis sa version modifiée par ICE-Cubed est montrée. Les contenus sont tous montrés en vitesse normale mais ils sont rejoués 3 fois chacun avant de passer au plan suivant. Entre chaque plan, les experts entrent individuellement leurs réponses au questionnaire via un formulaire « Google Forms ». Sur la figure 4.14 qui illustre les évaluations de 29 contenus distincts, nous observons une diversité de réponses. Ces contenus, issus de sources cinématographiques variées, soulignent la nature subjective de l'évaluation visuelle. Cependant, la figure 4.15, qui montre les évaluations des cinq experts ICE, révèle une convergence claire malgré l'absence de concertation entre eux.

Il est particulièrement notable qu'un consensus apparaisse entre les experts sur plusieurs critères, avec un accord marqué dans les sections Préférence et Immersion améliorée. Les experts ont très largement privilégié les versions inpaintées, comme l'illustre la distribution

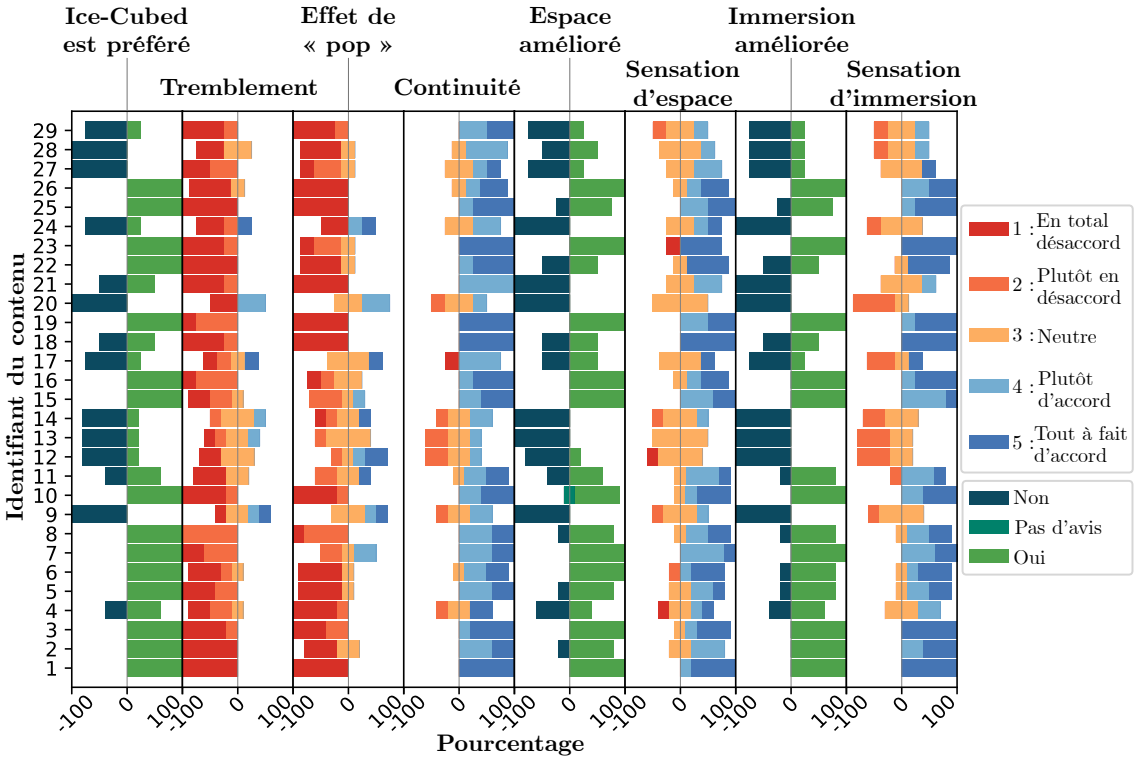


FIGURE 4.14 – Vingt-neuf contenus présentés aux cinq experts. Résultats de l'étude perceptuelle par rapport au contenu.

des couleurs, où les évaluations des contenus inpaintés dominent la zone positive (indiquant une amélioration) pour ces critères. Une tendance similaire est observée dans les évaluations liées à l'Espace amélioré et à la Perception spatiale, où les scènes inpaintées sont généralement perçues comme renforçant la perception spatiale, confirmant ainsi l'efficacité de ICE-Cubed pour élargir l'espace visuel.

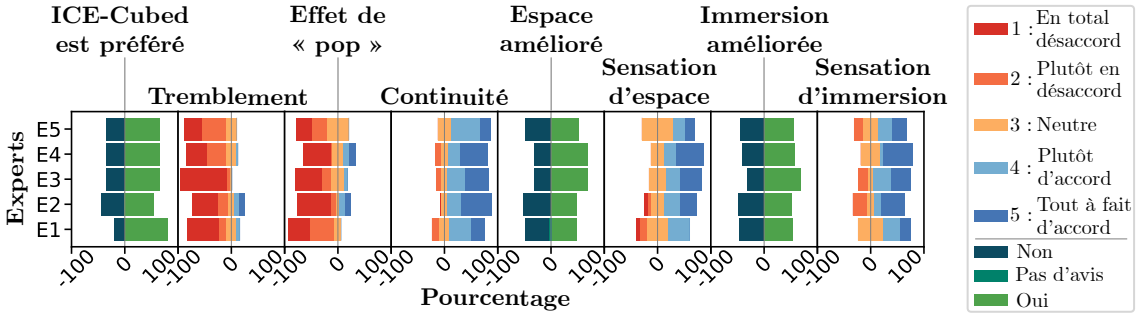


FIGURE 4.15 – Vingt-neuf contenus présentés aux cinq experts. Résultats de l'étude perceptuelle par expert.

À l'inverse, les critères liés au tremblement et aux effets de « pop » montrent une variabilité légèrement plus marquée. Bien que les évaluations de quelques contenus révèlent des divergences sur la figure 4.14, la tendance générale visible en figure 4.15 indique que ces artéfacts perceptuels n'étaient pas significativement perturbateurs, en particulier pour la majorité des contenus. Ce résultat souligne la robustesse de notre méthode dans la réduction des perturbations indésirables.

Les évaluations de la Continuité témoignent d'un fort niveau d'accord, suggérant que les versions inpaintées ont globalement préservé une intégration fluide avec le contenu original, élément essentiel pour maintenir la cohérence spatiale.

En résumé, les résultats indiquent que, malgré une certaine variabilité individuelle dans les réponses, un consensus substantiel se dégage sur l'efficacité globale de ICE-Cubed, notamment pour renforcer l'immersion, la perception spatiale et la continuité visuelle. Ces conclusions suggèrent que notre technique peut être appliquée efficacement à une large variété de contenus cinématographiques, offrant une expérience de visionnage enrichie tout en minimisant les perturbations perceptuelles.

4.6 Conclusion

Ce travail présente une solution conçue pour améliorer l'immersion d'un film en étendant son contenu dans la vision périphérique au sein d'une salle ICE. Étant donné que le contenu halluciné par l'IA doit être évité au cinéma afin de rester conforme à l'intention du réalisateur, nous avons proposé une méthode centrée sur la détection des éléments susceptibles de distraire s'ils étaient affichés en vision périphérique, suivie de l'effacement de ces éléments par une technique d'inpainting. Les bords d'écran ainsi inpaintés deviennent alors adaptés à l'affichage sur les écrans latéraux, afin d'élargir le champ de vision du spectateur. Des opérations morphologiques et des filtres temporels ont été intégrés à ICE-Cubed pour cibler des artéfacts spécifiques.

Plusieurs modèles de pointe ont été comparés objectivement et par inspection visuelle, et le modèle le plus performant a été réentrainé à l'aide d'un jeu de données construit sur mesure, améliorant ainsi sa précision. Ce jeu de données inclut des éléments nouveaux souvent sous-représentés dans les bases standard.

La validation de notre solution via une étude perceptuelle avec cinq experts ICE a montré qu'en moyenne, les experts préféraient la version inpaintée, la jugeant plus immersive tout en introduisant peu d'artéfacts perceptuels. Cela confirme que notre approche est adaptée à des applications réelles et ouvre la voie à de futures améliorations tant sur les aspects techniques que perceptifs des environnements cinématographiques basés sur ICE.

CONCLUSION

> Je te dédie ce jeu, à toi qui es arrivé jusqu'ici.
> Je te l'offre avec tout mon amour.
>
> *Benett Fody*

Devant l'avenir incertain du cinéma face aux différents modes de consommation du film et à son Ubérisation par les plateformes de vidéos à la demande, ICE apparait en 2017 comme une solution immersive afin de créer une expérience augmentée. Bien que le film soit étendu grâce aux panneaux ICE latéraux, des contraintes faisant partie intégrante de ICE viennent freiner la postproduction. Une première contrainte est la prise de décision plan par plan du film afin de décider d'appliquer tel ou tel traitement d'image optimal sur les bords du film en vue d'offrir la meilleure extension sur les panneaux latéraux pour générer l'effet immersif. Une deuxième contrainte réside dans le nombre réduit des traitements rudimentaires possibles et leur restriction au bord du film, qui limite la créativité des graphistes de postproduction.

Cette thèse vise à formaliser et à automatiser partiellement le processus de création des contenus immersifs du dispositif ICE, en s'appuyant sur l'analyse filmique, la modélisation de graphes sémantiques et la conception de modèles d'apprentissage profond afin d'améliorer l'efficacité ainsi que l'immersion que peut proposer le format ICE.

Le manuscrit est structuré en quatre chapitres : le premier retrace l'évolution du cinéma, les fondements de l'immersion et de l'IA. Le deuxième présente la formalisation du processus de création ICE et l'élaboration d'un graphe sémantique d'analyse filmique et de son ontologie ICE-HOLSS . Le troisième décrit le développement de modèles d'apprentissage profond destinés à prédire les traitements d'images immersifs optimaux, à partir de convolutions 3D (R-ICE) puis à partir d'attention (ICE-AGE). Enfin, le dernier chapitre présente ICE-Cubed dédié à la reconstruction et à l'enrichissement visuel des contenus immersifs. Il a ensuite été évalué au moyen d'une étude perceptuelle conduite auprès des experts ICE.

Le chapitre 1, par un historique du cinéma, nous a permis de comprendre ses points forts et la motivation derrière cette recherche constante de l'engagement et l'immersion du spectateur. Un état de l'art des autres formats similaires à ICE , tant dans le secteur de la recherche académique que dans le secteur de l'innovation commerciale, nous a permis de comprendre les directions qui peuvent intéresser à la fois les spectateurs et les distributeurs de films tout en étant pérennes. Une étude de l'immersion et de l'IA dans la littérature scientifique a mis en relief les points qui peuvent être améliorés dans la technique ICE.

Le chapitre 2 a été consacré à la formalisation du processus de création ICE, afin de comprendre les logiques de travail et les critères artistiques mobilisés par les graphistes lors de la conception des extensions latérales. Il a également présenté une étude approfondie de l'analyse filmique, fondée sur les éléments récurrents de la mise en scène et sur les principes largement partagés par les réalisateurs concernant la perception et les effets recherchés chez le spectateur. À partir de ces fondations, une modélisation des connaissances a été proposée sous la forme d'un graphe sémantique, puis d'une ontologie ICE-HOLSS dédiée à ICE, permettant d'organiser et de relier de manière structurée les relations entre les composantes filmiques, les intentions narratives et les effets perceptifs. Cette étape offre un socle conceptuel solide pour aborder, dans le chapitre suivant, la prédiction automatique des choix artistiques à l'aide de l'apprentissage profond.

Le chapitre 3 est dédié à la prédiction de la décision des graphistes pour chaque plan de film. Après l'état de l'art et l'étude de performances des modèles d'analyse vidéo, l'architecture RICE est mise au point, puis réentrainée selon les données de vérité terrain obtenues de ICE. Après un réagencement des données pour en faire des données d'entrainement viables ainsi que l'application d'une augmentation de données étudiée avec l'expert ICE, les hyper-paramètres ont permis de prédire à 71,72 % le bon traitement d'images à choisir pour optimiser l'immersion du spectateur, selon un plan aléatoire. Ce pourcentage reflète le taux d'accord entre les prédictions du modèle et les décisions des experts. Il est à noter qu'une part importante des prédictions incorrectes restantes correspond à des prédictions ne présentant pas d'incohérences visibles et pouvant être validées en l'état, bien que légèrement moins immersives. Ce travail a donné lieu à une publication dans le workshop international MMSP [158]. Une architecture repensée de zéro, en meilleure adéquation avec les problématiques de ICE, a été réalisée en s'inspirant des mécanismes d'attention et des modèles de *transformers* modernes, donnant lieu à ICE-AGE .

Le chapitre 4 se focalise sur la cohérence spatiale obligatoire entre la toile de cinéma et les panneaux ICE pour respecter l'immersion du spectateur. L'objectif de ce chapitre est de proposer aux graphistes ICE une solution ICE-Cubed à base d'apprentissage profond qui peut générer selon un plan donné, un nouveau traitement d'image à vocation d'immersion. Ce système détecte les éléments perturbant l'immersion des spectateurs afin de les effacer par une technique d'inpainting. Le résultat obtenu est superrésolu pour augmenter sa définition, puis affiché dans les panneaux. Pour réaliser cette solution, un état de l'art ainsi qu'une étude de performances des différents types de modèles de segmentation et d'inpainting video est réalisé. En ce qui concerne la détection des éléments, le logiciel SSPICE (Scene Segmentation Platform by ICE) a été créé sur mesure pour annoter rapidement des trames de bandes-annonces. Grâce à cela, 5 000 trames ont été segmentées, constituant ainsi le jeu de données DICE (Dataset of dIsturbing Cinematographic Elements) qui nous a permi de spécialiser notre IA en la réentrainant sur cette tâche de détection unique en son genre. La vérité terrain nous permet aussi de vérifier une précision quantitative selon des métriques de correspondance de masques. Pour ce qui est du rendu général, une étude perceptuelle a été conduite avec les cinq experts de rendu ICE. Ce travail a été valorisé par une publication internationale dans la conférence ACIVS [159].

