



POLYTECH  
NANTES

ÉCOLE POLYTECHNIQUE DE L'UNIVERSITÉ DE NANTES  
DÉPARTEMENT D'INFORMATIQUE

RAPPORT DE RECHERCHE ET DÉVELOPPEMENT

# Résumé vidéo et caractéristiques perceptuelles

*Rapport final*

**Josik SALLAUD & Nathan ROCHER**

**Février 2023**

encadré par Alexandre BRUCKERT & Patrick LE CALLET

— Équipe IPI —

LABORATOIRE DES SCIENCES DU NUMÉRIQUES DE NANTES

coordinateur : Vincent RICORDEL



## **Avertissement**

Toute reproduction, même partielle, par quelque procédé que ce soit, est interdite sans autorisation préalable.

Une copie par xérographie, photographie, photocopie, film, support magnétique ou autre, constitue une contrefaçon passible des peines prévues par la loi.

# Résumé vidéo et caractéristiques perceptuelles

## Rapport final

Josik SALLAUD & Nathan ROCHER

Catégories et descripteurs de sujets : **[Computing methodologies]**: Video summarization; **[Computing methodologies]**: Scene understanding; **[Computing methodologies]**: Activity recognition and understanding

Termes généraux : Résumé vidéo, Perception visuelle, Vision humaine, Apprentissage automatique, Apprentissage profond, Congruence visuelle inter-observateur, Mémorisation, Emotion

### Résumé

L'émergence des vidéos générées par des utilisateurs (User-Generated Content) ainsi que la quantité de vidéos proposées sur différentes plateformes augmentant exponentiellement ces dernières années rendent nécessaire leurs pré-visualisations sous forme de résumé vidéo pour permettre aux utilisateurs de s'y retrouver parmi une large collection de vidéos. De nombreuses méthodes permettent de générer automatiquement ces résumés de la vidéo, la plupart d'entre elles utilisent l'apprentissage profond et sont décrites dans ce rapport. Néanmoins, aucune d'entre elles ne se sert de l'influence de la composante perceptuelle humaine qui est essentielle pour générer un résumé vidéo attractif de vidéos générées par des utilisateurs. Nous proposons dans ce document, une méthode de génération automatique de résumé vidéo pour répondre à cette question en étudiant différentes caractéristiques perceptuelles (la congruence visuelle inter-observateurs, la mémorabilité et l'intensité émotionnelle). Ce travail est réalisé dans le cadre de notre dernière année d'étude d'école d'ingénieurs au sein de Polytech Nantes.

## **Remerciements**

Nous remercions Alexandre Bruckert, chercheur post-doctoral au Laboratoire des Sciences du Numérique de Nantes, pour son aide précieuse tout au long du projet ainsi que pour ses idées et ses retours sur ce rapport final.

# Table des matières

<b>1</b>	<b>Introduction</b>	<b>7</b>
1.1	Contexte . . . . .	7
1.2	Problématique . . . . .	8
1.3	Problèmes . . . . .	8
<b>2</b>	<b>État de l'art</b>	<b>9</b>
2.1	Résumé vidéo . . . . .	9
2.1.1	Définition . . . . .	9
2.1.2	Méthodes . . . . .	9
2.1.3	Conclusion sur ces méthodes . . . . .	12
2.2	Caractéristiques perceptuelles . . . . .	13
2.2.1	Congruence visuelle inter-observateurs . . . . .	13
2.2.2	Mémorabilité . . . . .	17
2.2.3	Intensité émotionnelle . . . . .	20
2.3	Conclusion . . . . .	22
<b>3</b>	<b>Proposition</b>	<b>24</b>
<b>4</b>	<b>Expérimentations et résultats</b>	<b>25</b>
4.1	Données utilisés . . . . .	25
4.2	Modèles utilisés . . . . .	26
4.2.1	Mémorabilité . . . . .	26
4.2.2	Congruence visuelle inter-observateurs . . . . .	26
4.2.3	Reconnaissance et intensité des émotions . . . . .	27
4.2.4	Résumé vidéo . . . . .	28
4.3	Trouver des profils-types . . . . .	28
4.3.1	Solution 1 . . . . .	28

4.3.2	Solution 2	29
4.4	Test de corrélation linéaire	33
4.4.1	Corrélation de l'IOVC	33
4.4.2	Corrélation de la mémorabilité	34
4.5	Test de Student	35
4.6	Test de similarité	37
4.7	Modèle de régression des scores	39
4.7.1	Réseau de neurones	39
4.7.2	Réseau de neurones avec couche LSTM	40
4.8	Modèle de classification binaire	44
4.8.1	Arbre de décision	44
4.9	Conclusion	46
<b>5</b>	<b>Conclusion</b>	<b>47</b>
5.1	Résumé du travail effectué	47
5.2	Enseignements	47
5.3	Perspectives de recherche	48
<b>A</b>	<b>Fiches de lecture</b>	<b>59</b>
A.1	Résumé vidéo	59
A.1.1	Video Summarization Using Deep Neural Networks : A Survey [AAM <sup>+</sup> 21b]	59
A.2	Congruence visuelle inter-observateurs	62
A.2.1	Prediction of the Inter-Observer Visual Congruency (IOVC) and Application to Image Ranking [LMBR11]	62
A.2.2	Factors Underlying Inter-Observer Agreement in Gaze Patterns : Predictive Modelling and Analysis [RB16]	63
A.2.3	Deep Learning For Inter-Observer Conguency Prediction [BLCLM19]	64
A.2.4	Inter-Observer Visual Congruency in Video-Viewing [YLZ <sup>+</sup> 21]	65
A.3	Mémorabilité	66
A.3.1	Memorability of natural scenes - the role of attention [MLM13]	67
A.3.2	Relative Spatial Features for Image Memorability [KYP13]	68

A.3.3	Understanding and Predicting Image Memorability at a Large Scale [KRTO15] . . . . .	69
A.3.4	Deep Learning for Image Memorability Prediction : the Emotional Bias [SHDGD18] . . . . .	70
A.3.5	Deep Learning for Predicting Image Memorability [SHDGD18] . . . . .	72
A.3.6	Embracing New Techniques in Deep Learning for Estimating Image Memorability [NB21] . . . . .	73
<b>A.4</b>	<b>Intensité émotionnelle . . . . .</b>	<b>74</b>
A.4.1	Facial Expression Recognition in Videos Using Dynamic Kernels [PRC20] . . . . .	74
A.4.2	Facial Expression Recognition Using Residual Masking Network [PVT21] . . . . .	75
A.4.3	Frame Attention Networks for Facial Expression Recognition in Videos [MPWQ19] . . . . .	76
A.4.4	POSTER - A Pyramid Cross-Fusion Transformer Network for Facial Expression Recognition [ZMC22] . . . . .	77
A.4.5	Recognition of Emotion Intensities Using Machine Learning Algorithms : A Comparative Study [MSJ19] . . . . .	79
A.4.6	Spatio-Temporal Facial Expression Recognition Using Convolutional Neural Networks and Conditional Random Fields [HM17] . . . . .	80
<b>B</b>	<b>Planification</b>	<b>83</b>
B.1	Phase 1 . . . . .	83
B.2	Phase 2 . . . . .	83
<b>C</b>	<b>Fiches de suivi</b>	<b>88</b>
<b>D</b>	<b>Auto-contrôle et auto-évaluation</b>	<b>103</b>



---

# Introduction

## 1.1 Contexte

Un résumé de vidéo se décompose en une série d'images ou de fragments extraits d'une vidéo permettant de révéler rapidement le contenu d'une vidéo à un utilisateur pour lui permettre de faire le choix de regarder ou non cette vidéo [AAM<sup>+21b</sup>]. Plus le résumé est intéressant ou attractif, plus le visiteur sera enclin à regarder la vidéo. Dans le cadre de *User-Generated Content* (UGC, i.e. contenu généré par des utilisateurs, souvent non-professionnels), la masse de données hébergées sur différentes plateformes (e.g. YouTube, Facebook...) rend ces prévisualisations essentielles pour rendre les vidéos plus attractives et pour simplifier le choix de l'audience confrontée à ce volume. Pour exemple, plus de 500 heures de vidéos sont mises en ligne chaque minute sur YouTube<sup>1</sup>, cela amène à une surcharge d'information liée aux nombreuses vidéos proposées. Cette surcharge informationnelle peut créer de la confusion ou de la frustra-

tion chez les utilisateurs. Le résumé vidéo facilite donc la navigation parmi une large collection de vidéos en lui permettant de choisir celle qu'il trouvera la plus pertinente. Le résumé vidéo augmentera l'engagement utilisateur et sa consommation de contenu. Le résumé vidéo est donc nécessaire pour éviter d'être submergé par une masse considérable d'informations.

Néanmoins, l'utilité des résumés vidéo reste diverse [AAM<sup>+21b</sup>] : il est possible de générer automatiquement des bandes-annonces de films ou de séries, de présenter les temps forts d'un événement (e.g. sportif, musical ou politique) ou encore de créer un résumé vidéo des activités principales ayant eu lieu les dernières 24 heures d'un enregistrement d'une caméra de vidéosurveillance.

Le résumé vidéo généré ainsi que son évaluation dépendent du type de vidéo qu'il s'agit de résumer. Dans le cas de vidéos *User-Generated Content*, le résumé généré doit être attractif et la composante perceptuelle des utilisateurs est essentielle. Dans ce projet, nous nous intéressons donc à différentes caractéristiques perceptuelles.

---

1. <https://blog.youtube/press/>

Dans le cadre de notre dernière année d'étude du cycle ingénieur en informatique à Polytech Nantes, nous réalisons un projet de recherche et de développement dans lequel ce sujet sera étudié. MM. Alexandre Bruckert et Patrick Le Callet sont les superviseurs de ce projet.

## 1.2 Problématique

Notre problématique est d'incorporer dans la réalisation de résumé de vidéo, des caractéristiques perceptuelles obtenues à partir d'utilisateurs ou d'indicateurs automatiques (caractéristiques inférées). Le procédé a pour but d'améliorer le résumé d'une vidéo pour inciter un utilisateur à regarder une vidéo basée sur son contenu et certaines caractéristiques perceptuelles.

## 1.3 Problèmes

La plupart des méthodes actuelles de résumé vidéo se basent sur des réseaux de neurones profonds [AAM<sup>+</sup>21b, HSK<sup>+</sup>22], qui rendent difficile l'interprétation des caractéristiques extraites. Pour cette raison, l'idée sera de confronter l'évolution des images jugées les plus intéressantes (grâce aux méthodes de résumé vidéo) avec l'évolution de certaines caractéristiques perceptuelles (inférées), afin de produire un résumé vidéo attrayant défini par des caractéristiques perceptuelles s'inspirant des mécanismes de la perception humaine.

De nombreuses caractéristiques perceptuelles centrées utilisateur existent. Néanmoins, l'objet de cette étude se

limite aux caractéristiques suivantes, jugées - subjectivement - les plus intéressantes : la congruence inter-observateurs [YLZ<sup>+</sup>21], la mémorabilité [MLM13], et l'intensité émotionnelle [MSJ19]. Ces différentes caractéristiques doivent être comprises en vue de leur intégration dans une méthode de génération de résumé vidéo. Dans un premier temps, ces caractéristiques sont définies, puis différentes méthodes de prédiction de ces dernières sont analysées afin d'établir un état de l'art. En effet, il existe des modèles, méthodes et algorithmes qui permettent de prédire ces réactions humaines via par exemple des méthodes d'apprentissage profond comme pour la congruence inter-observateur [BLCLM19].

# État de l'art

En premier lieu, l'état de l'art se focalise sur les résumés vidéo en section 2.1, en les définissant puis en détaillant les méthodes de génération utilisant l'apprentissage profond. Puis, en section 2.2 les caractéristiques perceptuelles sont définies et des méthodes de mesure et de prédiction sont introduites. Les caractéristiques concernées sont la congruence visuelle inter-observateurs, l'intensité émotionnelle et la mémorabilité.

## 2.1 Résumé vidéo

### 2.1.1 Définition

Les méthodes de résumé vidéo consistent en l'extraction de parties d'une vidéo pour permettre de visionner rapidement les points principaux d'une vidéo et de donner envie à un utilisateur de regarder cette vidéo. [AAM<sup>+21b</sup>, HSK<sup>+22</sup>]

Les résumés de vidéo existent sous deux formes : la première forme est appelée *Storyboard* qui est un en-

semble d'images sélectionnées de la vidéo originale, et la deuxième forme est appelé *video skim* qui est un ensemble de fragments vidéo sélectionnés de la vidéo originale. Un exemple des deux formes est montré dans la figure 2.1.

### 2.1.2 Méthodes

La littérature sur les méthodes de résumé vidéo utilisant l'apprentissage profond sépare les méthodes selon deux approches : unimodale et multimodale [AAM<sup>+21b</sup>]. L'approche unimodale utilise seulement l'aspect visuel des vidéos pour extraire des caractéristiques afin d'apprendre de manière faiblement supervisée, supervisée ou non supervisée. L'approche multimodale exploite les métadonnées textuelles disponibles et apprend la sémantique en augmentant la pertinence entre la sémantique du résumé et celle des métadonnées associées à la vidéo (titre, catégorie, description, etc...) L'approche unimodale est détaillée dans les sous-sections

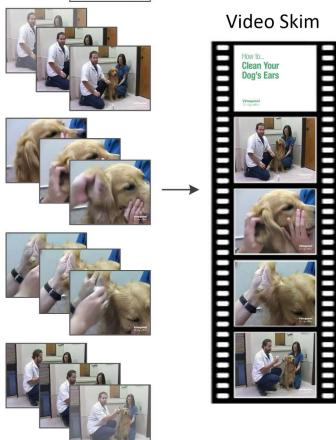
Original Video



Selected Frames (Storyboard)



Selected Fragments

FIGURE 2.1 – Storyboard et Video Skim [AAM<sup>+</sup>21b]

suivantes selon le mode d'apprentissage : supervisé en 2.1.2.1, non-supervisé en 2.1.2.2 et faiblement supervisé en 2.1.2.3. Puis, les méthodes fondées sur l'approche multimodale sont évoquées en sous-section 2.1.2.4.

### 2.1.2.1 Méthodes d'apprentissage supervisé

Les méthodes d'apprentissage supervisé permettant de générer des résumés vidéo entraînent leur modèle grâce à une vérité-terrain. Cette vérité-terrain est l'importance de chaque image annotée par des utilisateurs. Les utilisateurs jugeant subjectivement l'importance de chaque image, une même image pourra avoir des scores d'importance sensiblement différents selon 2 utilisateurs distincts. Cette vérité-terrain dépend totalement des utilisateurs ayant annoté les *frames* de la vidéo. Néanmoins, il existe différents types de méthodes d'apprentissage supervisé qui ont fait leur preuve dans la génération automatique de résumé vidéo.

Certaines méthodes cherchent à apprendre l'importance des images en modélisant les dépendances temporelles entre images. Le modèle prend en entrée la vidéo et les scores d'importance de chaque image selon les utilisateurs. Ces données servent à modéliser les dépendances temporelles entre images et à estimer l'importance des images. L'importance prédictive est comparé à l'importance de vérité-terrain pour améliorer le modèle. VASNet [FSA<sup>+</sup>18] est composé d'un mécanisme d'attention pour apprendre des caractéristiques qui donne de l'importance à une image de la vidéo et d'un réseau entièrement connecté à deux couches pour faire la régres-

sion sur les scores d’importance des images. SMN (Stacked Memory Network) [WWW<sup>+</sup>19] empile plusieurs couches de LSTM (long short-term memory) et plusieurs couches de mémoire, de manière hiérarchique pour obtenir le contexte temporel à long terme et utilise cette information afin d’estimer l’importance des images.

Des méthodes d’apprentissage supervisé apprennent à résumer une vidéo en trompant le discriminateur d’un GAN (Generative Adversarial Network, réseau antagoniste génératif) [AAM<sup>+</sup>21b], ce dernier essaie de distinguer un résumé généré par la machine d’un résumé généré par un humain. Le générateur du GAN reçoit en entrée la séquence des images vidéo et génère un score d’importance (ou poids) des images contenues dans la vidéo pour ensuite réaliser le résumé. L’entraînement se fait en alternant tour à tour une phase d’entraînement du générateur et une phase d’entraînement du discriminateur. Le résumé généré (poids pour chaque image) ainsi que les poids donnés par des utilisateurs pour un résumé optimal (vérité-terrain) sont fournis en entrée d’un discriminateur qui, en sortie, donne la similarité du résumé généré. Le modèle développé par [FTC19] utilise un GAN avec un réseau de neurones nommé *Ptr-Net* modifié permettant d’inférer le résultat pour des vidéos de taille variable. Ce modèle utilise un classifieur de type 3D-CNN en tant que discriminateur pour juger si un fragment vidéo provient du résumé de vérité-terrain ou d’un résumé créé par la machine.

### 2.1.2.2 Méthodes d’apprentissage non supervisé

Des méthodes apprennent à résumer une vidéo en trompant un discriminateur qui cherche à distinguer la vidéo originale d’une reconstruction de la vidéo grâce à un résumé [AAM<sup>+</sup>21a]. Ces méthodes sont fondées sur l’idée qu’un résumé représentatif doit permettre au spectateur d’inférer le contenu de la vidéo originale. Les GAN (Generative Adversarial Networks) apprennent à créer des résumés vidéo permettant une bonne reconstruction de la vidéo originale. Le générateur (générant le résumé) essaie de tromper le discriminateur qui cherche à distinguer la vidéo originale d’une vidéo reconstruite avec un résumé. La distinction n’est plus possible quand l’erreur de classification est quasi-égale pour la vidéo reconstruite et la vidéo originale, le générateur est alors considéré comme capable de créer un résumé vidéo fortement représentatif. Le forward GAN apprend à reconstruire la vidéo originale depuis le résumé vidéo et le backward GAN fait l’inverse. Le modèle AC-SUM-GAN [AAM<sup>+</sup>21a] intègre un modèle acteur-critique à un GAN. L’acteur et le critique participent à un jeu de sélection des images, ces choix donnent des récompenses en retour de la part du discriminateur. Le critique apprend à évaluer la sélection d’images et l’acteur apprend une politique de sélection d’images clés.

### 2.1.2.3 Méthodes d’apprentissage faiblement supervisé

Les méthodes d’apprentissage faiblement supervisé essaient d’atténuer la nécessité de disposer d’un grand

nombre de données de vérité-terrain annotées par l'homme. Au lieu de ne pas utiliser de vérité-terrain (apprentissage non-supervisé), ces méthodes utilisent des étiquettes dites faibles car moins coûteuses [AAM<sup>+</sup>21b]. Ces étiquettes peuvent être les métadonnées de la vidéo ou les annotations de vérité-terrain d'un petit sous-ensemble d'images de la vidéo qui permettraient d'apprendre à résumer une vidéo via l'apprentissage par renforcement et des fonctions de récompense adaptées. La méthode conçue par [PDW<sup>+</sup>17] utilise la métadonnée de titre des vidéos pour trouver des vidéos de la même catégorie que la vidéo à résumer. Ces vidéos sont utilisées pour extraire des caractéristiques avec un 3D-CNN afin d'apprendre les segments vidéo en commun de cette catégorie. Enfin, le meilleur résumé de la vidéo cible est celui qui est le plus représentatif de la catégorie. Cependant ces données là peuvent inclure des erreurs car les métadonnées des autres vidéos ajoutées par des humains peuvent être peu fiables. Il semble donc difficile de retrouver des vidéos d'une même catégorie. La pertinence des vidéos sélectionnées d'une même catégorie impacte donc fortement la génération d'un résumé vidéo vraiment représentatif de sa catégorie.