Ces travaux ont mis en avant le potentiel d'aide que peut permettre l'apprentissage profond dans le secteur du cinéma lorsqu'il est manipulé à bon escient. Ces algorithmes à base de réseaux de neurones ont bel et bien pour vocation de faire perdurer le cinéma et tous ses acteurs dans le temps via des innovations nécessaires face à un avenir incertain. Ils constituent une avancée vers le développement de solutions pour des milieux à application

et données particulièrement niches. Il demeure néanmoins que le domaine du cinéma reste très fermé et les données sont difficiles d'accès, sans parler des annotations des experts qui représentent une information d'une valeur inestimable.

4.7 Perspectives

Différents axes d'améliorations sont possibles pour ces travaux. Nous aborderons d'abord les perspectives à court terme, puis nous continuerons avec celles à moyen et long termes.

Dans un premier temps, compte tenu de la commercialisation actuelle de ICE, des données de vérité terrain apparaissent tous les mois. L'architecture R-ICE a été efficace, mais elle a été entrainée seulement sur une quarantaine de bandes-annonces, et il serait intéressant de la réentrainer sur des films récents annotés par les experts de CGR Cinémas.

Aussi, étant donné que la dynamique d'une bande-annonce diffère légèrement de celle d'un film, le plus intéressant serait d'inclure les données de films. En passant d'un contenu de 2 min 30 secondes à 2 heures 30 minutes en moyenne, on multiplierait la quantité de nos données par 60 et on augmenterait également sa diversité ainsi que sa qualité. Cependant, dû aux multiples versions d'un film, pouvoir exploiter toutes les données de films demeure un véritable défi. Ce dernier nécessiterait une synchronisation en aveugle entre les annotations textuelles de l'équipe ICE et les contenus vidéo disponibles. Seule une minorité des films n'a pas besoin de cette resynchronisation.

En ce qui concerne le moyen terme, la dimension de l'explicabilité peut être une piste très intéressante, notamment parce que les travaux ont une vocation industrielle, mais pas seulement. D'une part, les résultats ont pour vocation d'être vus par un public, et même si les graphistes travaillant sur ces projets sont particulièrement méticuleux, il faudra néanmoins arriver à convaincre les différents acteurs et décideurs du cinéma. L'explicabilité permet de démystifier l'IA mais aussi de renforcer sa fiabilité en gardant un droit de contrôle dessus. Il est ainsi possible d'identifier si l'inférence effectuée par le modèle est porteuse de sens. Les modèles comme Grad-CAM permettent de comprendre quels sont les pixels (ou trames) d'entrées les plus influents dans la décision finale de l'IA.

À long terme, les données de vérité terrain apparaissant en continu au rythme des sorties de cinéma promettent une potentielle base de données colossale et utilisant les nouveaux outils de cette thèse. Les dimensions de MLOps et leurs principes permettraient de planifier un modèle d'IA qui s'entraine selon les nouvelles sorties des films en ICE. L'IA pourrait ainsi se corriger selon les souhaits des graphistes avec les prédictions qui en découlent.

BIBLIOGRAPHIE

[1] G. Tissandier (rédacteur en chef), *La Nature. Revue des sciences et de leurs applications aux arts et à l'industrie*, 10e année, 2e sem., n°s 470–495, Paris : G. Masson, 1882, p. 357, https ://cnum.cnam.fr/redir ?4KY28, last accessed : Apr. 16, 2026.

[2] S. M. Anstis, « A chart demonstrating variations in acuity with retinal position, » *Vision Research*, vol. 14, no. 7, pp. 589–592, Jul. 1974.

[3] V. Nair and G. E. Hinton, « Rectified linear units improve restricted Boltzmann machines, » in Proceedings of the International Conference on Machine Learning (ICML), Madison, WI, Jun. 2010, pp. 807–814.

[4] B. Xu, N. Wang, T. Chen, and M. Li, « Empirical evaluation of rectified activations in convolutional network, » *arXiv preprint arXiv :1505.00853*, Nov. 2015.

[5] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, « Fast and accurate deep network learning by exponential linear units (ELUs), » *arXiv preprint arXiv :1511.07289*, Feb. 2016.

[6] M. Lee, « GELU activation function in deep learning : A comprehensive mathematical analysis and performance, » *arXiv preprint arXiv :2308.01033*, Aug. 2023.

[7] S. Elfwing, E. Uchibe, and K. Doya, « Sigmoid-weighted linear units for neural network function approximation in reinforcement learning, » *Neural Networks*, vol. 107, pp. 3–11, Jan. 2018.

[8] Y.-C. Cheng, C. H. Lin, H.-Y. Lee, J. Ren, S. Tulyakov, and M.-H. Yang, « InOut : Diverse image outpainting via GAN inversion, » in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, Jun. 2022, pp. 11421–11430.

[9] M. Azéma, « Prehistoric cave art : From image to graphic narration, » *Paragraph*, vol. 44, no. 3, pp. 379–393, Nov. 2021.

[10] D. E. Novy, « Computational immersive displays, » Ph.D. dissertation, Massachusetts Institute of Technology, 2013.

[11] G. Sadoul, « Histoire du cinéma mondial, des origines à nos jours », Paris, Flammarion, 719 p., p. 66-68, 1968

[12] T. Gunning, « An aesthetic of astonishment : Early film and the (in)credulous spectator, » *Art & Text*, vol. 34, no. 1, pp. 31–45, 1989.

[13] T. Porter and T. Duff, « Compositing digital images, » in Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), New York, NY, Jan. 1984, pp. 253–259.

[14] F. J. C. Blázquez, M. Cappello, G. Fontaine, J. T. Milla, and S. Valais, *Release Windows in Europe : A Matter of Time*. European Audiovisual Observatory, Council of Europe, 2019.

[15] B. Salvan, « La chronologie des médias face aux nouveaux usages : un changement de modèle inéluctable ? », PSL Research Paris, May 2018.

[16] W. Higbee and J. Vernon, « 'Que le carnage commence !' : French cinema, la chronologie des médias and the rise of the streamers, » *French Screen Studies*, vol. 24, no. 4, pp. 406–424, Oct. 2024.

[17] K. Kitsopanidou, « Innovation strategies of independent Parisian cinemas in the post-pandemic, platformized media environment, » *Création collective au cinéma*, no. 7, Feb. 2023.

[18] T. Graindorge, « Diffusion des films : Disney+ sort gagnant dans la nouvelle chronologie des médias, » *Le Point*, Jan. 2025, https ://www.lepoint.fr/economie/diffusion-des-films-disney-sort-gagnant-dans-la-nouvelle-chronologie-des-medias-29-01-2025-2581129_28.php, last accessed : Nov. 12, 2025.

[19] I. D. Lebocq, « Chronologie des médias : Canal+ revoit son engagement à la baisse, un tournant pour le financement du cinéma français, » *FreeNews*, Mar. 2025, https ://www.freenews.fr/actu-du-net/chronologie-des-medias-canal-revoit-son-engagement-a-la-baisse-un- tournant-pour-le-financement-du-cinema-francais, last accessed : Nov. 12, 2025.

[20] O. Ubertalli, « Cinéma français : les raisons du coup de force de Netflix, » *Le Point*, Apr. 2025, https ://www.lepoint.fr/economie/chronologie-des-medias-netflix-passe-a-l-attaque-1204-2025-2587186_28.php, last accessed : Nov. 12, 2025.

[21] A. Horton, « 'The full immersive experience' : The rise of cinema with an added twist, » *The Guardian*, Dec. 2024, https ://www.theguardian.com/film/2024/dec/19/immersive-experience-movies, last accessed : Nov. 12, 2025.

[22] S. Masunaga, « The casual moviegoer is a thing of the past. That's a big problem for Hollywood, » *Los Angeles Times*, Dec. 2024, https ://www.latimes.com/entertainment-arts/business/story/2024-12-02/the-casual-moviegoer-is-a-thing-of-the-past-thats-a- big-problem-for-hollywood, last accessed : Nov. 12, 2025.

[23] C. B. Weinberg, C. Otten, B. Orbach, J. McKenzie, R. Gil, D. C. Chisholm, and S. Basuroy, « Technological change and managerial challenges in the movie theater industry, » *Journal of Cultural Economics*, vol. 45, no. 2, pp. 239–262, Jun. 2021.

[24] « At these movie theaters, kids can play on a jungle gym next to their parents' seats, » *Los Angeles Times*, Mar. 2017, https ://www.latimes.com/business/hollywood/la-fi-ct-cinepolis-junior-20170302-story.html, last accessed : Nov. 12, 2025.

[25] R. Faughnder, « Rumbling seats. Virtual reality. Booze. How cinemas are adapting to uncertain future, » *Los Angeles Times*, Jun. 2017, last accessed : Nov. 12, 2025.

[26] S. Agrawal, A. Simon, S. Bech, K. Bærentsen, and S. Forchhammer, « Defining immersion : Literature review and implications for research on audiovisual experiences, » *Journal of the Audio Engineering Society*, vol. 68, no. 6, pp. 404–417, Jul. 2020.

[27] A. McMahan, « Immersion, engagement, and presence : A method for analyzing 3-D video games, » in *The Video Game Theory Reader*, M. J. P. Wolf and B. Perron, Eds. Routledge, 2004.

[28] M. Lombard and T. Ditton, « At the heart of it all : The concept of presence, » *Journal of Computer-Mediated Communication*, vol. 3, no. 2, Sep. 1997.

[29] M. Slater and S. Wilbur, « A framework for immersive virtual environments (FIVE) : Speculations on the role of presence in virtual environments, » *Presence : Teleoperators and Virtual Environments*, vol. 6, no. 6, pp. 603–616, Dec. 1997.

[30] L. Ermi and F. Mäyrä, « Fundamental components of the gameplay experience : Analysing immersion, » in *Proceedings of the International Conference on Digital Games Research (Worlds in Play)*, Vancouver, Canada, Jan. 2005.

[31] J.-N. Thon, « Immersion revisited : On the value of a contested concept, » Ph.D. dissertation, Universität Tübingen, 2008.

[32] A. Odeychuk, « Devising a method to adaptively determine human movement speed in crowd behavior simulation at extreme events, » *Eastern-European Journal of Enterprise Technologies*, vol. 2, pp. 66–74, Apr. 2025.

[33] J. Moreland and A. Cruz, « Colour perception with the peripheral retina, » *Optica Acta : International Journal of Optics*, vol. 6, no. 2, pp. 117–151, Apr. 1959.

[34] S. P. McKee and K. Nakayama, « The detection of motion in the peripheral visual field, » *Vision Research*, vol. 24, no. 1, pp. 25–32, 1984.

[35] M. De Murcia, « Étude et réalisation de solutions de communication en lumière visible en milieux industriels et professionnels, » Ph.D. dissertation, Université de Poitiers, 2024.

[36] C. M. Lachaud, « Les effets perturbateurs liés au scintillement subliminal des écrans cathodiques sur le contrôle oculomoteur lors du déplacement saccadique des yeux, » Ph.D. dissertation, Université de Provence, 1997.

[37] S. M. Berman, D. S. Greenhouse, I. L. Bailey, R. D. Clear, and T. W. Raasch, « Human electroretinogram responses to video displays, fluorescent lighting, and other high frequency sources, » *Optometry and Vision Science*, vol. 68, no. 8, pp. 645–662, Aug. 1991.

[38] J. Wagemans, J. H. Elder, M. Kubovy, S. E. Palmer, M. A. Peterson, M. Singh, and R. von der Heydt, « A century of Gestalt psychology in visual perception I : Perceptual grouping and figure-ground organization, » *Psychological Bulletin*, vol. 138, no. 6, pp. 1172–1217, Nov. 2012.

[39] P. Baudisch, N. Good, and P. Stewart, « Focus plus context screens : Combining display technology with visualization techniques, » in Proceedings of the Annual ACM Symposium on User Interface Software and Technology (UIST), Orlando, FL, 2001.

[40] P. Seuntiens, I. Vogels, and A. van Keersop, « Visual experience of 3D-TV with pixelated Ambilight, » in Proceedings of the Annual International Society for Presence Research (ISPR), Barcelona, Spain, Oct. 2007.

[41] « Philips AmbiLux PUS8901 review : Projector-laden 4K TV is mad, but brilliant, » *Pocket-lint*, Jun. 2016, https ://www.pocket-lint.com/tv/reviews/philips/137778-philips-ambilux-pus8901-review- projector-laden-4k-tv-is-mad- but-brilliant/, last accessed : Nov. 12, 2025.

[42] N. Kimura and J. Rekimoto, « ExtVision : Augmentation of visual experiences with generation of context images for a peripheral vision using deep neural network, » in Proceedings of the Annual Conference on Human Factors in Computing Systems (CHI), New York, NY, Apr. 2018, pp. 1–10.

[43] « Razer Ariana - Video Projection System, » *Razer*, Apr. 2018, https ://www.razer.com/concepts/project-ariana, last accessed : Nov. 12, 2025.

[44] B. R. Jones, H. Benko, E. Ofek, and A. D. Wilson, « IllumiRoom : Peripheral projected illusions for interactive experiences, » in Proceedings of the Annual ACM SIGCHI Conference on Human Factors in Computing Systems (CHI), Paris, France, Apr. 2013, pp. 869–878.

[45] P. Mills, A. Sheikh, G. Thomas, and P. Debenham, « Surround video, » in Proceedings of the Networked Electronic Media Summit (NEM Summit), Torino, Italy, Sep. 2011, pp. 55–63.

[46] B. Jones, R. Sodhi, M. Murdock, R. Mehra, H. Benko, A. Wilson, E. Ofek, B. MacIntyre, N. Raghuvanshi, and L. Shapira, « RoomAlive : Magical experiences enabled by scalable, adaptive projector-camera units, » in Proceedings of the Annual ACM Symposium on User Interface Software and Technology (UIST), Honolulu, HI, Oct. 2014, pp. 637–644.

[47] C. Cruz-Neira, D. J. Sandin, T. A. DeFanti, R. V. Kenyon, and J. C. Hart, « The CAVE : Audio visual experience automatic virtual environment, » *Communications of the ACM*, vol. 35, no. 6, pp. 64–73, Jun. 1992.

[48] R. Xiao and H. Benko, « Augmenting the field-of-view of head-mounted displays with sparse peripheral displays, » in Proceedings of the Annual ACM Conference on Human Factors in Computing Systems (CHI), San Jose, CA, May 2016.

[49] P. Lubos, G. Bruder, O. Ariza, and F. Steinicke, « Ambiculus : LED-based low-resolution peripheral display extension for immersive head-mounted displays, » in Proceedings of the ACM Virtual Reality International Conference (VRIC), New York, NY, Mar. 2016, pp. 1–4.

[50] A. M. Hashemian, A. Kitson, T. Nquyen-Vo, H. Benko, W. Stuerzlinger, and B. E. Riecke, « Investigating a sparse peripheral display in a head-mounted display for VR locomotion, » in Proceedings of the Annual IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Reutlingen, Germany, Mar. 2018, pp. 571–572.

[51] N. Kimura, M. Kono, and J. Rekimoto, « Deep dive : Deep-neural-network-based video extension for immersive head-mounted display experiences, » in Proceedings of the Annual ACM International Symposium on Pervasive Displays, Palermo, Italy, Jun. 2019, pp. 1–7.

[52] I. M. Pepperberg, « Grey parrot (Psittacus erithacus) numerical abilities : Addition and further experiments on a zero-like concept, » *Journal of Comparative Psychology*, vol. 120, no. 1, pp. 1–11, 2006.

[53] R. A. Gardner and B. T. Gardner, « Teaching sign language to a chimpanzee, » *Science*, vol. 165, no. 3894, pp. 664–672, Aug. 1969.

[54] F. Wario, B. Wild, R. Rojas, and T. Landgraf, « Automatic detection and decoding of honey bee waggle dances, » *PLOS One*, vol. 12, no. 12, p. e0188626, Dec. 2017.

[55] « MythBusters (2004 season), » *Wikipedia*, Jul. 2025, last accessed : Nov. 12, 2025.

[56] B. Meyer, « John McCarthy, » *Communications of the ACM Blog*, Oct. 2011. [Online]. Available : https ://cacm.acm.org/blogs/blog-cacm/138907-john-mccarthy/fulltext, last accessed : Jun. 03, 2026.

[57] R. Brooks, *Flesh and Machines : How Robots Will Change Us*. New York, USA : Pantheon Books, 2002.

[58] L. Tesler, « Adages and Coinages, » *nomodes.com*. [Online]. Available : https ://www.nomodes.com/larry-tesler-consulting/adages-and-coinages, last accessed : Jun. 03, 2026.

[59] N. Bostrom, « Nick Bostrom on Artificial Intelligence, » *Oxford Martin School Blog*, Sep. 2014. [Online]. Available : https ://www.oxfordmartin.ox.ac.uk/blog/nick-bostrom-on-artificial-intelligence, last accessed : Jun. 03, 2026.

[60] B. Zion, A. Barki, J. Grinshpon, L. Rosenfeld, and I. Karplus, « Retention of acoustic conditioning in St Peter's fish Sarotherodon galilaeus, » *Journal of Fish Biology*, vol. 78, no. 3, pp. 838–847, Mar. 2011.

[61] E. Thomas, « Descartes on the animal within, and the animals without, » *Canadian Journal of Philosophy*, vol. 50, no. 8, pp. 999–1014, Nov. 2020.

[62] K. Andrews and S. Monsó, « Animal cognition, » in *The Stanford Encyclopedia of Philosophy*, spring 2021 ed., E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2021.

[63] W. Smith, *Dictionary of Greek and Roman Biography and Mythology*, vol. 2. CC Little and J. Brown, 1867.

[64] L. Floridi and M. Chiriatti, « GPT-3 : Its nature, scope, limits, and consequences, » *Minds and Machines*, vol. 30, no. 4, pp. 681–694, Dec. 2020.

[65] R. Restrepo Echavarría, « ChatGPT-4 in the Turing Test, » *Minds and Machines*, vol. 35, no. 1, p. 8, Jan. 2025.

[66] F. Rosenblatt, « The perceptron : A probabilistic model for information storage and organization in the brain, » *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

[67] M. Minsky and S. A. Papert, *Perceptrons : An Introduction to Computational Geometry*. MIT Press, Sep. 2017.

[68] Suétone, *Vie des douze Césars : Vie de Vespasien*, XVIII, trad. J. C. Rolfe, Loeb Classical Library, 1914. Available : https ://penelope.uchicago.edu/thayer/e/roman/texts/suetonius/12caesars/vespasian*.html, last accessed : Jun. 2, 2026.

[69] Britannica Editors, Luddite, *Encyclopædia Britannica*. Available : https ://www.britannica.com/event/Luddite, Accessed : Jun. 2, 2026.

[70] Why did the Luddites protest ?, *The National Archives*, Oct. 2022. Available : https ://www.nationalarchives.gov.uk/education/resources/why-did-the-luddites-protest/, last accessed : Jun. 2, 2026.

[71] J. L. Zagorsky, What today's labor leaders can learn from the explosive rise and quick fall of the typesetters union, *The Conversation*, Jul. 2023. Available : https ://theconversation.com/what-todays-labor-leaders-can-learn-from-the-explosive-rise-and-quick-fall-of-the-typesetters-union-214519, last accessed : Jun. 2, 2026.

[72] Transforming the Waterfront, *National Museum of American History, Smithsonian Institution*. Available : https ://americanhistory.si.edu/explore/exhibitions/america-on-the-move/online/transforming-waterfront, last accessed : Jun. 2, 2026.

[73] A. Verdier-Molinié, Grève de la presse : pourquoi la CGT bloque Presstalis, *Fondation IFRAP*. Available : https ://www.ifrap.org/emploi-et-politiques-sociales/greve-de-la-presse-pourquoi-la-cgt-bloque-presstalis, last accessed : Jun. 2, 2026.

[74] R. Dédame, *Le Syndicat du livre-CGT et les travailleurs des imprimeries parisiennes*. Paris, France : Les Indes savantes, 2012.

[75] "Writers' Guild of America (WGA) Strike Resolution : AI Restrictions and Implications," *Fordham Journal of Corporate and Financial Law*, Oct. 2023. Available : https ://news.law.fordham.edu/jcfl/2023/10/25/writers-guild-of-america-wga-strike-resolution-ai-restrictions-and-implications/, last accessed : Jun. 03, 2026.

[76] The SAG-AFTRA Strike is Over, But the AI Fight in Hollywood is Just Beginning, *Center for Democracy and Technology*, Jan. 2024. Available : https ://cdt.org/insights/the-sag-aftra-strike-is-over-but-the-ai-fight-in-hollywood-is-just-beginning/, last accessed : Jun. 2, 2026.

[77] France Télévisions - Rédaction Culture, Après une pétition pour contester le doublage produit par l'IA, les comédiens français se mobilisent, *franceinfo*, mai 2024. Available : https ://www.franceinfo.fr/culture/cinema/apres-une-petition-pour-contester-le-doublage-produit-par-l-␣ia-les-comediens-francais-se-mobilisent_6547688.html, last accessed : Jun. 2, 2026.

[78] D. Lingemann, Protection des comédiens de doublage face à l'intelligence artificielle, Question écrite n° 6706, *Assemblée nationale*, 17e législature. Available : https ://www.assemblee-nationale.fr/dyn/17/questions/QANR5L17QE6706, last accessed : Jun. 2, 2026.

[79] J. Winterton, The 1984–85 miners' strike and technological change, *The British Journal for the History of Science*, vol. 26, no. 1, pp. 5–14, Mar. 1993.

[80] D. Metz, How the Miners' Strike Was Lost, *History Today*, Mar. 2024. Available : https ://www.historytoday.com/archive/feature/how-miners-strike-was-lost, last accessed : Jun. 2, 2026.

[81] History of Mining Strikes, *Special Collections and Archives, University of Kent*. [Online]. Available : https ://www.kent.ac.uk/library/special-collections/our-exhibitions/mining-in-kent/history-of-mining-strikes, last accessed : Jun. 2, 2026.

[82] A. D. Rasamoelina, F. Adjailia, and P. Sincak, « A review of activation function for artificial neural network, » in Proceedings of the IEEE World Symposium on Applied Machine Intelligence and Informatics (SAMI), Herl'any, Slovakia, Jan. 2020, pp. 281–286.

[83] C. Bishop, *Neural Networks for Pattern Recognition*, ser. Advanced Texts in Econometrics. Clarendon Press, 1995.

[84] S.-i. Amari, « Backpropagation and stochastic gradient descent method, » *Neurocomputing*, vol. 5, no. 4, pp. 185–196, Jun. 1993.

[85] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, « Learning representations by backpropagating errors, » *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

[86] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[87] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, « Gradient-based learning applied to document recognition, » *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[88] A. Krizhevsky, I. Sutskever, and G. E. Hinton, « ImageNet classification with deep convolutional neural networks, » in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25. Curran Associates, 2012.

[89] K. Simonyan and A. Zisserman, « Very deep convolutional networks for large-scale image recognition, » *arXiv preprint arXiv :1409.1556*, Sep. 2014.

[90] K. He, X. Zhang, S. Ren, and J. Sun, « Deep residual learning for image recognition, » in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 770–778.

[91] P. Bhoir and S. Kolte, « Sentiment analysis of movie reviews using lexicon approach, » in *Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, Madurai, India, 2015, pp. 1–6.

[92] J. Du, H. Xu, and X. Huang, « Box office prediction based on microblog, » *Expert Systems with Applications*, vol. 41, no. 4, Part 2, pp. 1680–1689, 2014.

[93] J. Krauss, S. Nann, D. Simon, K. Fischbach, and P. Gloor, « Predicting movie success and academy awards through sentiment and social network analysis, » in *Proceedings*

of the European Conference on Information Systems (ECIS), Galway, Ireland, 2008, p. Paper 116.

[94] S. Kumar, K. De, and P. P. Roy, « Movie recommendation system using sentiment analysis from microblogging data, » *IEEE Transactions on Computational Social Systems*, vol. 7, no. 4, pp. 915–923, 2020.

[95] C. Lee, D. Han, K. Han, and M. Yi, « Improving graph-based movie recommender system using cinematic experience, » *Applied Sciences*, vol. 12, no. 1493, 2022.

[96] J. Li, Y. Ding, Y. Shi, and J. Zhang, « Building a large annotation ontology for movie video retrieval, » *International Journal of Digital Content Technology and its Applications*, vol. 4, no. 5, pp. 74–81, 2010.

[97] A. Cannavò, A. Castiello, F. G. Pratticò, T. Mazali, and F. Lamberti, « Immersive movies : The effect of point of view on narrative engagement, » *AI & Society*, vol. 39, pp. 1811–1825, 2024.

[98] T. R. Gruber, « A translation approach to portable ontology specifications, » *Knowledge Acquisition*, vol. 5, no. 2, pp. 199–220, 1993.

[99] T. Berners-Lee, J. Hendler, and O. Lassila, « The Semantic Web : A new form of web content that is meaningful to computers will unleash a revolution of new possibilities, » in *Linking the World's Information : Essays on Tim Berners-Lee's Invention of the World Wide Web*, 2023, pp. 91–103.

[100] M. Uschold and M. Gruninger, « Ontologies and semantics for seamless connectivity, » *ACM SIGMOD Record*, vol. 33, no. 4, pp. 58–64, 2004.

[101] F. Manola, E. Miller, and B. McBride, « RDF primer, » *W3C Recommendation*, vol. 10, no. 1–107, p. 6, 2004.

[102] B. McBride, « The Resource Description Framework (RDF) and its vocabulary description language RDFS, » in *Handbook on Ontologies*, Springer, 2004, pp. 51–65.

[103] X. H. Wang, D. Q. Zhang, T. Gu, and H. K. Pung, « Ontology-based context modeling and reasoning using OWL, » in *Proceedings of the IEEE Annual Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, Orlando, FL, 2004, pp. 18–22.

[104] B. Bouchon-Meunier, *La Logique Floue*, ser. Que Sais-Je ? Humensis, 2013.

[105] R. Girshick, J. Donahue, T. Darrell, and J. Malik, « Rich feature hierarchies for accurate object detection and semantic segmentation, » in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, 2014, pp. 580–587.

[106] S. Sooksatra and S. Watcharapinchai, « A comprehensive review on temporal-action proposal generation, » *Journal of Imaging*, vol. 8, no. 8, p. 207, Aug. 2022.