#### 2.1.2.4 Méthodes multimodales

Pour résumer les vidéos, ces méthodes cherchent à ajouter des modalités supplémentaires en plus du flux visuel. Ces modalités peuvent être diverses : le flux audio, les sous-titres ou la transcription automatique (*Automatic Speech Recognition*), les métadonnées textuelles (titre

de la vidéo ou sa description), ou encore des données de contexte (commentaires des spectateurs) [YMCZ19, LKA<sup>+</sup>17].

La majorité des méthodes multimodales utilisent les métadonnées textuelles comme le travail [YMCZ19]. Ces méthodes prennent en entrée la séquence des images de la vidéo à résumer, la vérité-terrain (l'importance de chaque image selon des utilisateurs) et enfin les métadonnées de la vidéo. L'apprentissage est donc réalisé de manière supervisée. Ensuite le modèle estime l'importance de chaque image et génère un résumé vidéo. Le résumé généré est comparé à un résumé de vérité-terrain (donne un score de similarité) ainsi qu'aux métadonnées (analyse sémantique du résumé vidéo et des métadonnées). Cependant certains travaux de recherche cherchent à inclure d'autres types de données. Par exemple, dans [LKA<sup>+</sup>17], les auteurs cherchent à réaliser le résumé vidéo d'un match de football en combinant des données récoltées à partir de capteurs portés sur chaque joueur, permettant au modèle de cibler les actions les plus intéressantes d'un point de vue des caractéristiques de l'image mais aussi des capteurs.

#### 2.1.3 Conclusion sur ces méthodes

Les meilleurs résumés vidéo sont obtenus via des méthodes d'apprentissage supervisé et non-supervisé [AAM<sup>+</sup>21b]. Les meilleures méthodes d'apprentissage supervisé utilisent des mécanismes d'attention (VASNet [FSA<sup>+</sup>18], H-MAN, SUM-GDA, DASP, et CSNetsup) ou des réseaux de mémoire (SMN [WWW<sup>+</sup>19]) pour

capturer les dépendances temporelles. Les meilleures méthodes d'apprentissage non-supervisé utilisent des réseaux antagonistes génératifs (GAN) ainsi que des mécanismes d'attention (AU-SUM-GAN [AAM<sup>+</sup>21a], CSNet, CSNet+GL+RPE, SUM-GDAunsup, et SUM-GANAA).

Les autres méthodes supervisées et non supervisées sont moins performantes et les méthodes d'apprentissage faiblement supervisé ne permettent pas d'obtenir de résumés vidéos de qualité.

Néanmoins, l'évaluation de la performance de ces modèles de résumés vidéo se base sur des méthodes qui sont chronophages et subjectives puisqu'il s'agit de faire annoter des vidéos par des humains ou bien de faire des études utilisateurs.

## 2.2 Caractéristiques perceptuelles

Les modèles de résumés vidéo se reposant, pour leur évaluation comme pour leur entraînement, sur de nombreuses annotations humaines, il semble indispensable de prendre en compte l'impact de la vidéo sur le spectateur. Pour cela, les caractéristiques perceptuelles suivantes sont détaillées : la congruence inter-observateurs en 2.2.1, l'intensité émotionnelle en 2.2.3 et la mémorabilité en 2.2.2. Ces caractéristiques sont analysées dans l'objectif de leur utilisation pour générer des résumés vidéo.

### 2.2.1 Congruence visuelle inter-observateurs

#### 2.2.1.1 Définition

La congruence visuelle inter-observateurs (IOC ou IOVC, *inter-observer visual congruency*) reflète le degré de dispersion entre les zones d'attentions visuelles de différentes personnes observant une même image.

Selon la littérature sur la saillance visuelle, les observateurs d'une même image peuvent avoir des comportements oculaires considérablement différents [LMBR11]. Cette différence peut dépendre de l'image regardée ou bien de l'observateur. Les variations intrinsèques à l'observateur sont influencées par des aspects comme la culture, l'âge, la condition physique ou les expériences passées. Tandis que celles dépendant de l'image sont liées au contenu de l'image, comme la présence de visages qui attire fortement l'attention.

Une première analyse est réalisée par [RB16] sur 3 jeux de données et permet de mieux comprendre le score de l'IOVC (voir images triées par l'IOVC en figure 2.2). Un IOVC élevé implique généralement un petit nombre de régions à fort contraste. La présence de personnes ou de visages et la présence de texte donnent un IOVC élevé car les observateurs auront tendance à regarder à ces endroits précis de l'image. Tandis qu'un IOVC est faible si rien n'attire l'attention ou si trop de choses l'attirent et que le temps de visionnage est trop court. Il s'agit généralement de scènes encombrées ou de paysages.

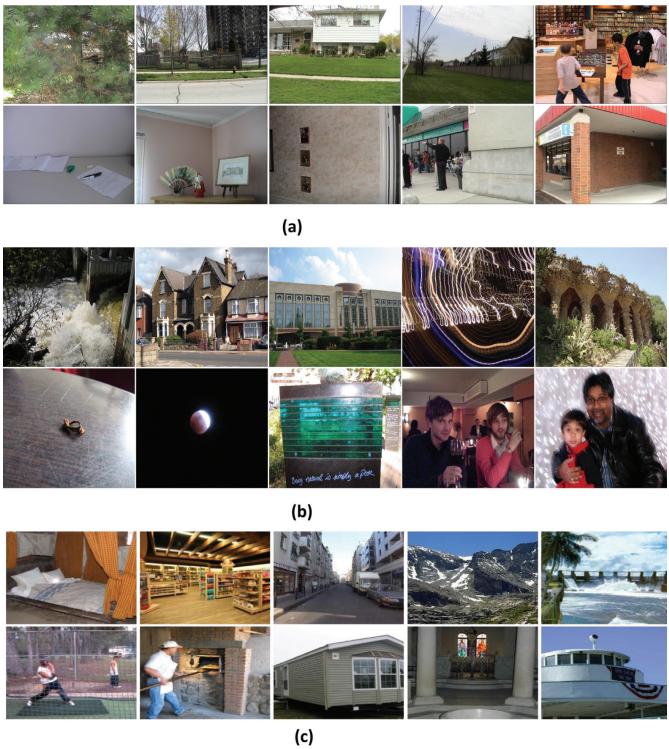


FIGURE 2.2 – Images triées par score de congruence inter-observateurs (le plus faible en haut à gauche et le plus fort en bas à droite). En (a) : jeu de Bruce/Toronto, (b) : jeu de Judd/MIT, (c) : jeu de mémorabilité [MLM13]). La figure provient de [RB16].

### 2.2.1.2 Mesure

La congruence inter-observateurs d'une image peut se mesurer sur des données d'eye tracking de plusieurs observateurs afin de servir de vérité-terrain sur un ensemble d'images observées. Le calcul de l'IOVC se fait grâce à la méthode *leave-one out* [LMBR11, RB16, BLCLM19, YLZ<sup>+</sup>21] aussi appelée *one-against-all*. La méthode se base sur les données de mouvements des yeux de n observateurs d'une même image. La congruence inter-observateurs est calculée pour chaque observateur (voir exemple pour un observateur en figure 2.3). Pour un observateur i, on superpose les zones de fixations du regard des autres observateurs (sans les zones de l'observateur i) et on applique une convolution Gaussien 2D sur ces zones afin de simuler l'angle visuel humain. L'écart-type de la convolution gaussienne estime la taille de la fovéa (1 degré dans [LMBR11], 2 degrés dans [YLZ<sup>+</sup>21]). Une zone de fixation est donc définie par un point et ses voisins représentant le degré d'angle visuel. Cela donne une carte de chaleur (heat map) avec la probabilité de fixation de chaque pixel.

Dans [LMBR11], cette carte est seuillée pour sélectionner seulement les 25% de l'image ayant la plus forte probabilité de fixation. Dès lors, il suffit de calculer le pourcentage de zones de fixation de l'observateur i qui sont incluses dans la carte seuillée.

Dans [BLCLM19], les zones de fixation de l'observateur i sont comparées à la carte de probabilité de fixation de chaque pixel avec des mesures de saillances telles que *AUC* (area under the curve) ou *NSS* (normalized scanpath saliency).

La congruence visuelle inter-observateurs d'une image est donc la moyenne entre les scores de congruence visuelle inter-observateurs de chaque observateur.

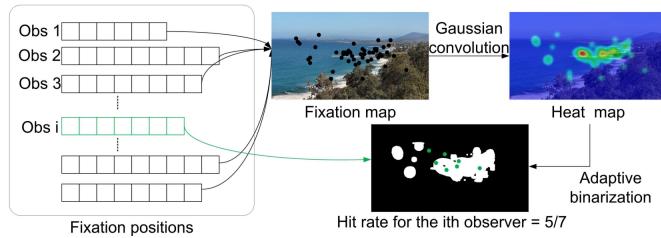


FIGURE 2.3 – Méthode leave-one out pour le calcul de la congruence inter-observateurs de l'observateur  $i$ , provient de [LMBR11]

La congruence inter-observateurs d'une vidéo est mesurable image par image mais il faut prendre en compte le fait qu'une même scène dure plusieurs images. Les observateurs peuvent regarder les mêmes zones saillantes d'une scène mais l'ordre du regard de ces zones variera pour chaque observateur [YLZ<sup>+21</sup>]. Calculer image par image donnerait donc un IOVC anormalement bas. Pour compenser ce problème, les zones de fixation de  $n$  images à suivre sont superposées (pour pallier au problème d'ordre) et le calcul se fait toutes les  $m$  images (pour éviter la redondance dans les vidéos ayant une trop haute fréquence), avec  $m \geq n$  pour s'assurer que tous les points de regards sont pris en compte.

### 2.2.1.3 Prédiction

Les travaux [LMBR11, RB16, BLCLM19] cherchent à prédire la congruence visuelle inter-observateurs d'images et le travail [YLZ<sup>+21</sup>] la congruence visuelle inter-observateurs de vidéos.

#### Prédiction de l'IOVC d'image

Le travail de recherche [LMBR11] se base sur des attributs qui impactent la variabilité inter-observateurs : la détection de visages, l'harmonie des couleurs, la profondeur de champ et la complexité de la scène (l'entropie, le nombre de régions et le nombre de contours). Ces 6 attributs sont calculés dans le but de simuler le comportement visuel humain lors de l'observation d'une image. Chaque image est donc représentée par un vecteur de caractéristiques de dimension 6. Ils utilisent un modèle de type *cluster-weighted* qui est une généralisation des modèles de mélanges Gaussiens. La complexité du modèle dépendant du nombre de clusters  $N$ , ils utilisent  $N = 9$  pour ne pas surajuster (*overfit*) afin d'obtenir un modèle robuste. Ils obtiennent une corrélation de Pearson  $r = 0.340$  sur le jeu de données Judd et al.. Ce jeu de données contient 1003 images de scènes en intérieur et extérieur, en format portrait ou paysage, ainsi que les données d'eye tracking de 15 observateurs durant 3 secondes en visionnage libre.

[RB16] se sert de l'analyse d'image de type *bottom-up* qui concerne l'influence des propriétés du contenu de l'image sur le regard. Ils proposent une mesure basée sur un histogramme : HoPS (Histogram of Predicted Salience). HoPS est un vecteur de caractéristiques basé

sur 12 algorithmes de saillance qui permet de représenter la variabilité entre les prédictions de saillance. Ils testent aussi les caractéristiques HoG (Histogram of Oriented Gradients) et Gist qui permettent d'obtenir une représentation grossière de la structure de la scène de l'image. De plus, ils se servent de l'analyse *top-down* de l'image qui concerne l'influence des perceptions haut-niveau ou des connaissances préalables sur le regard. Pour cela, des caractéristiques haut-niveau sont capturées via des réseaux de neurones convolutifs (DeepNet) (architecture BVLC Reference CaffeNet [Jia et al. 2014] basée sur l'architecture AlexNet [Krizhevsky et al. 2012] (et entraînée sur les données *ILSVRC12*)). Ils testent les caractéristiques et constatent que HoPS seul surpassé HoG et Gist. De plus, avec les caractéristiques HoPS + DeepNet, ils surpassent les précédents travaux en obtenant une corrélation  $r = 0.456$  sur Judd et al. (mais aussi  $r = 0.519$  sur le jeu de mémorabilité [MLM13] et  $r = 0.506$  sur le jeu Bruce/Toronto).

Un réseau de neurones profond est utilisé par [BLCLM19] pour prédire la congruence visuelle inter-observateurs d'image. Le réseau est composé d'un encodeur VGG19 (pré-entraîné sur ImageNet) et pour le décodeur, le réseau est composé de simples couches connectées pour faire la régression sur le score de l'IOVC. De plus, l'apprentissage par transfert (*transfer learning*) est utilisé pour pallier le manque de données. Ils surpassent [LMBR11, RB16] avec  $r = 0.611$  sur le jeu de données Judd et al. et surpassent [RB16] sur le jeu de mémorabilité ( $r = 0.537$ ), mais performent légèrement moins bien ( $r = 0.473$ ) que ce dernier sur le jeu Bruce/Toronto.

## Prédiction de l'IOVC de vidéo

Un réseau à double branches est utilisé par [YLZ<sup>+</sup>21] pour prédire l'IOVC d'une vidéo. Il y a une branche de contenu et une branche de flux. La branche de contenu prend une image de la vidéo en entrée et en extrait la sémantique. La branche flux extrait les caractéristiques de variation de mouvements entre images via le flux optique (estimé par FlowNet 2.0 [IMS<sup>+</sup>16]). Chaque branche a une structure similaire aux ResNet (réseaux de neurones résiduels) : une couche de convolution et quatre couches résiduelles. Les caractéristiques de mouvement entre images sont intégrées aux caractéristiques de contenu pour obtenir des informations d'attention.

Pour valider la performance, les auteurs ont établi un jeu de données de 12 clips vidéo de 5 minutes chacun grâce à la base *C.LIRIS-ACCEDE* contenant des films annotés avec les émotions procurées. Ils ont fait une expérience d'eye-tracking sur ces clips afin de mesurer l'IOVC.

Les auteurs évaluent leur modèle ainsi que les modèles de prédiction d'IOVC d'images [LMBR11, RB16, BLCLM19] sur ce nouveau jeu de données, en mesurant la corrélation de Spearman  $p$ . Ils obtiennent  $p = 0.693$  pour une division moyenne des jeux de données (d'entraînement, de test et de validation), et  $p = 0.573$  pour une division croisée des jeux de données. Leur méthode a été meilleure que les précédents modèles de prédiction sur ce nouveau jeu de données. Néanmoins, étant une méthode de prédiction d'IOVC de vidéos, les auteurs n'ont pas pu tester cette méthode sur les bases d'images annotées avec l'IOVC et testées par les méthodes de prédiction d'IOVC

d'images (Judd/MIT, Bruce/Toronto et jeu de mémorabilité).

Leur jeu de données leur permet de mesurer la corrélation entre l'IOVC et les émotions. La corrélation entre l'IOVC et l'*arousal* qui décrit le niveau d'énergie (e.g. excité ou endormi) corrèle à  $p = 0.43$  quand la *valence* (décrivant le degré positif de l'expérience, e.g. satisfait ou déçu) est positive et de 0.29 quand elle est négative. Une intrigue plaisante concentrera davantage l'attention des observateurs, donc un plus haut IOVC.

## 2.2.2 Mémorabilité

### 2.2.2.1 Définition

La mémorabilité de l'image est la faculté d'une image à être remémorée après un certain temps. Isola et al. [IXTO11] ont montré que la mémorabilité d'une image est congruente entre observateurs et est donc une propriété intrinsèque des images.

### 2.2.2.2 Mesure

Pour comparer la performance de modèles de prédiction de la mémorabilité, il est essentiel d'obtenir le score humain de mémorabilité afin d'établir une vérité-terrain.

Khosla et al. [KRTO15] ont pu établir le large jeu de données *LaMem* contenant 60000 images de sources diverses. Pour cela le protocole expérimental suivant a été utilisé (décrit en figure 2.4). Sur la plateforme web de crowdsourcing *Amazon's Mechanical Turk*, des images de ce jeu de données sont montrées à la suite à des parti-

cipants. Pour plus de précision, les participants échouant à détecter la même image répétée moins de 7 images après celle-ci ont été bloqués et leurs résultats non pris en compte. Le score de mémorabilité est donc calculé pour les autres sur base de leur capacité à se remémorer une image qu'ils ont vu au cours de l'expérience. Ce protocole permet d'obtenir environ 80 scores par image. Les 10 images les plus mémorables ainsi que les 10 les moins mémorables provenant de la source SUN sont visibles en figure 2.5. On peut y voir que les images les plus mémorables contiennent des humains ou bien un seul objet ou animal au centre. La salle de cinéma et le cimetière américain sont aussi très mémorables, il semble qu'un même objet simple répété de nombreuses fois (le fauteuil de cinéma et la croix) sur une image permettent d'augmenter la mémorabilité. Tandis que les images les moins mémorables sont des scènes en intérieur et des paysages contenant beaucoup de détails qui sont compliqués à retenir.

L'étude [SHDGD18] établit un jeu de données de 150 images annotées en utilisant le protocole expérimental de [KRTO15] sur 50 participants pour obtenir des scores de mémorabilité mais aussi sur les émotions procurées par l'image. Ces 150 images sont extraites du jeu de données *International Affective Picture System*. Le premier jour des images sont passées à la suite aux participants pour obtenir un score de mémorabilité des images (comme [KRTO15]), le second jour on passe des images pour obtenir un score d'émotions basé sur une échelle de mesure avec un système de pictogramme pour l'*arousal* (si l'image est stimulante) et la *valence* (si l'image procure une émotion positive ou négative).

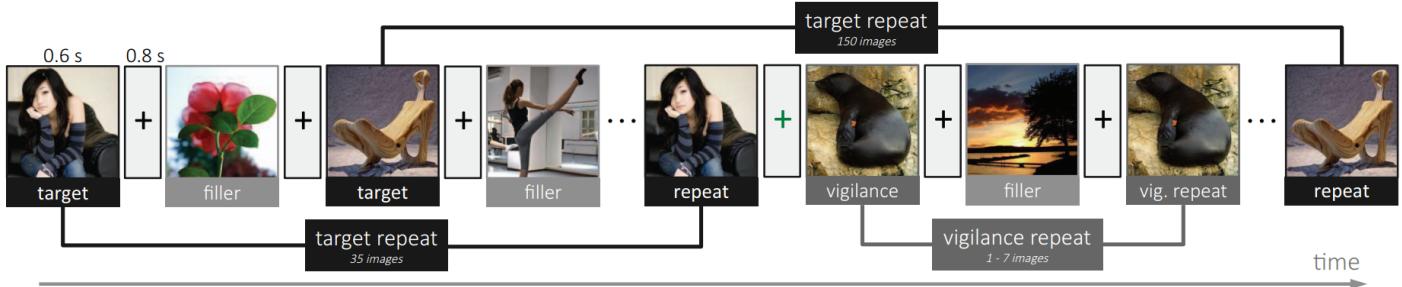


FIGURE 2.4 – Protocole expérimental de [KRTO15] pour mesurer la mémorabilité d'image par des humains (figure provenant de [KRTO15])

### 2.2.2.3 Modèles traditionnels de prédiction

L'étude [IXTO11] développe un modèle de prédiction de la mémorabilité utilisant l'ensemble des caractéristiques : Gist, SIFT, HOG, SSIM et les histogrammes Pixel et utilisent un classifieur SVR (Support Vector Regression) et obtiennent une corrélation de Spearman  $p = 0.46$ .