[107] B. Xiong, Y. Kalantidis, D. Ghadiyaram, and K. Grauman, « Less is more : Learning highlight detection from video duration, » in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, Jun. 2019, pp. 1258–1267.

[108] Y. Rui, A. Gupta, and A. Acero, « Automatically extracting highlights for TV baseball programs, » in *Proceedings of the ACM International Conference on Multimedia*, New York, NY, Oct. 2000, pp. 105–115.

[109] H. Xia and Y. Zhan, « A survey on temporal action localization, » *IEEE Access*, vol. 8, pp. 70477–70487, 2020.

[110] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, « A closer look at spatiotemporal convolutions for action recognition, » in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, Jun. 2018, pp. 6450–6459.

[111] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, « ActivityNet : A large-scale video benchmark for human activity understanding, » in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, Jun. 2015, pp. 961–970.

[112] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, *et al.*, « Scaling egocentric vision : The EPIC-KITCHENS dataset, » in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 720–736.

[113] B. Jia, T. Lei, S.-C. Zhu, and S. Huang, « EgoTaskQA : Understanding human tasks in egocentric videos, » in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, Oct. 2022.

[114] E. Vahdani and Y. Tian, « Deep learning-based action detection in untrimmed videos : A survey, » *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4302–4320, 2023.

[115] L. Fridman and Y. LeCun, « Transcript for Yann LeCun : Meta AI, open source, limits of LLMs, AGI & the future of AI, » *Lex Fridman Podcast*, Mar. 2024, https ://lexfridman.com/yann-lecun-3-transcript/, last accessed : Nov. 12, 2025.

[116] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, « Learning spatiotemporal features with 3D convolutional networks, » in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 4489–4497.

[117] J. Carreira and A. Zisserman, « Quo vadis, action recognition ? A new model and the Kinetics dataset, » in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 4724–4733.

[118] H. Alwassel, S. Giancola, and B. Ghanem, « TSP : Temporally-sensitive pretraining of video encoders for localization tasks, » in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Montreal, Canada, 2021.

[119] C. Feichtenhofer, H. Fan, J. Malik, and K. He, « SlowFast networks for video recognition, » in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, South Korea, Oct. 2019.

[120] International Telecommunication Union, « Subjective video quality assessment methods for multimedia applications, » *ITU Telecommunication Standardization Sector (ITU-T) Recommendation P.910*, Oct. 2023.

[121] C. Shorten and T. M. Khoshgoftaar, « A survey on image data augmentation for deep learning, » *Journal of Big Data*, vol. 6, no. 1, p. 60, Jul. 2019.

[122] C. Hughes, « Demystifying PyTorch's WeightedRandomSampler by example, » *Towards Data Science*, Aug. 2022, https ://towardsdatascience.com/demystifying-pytorchs- weightedrandomsampler-by-example-a68aceccb452, last accessed : Nov. 12, 2025.

[123] F. López, « Class imbalance : random sampling and data augmentation with imbalanced-learn, » *Towards Data Science*, Feb. 2021, https ://towardsdatascience.com/class-imbalance-random-sampling-and- data-augmentation-with-imbalanced-learn- 63f3a92ef04a, last accessed : Nov. 12, 2025.

[124] Z. S. Harris, « Distributional structure, » *WORD*, vol. 10, no. 2–3, pp. 146–162, Aug. 1954.

[125] S. Raschka, *Build a Large Language Model (From Scratch)*, Oct. 2024.

[126] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, « On the properties of neural machine translation : encoder–decoder approaches, » in *Proceedings of the Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST)*, Doha, Qatar, Oct. 2014, pp. 103–111.

[127] D. Bahdanau, K. Cho, and Y. Bengio, « Neural machine translation by jointly learning to align and translate, » in *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, CA, 2015.

[128] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, « Attention is all you need, » in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, Long Beach, CA, 2017.

[129] A. Rodriguez and A. Laio, « Clustering by fast search and find of density peaks, » *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[130] H. Tang, L. Ding, S. Wu, B. Ren, N. Sebe, and P. Rota, « Deep unsupervised key frame extraction for efficient video classification, » *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 19, no. 3, Feb. 2023.

[131] Z. Ying and A. Bovik, « 180-degree outpainting from a single image, » *arXiv preprint arXiv :2001.04568*, Jan. 2020.

[132] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, « PatchMatch : A randomized correspondence algorithm for structural image editing, » *ACM Transactions on Graphics*, vol. 28, no. 3, pp. 1–11, Jul. 2009.

[133] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, « Context encoders : Feature learning by inpainting, » in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016.

[134] N. van Noord and E. Postma, « Light-weight pixel context encoders for image inpainting, » *arXiv preprint arXiv :1801.05585*, Jan. 2018.

[135] V. V. Voronin, V. I. Marchuk, A. I. Sherstobitov, E. A. Semenischev, S. Agaian, and K. Egiazarian, « Image extrapolation for photo stitching using nonlocal patch-based inpainting, » in *Mobile Multimedia, Image Processing, Security, and Applications*, vol. 9120. SPIE, 2014.

[136] P. Gao, X. Yang, R. Zhang, J. Y. Goulermas, Y. Geng, Y. Yan, and K. Huang, « Generalised image outpainting with U-Transformer, » *Neural Networks*, Elsevier, Sep. 2022.

[137] T. Wu, C. Zheng, and T.-J. Cham, « PanoDiffusion : 360-degree panorama outpainting via diffusion, » in *Proceedings of the International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, 2023.

[138] B. Maggia, « Video outpainting using conditional generative adversarial networks, » Ph.D. dissertation, University of Oklahoma, 2021.

[139] C. Gao, A. Saraf, J.-B. Huang, and J. Kopf, « Flow-edge guided video completion, » in *Proceedings of the European Conference on Computer Vision (ECCV)*, Glasgow, United Kingdom, Sep. 2020.

[140] L. Dehan, W. Van Ranst, P. Vandewalle, and T. Goedeme, « Complete and temporally consistent video outpainting, » in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, New Orleans, LA, Jun. 2022, pp. 686–694.

[141] Z. Li, C.-Z. Lu, J. Qin, C.-L. Guo, and M.-M. Cheng, « Towards an end-to-end framework for flow-guided video inpainting, » in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, Apr. 2022.

[142] S. Lee, J. Lee, B. Kim, K. Kim, and J. Noh, « Video extrapolation using neighboring frames, » *ACM Transactions on Graphics*, vol. 38, no. 3, pp. 1–13, Jun. 2019.

[143] F. L. Kooi and M. Mosch, « Peripheral motion displays : tapping the potential of the visual periphery, » in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, San Francisco, CA, 2006, pp. 1604–1608.

[144] P. Zheng, D. Gao, D.-P. Fan, L. Liu, J. Laaksonen, W. Ouyang, and N. Sebe, « Bilateral reference for high-resolution dichotomous image segmentation, » *CAAI Artificial Intelligence Research (AIR)*, Jul. 2024.

[145] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, « A benchmark dataset and evaluation methodology for video object segmentation, » in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 724–732.

[146] N. Xu, L. Yang, Y. Fan, J. Yang, D. Yue, Y. Liang, B. Price, S. Cohen, and T. Huang, « YouTube-VOS : Sequence-to-sequence video object segmentation, » in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, Sep. 2018.

[147] K. C. K. Chan, S. Zhou, X. Xu, and C. C. Loy, « Investigating tradeoffs in real-world video super-resolution, » in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, 2022.

[148] L. Itti, C. Koch, and E. Niebur, « A model of saliency-based visual attention for rapid scene analysis, » *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[149] C. Rother, V. Kolmogorov, and A. Blake, « GrabCut — interactive foreground extraction using iterated graph cuts, » *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 309–314, Aug. 2004.

[150] K. He, G. Gkioxari, P. Dollár, and R. Girshick, « Mask R-CNN, » in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017.

[151] M. S. Lee, W. Shin, and S. W. Han, « TRACER : Extreme attention guided salient object tracing network, » in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vancouver, Canada, Jun. 2022.

[152] X. Tian, K. Xu, X. Yang, B. Yin, and R. W. H. Lau, « Learning to detect instance-level salient objects using complementary image labels, » *International Journal of Computer Vision*, vol. 130, pp. 729–746, 2022.

[153] Y. Zeng, P. Zhang, J. Zhang, Z. Lin, and H. Lu, « Towards high-resolution salient object detection, » in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, South Korea, Aug. 2019.

[154] X. Qin, H. Dai, X. Hu, D.-P. Fan, L. Shao, and L. Van Gool, « Highly accurate dichotomous image segmentation, » in *Proceedings of the European Conference on Computer Vision (ECCV)*, Tel Aviv, Israel, Jul. 2022.

[155] M. De Murcia, D. Traparic, L. Planchon, and F. Équilbey, Étude sur la satisfaction de l'instauration du menu à points au RU Génimi du Futuroscope, Xlim UMR CNRS 7252, Poitiers, Dec. 2024. Available : https ://hal.science/hal-04833729, last accessed : Jun. 03, 2026.

[156] R. Margolin, L. Zelnik-Manor, and A. Tal, « How to evaluate foreground maps, » in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, Jun. 2014, pp. 248–255.

[157] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, « Structure-measure : A new way to evaluate foreground maps, » in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, Aug. 2017.

[158] D. Traparic, M.-C. Larabi, and L. Bellatreche, « Towards automatic content generation for immersive cinema theater based on artificial intelligence, » in *Proceedings of the IEEE International Workshop on Multimedia Signal Processing*, Poitiers, France, Sep. 2023.

[159] D. Traparic, M. De Murcia, C. Larabi, and L. Bellatreche, « ICE-Cubed : Inpainting of cinematographic elements for intelligent context expansion, » in *Proceedings of the International Conference on Advanced Concepts for Intelligent Vision Systems (ACIVS)*, Tokyo, Japan, Jul. 2025.

SOURCES GRAPHIQUES ET MULTIMÉDIAS

[M1] David Traparic, figure inspirée du drapeau de la liberté d'expression, lui-même conçu à partir de la clé de chiffrement AACS utilisée par les HD DVD et les disques Blu-ray.

[M2] Sylvain « Sylvqin » Szewczyk, « La seule vidéo sur les croix de pharmacie (oui j'en ai une), » YouTube, 20 oct. 2024, https ://www.youtube.com/watch ?v=ghh-28ln-z4, consulté le 16 avr. 2026.

[M3] Industrial Light & Magic, « The Virtual Production of The Mandalorian Season One, » YouTube, https ://www.youtube.com/watch ?v=gUnxzVOs3rk, consulté le 16 avr. 2026.

[M4] L. Poyet (grav.), G. Tissandier, « Le Théâtre Optique de M. Reynaud, » *La Nature*, juil. 1892.

[M5] *Potion Craft*, développé par niceplay games, édité par TinyBuild, 13 déc. 2022.

[M6] A. Odeychuk, « Devising a method to adaptively determine human movement speed in crowd behavior simulation at extreme events, » *Eastern-European Journal of Enterprise Technologies*, vol. 2, n° 3(134), pp. 66–74, avr. 2025, https ://www.researchgate.net/publication/391422046, consulté le 16 avr. 2026.

[M7] Tobii, « The Eye » *Tobii Developer Portal*, https ://developer.tobii.com/xr/learn/eye-behavior/the-eye/, consulté le 16 avr. 2026.

[M8] David Traparic, figure inspirée de Cyberpunk : Edge Runners (Rafal Jaki, Mike Pondsmith), Super Bomberman 2 (Produce !, SNES, 1994), Street Fighter 2 (Capcom, SNES, 1991)

[M9] David Traparic, figure avec éléments graphiques de CGP Grey, « You Don't Understand How AI Learns », https ://youtu.be/R9OHn5ZF4Uo, consulté le 29 avr. 2026.

[M10] Brian Amberg, https ://en.wikipedia.org/wiki/Convolution#/media/File :Convolution_of_box_signal_with_itself2.gif, consulté le 22 avr. 2026.

[M11] Noriko Tomuro, CSC 578 Neural Networks and Deep Learning, p. 7, https ://www.slideserve.com/juanita/neural-networks, consulté le 22 avr. 2026.

[M12] David Traparic, figure composée de The Lion King (Westwood Studios, SNES, 1994)

[M13] David Traparic, figure composée de Resident Evil (Capcom, 1996), FrostPunk (11 Bit Studios, 2018), Martin Lagrange (programmation informatique ICE)

[M14] Laurent Bernal, 2016, http ://blog.ac-versailles.fr/monsieurteste/index.php/post/07/11/2016/Le-cadrage-%3A-les-%C3%A9chelles-de-plan, consulté le 22 avr. 2026

[M15] Sam Esmail, Mr. Robot, 2015

[M16] David Traparic, figure composée de Clair Obscur : Expedition 33 (Sandfall Interactive, 2025), Harry Potter et les Reliques de la Mort, partie 2 (David Yates, 2011)

[M17] Charles Raymond, « CGR rejoue ses classiques », 2025

[M18] Getting Over It with Bennett Foddy, Bennett Foddy, 2017

[M19] David Traparic, figure inspirée de Étienne « le bolideur » Nelsom - Vega Missyl, Journal télévisé Yvelines Première, 2010

[M20] David Traparic, figure avec éléments graphiques de Mario Bros 1 et Mario Bros World

[M21] David Traparic, figure avec éléments graphiques de « Noël, Ramadan... 2008, quand tu te fous de tout ! », archive INA, France 2, 2008, de « Donjons et Dragons : L'Honneur des voleurs », 2023, ainsi que de « Mon colloque serait-il une sacoche ? ? », YouTube, 2009

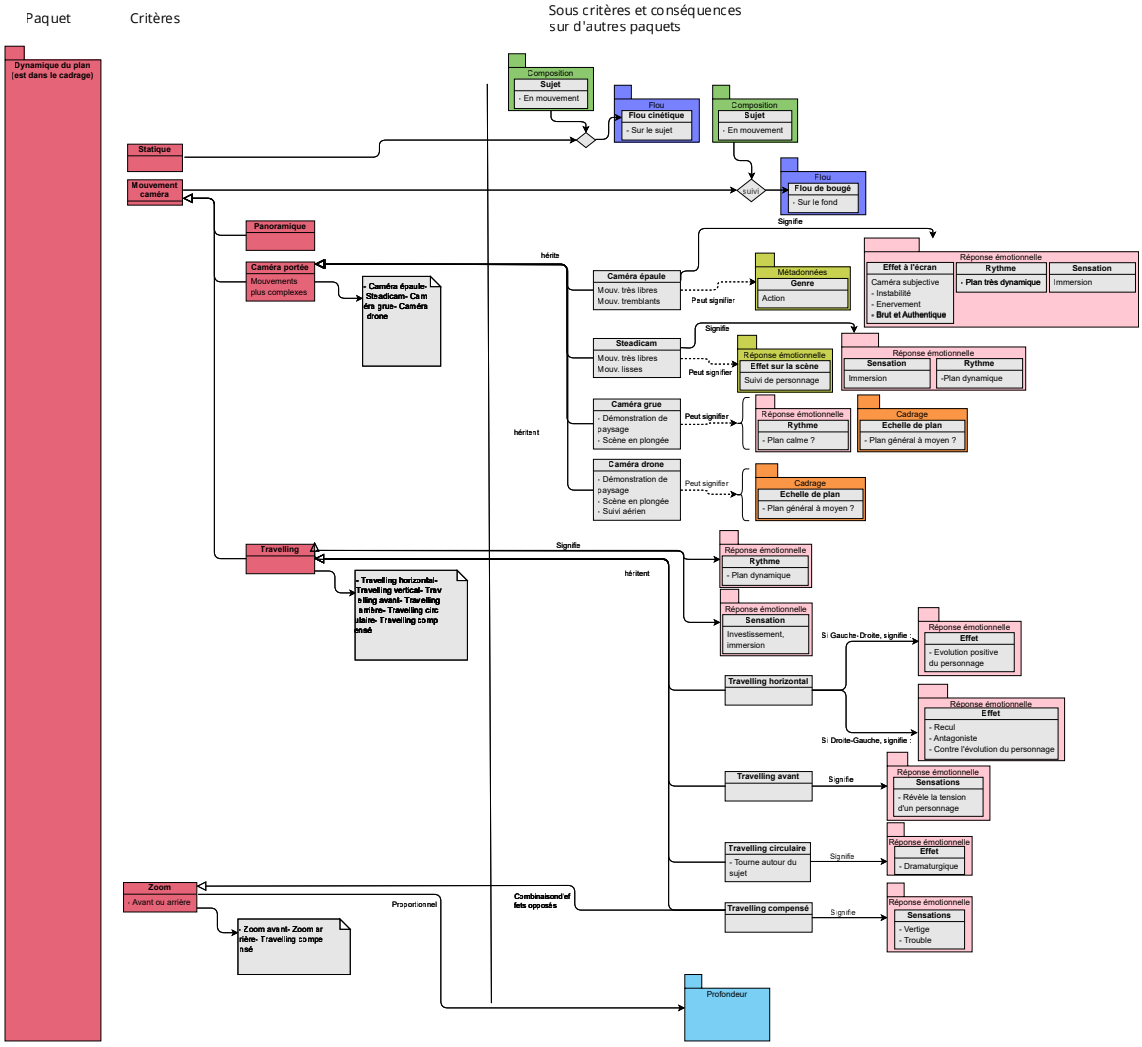
[M22] Traduction par David Traparic d'une figure de S. J. Huston, H. G. Krapp, dans « Visuomotor Transformation in the Fly Gaze Stabilization System », PLOS Biology 6(7) : e173, 2008, avec ajout d'éléments graphiques de « Kestrel 3D - Faster Than Light », par Léo Requier, Artstation, 2024, https ://www.artstation.com/artwork/dKPN8X, consulté le 27 avr. 2026

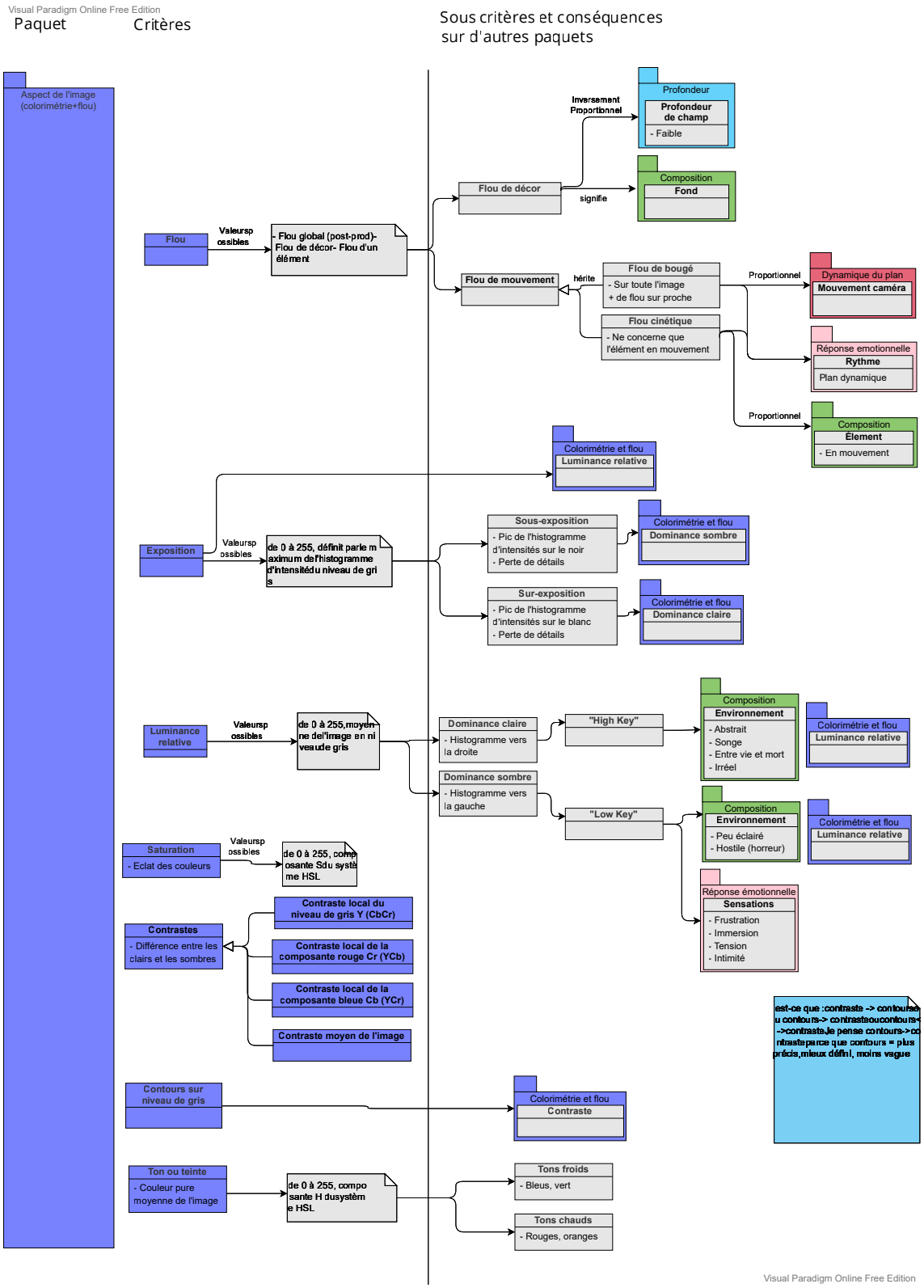
[M23] David Traparic, éléments graphiques de composition de la figure, de gauche à droite : « L'Odyssée de l'écriture » par Arte, Cœur de Super Bomberman 2 (Produce !, 1994, SNES), logo Rolling Stones, Mass Effect (BioWare, 2007, PC), vache de Stardew Valley (Concerned Ape, 2016, PC), seau de Minecraft (Mojang, 2011, PC), photographie du pont de Trebinje (David Traparic, 2018), peinture de Léa Varin instagram.com/ladameauxportes/.

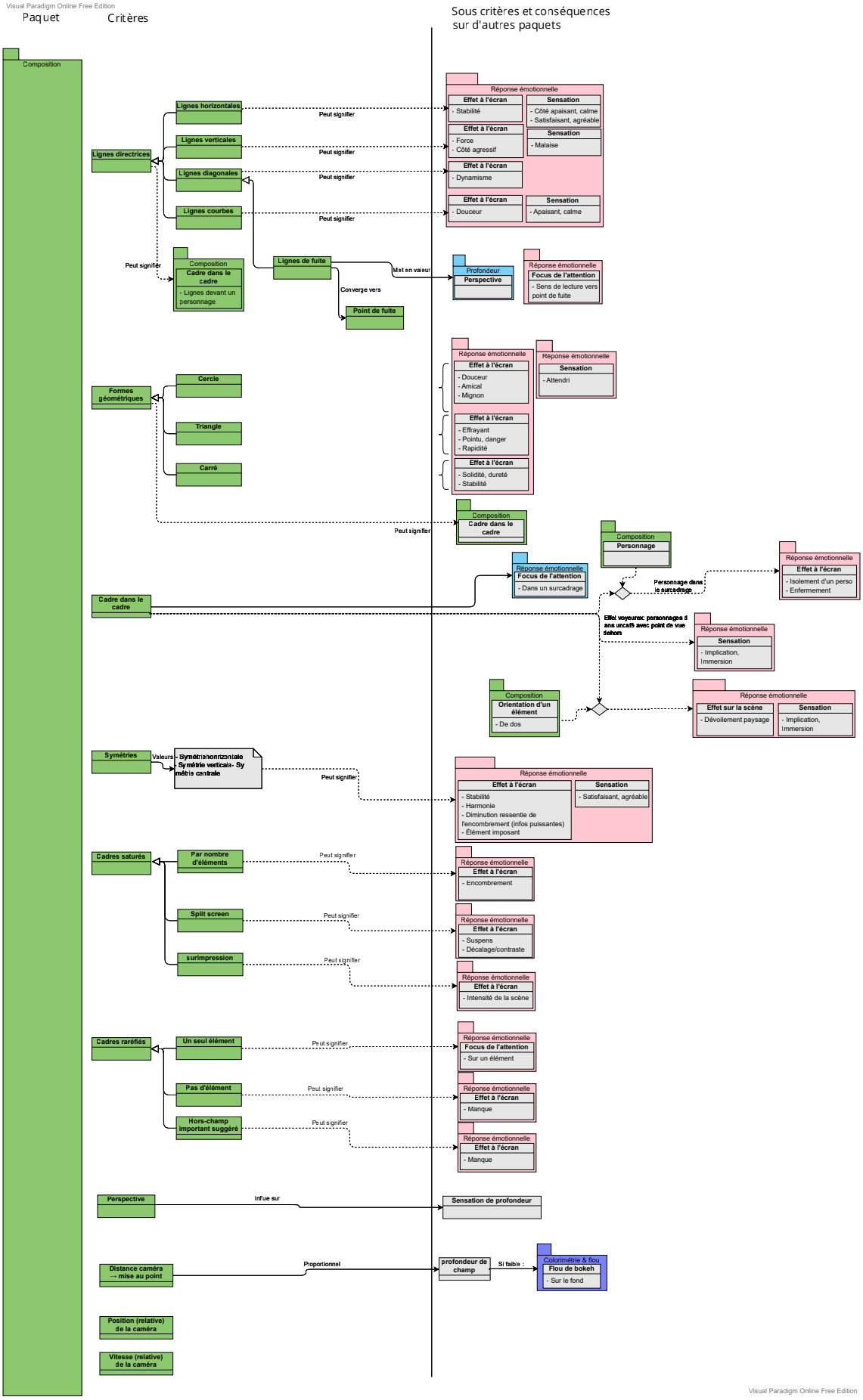
[M24] David Traparic, Demonstration of SSPICE, https ://youtu.be/tgiuxlkn01E, consulté le 28 avr. 2026.