L'étude [MLM13] prouve que la mémorabilité d'une image et l'attention visuelle sont liées, à travers une expérience d'eye-tracking sur 17 volontaires ayant regardé 135 images extraites du jeu de données *SUN* composé de 2222 images annotées avec score de mémorabilité. L'expérience d'eye-tracking permet de mesurer l'attention visuelle à travers la congruence inter-observateurs et la durée de fixation. Les 20 images les plus mémorables sont significativement fixées plus longtemps que les 20 moins mémorables. Les 2 classes d'images les moins mémorables ont une congruence inter-observateurs significati-

vement moins élevée que la classe d'images les plus mémorables. Leur modèle utilise donc deux caractéristiques liées à l'attention : la couverture moyenne des cartes de saillance et la visibilité selon le contraste des structures de l'image. Ils utilisent aussi les caractéristiques de [IXTO11] pour prédire la mémorabilité. Puis, ils utilisent un classifieur SVR. Avec ces 2 caractéristiques + SIFT, HOG, SSIM et Pixel, ils surpassent [IXTO11] en obtenant  $p = 0.479$ .

L'étude [KYP13] cherche à remplacer certaines caractéristiques (Object, Scene et Attributes) en développant deux nouvelles caractéristiques spatiales, corrélées à la mémorabilité. Il s'agit de la zone d'objet pondérée (WOA, Weighted Objet Area) prenant en compte la position et la taille des objets et le rang relatif de la zone (RAR, Relative Area Rank) capturant le caractère inhabituel relatif de la taille des objets. Avec les caractéristiques WOA + RAR + Scene + Attributes + celles de [IXTO11] et un classifieur SVR, ils obtiennent  $p = 0.58$ .

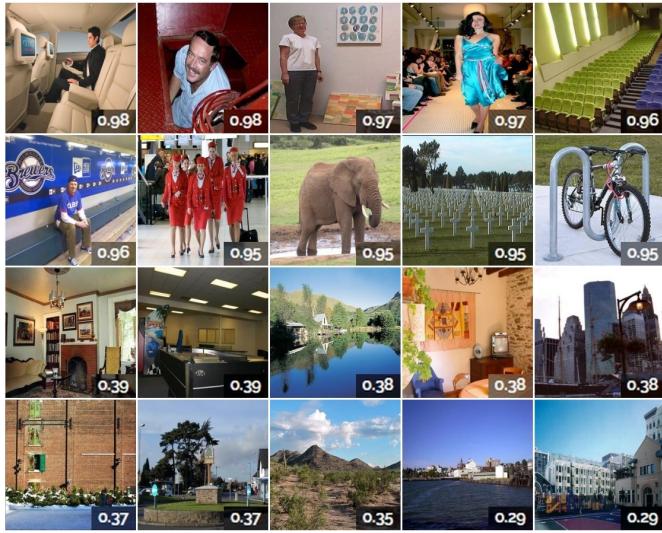


FIGURE 2.5 – Images classées par ordre décroissant de mémorabilité. En haut : les 10 images les plus mémorables, en bas : les 10 images les moins mémorables, provenant du jeu de données SUN utilisé dans le jeu de mémorabilité LaMem [KRTO15]

#### 2.2.2.4 Modèles de prédiction utilisant l'apprentissage profond

L'étude [KRTO15] utilise l'apprentissage profond et a montré que des caractéristiques *fine-tuned* surpassent toutes les autres par une large avance, atteignant une corrélation  $p = 0.64$ , proche de celle humaine ( $p = 0.68$ ). Pour cela, ils ont conçu un très large jeu de données de mémorabilité (60 000 images de sources diverses). Puis, ils ont développé le réseau MemNet utilisant un réseau

de neurones convolutifs hybride [ZLX<sup>+</sup>14] pré-entraîné avec les jeux de données *ILSVRC 2012* et *Places* (3,5 millions d'images).

L'étude [BCPDSLC16] développe le modèle *MemoNet* utilisant le modèle *GoogleNet* qui est l'état de l'art de 2014 pour la classification d'images sur le large jeu de données *ImageNet*. Pour obtenir le modèle MemoNet, ils ont *fine-tune* le modèle *GoogleNet* pour l'adapter à la tâche de prédiction de la mémorabilité. Le *fine-tuning* consiste en le fait de pré-entraîner un réseau de neurones à convolution (*GoogleNet*) sur un large jeu de données externe (*ImageNet*) et de peaufiner (*fine-tune*) le réseau pré-entraîné en poursuivant l'entraînement sur les données cibles pour s'adapter à une tâche spécifique de classification (prédiction de la mémorabilité). Leur meilleur modèle a été obtenu avec 30000 itérations d'entraînement. Ils obtiennent une corrélation  $p = 0.636$ , comparable à [KRTO15] sur le jeu de données composé de 2222 images annotées et collectées par [IXTO11] provenant du jeu de données SUN [XHE<sup>+</sup>10]. De plus, sur le nouveau jeu de données des 150 images annotées avec la mémorabilité et les émotions, ils obtiennent  $p = 0.251$ . Cette faible performance est expliquée par le biais émotionnel. Ce biais indique que la plus haute performance de prédiction de la mémorabilité est obtenue pour les images suscitant des émotions négatives stimulantes (*valence* et *arousal* élevées), tandis que les images procurant des émotions neutres ou positives et peu stimulantes rendent la prédiction moins fiable (*valence* moyenne à élevée et *arousal* moyen à bas).

[SHDGD18] utilise des caractéristiques visuelles et sé-

mantiques et traite le problème comme un problème de classification (et non de régression comme [KRT015]) en divisant le jeu de données de mémorabilité LaMem [KRT015] en 4 classes de 15000 images. Ils utilisent le réseau VGG16 pré-entraîné sur ImageNet pour extraire des caractéristiques visuelles. De plus, ils utilisent des caractéristiques sémantiques avec un modèle de légende d'image (IC, *image captioning*). Le modèle d'IC utilise un réseau de neurones à convolution et un LSTM (long short-term memory recurrent network). Le classifieur contient une branche pour les caractéristiques du VGG16 et une pour les caractéristiques IC. Une branche est composée d'un ou plusieurs perceptrons multicouche (MLP, multilayer perceptron). Une couche fusionne les deux branches, puis des couches de MLP et enfin un *softmax* est appliqué pour obtenir les probabilités de chacune des 4 classes. Ils surpassent l'état de l'art (MemNet [KRT015]) en obtenant une corrélation de Spearman  $p = 0.72$  sur le jeu de données LaMem [KRT015].

*ResMem* [NB21] s'intéresse à l'usage de la technique de réseau de neurones convolutif résiduel. Introduit par les modèles *ResNet* [HZRS15], ce type de connexion entre les caractéristiques extraites et les caractéristiques en entrée du module permettent aux modèles d'obtenir de meilleurs résultats sur un ensemble de tâches variées comme la classification ou la segmentation. Ainsi, en couplant ce type de connexion et en utilisant un réseau de type *AlexNet* [KSH17] en parallèle pour l'extraction de caractéristiques liées à l'image, les auteurs parviennent à obtenir un meilleur score de corrélation sur une combinaison des jeux de données LaMem [KRT015]

et MemCat [GW19] que le modèle MemNet [KRT015] (ré-entraîné sur le même jeu de données pour valider les résultats). L'article montre aussi que l'utilisation de différentes techniques comme le *fine-tuning*, et l'utilisation d'une branche de segmentation d'image en parallèle de l'extraction des caractéristiques d'image permettent au modèle d'obtenir de meilleurs résultats au prix des performances lors de l'entraînement ou de l'inférence.

## 2.2.3 Intensité émotionnelle

### 2.2.3.1 Définition

L'intensité émotionnelle cherche à connaître l'émotion d'une personne en analysant son visage. Chaque mouvement du visage est répertorié en *Action Unit* (AU). Par exemple : un sourire ou des yeux plissés. L'intensité émotionnelle peut être utilisée dans de nombreux contextes : sécurité avec reconnaissance faciale, détection d'autisme, système de recommandation, et retour automatique d'expérience utilisateur sur un produit en temps réel lors de son utilisation. Les travaux de recherche se focalisent notamment sur les 7 expressions identifiées en figure 2.6.



FIGURE 2.6 – Classes d’émotions généralement utilisées, provient de [ZMC22]

### 2.2.3.2 Méthodes de classification

Deux méthodes de classification sont présentes dans les travaux de recherche : la première est l’utilisation d’image seule, et la seconde est l’utilisation de séquence d’images d’une personne réagissant à un évènement. La deuxième méthode permet d’ajouter du contexte pour permettre la précision de la classification. Certains algorithmes utilisent des caractéristiques de l’image du visage de la personne, et d’autres utilisent des caractéristiques obtenues via la représentation géométrique du marquage du visage.

### 2.2.3.3 Méthodes de classification via vidéo

Fonctionnant avec des vidéos (ou séquence d’images) pour utiliser le contexte de la temporalité, [MSJ19] utilise des techniques de *Machine Learning* : dans un premier temps, un algorithme d’extraction de caractéristiques est appliqué sur chaque image individuellement, puis un second algorithme permet la classification de toutes les caractéristiques extraites. Pour l’extraction des caractéristiques : *Gabor Features*, *Histogram of Oriented Gradients*, et *Local Binary Pattern* furent testés et pour les algorithmes de classification : *Support Vector Machine*, *Random Forest*, et *k-Nearest Neighbors* furent testés. Une fois toutes les combinaisons entraînées et testées, les meilleurs résultats obtenus utilisent la méthode *Local Binary Pattern + Support Vector Machine*.

[HM17] s’intéresse aux relations entre les caractéristiques extraites de chaque image de la séquence pour établir un lien permettant d’améliorer la classification de la séquence. Pour cela, un encodeur *Inception-ResNet* est appliqué sur chaque image individuellement pour extraire les caractéristiques, puis un réseau de neurones *Conditional Random Fields* permet d’apprendre les relations entre les caractéristiques à travers les différentes images. Il est montré dans cet article de recherche, que l’utilisation de séquence d’images permet de renforcer la précision du modèle dans la classification des émotions. Pour l’entraînement de ce réseau, plusieurs jeux de données sont utilisés : Ck+ [LCK<sup>+</sup>10], MMI [PVRM05], et FERA [BMS11]. Ces jeux de données contiennent des vidéos de personnes seules réagissant à un évènement avec la catégorie des émotions identifiées au cours du temps.

*Emotion-FAN* [MPWQ19] s'intéresse aux relations entre plusieurs images d'une vidéo en utilisant : premièrement, un module d'attention est appliqué aux images seules pour extraire des caractéristiques, puis en utilisant un module d'attention sur l'ensemble des caractéristiques résultant de la première étape pour pouvoir analyser les caractéristiques afin de classifier l'émotion de la séquence d'images.

Utilisant des caractéristiques obtenues via la représentation géométrique du marquage du visage, [PRC20] s'intéresse au déplacement des points du marquage du visage et au flux optique autour de chaque point pour en extraire des caractéristiques du visage sur chaque image de la vidéo. Plusieurs noyaux de classification sont testés pour en extraire la classification via un *Support Vector Machine*.

#### 2.2.3.4 Méthodes de classification via image unique

Utilisant des caractéristiques extraites d'image unique, *POSTER* [ZMC22] fonctionne avec une combinaison d'extraction de caractéristiques de l'image, et une représentation géométrique du marquage du visage. Les deux branches sont fusionnées dans un module d'attention en inversant les clés des caractéristiques permettant au modèle de les fusionner et donc d'utiliser ces deux types de caractéristiques pour inférer l'émotion de l'image actuelle.

[PVT21] s'intéresse à un moyen de forcer le réseau à favoriser des zones précises du visage comme les sourcils, les yeux ou la bouche en utilisant des blocs similaires à l'architecture d'un U-Net et complété par un mo-

dule d'attention permettant de contrôler l'importance des caractéristiques mises en évidence. Partant du constat que la représentation géométrique du visage est généralement mal prédite dans des conditions qui ne sont pas optimales (par exemple : User-Generated Content), le réseau est forcé d'utiliser les bonnes parties du visage pour inférer la classe correcte.

## 2.3 Conclusion

Afin d'inférer les différentes caractéristiques perçues étudiées, différents modèles sont retenus.

Pour la congruence visuelle inter-observateurs (IOVC), le modèle proposé par Bruckert et al. [BLCLM19] permettant de prédire l'IOVC d'une image est le plus robuste car a été testé sur 4 jeux de données différents (Judd/MIT, Bruce/Toronto, Mémorabilité et CAT2000) et obtient de meilleures performances que les autres recherches (mais performe légèrement moins bien que [RB16] sur le jeu Bruce/Toronto). De plus, le code source pourra être fourni par M. Bruckert, superviseur de ce projet. Par ailleurs, le réseau de neurones proposé par Jiaomin Yue et al. [YLZ<sup>+</sup>21] obtient de bonnes performances et permettrait de prédire l'IOVC de vidéos, néanmoins, le code source n'est - a priori - pas disponible.

Pour la prédiction de la mémorabilité, le modèle proposé par l'étude [SHDGD18] surpasse nettement le réseau MemNet [KRTO15]. Néanmoins il n'apparaît pas que le code de ce modèle soit public. Pour cela, il était essentiel d'élargir nos recherches d'articles avec code. ReSMem [NB21] est un modèle de prédiction de la mémora-

bilité d’images dont le code est public et simple d’utilisation, sous forme de [bibliothèque Python](#) où il est possible d’utiliser le modèle pré-entraîné.

Pour la reconnaissance des émotions, nous retiendrons le modèle proposé par Pham et al. [PVT21] proposant la technique la plus récente, obtenant de bons résultats, et ayant le code source ainsi que les fichiers du pré-entraînement disponibles sur le [GitHub du projet](#).

Pour pouvoir comparer les résultats de notre réseau avec une méthode de résumé vidéo automatique, nous utiliserons le modèle AC-SUM-GAN [AAM<sup>+</sup>21a] qui est une méthode d’apprentissage non-supervisé qui obtient des résultats intéressants et met à disposition le code source et des poids du réseau disponible sur le [dépôt GitHub du projet](#). Nous pourrons aussi utiliser le modèle d’apprentissage supervisé VASNet [FSA<sup>+</sup>18] dont le code est public sur le [dépôt du projet](#) et dont le modèle obtient aussi de bonnes performances.



---

# Proposition

Lors de notre étude de l'état de l'art nous n'avons pas rencontré d'étude effectuant un travail de résumé de vidéo prenant en compte des caractéristiques inférées s'inspirant des mécanismes de la perception humaine.

Ce projet vise à intégrer des caractéristiques perceptuelles dans la génération de résumé vidéo.

Dans un premier temps, il s'agira d'analyser les bases de données de résumé vidéo à notre disposition à travers ces différentes caractéristiques. Nous essaierons de trouver des profils-types de vidéos en étudiant l'évolution du score de chaque caractéristique au cours des vidéos, à travers des méthodes de réduction de dimensions et de clustering. Puis, nous ferons des tests de corrélation linéaire avec les tests de Pearson et de Spearman ainsi que des tests de similarité en utilisant la distance de Wasserstein. Enfin, nous pourrons faire un test d'indépendance avec le test de Student pour tester la différence entre le score des caractéristiques sur les frames sélectionnées et sur les frames non sélectionnées dans le résumé vidéo.

Après avoir analyser les différentes caractéristiques

perceptuelles sur nos bases de données, il s'agira de proposer des modèles de prédiction de l'importance des images pour établir un résumé vidéo. Nous testerons deux approches, l'une où l'on souhaitera prédire le score d'importance de chaque image de la vidéo et l'autre où nous souhaiterons classifier si une image a été sélectionné ou non dans le résumé vidéo. Dans le cas de prédiction de score, il suffirait de seuiller le score pour sélectionner les image à utiliser dans le résumé vidéo final.

# Expérimentations et résultats

## 4.1 Données utilisés

Le jeu de données **TVSum** est composé de 50 vidéos courtes (moyenne de 4 minutes) dont la vérité-terrain est sous forme d'annotations d'importance de chaque frame, donnée par 20 personnes. Les notes vont de 1 à 5. La vérité-terrain utilisée durant nos expérimentations est la moyenne par frame des scores donnés par les 20 utilisateurs.

Le jeu de données **SumMe** est composé de 25 vidéos courtes (moyenne de 2 minutes) dont la vérité-terrain est proposée en format MatLab et est calculée depuis les annotations d'au moins 15 personnes.

Le jeu de données **VSumm** est composé de 50 vidéos courtes de vidéos de YouTube ainsi que de 50 vidéos courtes provenant d'Open-Video (moyenne de 1 minute 30). La vérité-terrain proposée est la simple sélection de frames des utilisateurs (et non un score par frame). 10 vidéos provenant de YouTube sont au format FLV et n'ont pas été utilisées dans les expérimentations puisque la fré-

quence d'images indiquée dans les métadonnées et utilisée pour établir leur vérité-terrain ne correspondait pas à la réalité des vidéos, la fréquence étant en réalité bien moindre.

Le jeu de données **VISIOCITY** est composé de 67 vidéos allant de 14 à 121 minutes avec une durée moyenne 55 minutes. Néanmoins, les liens de ce jeu de données n'étaient plus actifs et semblaient non maintenus. Un contact au chercheur à l'origine de l'établissement du jeu de données a été tenté, mais nous n'avons jamais eu de réponses.

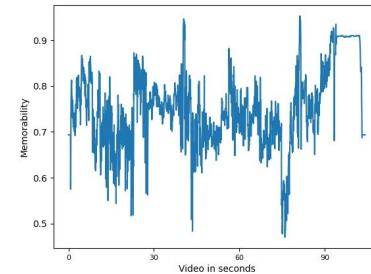
Le jeu de données **MED Summaries** est composé de 160 vidéos avec annotations. Les données n'étant plus accessible sur internet, nous avons contacté l'auteur du jeu de données. Malheureusement, l'auteur ne pouvait pas nous fournir les vidéos car ils ne les possèdent pas et étaient seulement disponibles pour les participants du challenge *Trecvid MED 2011*.

## 4.2 Modèles utilisés

### 4.2.1 Mémorabilité

Pour inférer la mémorabilité, le modèle ResMem [NB21] a été utilisé, pour sa simplicité et son efficacité. En effet, il s'agit d'un modèle disponible sous forme de bibliothèque Python, de plus ses performances étaient équivalentes à l'état de l'art.

Analyse qualitative : sur la courbe d'évolution d'une vidéo (voir figure 4.1a) à 75 secondes de la vidéo, le score prédit de mémorabilité chute drastiquement puis il y a un pic très haut, puis une légère chute vers un score tout de même assez haut. La chute sévère (figure ) correspond à un basculement vers un plan large très sombre où l'on aperçoit une personne sur un vélo mais l'on n'arrive pas à voir son visage. Puis le pic très haut (figure ) correspond à la transition avec le logo de l'émission très simple qui est composé de deux lettres, le texte augmentant fortement la mémorabilité en théorie. Enfin, une seconde chute vers un score tout de même haut de mémorabilité (figure ) où l'on voit la présentatrice au centre de la vidéo dans un plan assez rapproché : la mémorabilité semble bien augmenter avec la reconnaissance de visages.