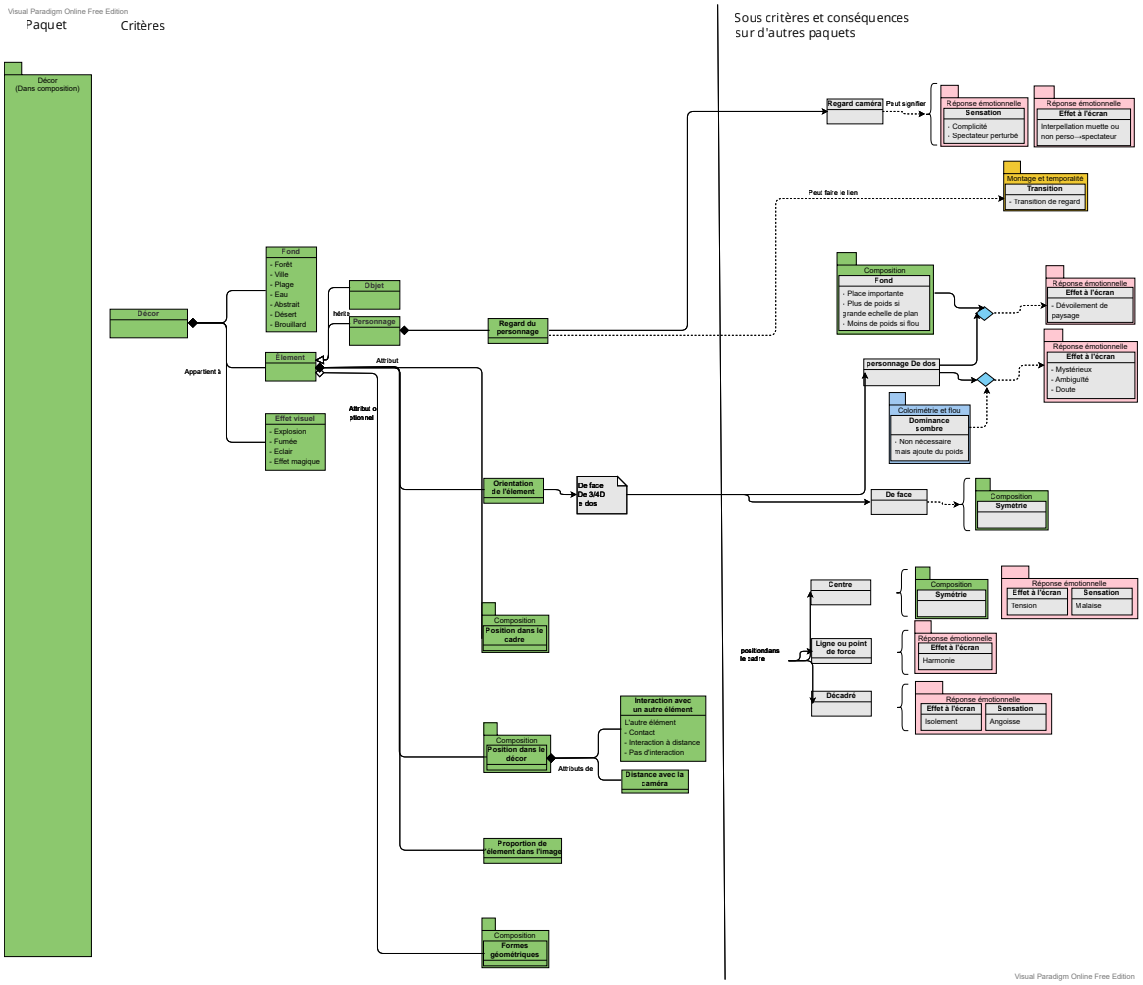
ANNEXE A

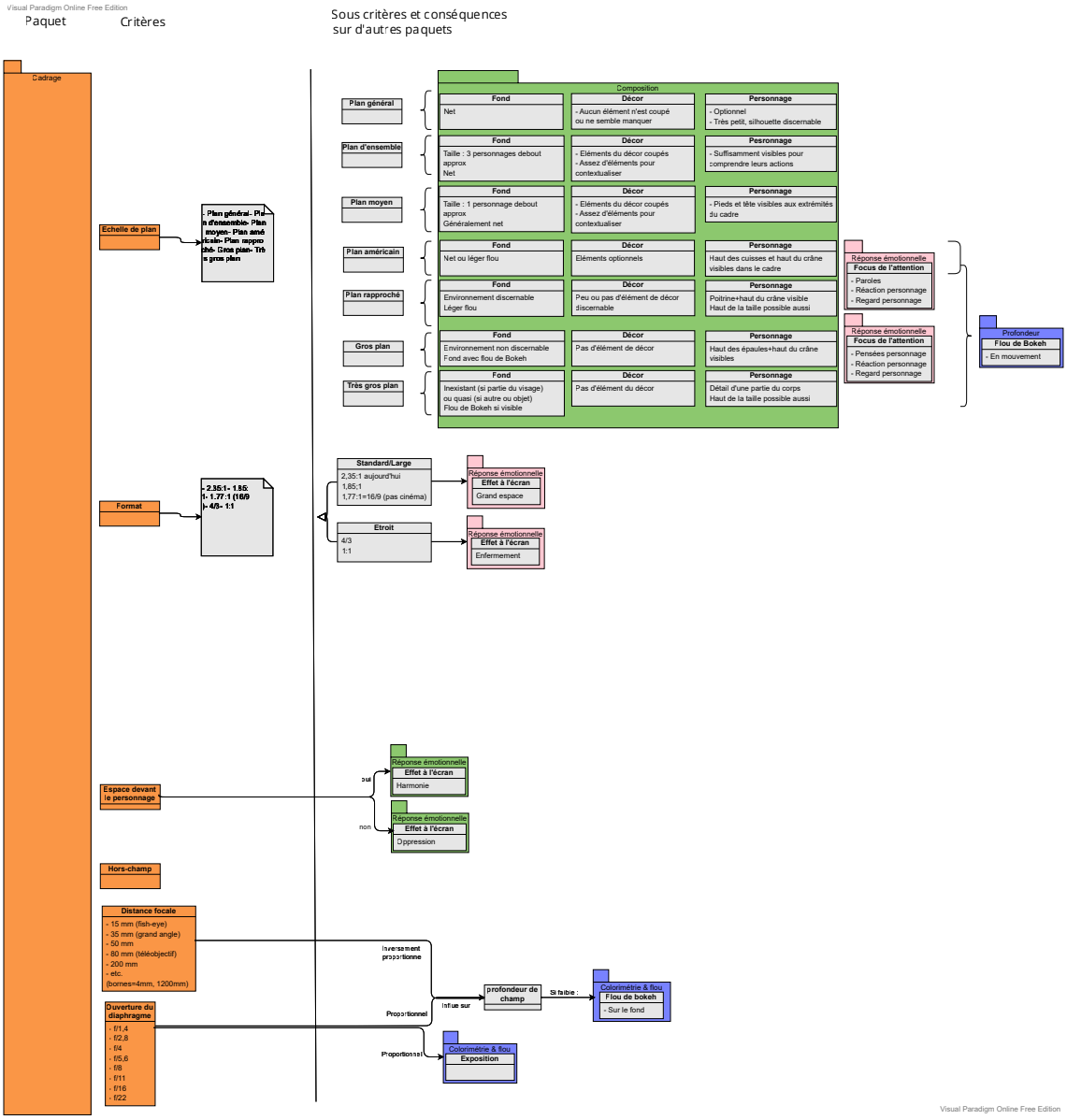
ANNEXES

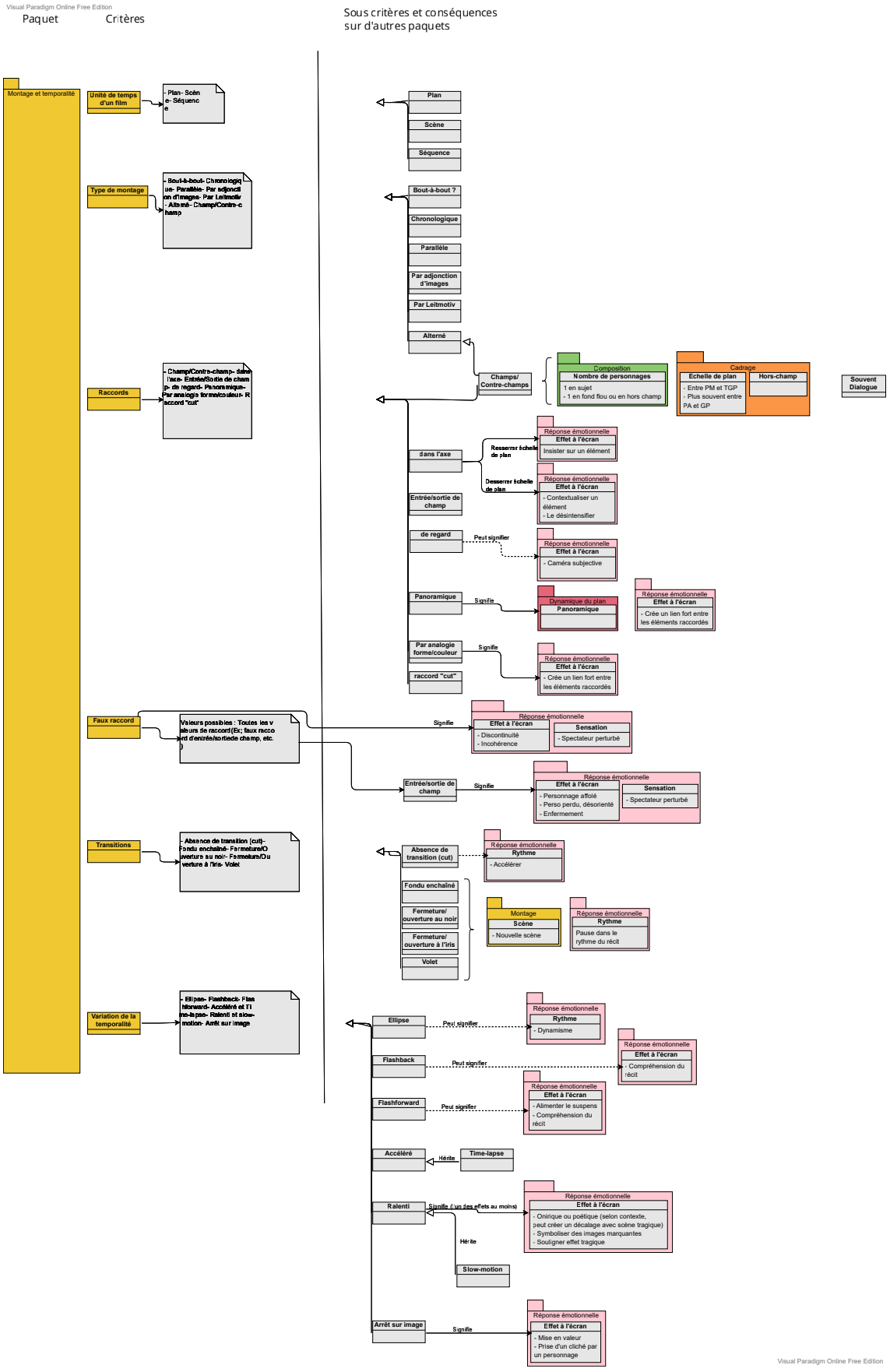

Paquet
Critères
Sous critères et conséquences sur d'autres paquets
Dynamique du plan (est dans le cadrage)
Statique
Mouvement caméra
Panoramique
Caméra portée
Mouvements plus complexes
- Caméra épaule- Steadicam- Caméra grue- Caméra drone
Travelling
- Travelling horizontal- Traveling vertical- Traveling avant- Travelling arrière- Traveling circulaire- Travelling compensé
Zoom
- Avant ou arrière
- Zoom avant- Zoom arrière- Traveling compensé
Composition
Sujet
- En mouvement
Flou
Flou cinétique
- Sur le sujet
Composition
Sujet
- En mouvement
Flou
Flou de bougé
- Sur le fond
Suivi
Signifie
hérite
Caméra épaule
Mouv. très libres
Mouv. tremblants
Métadonnées
Genre
Action
Peut signifier
Réponse émotionnelle
Effet à l'écran
Caméra subjective
- Instabilité
- Enervement
- Brut et Authentique
Rythme
- Plan très dynamique
Sensation
Immersion
Steadicam
Mouv. très libres
Mouv. lisses
Réponse émotionnelle
Effet sur la scène
Suivi de personnage
Réponse émotionnelle
Sensation
Immersion
Rythme
-Plan dynamique
Caméra grue
- Démonstration de paysage
- Scène en plongée
Réponse émotionnelle
Rythme
- Plan calme ?
Cadrage
Echelle de plan
- Plan général à moyen ?
héritant
Caméra drone
- Démonstration de paysage
- Scène en plongée
- Suivi aérien
Cadrage
Echelle de plan
- Plan général à moyen ?
Réponse émotionnelle
Rythme
- Plan dynamique
Réponse émotionnelle
Sensation
Investissement, immersion
Si Gauche-Droite, signifie :
Réponse émotionnelle
Effet
- Evolution positive du personnage
Travelling horizontal
Si Droite-Gauche, signifie :
Réponse émotionnelle
Effet
- Recul
- Antagoniste
- Contre l'évolution du personnage
Travelling avant
Signifie
Réponse émotionnelle
Sensations
- Révèle la tension d'un personnage
Travelling circulaire
- Tourne autour du sujet
Signifie
Réponse émotionnelle
Effet
- Dramaturgique
Travelling compensé
Signifie
Réponse émotionnelle
Sensations
- Vertige
- Trouble
Proportionnel
Combinaisond'effets opposés
Profondeur


Visual Paradigm Online Free Edition
Paquet
Critères
Sous critères et conséquences sur d'autres paquets
Aspect de l'image (colorimétrie+flou)
Flou
Valeursp ossibles
- Flou global (post-prod)- Flou de décor- Flou d'un élément
Flou de décor
Inversement Proportionnel
Profondeur
Profondeur de champ
- Faible
signifie
Composition
Fond
Flou de mouvement
hérite
Flou de bougé
- Sur toute l'image + de flou sur proche
Flou cinétique
- Ne concerne que l'élément en mouvement
Proportionnel
Dynamique du plan
Mouvement caméra
Réponse emotionnelle
Rythme
Plan dynamique
Proportionnel
Composition
Élement
- En mouvement
Colorimétrie et flou
Luminance relative
Exposition
Valeursp ossibles
de 0 à 255, définit parle maximum del'histogramme d'intensitédu niveau de gris
Sous-exposition
- Pic de l'histogramme d'intensités sur le noir
- Perte de détails
Colorimétrie et flou
Dominance sombre
Sur-exposition
- Pic de l'histogramme d'intensités sur le blanc
- Perte de détails
Colorimétrie et flou
Dominance claire
Luminance relative
Valeursp ossibles
de 0 à 255,moyenne del'image en niveaude gris
Dominance claire
- Histogramme vers la droite
"High Key"
Composition
Environnement
- Abstrait
- Songe
- Entre vie et mort
- Irréel
Colorimétrie et flou
Luminance relative
Dominance sombre
- Histogramme vers la gauche
"Low Key"
Composition
Environnement
- Peu éclairé
- Hostile (horreur)
Colorimétrie et flou
Luminance relative
Réponse émotionnelle
Sensations
- Frustration
- Immersion
- Tension
- Intimité
Saturation
- Eclat des couleurs
Valeursp ossibles
de 0 à 255, composante Sdu système HSL
Contrastes
- Différence entre les clairs et les sombres
Contraste local du niveau de gris Y (CbCr)
Contraste local de la composante rouge Cr (YCb)
Contraste local de la composante bleue Cb (YCr)
Contraste moyen de l'image
est-ce que :contraste -> contoursou contours-> contrasteoucontours<-contraste.Je pense contours->contrasteparce que contours = plus précis,mieux défini, moins vague
Contours sur niveau de gris
Colorimétrie et flou
Contraste
Ton ou teinte
- Couleur pure moyenne de l'image
de 0 à 255, composante H dusystème HSL
Tons froids
- Bleus, vert
Tons chauds
- Rouges, oranges
Visual Paradigm Online Free Edition


Visual Paradigm Online Free Edition
Paquet
Critères
Sous critères et conséquences sur d'autres paquets
Composition
Lignes horizontales
Lignes verticales
Lignes directrices
Lignes diagonales
Lignes courbes
Peut signifier
Réponse émotionnelle
Effet à l'écran
- Stabilité
Sensation
- Côté apaisant, calme
- Satisfaisant, agréable
Effet à l'écran
- Force
- Côté agressif
Sensation
- Malaise
Effet à l'écran
- Dynamisme
Effet à l'écran
- Douceur
Sensation
- Apaisant, calme
Composition
Cadre dans le cadre
- Lignes devant un personnage
Lignes de fuite
Point de fuite
Converge vers
Met en valeur
Profondeur
Perspective
Réponse émotionnelle
Focus de l'attention
- Sens de lecture vers point de fuite
Formes géométriques
Cercle
Triangle
Carré
Réponse émotionnelle
Effet à l'écran
- Douceur
- Amical
- Mignon
Réponse émotionnelle
Sensation
- Attendri
Effet à l'écran
- Effrayant
- Pointu, danger
- Rapidité
Effet à l'écran
- Solidité, dureté
- Stabilité
Composition
Cadre dans le cadre
Cadre dans le cadre
Réponse émotionnelle
Focus de l'attention
- Dans un surcadrage
Composition
Personnage
Personnage dans le surcadrage
Réponse émotionnelle
Effet à l'écran
- Isolement d'un perso
- Enfermement
Effet voyeurs: personnages dans un cadre avec point de vue séhors
Réponse émotionnelle
Sensation
- Implication, immersion
Composition
Orientation d'un élément
- De dos
Réponse émotionnelle
Effet sur la scène
- Dévoilement paysage
Sensation
- Implication, immersion
Symétries
Valeurs
- Symétrie horizontale
- Symétrie verticale
- Symétrie centrale
Réponse émotionnelle
Effet à l'écran
- Stabilité
- Harmonie
- Diminution ressentie de l'encombrement (infos puissantes)
- Élément imposant
Sensation
- Satisfaisant, agréable
Cadres saturés
Par nombre d'éléments
Split screen
surimpression
Réponse émotionnelle
Effet à l'écran
- Encombrement
Réponse émotionnelle
Effet à l'écran
- Suspens
- Décalage/contraste
Réponse émotionnelle
Effet à l'écran
- Intensité de la scène
Cadres raréfiés
Un seul élément
Pas d'élément
Hors-champ important suggéré
Réponse émotionnelle
Focus de l'attention
- Sur un élément
Réponse émotionnelle
Effet à l'écran
- Manque
Réponse émotionnelle
Effet à l'écran
- Manque
Perspective
Influe sur
Sensation de profondeur
Distance caméra → mise au point
Proportionnel
profondeur de champ
Si faible :
Colorimétrie & flou
Flou de bokeh
- Sur le fond
Position (relative) de la caméra
Vitesse (relative) de la caméra
Visual Paradigm Online Free Edition

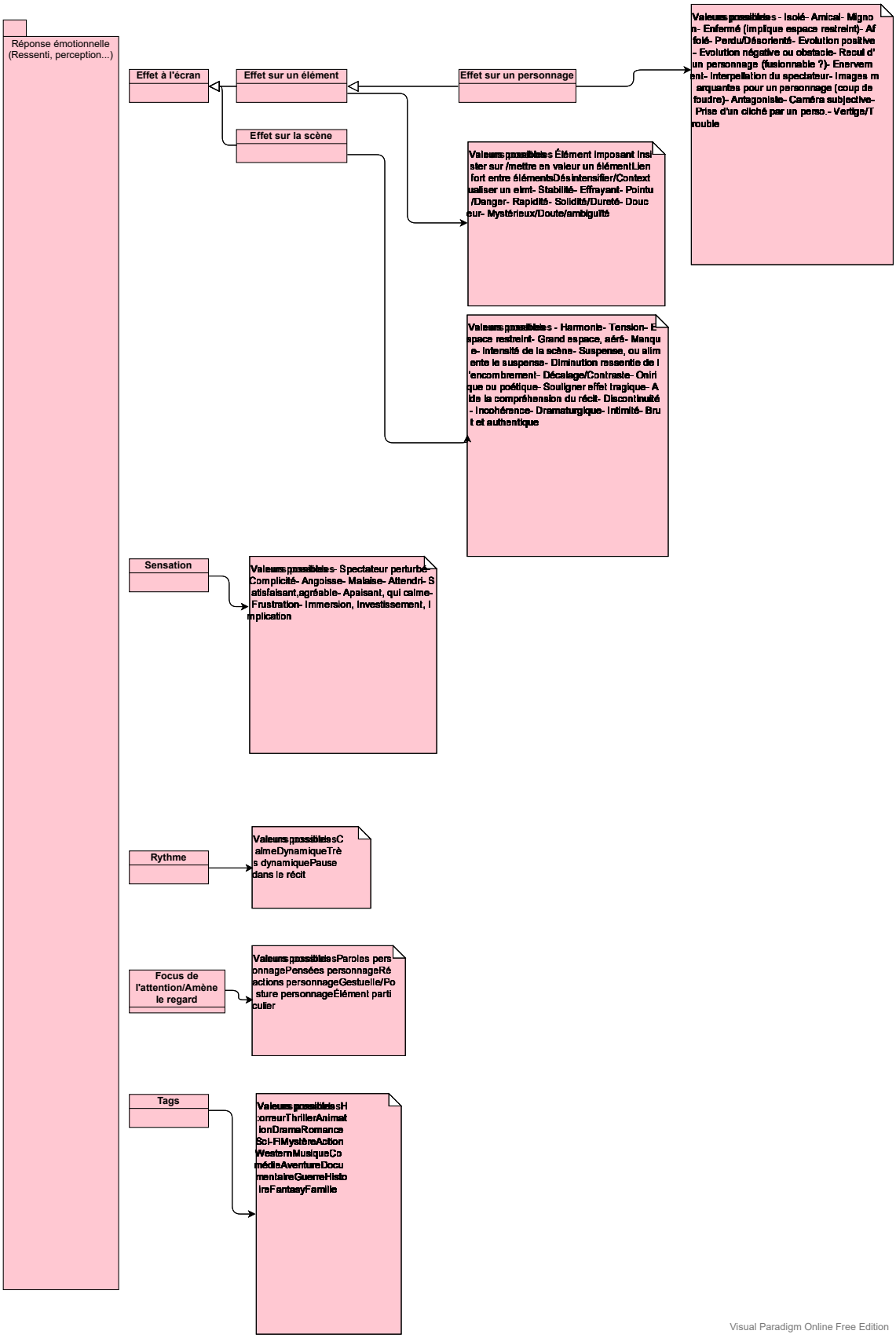

Visual Paradigm Online Free Edition
Paquet
Critères
Sous critères et conséquences
sur d'autres paquets
Décor
(Dans composition)
Décor
Fond
- Forêt
- Ville
- Plage
- Eau
- Abstrait
- Désert
- Brouillard
Objet
Personnage
hérit
Élement
Appartient à
Attribut
Attribut o
ptionnel
Effet visuel
- Explosion
- Fumée
- Eclair
- Effet magique
Regard du
personnage
Orientation
de l'élement
De face
De 3/4D
e dos
Composition
Position dans le
cadre
Composition
Position dans le
décor
Attributs de
Interaction avec
un autre élément
L'autre élément
- Contact
- Interaction à distance
- Pas d'interaction
Distance avec la
caméra
Proportion de
élement dans l'image
Composition
Formes
géométriques
Regard caméra
Peut signifier
Réponse émotionnelle
Sensation
- Complicité
- Spectateur perturbé
Réponse émotionnelle
Effet à l'écran
Interpellation muette ou
non perso→spectateur
Peut faire le lien
Montage et temporalité
Transition
- Transition de regard
Composition
Fond
- Place importante
- Plus de poids si
grande echelle de plan
- Moins de poids si flou
Réponse émotionnelle
Effet à l'écran
- Dévoilement de
paysage
personnage De dos
Réponse émotionnelle
Effet à l'écran
- Mystérieux
- Ambiguité
- Doute
Colorimétrie et flou
Dominance
sombre
- Non nécessaire
mais ajoute du poids
De face
Composition
Symétrie
Centre
Composition
Symétrie
Réponse émotionnelle
Effet à l'écran
Sensation
Tension
Malaise
positiondans
le cadre
Ligne ou point
de force
Réponse émotionnelle
Effet à l'écran
Harmonie
Décadré
Réponse émotionnelle
Effet à l'écran
Sensation
Isolement
Angoisse
Visual Paradigm Online Free Edition


Visual Paradigm Online Free Edition
Paquet
Critères
Sous critères et conséquences
sur d'autres paquets
Cadrage
Echelle de plan
- Plan général- Plan d'ensemble- Plan moyen- Plan américain- Plan rapproché- Gros plan- Très gros plan
Format
- 2.35:1 - 1.85:1- 1.77:1 (16/9)- 4/3- 1:1
Espace devant le personnage
Hors-champ
Distance focale
- 15 mm (fish-eye)
- 35 mm (grand angle)
- 50 mm
- 80 mm (téléobjectif)
- 200 mm
- etc.
(bornes=4mm, 1200mm)
Ouverture du diaphragme
- f/1,4
- f/2,8
- f/4
- f/5,6
- f/8
- f/11
- f/16
- f/22
Composition
Plan général
Fond
Net
Décor
- Aucun élément n'est coupé ou ne semble manquer
Personnage
- Optionnel
- Très petit, silhouette discernable
Plan d'ensemble
Fond
Taille : 3 personnages debout approx
Net
Décor
- Eléments du décor coupés
- Assez d'éléments pour contextualiser
Pesronnage
- Suffisamment visibles pour comprendre leurs actions
Plan moyen
Fond
Taille : 1 personnage debout approx
Généralement net
Décor
- Eléments du décor coupés
- Assez d'éléments pour contextualiser
Personnage
- Pieds et tête visibles aux extrémités du cadre
Plan américain
Fond
Net ou léger flou
Décor
Eléments optionnels
Personnage
Haut des cuisses et haut du crâne visibles dans le cadre
Plan rapproché
Fond
Environnement discernable
Léger flou
Décor
Peu ou pas d'élément de décor discernable
Personnage
Poitrine+haut du crâne visible
Haut de la taille possible aussi
Gros plan
Fond
Environnement non discernable
Fond avec flou de Bokeh
Décor
Pas d'élément de décor
Personnage
Haut des épaules+haut du crâne visibles
Très gros plan
Fond
Inexistant (si partie du visage) ou quasi (si autre ou objet)
Flou de Bokeh si visible
Décor
Pas d'élément du décor
Personnage
Détail d'une partie du corps
Haut de la taille possible aussi
Réponse émotionnelle
Focus de l'attention
- Paroles
- Réaction personnage
- Regard personnage
Réponse émotionnelle
Focus de l'attention
- Pensées personnage
- Réaction personnage
- Regard personnage
Profondeur
Flou de Bokeh
- En mouvement
Standard/Large
2,35:1 aujourd'hui
1,85:1
1,77:1=16/9 (pas cinéma)
Réponse émotionnelle
Effet à l'écran
Grand espace
Etroit
4/3
1:1
Réponse émotionnelle
Effet à l'écran
Enfermement
oui
Réponse émotionnelle
Effet à l'écran
Harmonie
non
Réponse émotionnelle
Effet à l'écran
Oppression
Inversement proportionne
Proportionnel
Proportionnel
influe sur
profondeur de champ
Si faible :
Colorimétrie & flou
Flou de bokeh
- Sur le fond
Colorimétrie & flou
Exposition
Visual Paradigm Online Free Edition