(a) Evolution de la mémorabilité



(b) Chute de mémorabilité (c) Pic de mémorabilité



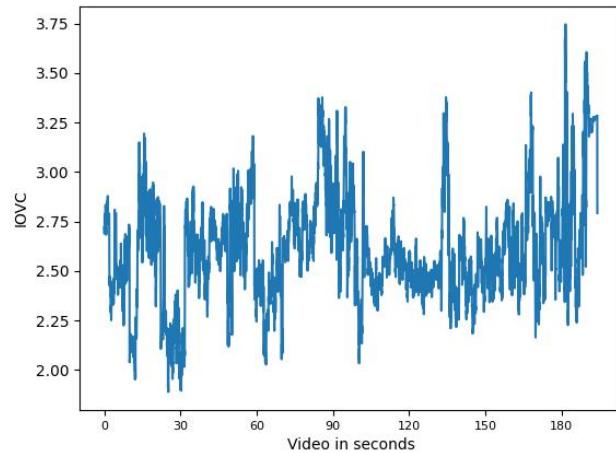
(d) Après la seconde chute

FIGURE 4.1 – Exemple de chutes et pic du score de mémorabilité

### 4.2.2 Congruence visuelle inter-observateurs

Pour inférer la congruence visuelle inter-observateurs, M. Bruckert, auteur d'un modèle d'IOVC utilisant l'apprentissage profond [BLCLM19], a pu entraîner le modèle développé durant ce travail de recherche sur moins

de données qu'originalement. Il s'agit donc d'une variante du modèle [BLCLM19] qui a été utilisé pour inférer cette caractéristique.



Analyse qualitative de la congruence visuelle inter-observateurs inférée avec notre modèle : plusieurs chutes de score de l'IOVC interviennent au cours de la vidéo 4.2a, prenons l'exemple de la chute d'IOVC vers 20 secondes. Juste avant la chute, l'IOVC est assez haut, on a une dame avec ses chiens (figure 4.2b), ce qui fait qu'on aura tendance à regarder ces différents visages (principalement la personne, puis les animaux de compagnie). Puis on a une chute de score d'IOVC (figure 4.2c), correspondant à un plan large devant un commerce. Cette fois, on n'arrive pas à remarquer une personne en particulier, et il n'y a pas de point d'intérêt, on ne sait pas où regarder, donc la congruence visuelle inter-observateurs est logiquement beaucoup plus faible et provoque une chute entre ces deux plans.

(a) Vers 20 secondes on constate une chute du score prédit de congruence visuelle inter-observateurs



(b) IOVC assez haut avant la chute

(c) Chute de l'IOVC

FIGURE 4.2 – Exemple de chute du score d'IOVC

### 4.2.3 Reconnaissance et intensité des émotions

Pour inférer la reconnaissance d'émotions et leur intensité, nous avons utilisé le modèle *Facial Expression Re-*

*cognition using Residual Masking Network* [PVT21]. Le réseau et les poids d’entraînement étant disponibles sur le dépôt GitHub du projet nous a permis de nous concentrer sur la création des caractéristiques extraites. Ce modèle prenant en entrée une image, est capable dans un premier temps d’extraire la position des différents visage, puis dans un second temps, d’inférer leur score d’intensité émotionnelle. Comme dans une image nous pouvons avoir N visages, nous avons fait le choix d’obtenir un nombre fixe de caractéristiques. Pour cela, pour chaque image, nous extrayons le nombre de visages détectés, l’intensité maximale sur tous les visages et sur toutes les émotions, et le score d’intensité maximale pour chaque émotion.

#### 4.2.4 Résumé vidéo

AC-SUM-GAN [AAM<sup>+</sup>21a] aurait pu être utilisé, malheureusement aucun fichier du modèle pré-entraîné par les auteurs n’était fourni. Il a donc fallu entraîné nous-même, nous avons pu commencer mais le serveur à notre disposition s’éteignait dès que nous n’étions plus connectés et l’entraînement prenait environ heures. Il a été préférable de trouver un autre modèle où les auteurs fourniraient le modèle déjà entraîné.

PGL-SUM [ABMP21], des mêmes auteurs que AC-SUM-GAN, est un modèle de génération de résumé vidéo, lui aussi entraîné sur TVSum et SumMe et a été utilisé afin d’inférer les frames à sélectionner pour les résumés vidéos de ces deux bases de données.

Les auteurs proposent d’évaluer leur modèle avec le

F-Score, nous l’utilisons pour tester les deux modèles sur TVSum : pour le premier modèle (Table III) les splits donnent respectivement 62.59%, 61.79%, 60.67%, 62.76% et **65.37%**. Pour le second (Table IV) : 60.08% 60.40%, 59.47%, 60.53%, 64.72%.

Sur SumMe, Table III : 52.98%, 58.52%, 58.88%, 55.34% et 59.67% et Table IV : 53.72%, 53.13%, **66.41%**, 58.55% et 46.39%.

On utilise donc le meilleur split afin d’inférer sur les bases entières de vidéos. On obtient un F-Score pour toutes les vidéos de SumMe de 55.04% (avec le 3ème split de la table IV), sur TVSum de 62.11% (avec le 5ème split de la table III).

### 4.3 Trouver des profils-types

Il s’agit de trouver des profils-types de vidéos à partir de l’évolution des scores des caractéristiques inférées au cours des vidéos.

Il faut d’abord pouvoir comparer les vidéos entre elles, deux solutions s’offrent à nous, mettre toutes les vidéos sur un même nombre de frames 4.3.1 ou bien créer des fenêtres vidéos de même taille 4.3.2.

#### 4.3.1 Solution 1

Les vidéos doivent être comparables, dans cette méthode, la durée des vidéos est transposée vers une durée beaucoup plus grande, en mettant toutes les vidéos sur le même nombre de frames (certaines vidéos font 1 minute et d’autres 5 minutes). Pour cela, on peut répéter

N fois chaque frame pour atteindre une cible beaucoup plus grande (par exemple 1 million de frames). Exemple : si la vidéo fait 5 minutes,  $5 * 60 * 24 = 7200$  frames,  $\text{ceil}[1000000/7200] = 139$  , on répète 139 fois le score de chaque frame. Puis, on ne prend pas en compte les frames dépassant 1 million, donc  $7200 * 139 = 1 000 800$  frames, on a  $800/1000800 = 0.08\%$  de perte. Soit  $0.08\% * 7200 = 5.76$  frames non prises en compte à la fin de la vidéo. Appliqué sur les 50 vidéos de TVSum, un clustering de la mémorabilité inférée de chaque vidéo, utilisant l'algorithme des k-moyennes et cherchant à définir 3 clusters, on a les 3 clusters aux évolutions différentes suivants, voir en figure 4.3. Le problème de cette méthode est qu'elle répète le score des frames de nombreuses fois donc les courbes sont très plates pendant longtemps et les augmentations ou baisses sont donc très droites. Bien qu'ayant lissé les courbes pour essayer de pallier ce problème, cette méthode ne semble pas la meilleure. C'est pour ces raisons et grâce à l'apport d'une nouvelle solution présentée dans la partie suivante, que nous n'avons pas continué plus en détails à chercher des profils-types avec cette méthode.

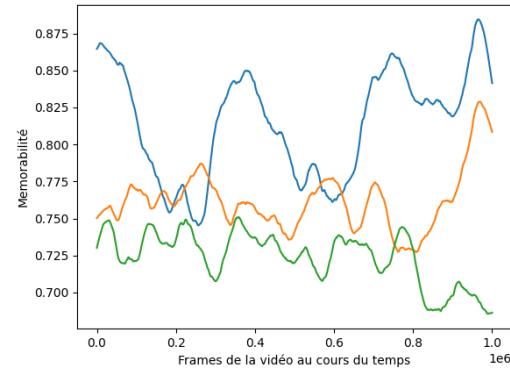
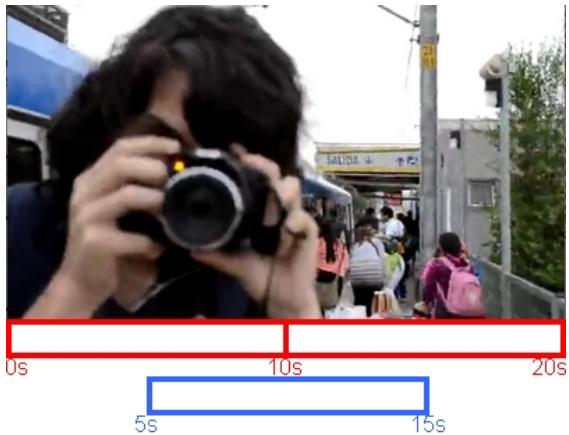


FIGURE 4.3 – 3 clusters types d'évolution de la mémorabilité au cours de vidéo - TVSum

#### 4.3.2 Solution 2

Pour pouvoir comparer les différentes vidéos, la seconde solution est de diviser les vidéos en fenêtres vidéos de même taille 4.4. On peut comparer des fenêtres vidéos en découplant chaque vidéo en période de N secondes. Il est aussi possible avec un décalage entre deux fenêtres très faible de n'avoir aucune perte et donc de prendre en compte toutes les frames.

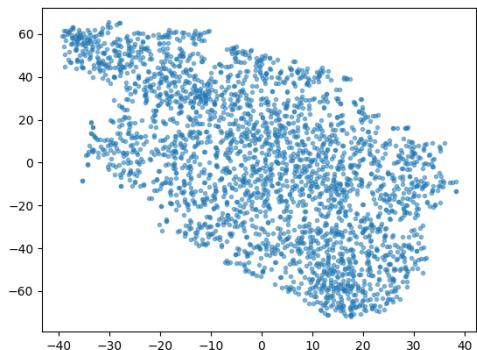


représente une fenêtre de plusieurs secondes (ici 25 secondes).

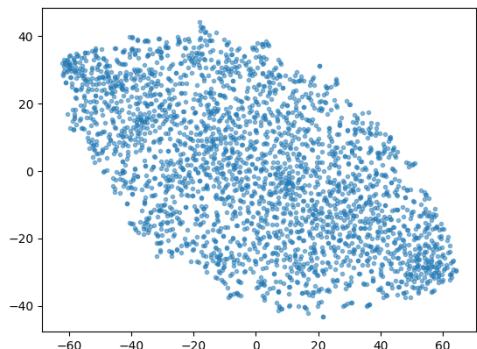
FIGURE 4.4 – Fenêtres (rectangles) de 10 secondes avec un pas de 5 secondes sur une vidéo de 20 secondes

Ensuite pour pouvoir réaliser un clustering, il faut réduire le nombre de dimensions avec des méthodes telles que l'analyse en composantes principales ou T-SNE (*T-distributed Stochastic Neighbor Embedding*). En effet, il y a autant de dimensions que de frames dans une fenêtre vidéo (par exemple une fenêtre de 10 secondes aura  $10 * 24 = 240$  dimensions).

Ces réductions de dimensions sont appliquées aux vidéos de la base TVSum, plus précisément à des fenêtres vidéos de 25 secondes avec un pas de 5 secondes. Que ce soit avec l'ACP ou T-SNE, les données réduites en 2D de l'évolution des caractéristiques d'IOVC et de mémorabilité au cours des fenêtres vidéos ne donnent pas vraiment d'indications sur la présence de clusters clairs. Avec T-SNE, le nuage de points de l'IOVC est visible en figure 4.5a et pour la mémorabilité en figure 4.5b où un point



(a) IOVC



(b) Mémorabilité

FIGURE 4.5 – Réduction à deux dimensions avec T-SNE  
- TVSum

Comme aucun cluster ne semble émerger, l'idée est

désormais d'utiliser les données brutes des fenêtres vidéos et d'utiliser non plus la mesure de distance euclidienne entre les points 2D du nuage mais la distance de déformation temporelle dynamique *Dynamic Time Warping* (exemple en figure 4.6) sur les données brutes car cette distance prend en compte le décalage temporelle entre deux courbes.

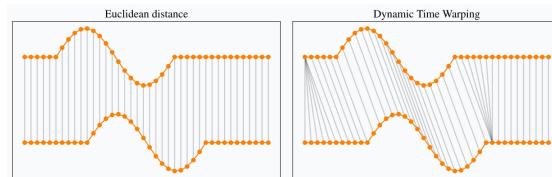


FIGURE 4.6 – Distance Euclidienne et distance de déformation temporelle dynamique

Source : Romain Tavenard <sup>1</sup>

L'algorithme des k-moyennes cherche à minimiser la somme des distances euclidiennes au carré entre chaque point et le centroïde (centre d'un cluster). Une variante, proposée par la bibliothèque TSLearn <sup>2</sup> permet de réaliser un clustering sur des séries temporelles et ainsi d'utiliser une distance différente de la distance Euclidienne, à savoir la distance de déformation temporelle dynamique évoquée ci-dessus. Un clustering est appliqué sur l'évolution de l'IOVC et de la mémorabilité au cours de fenêtres vidéos de la base de vidéos TVSum. Ce clustering utilise les k-moyennes et la distance de déformation temporelle dynamique, les 3 clusters obtenus pour l'IOVC sont très plats, aucune évolution ou baisse. Tandis que les 3 clus-

1. <https://rtavenar.github.io/blog/dtw.html>

ters de la mémorabilité, dans le cluster 1 la mémorabilité augmente au cours de la fenêtre vidéo et dans le cluster 2 la mémorabilité baisse légèrement. Il semble qu'il y a davantage de tendances dans la mémorabilité que pour l'IOVC mais pas assez pour voir un profil-type clair.

Mais le fléau de la dimension (*curse of dimensionality*) fait que plus il y a de dimensions plus il faudrait de données. Ici la dimension de chaque fenêtre est de 600 car de 25 secondes ( $25 * 24$ ). Sans chevauchement, on a 563 fenêtres, et avec un pas de 5 secondes (celui utilisé) on arrive à 2714 fenêtres. Malheureusement c'est encore trop peu de fenêtres par rapport aux nombres de dimensions, il faudrait des millions de fenêtres et donc plusieurs milliers de vidéos plutôt que les 50 vidéos de TVSum.

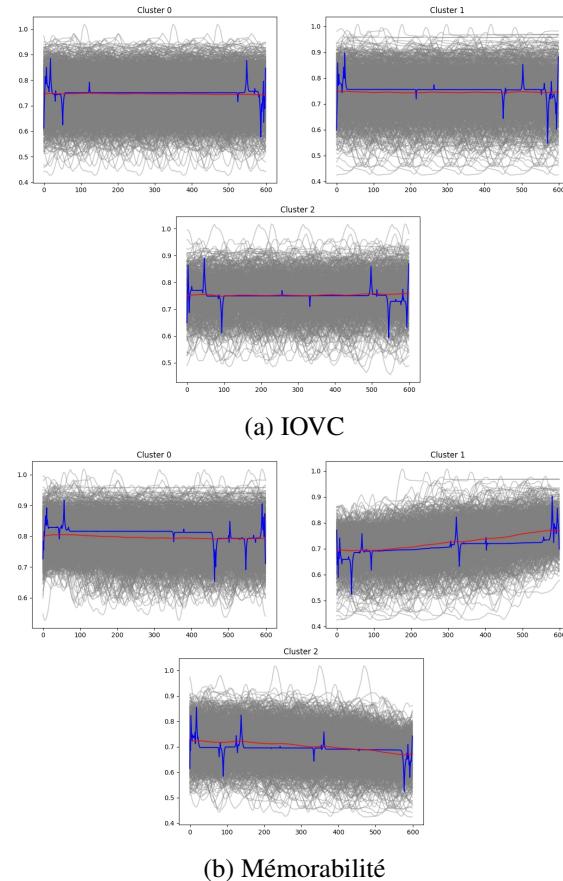


FIGURE 4.7 – Clusters utilisant la distance de déformation temporelle dynamique - TVSum

2. [https://tslearn.readthedocs.io/en/stable/gen\\_modules/clustering/tslearn.clustering.TimeSeriesKMeans.html](https://tslearn.readthedocs.io/en/stable/gen_modules/clustering/tslearn.clustering.TimeSeriesKMeans.html)

## 4.4 Test de corrélation linéaire

Sur la mémorabilité et la congruence visuelle inter-observateurs nous avons un score de ces caractéristiques par frame. Nous pouvons donc tester la corrélation de ces deux caractéristiques avec la vérité-terrain. Nous faisons ce processus par vidéo et nous testons la corrélation sur deux groupes distincts de frames : les frames utilisées dans le résumé vidéo ainsi que les frames non sélectionnées.

### 4.4.1 Corrélation de l'IOVC

Sur la base TVSum (voir figure 4.8) : pour les corrélation de Pearson et Spearman : Il y a une grande variance dans l'IOVC des frames non sélectionnées, la variance reste grande pour les frames sélectionnées mais la moyenne semble avoir augmenter.

Sur la base SumMe : avec Pearson, l'IOVC est concentré autour de zéro (pas de corrélation) sur les frames non sélectionnées. Mais la variance est plus grande pour les frames sélectionnées où 7 vidéos corrèlent inversement (entre -0.3 et -0.5) et 4 vidéos corrèlent entre 0.4 et 0.5. Avec Spearman, idem il y a une concentration autour de 0.15, pas d'extrême corrélation (ni inverse) pour les frames non sélectionnées. Sur les frames sélectionnées, les vidéos sont assez concentrées en zéro (pas de corrélation) mais 4 vidéos corrèlent inversement entre -0.3 et -0.6 et 3 vidéos corrèlent entre 0.4 et 0.6.

La corrélation linéaire entre la congruence visuelle inter-observateurs et la vérité-terrain n'augmente que très légèrement (environ 0.12 sur TVSum et 0.03 sur SumMe)

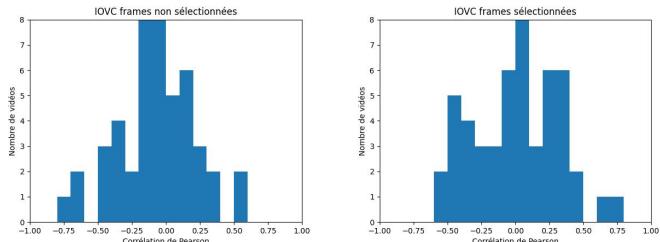
entre la corrélation sur les frames sélectionnées et celles non sélectionnées dans le résumé vidéo (voir tableaux 4.1 et 4.2). Il n'y a donc pas ou presque de corrélation linéaire entre cette caractéristique et les scores d'importance des frames.

Corrélation	Frames sélectionnées	Frames non sélectionnées
Pearson	0.0091	-0.0779
Spearman	0.0072	-0.0756

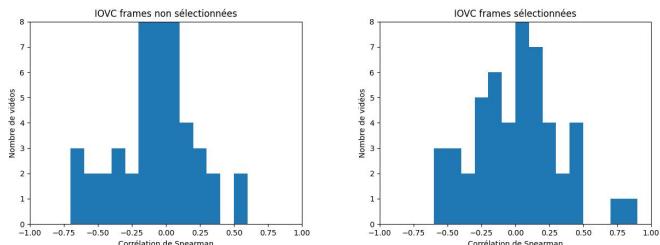
TABLE 4.1 – Corrélation moyenne de l'IOVC sur TV-Sum

Corrélation	Frames sélectionnées	Frames non sélectionnées
Pearson	-0.0004	-0.0011
Spearman	0.0126	-0.0169

TABLE 4.2 – Corrélation moyenne de l'IOVC sur SumMe



(a) Corrélation de Pearson des frames non sélectionnées (b) Corrélation de Pearson des frames sélectionnées



(c) Corrélation de Spearman des frames non sélectionnées (d) Corrélation de Spearman des frames sélectionnées

FIGURE 4.8 – Histogramme de corrélation entre l’IOVC et la vérité-terrain des vidéos de TVSum

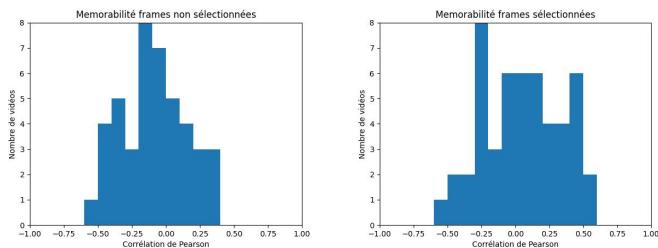
#### 4.4.2 Corrélation de la mémorabilité

Sur la base TVSum 4.9 : avec la corrélation de Pearson, la concentration est autour de -0.15 sur les frames non sélectionnées et il y a une plus grande variance sur les frames sélectionnées avec 8 vidéos qui anti corrèlent (inférieures à -0.2) et 8 qui corrèlent à plus de 0.4. Avec Spearman, la concentration est autour de -0.1 sur les

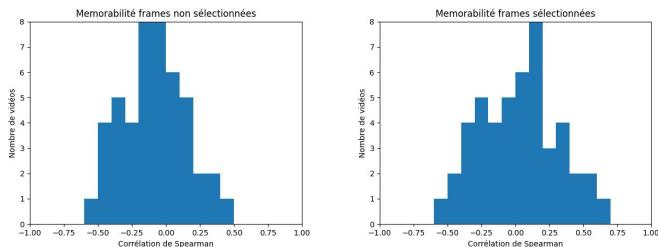
frames non sélectionnées et autour de 0.15 sur les frames sélectionnées.

Sur la base SumMe : avec la corrélation de Pearson, la mémorabilité est très concentrée autour de 0 sur les frames non sélectionnées et a une plus grande variance sur les frames sélectionnées (qui corrèlent ou anti corrèlent un peu plus). Tandis qu’avec Spearman, la mémorabilité est très concentrée autour de 0 sur les frames non sélectionnées. Sur les frames sélectionnées, la distribution a deux pics, l’un concentré sur 0.15 et l’autre sur -0.4, 5 vidéos semblent anti-corréler (inférieurs à -0.3) et 6 autres corrèlent faiblement (supérieurs à 0.2).

La corrélation linéaire entre la mémorabilité et la vérité-terrain n’augmente que très légèrement (0.14 environ sur TVSum et 0.03 sur SumMe) entre la corrélation sur les frames sélectionnées et celles non sélectionnées dans le résumé vidéo (voir tableaux 4.3 et 4.4). Il n’y a donc pas ou presque de corrélation linéaire entre la mémorabilité et les scores d’importance des frames.



(a) Corrélation de Pearson des frames non sélectionnées (b) Corrélation de Pearson des frames sélectionnées



(c) Corrélation de Spearman des frames non sélectionnées (d) Corrélation de Spearman des frames sélectionnées

FIGURE 4.9 – Histogramme de corrélation entre la mémorabilité et la vérité-terrain des vidéos de TVSum

Corrélation	Frames sélectionnées	Frames non sélectionnées
Pearson	0.0556	-0.0972
Spearman	0.0334	-0.0908

TABLE 4.3 – Corrélation moyenne de la mémorabilité sur TVSum

Corrélation	Frames sélectionnées	Frames non sélectionnées
Pearson	-0.0154	-0.0343
Spearman	-0.006	-0.0538

TABLE 4.4 – Corrélation moyenne de la mémorabilité sur SumMe

## 4.5 Test de Student

Le test de Student permet de tester l’indépendance entre 2 distributions. Pour chaque caractéristique, les deux distributions d’intérêt sont la distribution du score de la caractéristique sur les frames sélectionnées pour le résumé vidéo ainsi que sur la distribution du score sur les frames non gardées dans le résumé vidéo.

TVSum et SumMe proposent des vidéos dont chaque frame est annotée et représente le score moyen des importances données par des utilisateurs. Nous allons garder au moins 10% des frames des plus importantes pour le résumé vidéo. Pour sélectionner les frames, on garde celles ayant une valeur d’importance supérieure la valeur au 10ème percentile, il est donc possible d’avoir plus de 10% de frames dans le résumé vidéo.

Le test de Student est réalisé grâce à une fonction de la bibliothèque SciPy<sup>3</sup> et la p-valeur rentrée nous permet de constater si les deux échantillons sont significativement différents ou non.

Ce test d’indépendance est donc réalisé sur chacune des caractéristiques : l’intensité émotionnelle maximale, le nombre de visages reconnus, les 6 émotions

3. <https://docs.scipy.org/doc/scipy/tutorial/general.html>

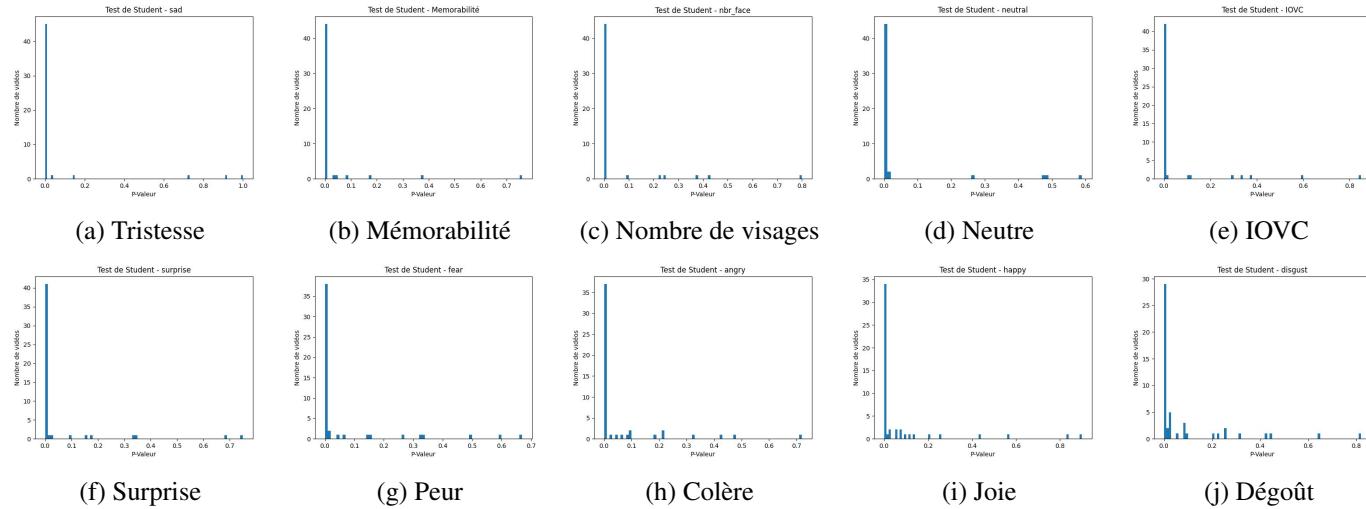


FIGURE 4.10 – Test de Student entre les frames sélectionnées et non sélectionnées de chaque vidéo de TVSum

(neutre, joie, colère, tristesse, peur, dégoût et surprise), la congruence visuelle inter-observateurs et la mémorabilité. Le résultat est présent en figure 4.10 où sont présents pour chaque caractéristique un histogramme de répartition des p-valeurs par vidéo.

Si l'on prend un seuil  $\alpha = 0.01$ , alors pour chaque caractéristique, on compte le nombre de vidéos ayant une  $p - valeur \leq 0.01$ . Les caractéristiques dont les distributions sont les plus significativement indépendantes entre les scores sur les frames sélectionnées pour le résumé vidéo et celles non sélectionnées sont (voir histogrammes 4.10) dans l'ordre : Tristesse > Mémorabilité = Nombre de visages = Neutre > IOVC > Surprise > Peur > Colère > Joie > Dégout.

Cela ajoute une preuve de l'utilité de chacune des caractéristiques inférées dans le but de prédire un score d'importance des frames. En effet, pour une grande majorité des vidéos, la distribution des scores des caractéristiques est significativement différente sur les frames sélectionnées ou non dans le résumé vidéo.

Si on regarde les vidéos dont la p-valeur est supérieure à 0.1, pour chacune des 3 caractéristiques principales (mémorabilité, IOVC et probabilité émotionnelle), nous ne retrouvons pas les mêmes vidéos. Ces caractéristiques semblent donc complémentaires.

Leur complémentarité et leur utilité est une indication très utile sur le potentiel de prédiction de nos caractéristiques, dans le but de développer un modèle de prédiction

du score d'importance de chaque frame.

## 4.6 Test de similarité

Pour le test de similarité, nous avons utilisé la distance de Wasserstein [BHK22]. Cette distance permet de calculer la similarité entre deux distributions. Nous avons pour cela utilisé les données d'IOVC et de mémorabilité inférées sur le jeu de donnée TVSum. On peut voir en figure 4.11 la matrice de Wasserstein des distances de distribution de l'IOVC et en figure 4.12 la matrice de Wasserstein de distribution des distances de mémorabilité.

Ces deux matrices utilisent le code couleur JET, quand la distribution des deux vidéos est plus éloignée, alors la couleur sera plus tournée vers le rouge, et quand deux vidéos se ressemblent, la couleur de la case sera plus tournée vers le bleu.

Dans l'ensemble, les distributions de l'IOVC et de la mémorabilité semblent similaires puisque le fond bleu est clairement visible. Seules quelques vidéos se démarquent et sont plus différentes des autres.

En analysant la matrice de Wasserstein de l'IOVC, nous regardons deux vidéos : *3eYKfiOEJNs* en comparaison de la vidéo *vdmoEJ5YbrQ*. La première est une vidéo de reportage sur le toilettage canin filmée avec une caméra posée sur trépieds avec des mouvements de caméra lents, avec beaucoup d'animaux et au minimum une personne présente sur pratiquement la totalité de la vidéo alors que la deuxième vidéo est une vidéo filmée avec un téléphone avec beaucoup de mouvement de caméra très rapide ou peu de personnes apparaissent à l'image. On

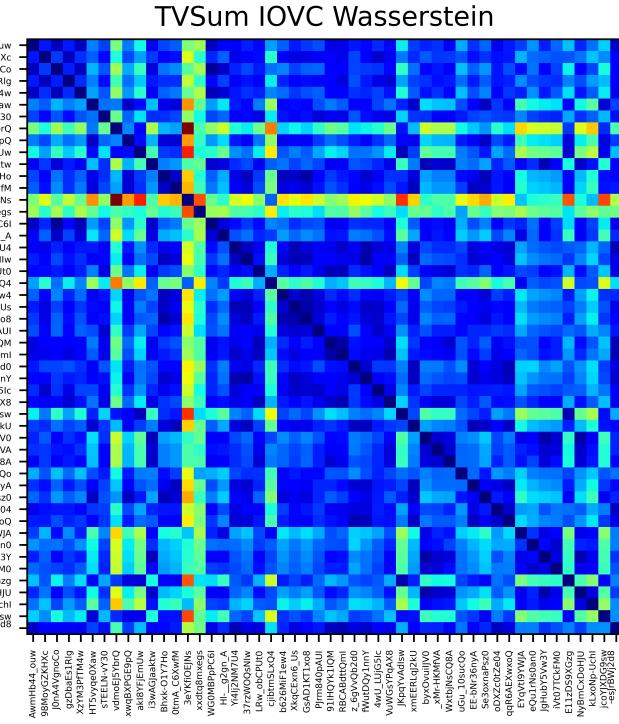


FIGURE 4.11 – Matrice de Wasserstein sur l'IOVC

TVSum Memorability Wasserstein

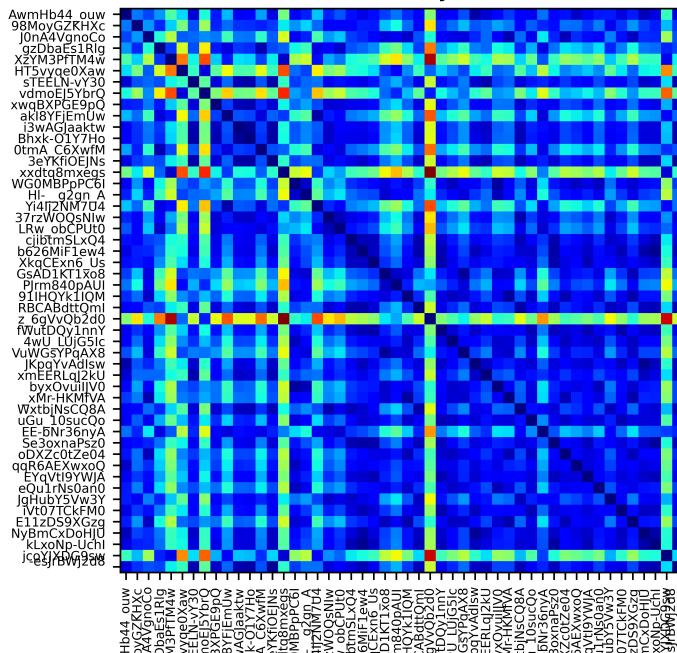


FIGURE 4.12 – Matrice de Wasserstein sur la mémorabilité

peut supposer que le regard aura plus tendance à être similaire lorsque la vidéo est stable et bien cadré comparé à une vidéo avec beaucoup de mouvements de caméra.

De la même manière avec la matrice de Wasserstein de la mémorabilité, nous regardons deux vidéos : *z\_6gVvQb2d0* en comparaison de la vidéo *3eYKfiOEJNs*. La première est une vidéo filmée sur téléphone, avec beaucoup de mouvements de caméra, d'un défilé de

joueurs de sport encadré par la police et des services de secours. La deuxième vidéo est la même que dans le cas de l'IOVC. Nous supposons dans ce cas, que le reportage canin ayant des images plus stables sur de longues périodes avec un sujet principal bien identifié augmente la mémorabilité.

Ces deux matrices de Wasserstein montrent bien que pour ces deux caractéristiques inférées ces vidéos sont relativement similaires. Seulement quelques vidéos se démarquent, mais on peut en retenir que pour permettre à un modèle de généraliser sur l'apprentissage de cette base à partir de ces deux caractéristiques cela va être compliqué car les vidéos ne sont pas suffisamment diversifiées.

## 4.7 Modèle de régression des scores

Dans cette partie sont détaillées les modèles de réseaux de neurone qui ont été développés afin de faire de la régression sur les scores d'importance des images des vidéos. Nous avons pour cela inféré ces scores en fournissant aux modèles des séquences de caractéristiques. Nous utilisons les 11 caractéristiques définies plus haut. L'usage de séquence de caractéristiques d'image a du sens sur notre domaine d'application car cela permet de donner du contexte environnant les caractéristiques de l'image dont nous souhaitons prédire le score d'importance.

Nous avons testé différentes tailles de séquences pour chaque modèle. Nous avons fixé aléatoirement 10 vidéos (20% du jeu de données) qui constitueront notre jeu de validation à travers l'entraînement des différents modèles.

### 4.7.1 Réseau de neurones

Nous avons tenté d'inférer les scores d'importance des images de la vidéo via un perceptron multi-couche. Pour cela, nous avons créé une architecture de trois couches entièrement connectées de taille 512 puis 256 puis 64 neurones, et enfin une neurone pour prédire le score d'importance de l'image. Une représentation du réseau de neurones est disponible en figure 4.13. Nous appliquons une fonction d'activation ReLU [Aga18] sur la sortie de chaque couche cachée.

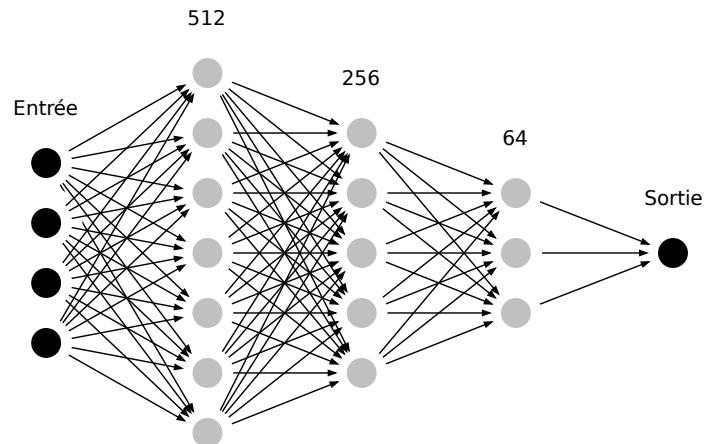


FIGURE 4.13 – Architecture du réseau de neurones

En entrée du réseau, nous envoyons des séquences de caractéristiques de plusieurs images à la suite. Nous avons 11 caractéristiques par image. Nous avons testé différentes tailles de séquences. En fonction de la taille de la séquence, la taille d'entrée du réseau est donc différente. Pour connaître la taille d'entrée du réseau, il suffit de calculer  $11 \times S$ .  $S$  étant la taille de la séquence.

Lors de l'entraînement, nous utilisons pour l'optimisation : Adam [KB14] avec un taux d'apprentissage de  $1 \times 10^{-5}$ . Comme nous cherchons à faire la régression du score d'importance des images, nous utilisons comme fonction de perte : l'erreur quadratique moyenne (ou Mean Squared Error en anglais) [SW10]. Nous avons fait le choix d'utiliser la métrique  $R^2$  (coefficient de détermination) pour mesurer la précision des modèles. La valeur de cette métrique évolue entre -moins l'infini et 1. Entre

0 et 1, cette métrique indique que deux variables (prédition et réel) sont liées.

On peut voir en figure 4.14a la fonction de perte sur le jeu d'entraînement des différents modèles entraînés en fonction de la taille de séquence. Les modèles convergent très rapidement vers une valeur de perte faible, et obtiennent très rapidement des scores  $R^2$  assez important. On peut aussi noter que le score de  $R^2$  augmente légèrement plus pour les séquences de tailles plus importantes. Cependant, nous pouvons voir en figure 4.14b que sur le jeu de validation, les résultats sont différents. La fonction de perte progresse lentement vers des valeurs faibles mais qui reste assez importantes pour une fonction de perte d'erreur quadratique moyenne. On peut voir qu'au fur et à mesure de l'entraînement, le score  $R^2$  des différents modèles se dégrade par rapport aux métriques  $R^2$  du jeu d'entraînement, cela est dû à un sur-apprentissage des données d'entraînement car on peut voir pour les mêmes époques, que la fonction de perte du jeu de validation augmente aussi. La taille des séquences ne change pas le score  $R^2$  qui ne dépasse pas les 0.135 (contre 0.379 sur le jeu d'entraînement). Les scores  $R^2$  maximum atteint par chaque modèle avec la taille des séquences est disponible en table 4.5.