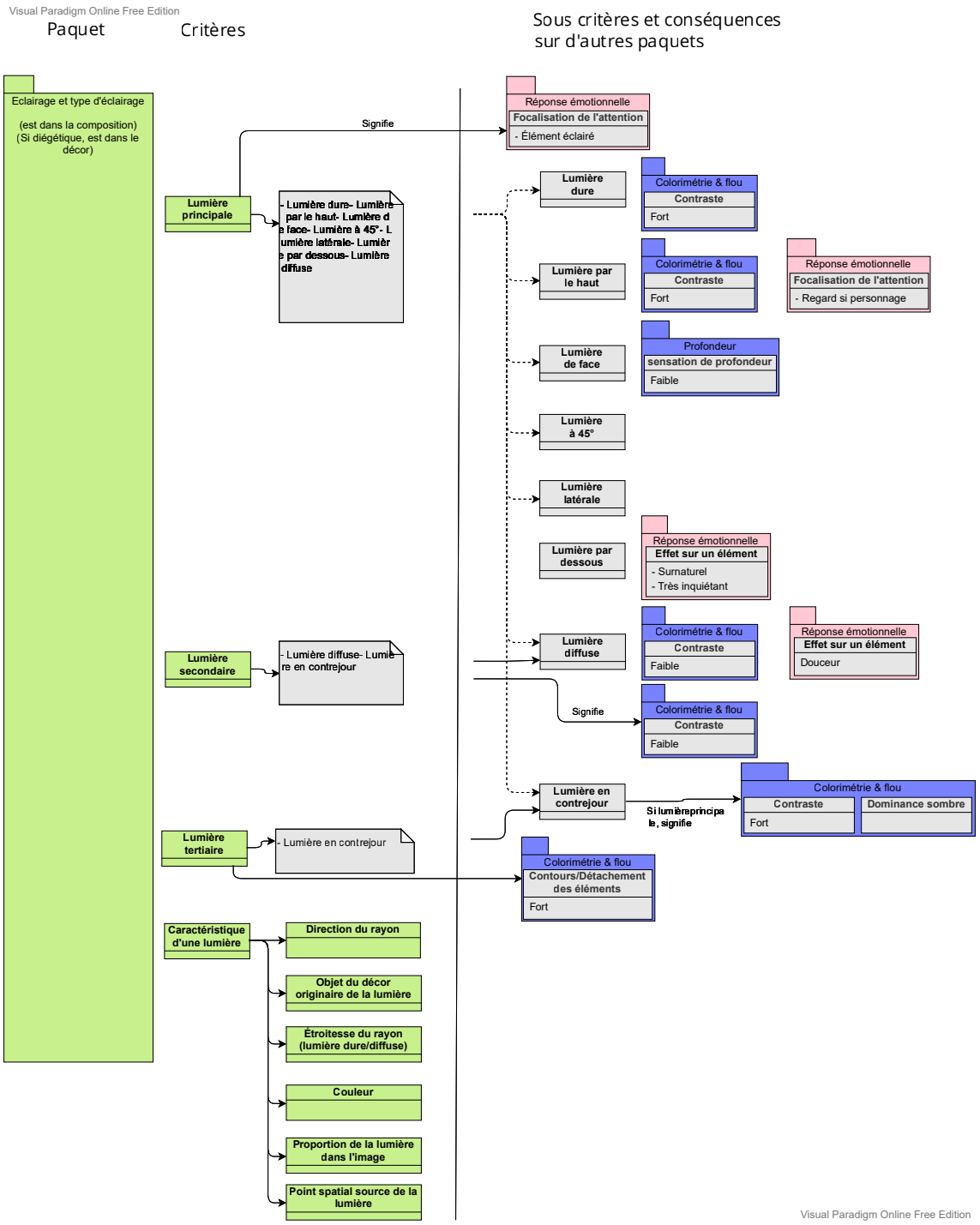

Visual Paradigm Online Free Edition
Paquet
Critères
Sous critères et conséquences sur d'autres paquets
Montage et temporalité
Unité de temps d'un film
- Plan- Scène- Séquence
Type de montage
- Bout-à-bout- Chronologique- Parallèle- Par adjonction d'images- Par Leitmotiv- Alterné- Champ/Contre-champ
Raccords
- Champ/Contre-champ- Ligne d'axe- Entrée/Sortie de champ- Point de regard- Panoramique- Par analogie forme/couleur- Raccord "cut"
Faux raccord
Valeurs possibles : Toutes les valeurs de raccord (Ex; faux raccord d'entrée/sortie de champ, etc.)
Transitions
- Absence de transition (cut)- Fondu enchaîné- Fermeture/Ouverture au noir- Fermeture/Ouverture à l'iris- Volet
Variation de la temporalité
- Ellipse- Flashback- Flashforward- Accéléré et Time-lapse- Ralenti et slow-motion- Arrêt sur image
Plan
Scène
Séquence
Bout-à-bout ?
Chronologique
Parallèle
Par adjonction d'images
Par Leitmotiv
Alterné
Champs/ Contre-champs
Composition
Nombre de personnages
1 en sujet
- 1 en fond flou ou en hors champ
Cadrage
Echelle de plan
- Entre PM et TGP
- Plus souvent entre PA et GP
Hors-champ
Souvent Dialogue
dans l'axe
Resserrer échelle de plan
Desserrer échelle de plan
Réponse émotionnelle
Effet à l'écran
Insister sur un élément
Réponse émotionnelle
Effet à l'écran
- Contextualiser un élément
- Le désintensifier
Entrée/sortie de champ
de regard
Peut signifier
Réponse émotionnelle
Effet à l'écran
- Caméra subjective
Panoramique
Signifie
Dynamique du plan
Panoramique
Réponse émotionnelle
Effet à l'écran
- Crée un lien fort entre les éléments raccordés
Par analogie forme/couleur
raccord "cut"
Signifie
Réponse émotionnelle
Effet à l'écran
- Crée un lien fort entre les éléments raccordés
Signifie
Réponse émotionnelle
Effet à l'écran
- Discontinuité
- Incohérence
Sensation
- Spectateur perturbé
Entrée/sortie de champ
Signifie
Réponse émotionnelle
Effet à l'écran
- Personnage affolé
- Perso perdu, désorienté
- Enfermement
Sensation
- Spectateur perturbé
Absence de transition (cut)
Réponse émotionnelle
Rythme
- Accélérer
Fondu enchaîné
Fermeture/ ouverture au noir
Fermeture/ ouverture à l'iris
Volet
Montage
Scène
- Nouvelle scène
Réponse émotionnelle
Rythme
Pause dans le rythme du récit
Ellipse
Peut signifier
Réponse émotionnelle
Rythme
- Dynamisme
Flashback
Peut signifier
Réponse émotionnelle
Effet à l'écran
- Compréhension du récit
Flashforward
Peut signifier
Réponse émotionnelle
Effet à l'écran
- Alimenter le suspens
- Compréhension du récit
Accéléré
Hérite
Time-lapse
Ralenti
Signifie (l'un des effets au moins)
Réponse émotionnelle
Effet à l'écran
- Onirique ou poétique (selon contexte, peut créer un décalage avec scène tragique)
- Symboliser des images marquantes
- Souligner effet tragique
Hérite
Slow-motion
Arrêt sur image
Signifie
Réponse émotionnelle
Effet à l'écran
- Mise en valeur
- Prise d'un cliché par un personnage
Visual Paradigm Online Free Edition

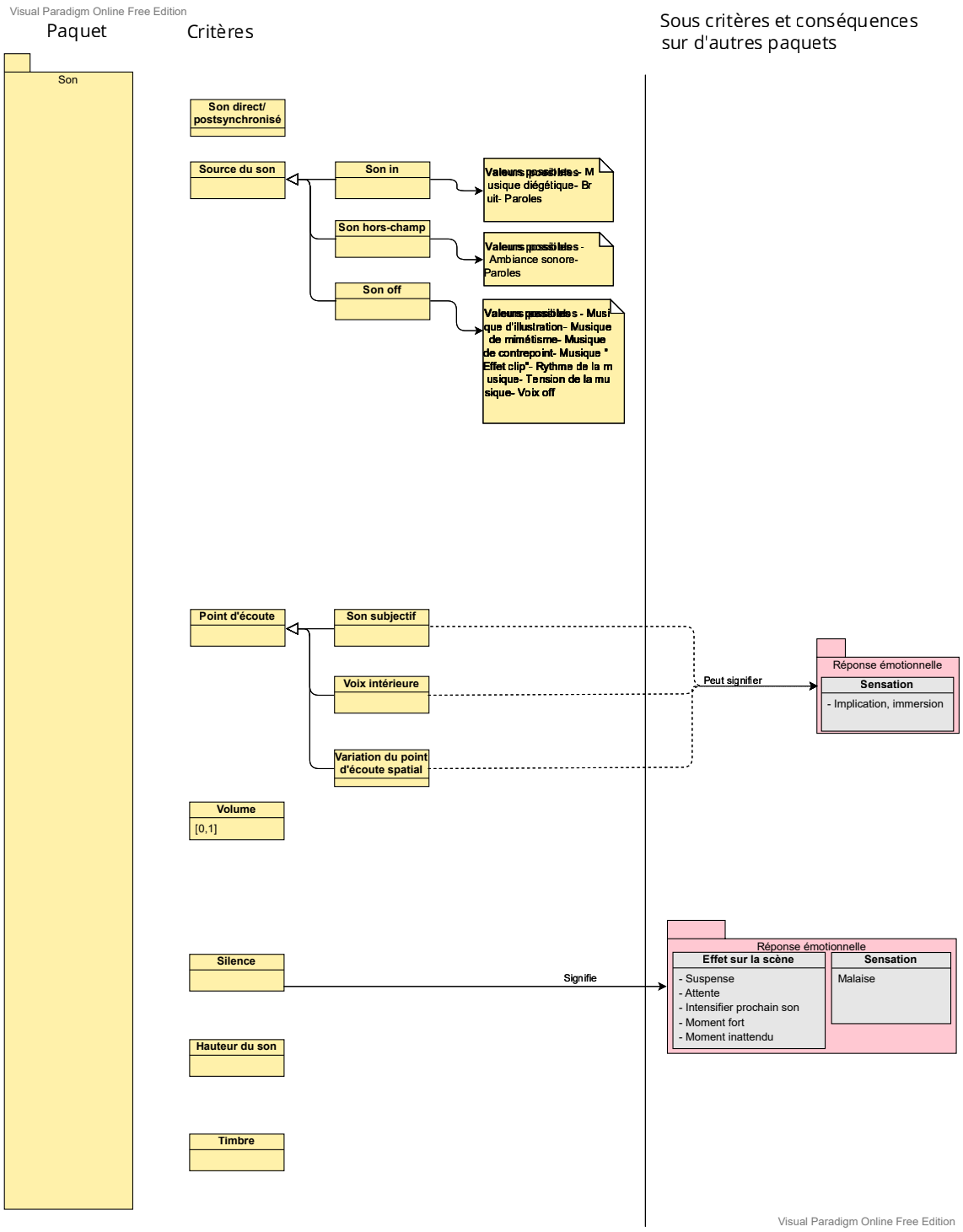
Visual Paradigm Online Free Edition

Paquet Critères



Visual Paradigm Online Free Edition


Visual Paradigm Online Free Edition
Paquet
Critères
Sous critères et conséquences sur d'autres paquets
Eclairage et type d'éclairage
(est dans la composition)
(Si diégétique, est dans le décor)
Signifie
Réponse émotionnelle
Focalisation de l'attention
- Élément éclairé
Lumière principale
- Lumière dure- Lumière par le haut- Lumière d e face- Lumière à 45°- L umière latérale- Lumièr e par dessous- Lumière diffuse
Lumière dure
Colorimétrie & flou
Contraste
Fort
Lumière par le haut
Colorimétrie & flou
Contraste
Fort
Réponse émotionnelle
Focalisation de l'attention
- Regard si personnage
Lumière de face
Profondeur
sensation de profondeur
Faible
Lumière à 45°
Lumière latérale
Lumière par dessous
Réponse émotionnelle
Effet sur un élément
- Surnaturel
- Très inquiétant
Lumière secondaire
- Lumière diffuse- Lumiè re en contrejour
Lumière diffuse
Colorimétrie & flou
Contraste
Faible
Réponse émotionnelle
Effet sur un élément
Douceur
Signifie
Colorimétrie & flou
Contraste
Faible
Lumière en contrejour
Si lumièreprincipa le, signifie
Colorimétrie & flou
Contraste
Dominance sombre
Fort
Lumière tertiaire
- Lumière en contrejour
Colorimétrie & flou
Contours/Détachement des éléments
Fort
Caractéristique d'une lumière
Direction du rayon
Objet du décor originaire de la lumière
Étroitesse du rayon (lumière dure/diffuse)
Couleur
Proportion de la lumière dans l'image
Point spatial source de la lumière
Visual Paradigm Online Free Edition


Visual Paradigm Online Free Edition
Paquet
Critères
Sous critères et conséquences sur d'autres paquets
Son
Son direct/ postsynchronisé
Source du son
Son in
Valeurs possibles- Musique diégétique- Bruit- Paroles
Son hors-champ
Valeurs possibles - Ambiance sonore- Paroles
Son off
Valeurs possibles - Musique d'illustration- Musique de mimétisme- Musique de contrepoint- Musique " Effet clip"- Rythme de la musique- Tension de la musique- Voix off
Point d'écoute
Son subjectif
Voix intérieure
Variation du point d'écoute spatial
Peut signifier
Réponse émotionnelle
Sensation
- Implication, immersion
Volume
[0,1]
Silence
Signifie
Réponse émotionnelle
Effet sur la scène
- Suspense
- Attente
- Intensifier prochain son
- Moment fort
- Moment inattendu
Sensation
Malaise
Hauteur du son
Timbre
Visual Paradigm Online Free Edition