Nous avons aussi essayé de modifier l'architecture du réseau en augmentant la couche de taille 512 en taille 1024 pour rendre le réseau plus profond mais cela a eu pour conséquence de réduire la précision du modèle. De la même façon, nous avons aussi essayé de désactiver temporairement certains neurones du modèle pour réduire le sur-entraînement du réseau grâce à une couche

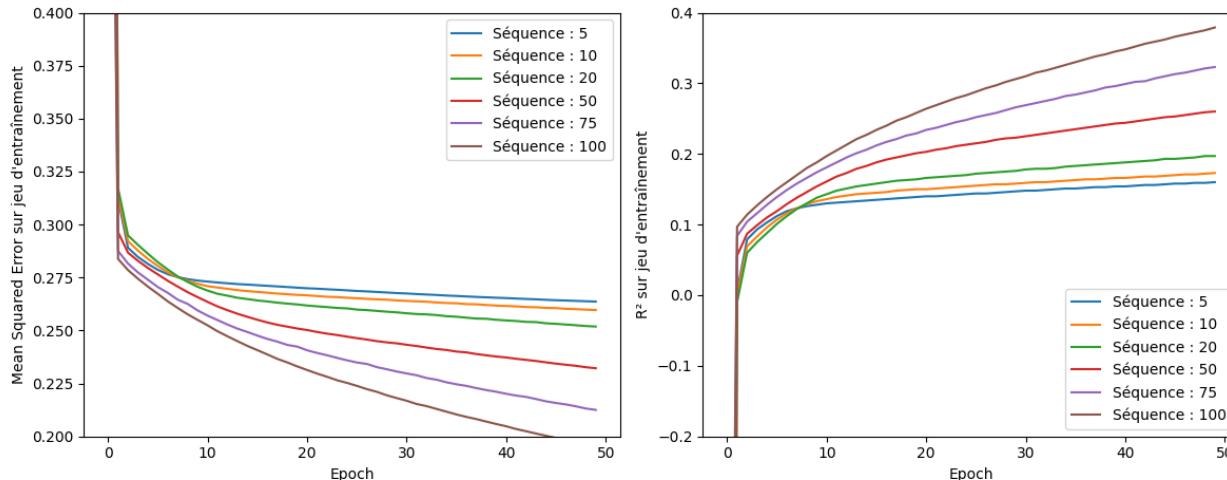
Dropout [SHK<sup>+</sup>14] configurée à 15% (placée entre la couche de taille 512 et la couche de taille 256) mais cette expérience a aussi eu pour conséquence de réduire légèrement la précision des modèles.

Taille de la séquence	Score $R^2$
Séquence = 5	0.135
Séquence = 10	0.132
Séquence = 20	0.132
Séquence = 50	0.131
Séquence = 75	0.130
Séquence = 100	0.114

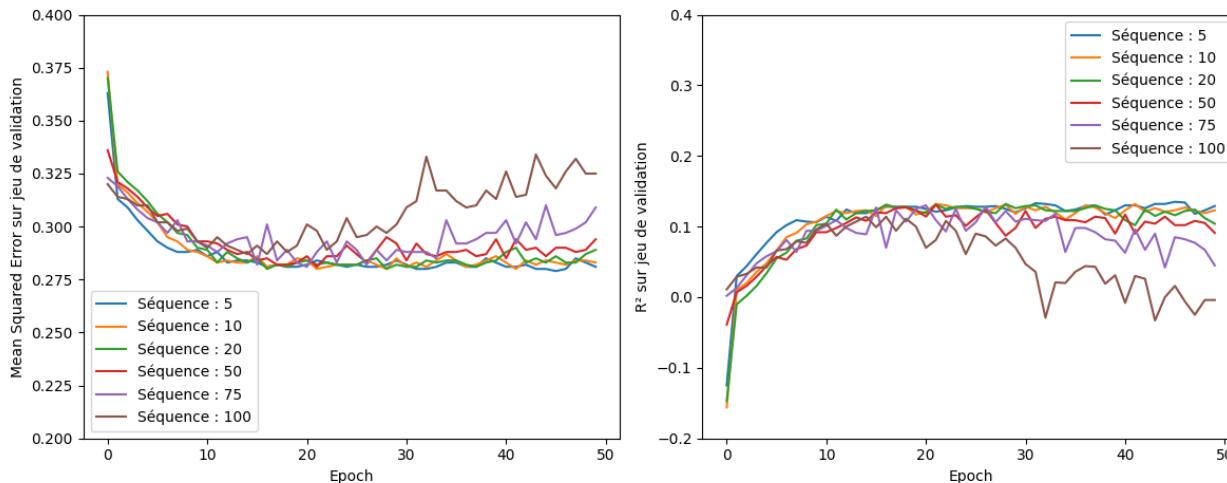
TABLE 4.5 – Score  $R^2$  du réseau de neurones sur le jeu de validation

## 4.7.2 Réseau de neurones avec couche LSTM

Nous avons testé une deuxième architecture composée d'une couche LSTM [HS97] de taille 512 en entrée du réseau. Cette couche est faite pour retenir l'information dans les séquences de caractéristique et pourrait donc permettre à notre réseau d'obtenir de meilleurs résultats. Le réseau, visible en figure 4.15, est composé par la suite d'une couche entièrement connectée de taille 256, puis 64, puis une dernière neurone pour prédire le score d'importance de l'image. Entre chaque couche du réseau, une fonction d'activation ReLU est appliquée sur le vecteur de sortie.



(a) Résultat sur jeu d'entraînement



(b) Résultat sur jeu de validation

FIGURE 4.14 – Résultat d'entraînement du réseau de neurone sur la base de donnée TVSum

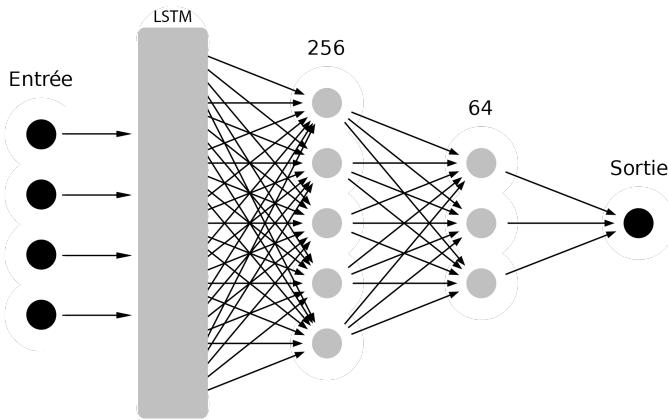


FIGURE 4.15 – Architecture du réseau de neurones avec couche LSTM

Contrairement au modèle sans couche LSTM, le modèle ne sur-entraîne pas les données d'entraînement au bout 17 époques, comme on peut le voir en figure 4.16b la fonction de perte sur le jeu de validation reste stable jusque vers 35 époques. Cela a pour conséquence que le modèle a de moins bons résultats sur le jeu d'entraînement visible en figure 4.16a, mais obtient des résultats sensiblement plus précis sur le jeu de validation. Comme on peut le voir en table 4.6 les résultats sont légèrement meilleurs que sans la couche LSTM.

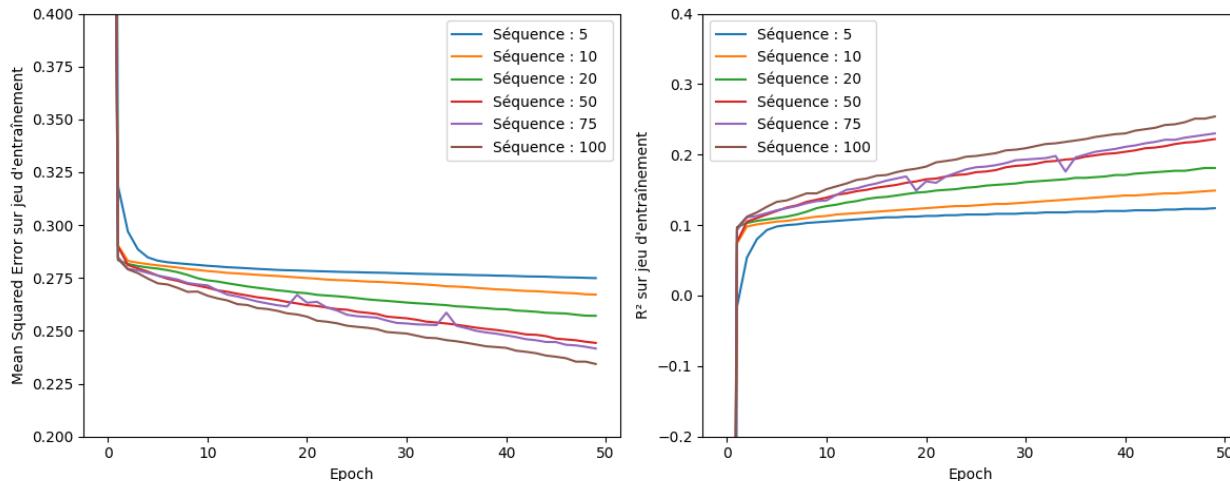
De la même manière qu'avec le réseau précédent 4.7.1, nous avons tenté d'améliorer les performances du modèle en augmentant la taille de la couche du LSTM en l'augmentant à 1024 mais ceci n'augmenta pas la précision du modèle, et de la même manière, nous avons ajouté une

couche de Dropout mais cela n'a pas aidé le modèle à améliorer la précision.

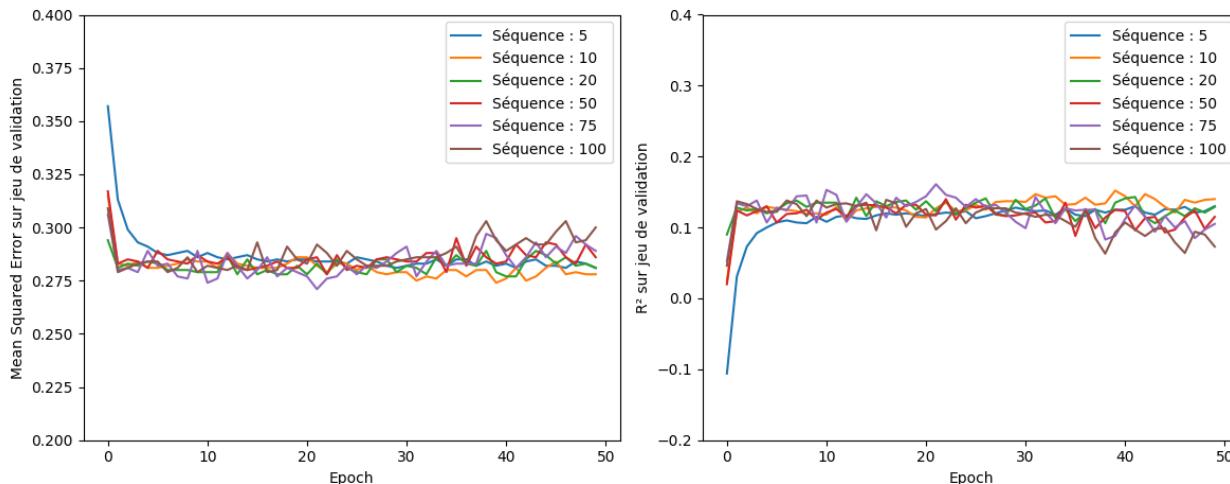
On peut en conclure que la couche LSTM ajoute de l'intérêt dans notre domaine d'application composé de séquence de caractéristiques mais que nos caractéristiques ne sont pas suffisantes pour permettre une inférence de l'importance des images tout en obtenant une précision intéressante. On peut cependant noter que même si les performances ne sont pas au rendez-vous, la faible valeur de la mesure  $R^2$  indique une faible corrélation entre nos caractéristiques et le score d'importance.

Taille de la séquence	Score $R^2$
Séquence = 5	0.130
Séquence = 10	0.152
Séquence = 20	0.143
Séquence = 50	0.140
Séquence = 75	0.161
Séquence = 100	0.139

TABLE 4.6 – Score  $R^2$  du réseau de neurones avec LSTM sur le jeu de validation



(a) Résultat sur jeu d'entraînement



(b) Résultat sur jeu de validation

FIGURE 4.16 – Résultat d'entraînement du réseau de neurones avec LSTM sur la base de données TVSum

## 4.8 Modèle de classification binaire

Dans cette partie sont détaillés les modèles de classification des scores à partir des caractéristiques. Les modèles cherchent à prédire si la frame est sélectionnée ou non dans le résumé vidéo.

### 4.8.1 Arbre de décision

L’arbre de décision est un algorithme d’apprentissage automatique bien connu et généralement performant pour faire de la classification. De plus, l’avantage est qu’il soit à boîte blanche, c’est-à-dire qu’il est possible de savoir comment la classification est réalisée. Il est possible de savoir à chaque noeud de l’arbre quelle condition a été mise sur quelle variable pour séparer les données en deux noeuds. Enfin, la classification est faite sur les feuilles de l’arbre qui indiquent la classe attribuée.

Dans notre cas, la classification permet de donner directement les frames à utiliser pour le résumé. L’arbre de décision *DecisionTreeClassifier* de la bibliothèque scikit-learn<sup>4</sup> a été utilisé pour développer notre modèle.

Les deux classes sont très déséquilibrées, en effet dans notre jeu d’entraînement, composé de 40 vidéos de TV-Sum, nous avons 290 964 frames au total dont 32 099 (soit 11%) seulement sont sélectionnées dans les résumés vidéo (classe 1) et 258 865 frames (soit 89%) ne sont pas sélectionnées (classe 0). Notre jeu de test est composé de 10 vidéos, plus précisément de 61 389 frames dont 54 417 non sélectionnées (soit 89%) et 6 972 sélectionnées (soit

11%).

Le jeu d’entraînement est donc équilibré, pour cela, les features de classe 0 sont mélangées aléatoirement en  $\text{floor}(258865/32099) = 8$  splits composés de 32 099 features de classe 0 auquel on ajoute les 32 099 features de classe 1. Les matrices de confusion sont visibles en figure 4.18. On peut voir que le taux de faux positif est énorme, puisque le jeu de test est lui aussi déséquilibré car on teste sur des vidéos entières dont seulement 11% des frames sont sélectionnées. Les courbes ROC (*receiver operating characteristic*) sont visibles en figure 4.17, le modèle obtient une précision (moyenne des aires en dessous la courbe ROC ou AUC) de 0.625, ainsi qu’une robustesse (écart-type des AUC) de 0.0036.

Si on teste notre premier split équilibré de données d’entraînement sur notre jeu de test équilibré (divisé en  $\text{floor}(54417/6972) = 7$  splits de 6 972 features de classe 0 et 6 972 de classe 1), on a un taux de faux positif réduit et on arrive beaucoup mieux à prédire les frames sélectionnées. En effet, le taux de vrai positif est plus élevé, voir matrices de confusion en figure 4.20. Le modèle a une précision de 0.626 (moyenne des AUC) et une robustesse de 0.0038 (écart-type des AUC), voir courbes ROC en figure 4.19.

4. <https://scikit-learn.org>

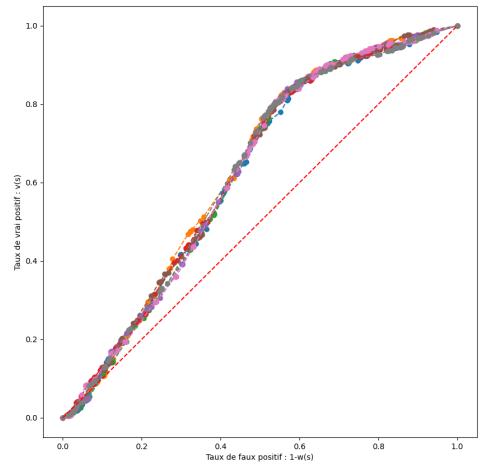


FIGURE 4.17 – Courbes ROC des 8 splits du jeu d’entraînement, la diagonale représente une classification aléatoire

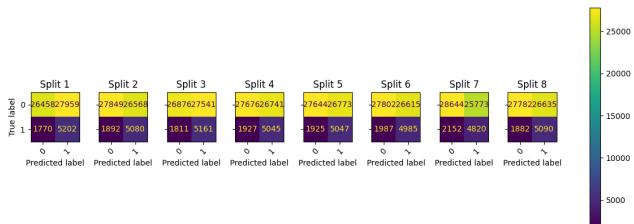


FIGURE 4.18 – Matrices de confusion des 8 splits du jeu d’entraînement sur un jeu de test déséquilibré

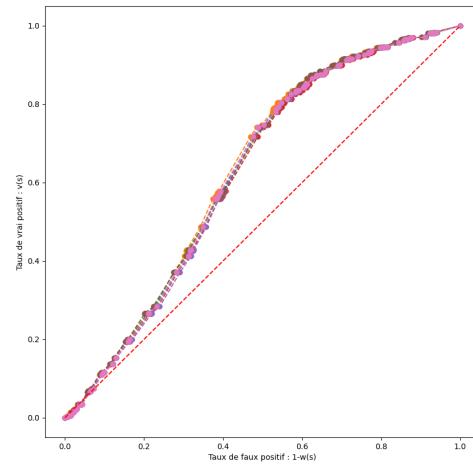


FIGURE 4.19 – Courbes ROC des 7 splits du jeu de test équilibré, la diagonale représente une classification aléatoire

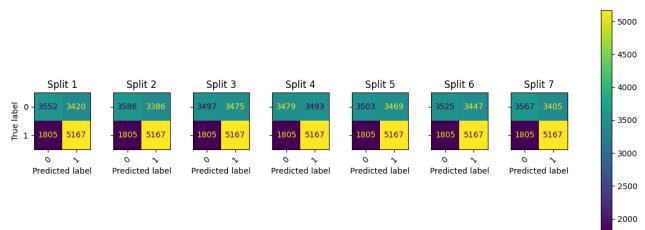


FIGURE 4.20 – Matrices de confusion du 1er split d’entraînement sur les 7 splits du jeu de test équilibré

## 4.9 Conclusion

À travers l'analyse des bases, nous avons pu constater l'utilité des différentes caractéristiques (mémorabilité, congruence visuelle inter-observateurs et reconnaissance d'émotions) en termes de prédition du score d'importance de chaque frame, utilisé pour générer un résumé vidéo. Enfin, nous avons développé plusieurs modèles d'apprentissage automatique et d'apprentissage profond dans le but de faire la régression du score d'importance et de classifier les frames sélectionnées ou non dans le résumé vidéo.

Pour le modèle de classification, de par le déséquilibre des classes entre le nombre de frames sélectionnées ou non dans le résumé, il semble difficile de classifier ces frames sur des données réelles de vidéos (car non équilibrées).

Pour les modèles de régression, nous avons pu montrer que l'ajout d'une couche LSTM permet d'améliorer sensiblement les résultats, mais ce n'est pas encore suffisant pour obtenir un résultat intéressant. La métrique  $R^2$  montre qu'il existe un lien entre la vérité terrain et nos caractéristiques mais ce lien est assez faible.

Dans les deux cas, les 11 caractéristiques utilisées ne sont pas suffisantes pour permettre d'inférer l'importance de chaque image ou de les classifier pour choisir si elles feront partie du résumé vidéo ou non même si une faible corrélation entre la vérités terrains et ces caractéristiques existe.



---

# Conclusion

Pour conclure notre travail, nous résumons les principales étapes de ce dernier. Puis, nous discutons des enseignements que nous a apportés ce projet. Enfin nous proposons quelques pistes pour de futurs travaux de recherche et développement.

## 5.1 Résumé du travail effectué

Dans un premier temps nous avons effectué un état de l'art sur les différentes méthodes de résumé vidéo existantes et avons étudié différentes caractéristiques perceptuelles. Ces caractéristiques perceptuelles étaient : la memorabilité, la congruence visuelle inter-observateurs et la reconnaissance d'émotions. Ces caractéristiques nous avaient parues intéressantes à prendre en compte pour produire de meilleurs résumés vidéo.

Pour faire cela, nous avons dans un premier temps analysé les bases de données de résumé vidéo à notre disposition afin de chercher à savoir si ces caractéristiques avaient une influence sur le score d'importance attribué

par des utilisateurs sur chaque frame des vidéos (score utilisé pour générer un résumé vidéo).

Enfin nous avons testé différents modèles d'apprentissage automatique selon deux approches : la régression du score d'importance et la classification des frames sélectionnées dans le résumé vidéo. Ces modèles ne semblent cependant pas suffisamment précis pour permettre la génération d'un résumé vidéo basé sur ces 11 caractéristiques.

## 5.2 Enseignements

Ce travail nous a permis d'avoir une idée plus précise du monde de la recherche. En effet, pour réaliser l'état de l'art nous avons dû chercher et croiser les sources, faire des fiches de lecture d'articles, vérifier les informations et enfin, critiquer les travaux de recherche. Ce travail fut enrichissant grâce au suivi de M. Bruckert qui a pu nous accompagner et nous conseiller tout au long du projet.

### 5.3 Perspectives de recherche

En perspective, d'autres modèles plus évoluées d'apprentissage profond pourraient être utilisés. Notamment avec l'usage de méthodes plus récentes comme des modules d'attention qui permettraient de mieux constituer une représentation du contexte et donc potentiellement d'obtenir un meilleur résultat.

Il pourrait aussi être intéressant d'élargir le travail à d'autres caractéristiques perceptuelles que les trois étudiées dans ce travail, notamment à travers l'usage d'autres caractéristiques extraites de l'image, ou de l'audio.

Il serait bénéfique de réaliser le travail d'analyse sur d'autres bases de données de résumé vidéo telles que *VISIOCITY* qui pourrait être utilisée. A la fois en termes d'analyse des bases pour pouvoir les comparer, mais aussi en termes d'apprentissage du modèle afin d'avoir des bases de vidéos diversifiées.

# Bibliographie

- [AAM<sup>+</sup>21a] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, and Ioannis Patras. Ac-sumgan : Connecting actor-critic and generative adversarial networks for unsupervised video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8) :3278–3292, 2021. [11](#), [13](#), [23](#), [28](#)
- [AAM<sup>+</sup>21b] Evlampios E. Apostolidis, Eleni Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, and Ioannis Patras. Video summarization using deep neural networks : A survey. *CoRR*, abs/2101.06072, 2021. [5](#), [7](#), [8](#), [9](#), [10](#), [11](#), [12](#), [56](#), [59](#), [88](#)
- [ABMP21] Evlampios Apostolidis, Georgios Ballouras, Vasileios Mezaris, and Ioannis Patras. Combining global and local attention with positional encoding for video summarization. In *2021 IEEE International Symposium on Multimedia (ISM)*, pages 226–234, December 2021. [28](#)
- [Aga18] Abien Fred Agarap. Deep learning using rectified linear units (relu), 2018. [39](#)
- [BCPDSLC16] Yoann Baveye, Romain Cohendet, Matthieu Perreira Da Silva, and Patrick Le Callet. Deep learning for image memorability prediction : The emotional bias. In *Proceedings of the 24th ACM International Conference on Multimedia, MM ’16*, page 491–495, New York, NY, USA, 2016. Association for Computing Machinery. [19](#)
- [BHK22] Fynn Bachmann, Philipp Hennig, and Dmitry Kobak. Wasserstein t-sne, 2022. [37](#)
- [BI15] Ali Borji and Laurent Itti. Cat2000 : A large scale fixation dataset for boosting saliency research, 2015. [65](#)
- [BLCLM19] Alexandre Bruckert, Yat Hong Lam, Marc Christie, and Olivier Le Meur. Deep learning for inter-observer congruency prediction. In *ICIP 2019 - IEEE International Conference on Image Processing*, pages 3766–3770, Taipei, Taiwan, September 2019. IEEE. [5](#), [8](#), [14](#), [15](#), [16](#), [22](#), [26](#), [27](#), [64](#), [89](#)
- [BMS11] Tanja Bänziger, Marcello Mortillaro, and Klaus Scherer. Introducing the geneva multimodal expression corpus for

- experimental research on emotion perception. *Emotion (Washington, D.C.)*, 12:1161–79, 11 2011. 21, 80
- [BQSM16] C. Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M. Martinez. Emotionet : An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5562–5570, 2016. 76
- [BZCFZ16] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ACM International Conference on Multimodal Interaction (ICMI)*, 2016. 76, 78
- [CDDE19] Romain Cohendet, Claire-Helene Demarty, Ngoc Q. K. Duong, and Martin Engilberge. Videomem : Constructing, analyzing, predicting short-term and long-term video memorability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [DHKC<sup>+</sup>20] Alba García Seco De Herrera, Rukiye Savran Kiziltepe, Jon Chamberlain, Mihai Gabriel Constantin, Claire-Hélène Demarty, Faiyaz Doctor, Bogdan Ionescu, and Alan F. Smeaton. Overview of mediaeval 2020 predicting media memorability task : What makes a video memorable ? 2020. 89
- [FSA<sup>+</sup>18] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Summarizing videos with attention, 2018. 10, 12, 23
- [FTC19] Tsu-Jui Fu, Shao-Heng Tai, and Hwann-Tzong Chen. Attentive and adversarial learning for video summarization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1579–1587, 2019. 11
- [GW19] Lore Goetschalckx and Johan Wageman. Memcat : a new category-based image set quantified on memorability. *PeerJ*, 7:e8169, 12 2019. 20, 73
- [HL08] Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, MIR '08*, page 39–43, New York, NY, USA, 2008. Association for Computing Machinery. 73
- [HM17] Behzad Hasani and Mohammad H. Mahoor. Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields. In *2017 12th IEEE International Conference on Computer Vision (ICCV)*, October 2017. 76

- rence on Automatic Face Gesture Recognition (FG 2017), pages 790–795, 2017. 6, 21, 80, 89
- [HS97] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9 :1735–80, 12 1997. 40
- [HSK<sup>+</sup>22] Sunil S Harakannanavar, Shaik Roshan Sameer, Vikash Kumar, Sunil Kumar Behera, Adithya V Amberkar, and Veena I. Puranikmath. Robust video summarization algorithm using supervised machine learning. *Global Transitions Proceedings*, 3(1) :131–135, 2022. International Conference on Intelligent Engineering Approach(ICIEA-2022). 8, 9
- [HZRS15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 20, 73
- [IMS<sup>+</sup>16] Eddy Ilg, Nikolaus Mayer, Tonmoy Sainia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0 : Evolution of optical flow estimation with deep networks. *CoRR*, abs/1612.01925, 2016. 16
- [IXTO11] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 145–152, 2011. 17, 18, 19, 67, 68, 71
- [JDT12] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. 2012.
- [KB14] Diederik P. Kingma and Jimmy Ba. Adam : A method for stochastic optimization, 2014. 39
- [KRT015] Aditya Khosla, Akhil S. Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2390–2398, 2015. 6, 17, 18, 19, 20, 22, 56, 69, 72, 73, 89
- [KSH17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6) :84–90, may 2017. 20, 73
- [KYP13] Jongpil Kim, Sejong Yoon, and Vladimir Pavlovic. Relative spatial features for image memorability. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM ’13, page 761–764, New York, NY, USA, 2013. Association for Computing Machinery. 5, 18, 68, 89
- [LCK<sup>+</sup>10] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar,

- and Iain Matthews. The extended cohn-kanade dataset (ck+) : A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101, 2010. [21](#), [80](#)
- [LDD17] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593. IEEE, 2017. [78](#)
- [LKA<sup>+</sup>17] Yujie Li, Atsunori Kanemura, Hideki Asoh, Taiki Miyanishi, and Motoaki Kawanabe. Extracting key frames from first-person videos in the common space of multiple sensors. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3993–3997, 2017. [12](#)
- [LLS<sup>+</sup>15] Mengyi Liu, Shaoxin Li, Shiguang Shan, Ruiping Wang, and Xilin Chen. Deeply learning deformable facial action parts model for dynamic expression analysis. In Daniel Cremers, Ian Reid, Hideo Saito, and Ming-Hsuan Yang, editors, *Computer Vision – ACCV 2014*, pages 143–157, Cham, 2015. Springer International Publishing. [81](#)
- [LMBR11] Olivier Le Meur, Thierry Baccino, and Aline Roumy. Prediction of the Inter-Observer Visual Congruency (IOVC) and Application to Image Ranking. In *ACM Multimedia*, Phoenix, United States, November 2011. [5](#), [13](#), [14](#), [15](#), [16](#), [56](#), [62](#), [89](#)
- [MCM16] Ali Mollahosseini, David Chan, and Mohammad H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, 2016. [82](#)
- [MH10] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM ’10, page 83–92, New York, NY, USA, 2010. Association for Computing Machinery. [73](#)
- [MHM17] Ali Mollahosseini, Behzad Hassani, and Mohammad H. Mahoor. Affectnet : A database for facial expression, valence, and arousal computing in the wild. *CoRR*, abs/1708.03985, 2017. [78](#)
- [MLM13] Matei Mancas and Olivier Le Meur. Memorability of natural scenes : the role of attention. In *ICIP*, Sydney, Australia,

- September 2013. [5](#), [8](#), [14](#), [16](#), [18](#), [56](#), [67](#)
- [MMP12] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava : A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415, 2012. [73](#)
- [MPWQ19] Debin Meng, Xiaojiang Peng, Kai Wang, and Yu Qiao. Frame attention networks for facial expression recognition in videos. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3866–3870, 2019. [6](#), [22](#), [76](#), [89](#)
- [MSJ19] Dhwani Mehta, Mohammad Faridul Haque Siddiqui, and Ahmad Y. Javid. Recognition of emotion intensities using machine learning algorithms : A comparative study. *Sensors*, 19(8), 2019. [6](#), [8](#), [21](#), [79](#)
- [NB21] Coen D. Needell and Wilma A. Bainbridge. Embracing new techniques in deep learning for estimating image memorability. *CoRR*, abs/2105.10598, 2021. [6](#), [20](#), [22](#), [26](#), [73](#)
- [PDW<sup>+</sup>17] Rameswar Panda, Abir Das, Ziyan Wu, Jan Ernst, and Amit K. Roy-Chowdhury. Weakly supervised summarization of web videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3677–3686, 2017. [12](#)
- [PRC20] Nazil Perveen, Debaditya Roy, and Krishna Mohan Chalavadi. Facial expression recognition in videos using dynamic kernels. *IEEE Transactions on Image Processing*, 29 :8316–8325, 2020. [6](#), [22](#), [74](#), [89](#)
- [PVRM05] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *2005 IEEE International Conference on Multimedia and Expo*, pages 5 pp.–, 2005. [21](#), [80](#)
- [PVT21] Luan Pham, The Huynh Vu, and Tuan Anh Tran. Facial expression recognition using residual masking network. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4513–4519, 2021. [6](#), [22](#), [23](#), [28](#), [75](#)
- [RB16] Shafin Rahman and Neil D. B. Bruce. Factors underlying inter-observer agreement in gaze patterns : Predictive modelling and analysis. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications, ETRA '16*, page 155–162, New York, NY, USA, 2016. Association for Computing Machinery. [5](#), [13](#), [14](#), [15](#), [16](#), [22](#), [56](#), [63](#), [89](#)

- [SGM09] Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Facial expression recognition based on local binary patterns : A comprehensive study. *Image and Vision Computing*, 27(6) :803–816, 2009. 81
- [SHDGD18] Hammad Squalli-Houssaini, Ngoc Q. K. Duong, Marquant Gwenaelle, and Claire-Helene Demarty. Deep learning for predicting image memorability. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2371–2375, 2018. 6, 17, 19, 22, 70, 72, 89
- [SHK<sup>+</sup>14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout : A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56) :1929–1958, 2014. 40
- [SW10] Claude Sammut and Geoffrey I. Webb, editors. *Mean Absolute Error*, pages 652–652. Springer US, Boston, MA, 2010. 39
- [Toy] ToyFight. Hume ai.
- [WWW<sup>+</sup>19] Junbo Wang, Wei Wang, Zhiyong Wang, Liang Wang, Dagan Feng, and Tieniu Tan. Stacked memory network for video summarization. In *Proceedings of the 27th ACM International Conference* [XHE<sup>+</sup>10] on *Multimedia*, MM ’19, page 836–844, New York, NY, USA, 2019. Association for Computing Machinery. 11, 12
- [YLZ<sup>+</sup>21] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database : Large-scale scene recognition from abbey to zoo. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010. 19, 71
- [YMCZ19] Jiaomin Yue, Qiang Lu, Dandan Zhu, Xiongkuo Min, Xiao-Ping Zhang, and Guangtao Zhai. Inter-observer visual congruency in video-viewing. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5, 2021. 5, 8, 14, 15, 16, 22, 65, 89
- [ZLX<sup>+</sup>14] Yitian Yuan, Tao Mei, Peng Cui, and Wenwu Zhu. Video summarization by learning deep side semantic embedding. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(1) :226–237, 2019. 12
- [ZLX<sup>+</sup>14] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Ad-*

- vances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 19, 70
- [ZMC22] Ce Zheng, Matias Mendieta, and Chen Chen. Poster : A pyramid cross-fusion transformer network for facial expression recognition, 2022. 6, 21, 22, 56, 77
- [ZMM15] Xiao Zhang, Mohammad H. Mahoor, and Seyed Mohammad Mavadati. Facial expression recognition using l<sub>p</sub>-norm mkl multiclass-svm. 2015. 81, 82
- [ZX18] Haimin Zhang and Min Xu. Recognition of emotions in user-generated videos with kernelized features. *IEEE Transactions on Multimedia*, 20(10) :2824–2835, 2018. 89

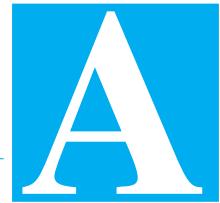
# Table des figures

2.1	Storyboard et Video Skim [AAM <sup>+</sup> 21b]	10
2.2	Images triées par score de congruence inter-observateurs (le plus faible en haut à gauche et le plus fort en bas à droite). En (a) : jeu de Bruce/Toronto, (b) : jeu de Judd/MIT, (c) : jeu de mémorabilité [MLM13]). La figure provient de [RB16].	14
2.3	Méthode leave-one out pour le calcul de la congruence inter-observateurs de l'observateur $i$ , provient de [LMBR11]	15
2.4	Protocole expérimental de [KRTO15] pour mesurer la mémorabilité d'image par des humains (figure provenant de [KRTO15])	18
2.5	Images classées par ordre décroissant de mémorabilité. En haut : les 10 images les plus mémorables, en bas : les 10 images les moins mémorables, provenant du jeu de données SUN utilisé dans le jeu de mémorabilité LaMem [KRTO15]	19
2.6	Classes d'émotions généralement utilisées, provient de [ZMC22]	21
4.1	Exemple de chutes et pic du score de mémorabilité	26
4.2	Exemple de chute du score d'IOVC	27
4.3	3 clusters types d'évolution de la mémorabilité au cours de vidéo - TVSum	29
4.4	Fenêtres (rectangles) de 10 secondes avec un pas de 5 secondes sur une vidéo de 20 secondes	30
4.5	Réduction à deux dimensions avec T-SNE - TVSum	31
4.6	Distance Euclidienne et distance de déformation temporelle dynamique	31
4.7	Clusters utilisant la distance de déformation temporelle dynamique - TVSum	32
4.8	Histogramme de corrélation entre l'IOVC et la vérité-terrain des vidéos de TVSum	34
4.9	Histogramme de corrélation entre la mémorabilité et la vérité-terrain des vidéos de TVSum	35
4.10	Test de Student entre les frames sélectionnées et non sélectionnées de chaque vidéo de TVSum	36
4.11	Matrice de Wasserstein sur l'IOVC	37
4.12	Matrice de Wasserstein sur la mémorabilité	38
4.13	Architecture du réseau de neurones	39
4.14	Résultat d'entraînement du réseau de neurone sur la base de donnée TVSum	41
4.15	Architecture du réseau de neurones avec couche LSTM	42

4.16 Résultat d'entraînement du réseau de neurones avec LSTM sur la base de données TVSum . . . . .	43
4.17 Courbes ROC des 8 splits du jeu d'entraînement, la diagonale représente une classification aléatoire . . . . .	45
4.18 Matrices de confusion des 8 splits du jeu d'entraînement sur un jeu de test déséquilibré . . . . .	45
4.19 Courbes ROC des 7 splits du jeu de test équilibré, la diagonale représente une classification aléatoire . . . . .	45
4.20 Matrices de confusion du 1er split d'entraînement sur les 7 splits du jeu de test équilibré . . . . .	45
B.1 Phase 1 - Planification prévisionnelle . . . . .	85
B.2 Phase 1 - Planning effectif . . . . .	85
B.3 Phase 2 - Planification prévisionnelle . . . . .	86
B.4 Phase 2 - Planning effectif . . . . .	87
D.1 Points à contrôler à l'issue de la phase I . . . . .	104
D.2 Points à contrôler à l'issue de la phase II . . . . .	105

# Liste des tableaux

4.1	Corrélation moyenne de l'IOVC sur TVSum . . . . .	33
4.2	Corrélation moyenne de l'IOVC sur SumMe . . . . .	33
4.3	Corrélation moyenne de la mémorabilité sur TVSum . . . . .	35
4.4	Corrélation moyenne de la mémorabilité sur SumMe . . . . .	35
4.5	Score R <sup>2</sup> du réseau de neurones sur le jeu de validation . . . . .	40
4.6	Score R <sup>2</sup> du réseau de neurones avec LSTM sur le jeu de validation . . . . .	42
A.1	ResMem - Résultats des modèles testés . . . . .	74
A.2	Résultats des méthodes testées . . . . .	80
A.3	Résultats des entraînements par jeux de données . . . . .	81
A.4	Résultats des entraînements par jeux de données croisés . . . . .	82
C.1	Avancement du projet par rapport au temps de travail théorique minimal (respectivement haut) . . . . .	102



---

# Fiches de lecture

## A.1 Résumé vidéo

Nous allons voir dans cette partie des articles concernant les méthodes de résumé vidéo.

### A.1.1 Video Summarization Using Deep Neural Networks : A Survey [AAM<sup>+</sup>21b]

Publication : 15 janvier 2021

**Introduction** Ce papier de recherche fournit une enquête exhaustive sur les méthodes d'apprentissage profond pour créer des résumés vidéos génériques.

**Contexte** Aujourd'hui, il y a beaucoup de contenus vidéos sur les sites web d'hébergement vidéo (environ 500h de video YouTube mis en ligne par minute estimée pour 2021) mais il existe d'autres plateformes (e.g., Dailymotion et Vimeo), réseaux sociaux (e.g., Facebook, Twitter, and Instagram) ainsi que les médias et journaux...

Problèmes : comment un utilisateur peut-il naviguer efficacement à travers autant de contenu afin de trouver la video recherchée ? Le résumé video permet de faciliter la navigation parmi une large collection de vidéos donc augmente l'engagement utilisateur et sa consommation de contenu. L'utilité des résumés vidéo est variée : générer des bandes-annonces de film ou série, ; présenter les temps forts d'un événement et créer un résumé vidéo des activités principales ayant eu lieu par exemple les dernières 24 heures d'un enregistrement d'une caméra de vidéosurveillance.

**Méthodes** La bibliographie sur les résumés vidéos avec réseaux neuronaux profonds divisée en deux :

1. Approche **unimodale** utilisant seulement l'aspect visuel des vidéos pour extraire des caractéristiques et apprendre d'une manière (faiblement) supervisée ou non supervisée.
2. Approche **multimodale** exploitant les métadonnées textuelles disponibles et apprenant la sémantique en

augmentant la pertinence entre la sémantique du résumé et celle des métadonnées associées à la vidéo (titre, catégorie, description etc...)

**Format du résumé vidéo** Le résumé vidéo peut prendre la forme de l'un des deux formats principaux en contenant un ensemble représentatif :

- D'images de la vidéo (video key frames) : format appelé **video storyboard**
- De fragments vidéo (video key fragments) : format appelé **video skim**

Cet ensemble d'images ou de fragments vidéo sont apposés chronologiquement et forment une courte vidéo. Souvent plus intéressant et divertissant pour le spectateur de regarder un résumé de type *skim* car il permet d'inclure de l'audio et du mouvement grâce aux fragments vidéo. Tandis que les résumés de type *storyboard* ne sont pas contraints par des problèmes de synchronisation ou de timing, ils sont donc davantage flexibles en termes d'organisation de données pour la navigation.

La longueur  $L$  d'un résumé vidéo doit dépendre du temps  $T$  de la vidéo complète. Cette longueur doit être égale à  $L = p * T$ , avec  $p$  un pourcentage, généralement 15 %. Cela donne donc une contrainte de temps à laquelle vient s'ajouter un poids d'importance de chaque image ou fragment de la vidéo, cela ramène donc le problème d'optimisation au problème du sac-à-dos (maximiser le nombre d'images ou fragments importants dans un temps donné).

**Méthodes d'entraînement** : Supervisée, Non-supervisée, Faiblement supervisée

**Méthodes supervisées** Apprendre l'importance des images en modélisant les dépendances temporelles entre les images Utilise des unités de LSTM (long short-term memory) pour modéliser : - Dépendance temporelle entre images - L'importance des images est estimée par un perceptron multicouche MLP. - diversité visuelle accrue avec le processus ponctuel déterminant Apprendre l'importance des images en modélisant la structure spatiotemporelle de la vidéo

Apprendre à résumer une vidéo en trompant un discriminateur essayant de distinguer un résumé généré par la machine d'un généré par l'humain Le « résumeur » (qui fait office de générateur du GAN) reçoit en entrée la séquence des images vidéo et génère un résumé en calculant l'importance des images. Dans les méthodes supervisées basées sur les GAN, l'entraînement se fait d'une manière contradictoire (avec vérité-terrain). Le résumé généré (poids pour chaque image) ainsi que les poids donnés par des utilisateurs pour un résumé optimal sont donnés en entrée d'un discriminateur qui, en sortie, donne la similarité de l'entraînement de la création d'un résumé.

**Méthodes non-supervisées** Pas de vérité-terrain. Apprendre à résumer une vidéo en trompant un discriminateur qui cherche à distinguer la vidéo originale d'une reconstruction basée sur un résumé Se base sur l'idée qu'un résumé représentatif doit permettre au spectateur d'inférer le contenu de la vidéo originale. Les GAN (Ge-

nerative Adversial Networks) apprennent à créer des résumés vidéo permettant une bonne reconstruction de la vidéo originale. Le *résumeur* essaie de tromper le *discriminateur* qui cherche à distinguer la vidéo reconstruite grâce à un résumé de la vidéo originale. La distinction n'est plus possible quand l'erreur de classification est quasi-égale pour la vidéo reconstruite et la vidéo originale, le « résumeur » est alors considéré comme capable de créer un résumé vidéo fortement représentatif. Le forward GAN apprend à reconstruire la vidéo originale depuis le résumé vidéo et le backward GAN fait l'inverse.

Apprendre à résumer une vidéo en ciblant les propriétés spécifiques désirées pour le résumé. Apprentissage par renforcement avec des fonctions de récompenses quantifiant l'existence de caractéristiques désirées dans le résumé généré. Construire un résumé orienté-objet modélisant le mouvement clé des objets visuels importants. Préserve les informations sémantiques et de mouvement à grain fin de la vidéo. Étape de prétraitement qui vise à trouver les objets importants et leurs mouvements clés. Représente la vidéo entière en créant des clips de mouvement d'objets super-segmentés. Génère des résumés qui montrent les objets représentatifs de la vidéo et les mouvements clés de chacun des objets.

**Méthodes faiblement supervisées** Au lieu de ne pas utiliser de données de vérité-terrain, utilise des étiquettes dites *faibles* et moins coûteuses. Etiquettes comme : les métadonnées de la vidéo pour la catégorisation de la vidéo et le résumé par catégorie, ou les annotations des vérités-terrain pour un petit sous-ensemble d'images pour

l'apprentissage du résumé par l'apprentissage par renforcement clairsemé et les fonctions de récompense adaptées). Exemple : vidéo avec un titre "homme qui cuisine" : grâce au titre, on a des vidéos relatives qui vont servir d'entraînement pour apprendre les segments vidéo en commun (3D CNN).

**Approches multimodales** Modalités supplémentaires (en plus du flux visuel). Pour apprendre à résumer : flux audio, les sous-titres de la vidéo ou les transcriptions générées avec reconnaissance vocale, toutes les métadonnées textuelles disponibles (titre et/ou résumé de la vidéo), ou d'autres comme des données de contexte, catégorie de la vidéo, description humaine du contenu de la vidéo. Extrait la sémantique de haut niveau sur le contenu visuel grâce à des réseaux CNN/DCNN pré-entraînés et apprend de manière supervisée en maximisant la similarité sémantique entre le résumé et les métadonnées contextuelles. Résumé généré comparé à (score de similarité) un résumé vérité-terrain et les métadonnées de la vidéo.

**Méthodes d'évaluation** Méthodes d'évaluation :  
— Qualitative basées sur des études d'utilisateurs  
— Quantitative basées sur des vérités-terrain

Ces méthodes prennent du temps (annoter des vidéos ou faire des études utilisateurs) et dépendent donc d'utilisateurs humaines (subjectivité).

**Résultats** Les meilleurs résumés vidéo sont obtenus via les méthodes d'apprentissages :

1. **Supervisées** utilisant des mécanismes d'attention (VASNet, H-MAN, SUM-GDA, DASP, et CSNet-sup) ou des réseaux de mémoire (SMN) pour capturer les dépendances temporelles.
2. **Non supervisées** utilisant des GAN ainsi que des mécanismes d'attention ((CSNet, CSNet+GL+RPE, SUM-GDAunsup, et SUM-GANAA)

Tandis que les méthodes multimodales ainsi que les méthodes d'apprentissage faiblement supervisées ferment très mal.

#### Code Supervisées

1. **PGL-SUM**, 2021 [\[Combining Global and Local Attention with Positional Encoding for Video Summarization\]](#)
2. **DSNet**, 2020 [\[DSNet: A Flexible Detect-to-Summarize Network for Video Summarization\]](#)
3. **VASNet**, 2021 [\[Video Summarization with Attention\]](#)

#### Non-supervisées

1. **CA-SUM**, 2022 [\[Summarizing Videos using Concentrated Attention and Considering the Uniqueness and Diversity of the Video Frames\]](#)
2. **AC-SUM-GAN**, 2020 [\[Connecting Actor-Critic and Generative Adversarial Networks for Unsupervised Video Summarization\]](#)

## A.2 Congruence visuelle inter-observateurs

Cette partie rend compte de différents articles ayant étudié la congruence visuelle inter-observateurs.

### A.2.1 Prediction of the Inter-Observer Visual Congruency (IOVC) and Application to Image Ranking [[LMBR11](#)]

Publication : 28 novembre 2011

**Prédiction** : La méthode prédit la congruence visuelle inter-observateurs (IOVC) grâce aux caractéristiques visuelles de bas-niveau ainsi que des mesures d'eye tracking. La vérité-terrain est paramétrée en calculant l'IOVC à partir des données de l'eye tracking. La mesure est un score de 0 (congruence minimale) à 1 (congruence maximale)

**Jeu de données** : 15 observateurs de 18 à 35 ans ont réalisé un eye tracking sur la base Judd et al.'s : 1003 images au contenu varié, différentes résolutions, orientations (paysage ou portrait). Grâce à la mesure de l'IOVC, on obtient sur ce jeu de données une dispersion moyenne de 72% et une dispersion médiane de 76%.

**Mesure de la congruence inter-observateur** : Mesure avec l'approche *one-against-all* aussi appelée *leave one out*. Pour un observateur N, on calcule une distribution 2D des zones de fixation des N-1 observateurs et une

convolution gaussien 2D, on obtient une carte représentant la probabilité de fixation de chaque pixel. 25% de l'image où la fixation est la plus importante est gardée. Enfin, on calcule la correspondance entre les fixations visuelles de l'observateur N et la carte seuillée.

**Attributs visuels impactant la variabilité inter-observateur** : La détection de visages, l'harmonie des couleurs, la profondeur de champ, la complexité de la scène (entropie, nombre de régions, quantité de contours). Ces attributs sont calculés pour simuler le comportement visuel humain.

**Modèle d'apprentissage** : L'apprentissage utilise les attributs visuels pour prédire l'IOVC. Le modèle utilisé est un cluster-weighted model (CWM), qui est une généralisation de Gaussian mixture.

**Résultat** : La méthode est évaluée qualitativement, mais aussi quantitativement : le critère IOVC surpassé la mesure de Feature Congestion. Performe bien pour prédire la dispersion d'observateurs seulement dans une tâche de visionnage libre.

### A.2.2 Factors Underlying Inter-Observer Agreement in Gaze Patterns : Predictive Modelling and Analysis [RB16]

Publication : 14 mars 2016

**Problème abordé** : Evaluer à quel point différentes caractéristiques contribuent à la variabilité du regard selon les observateurs (IOC, congruence inter-observateurs)

- Valeur de l'IOC haute : contient généralement un petit nombre de régions (concentrées ou séparées) à fort contraste, la présence de personnes ou de visages et la présence de texte.
- Valeur de l'IOC basse : si rien n'attire l'attention ou si trop de choses l'attire et que le temps de visionnage est trop court. Généralement des scènes encombrées et des paysages.

Accompli en considérant la corrélation entre des caractéristiques dérivées d'images et l'IOC, et basé sur la capacité de caractéristiques plus complexes pour prédire l'IOC avec un modèle de régression.

**Caractéristiques simples liés à la complexité d'une image** *L'entropie* (mesurant l'aléatoire d'une image), *le désordre visuel* et *la taille d'une image JPEG* (liée au désordre de l'image) sont des caractéristiques ayant une corrélation inverse avec la congruence inter-observateurs.

**Caractéristiques plus complexes : analyse de l'image Bottom-Up et Top-Down**

**Bottom- up** : influence du regard par les propriétés du contenu de l'image.

*Gist* et *HoG* (histogram of oriented gradients) donnent une représentation grossière de la structure de la scène.

*HoPS* (*Histogram of Predicted Salience*) : permet de prédire l'IOC et présente un vecteur de caractéristiques

contenant des histogrammes concaténés dérivés d'images de saillance basées sur 12 algorithmes (produisant des prédictions différentes). HoPS a de très bonnes capacités à représenter la variabilité dans la saillance prédictive et aussi la cohérence et les points communs entre les algorithmes.

**Top-Down** : influence du regard par des perceptions haut-niveau ou des connaissances préalables. Capture des caractéristiques haut-niveau avec des CNN (réseaux de neurones convolutifs) (avec l'architecture BVLC Reference CaffeNet [Jia et al. 2014] basée sur l'architecture AlexNet [Krizhevsky et al. 2012] (et entraînée sur les données ILSVRC12).

**Apprentissage** Les vecteurs de caractéristiques HoG, GIST, HoPS et DeepNet sont utilisés pour apprendre un modèle de régression prédisant l'IOC.

Un modèle « support vector regression » (SVR) est appliqué pour prédire l'IOC en utilisant un kernel RBF dans le modèle de régression.

**Résultats principaux** Les caractéristiques de HoPS surpassent les caractéristiques de HoG [Felzenszwalb et al. 2010] et Gist [Oliva and Torralba 2001] pour la prédition de l'IOC.

Les caractéristiques de HoPS performent mieux que LeMeur et al. 2011 avec une corrélation de Pearson  $r = 0.430$  sur le Judd Dataset (LeMeur :  $r = 0.340$ ). De plus, avec une combinaison DeepNet (1L) + HoPS, la performance est encore meilleure avec  $r = 0.456$ .

Les caractéristiques de *HoPS* donnent une idée de la *complexité* de l'image, celles de *DeepNet* incluent des *regroupements selon les objets* présents.

L'intérêt des différents types de caractéristiques (Bottom-up et Top-Down) pour la prédiction dépend du type de contenu des images. - HoPS prédit bien sur les images simples sans visage. - Images complexes : la détection d'objets de haut-niveau fonctionnent mieux. Combiner les deux types de caractéristiques rend la prédiction efficace sur tous types d'images.

### A.2.3 Deep Learning For Inter-Observer Conguency Prediction [BLCLM19]

Publication : Septembre 2019 à l'IEEE International Conference on Image Processing (ICIP)

**Introduction** Introduction d'un réseau de neurone convolutionnel pour prédire le scores IOC d'image. Le score IOC correspond à la congruence visuelle entre différents observateurs sur une même image.

Dans le papier, 2 méthodes de génération des données de vérités terrain sont utilisé : AUC-Borji (area under curve) et NSS (normalized scanpath saliency). Le premier est une valeur comprise entre 0 et 1 qui plus la valeur est importante est plus le score a un haut niveau de congruence. Le deuxième est compris entre 0 et  $+\infty$  qui de la même manière indique un haut score de congruence.

**Jeux de données** Les deux jeux de données sont disponibles ici : <http://saliency.mit.edu/>

**Judd/MIT** Ce jeu de données contient 1003 images d'intérieurs et d'extérieurs avec des résolutions différentes ainsi que les données de 15 observateurs par image (âges de 18 à 35 ans).

**CAT2000 [BI15]** Ce jeu de données contient 4000 images de 20 catégories différentes à une résolution de 1920\*1080px et les données de 24 observateurs par image (âges de 18 à 27 ans).

**Architecture** Le modèle est un réseau de neurone profond utilisant comme encodeur d'image VGG19 pré-entraîné sur ImageNet pour l'extraction des features de l'image en entrée. Et pour la partie décodeur, le réseau est composé de simples couches connectées pour faire la régression sur le score IOC.

**Résultat** Le modèle surpassé des techniques plus basiques comme l'utilisation d'un SVM, Random Forest, Perceptron, et deux autres techniques. Les résultats de corrélation des résultats du modèle avec des vérités terrains sont : environ de 61.1% sur le jeu de données MIT300 et 64.2% sur le jeu de données CAT2000.

#### A.2.4 Inter-Observer Visual Congruency in Video-Viewing [YLZ<sup>+</sup>21]

Publication : Décembre 2021 à l'International Conference on Visual Communications and Image Processing

(VCIP)

**Contexte** Congruence visuelle inter-observateurs (IOVC) : différences individuelles dans l'attention visuelle entre observateurs - degré de dispersion entre les zones d'attention visuelle de différentes personnes observant un même stimulus.

Facteurs influençant le comportement du regard : l'âge, la culture et la condition physique.

Les recherches actuelles se concentrent sur les « regions of interest » (ROIs), de nombreux modèles proposés pour prédire les régions saillantes. L'IOVC est un indicateur auxiliaire pour mesurer la généralisation et la représentativité des bases de données de regards et méthodes de prédiction.

**Mesure pour calculer l'IOVC d'une vidéo** Le calcul de l'IOVC est basé sur les cartes de densité de fixation (FDM) des observateurs. Pour simuler l'angle visuel humain, une convolution (kernel Gaussien) est utilisée sur les cartes de fixation pour obtenir les FDMs. Les FDMs générées sont des cartes de niveau de gris, les régions claires sont les zones saillantes.

Méthode utilisée : *leave-one*, on calcule le degré de congruence d'un observateur grâce aux valeurs de tous les autres observateurs (excepté lui-même).

Pour une même scène, les observateurs peuvent se concentrer sur les mêmes zones, mais avec une différence dans l'ordre d'observation. Cet ordre donnera un IOVC anormalement bas. Pour compenser ce problème,

les données de mouvement oculaire de n images sont superposées, et le calcul est effectué toutes les m images.

Une expérience de eye-tracking lors de visionnage de films permet construire un ensemble de données sur les mouvements oculaires dynamiques.

**Résultats de l'IOVC sur le jeu de données** Les images contenant un seul sujet avec un fond propre ont des scores IOVC plus élevés. Les images dont l'IOVC est faible sont généralement des arrière-plans et ne contiennent pas de sujet.

**Réseau prédisant la valeur de l'IOVC** Les informations de mouvement entre images peuvent influencer l'attention visuelle et être utilisées pour prédire la saillance. Et les informations basées sur la saillance permettent de prédire l'IOVC.

Structure à double branche pour combiner les informations sur le mouvement, une branche de contenu et une branche de flux.

- la branche du contenu (branche principale) prend l'image vidéo en entrée pour extraire les caractéristiques du contenu de l'image. (sémantique visuelle)
- la branche flux extrait les caractéristiques de variation entre les images via le flux optique (estimé par FlowNet 2.0)

Structures des deux branches similaires au ResNet (une couche de convolution et quatre couches résiduelles).

Les caractéristiques entre les images extraites par la branche de flux sont intégrées à la caractéristique de contenu pour obtenir des informations sur l'attention. La

loss de type "smooth L1" est utilisée entre la cohérence prédite et la vérité-terrain pour entraîner le modèle.

**Corrélation entre IOVC et l'émotion** Les données utilisées sont des clips vidéo annotées avec des émotions définies par la valence et l'arousal. La valence décrit le degré positif de l'expérience psychologique (par exemple, satisfait et déçu). L'arousal décrit le niveau d'énergie (par exemple, excité et endormi).

Spearman Rank Order Correlation Coefficient pour mesurer le degré de corrélation entre les courbes d'émotion et d'IOVC. Le SROCC entre l'arousal et l'IOVC est de 0.43 quand valence est positive et de 0.29 quand elle est négative. Quand l'intrigue est plaisante, l'attention des observateurs est davantage concentrée, donc un plus haut IOVC.

**Résultats du modèle de prédiction** Le SROCC entre l'IOVC prédit et la vérité terrain sur l'ensemble de validation est utilisé pour évaluer l'efficacité de la méthode de prédiction. La méthode obtient un SROCC de 0.693 pour une division moyenne des jeux de données (d'entraînement, test et validation), et de 0.573 pour une division croisée. Leur méthode a été meilleure que les précédents modèles de prédiction.

### A.3 Mémorabilité

Nous allons voir dans cette partie des articles concernant la caractéristique de la mémorabilité.

### A.3.1 Memorability of natural scenes - the role of attention [MLM13]

Publication : Septembre 2013 à l'IEEE International Conference on Image Processing

**Problème abordé** : L'article cherche à prouver que le comportement visuel dépend de la mémorabilité des images. Mémorabilité de l'image : faculté d'une image à être remémorée après un certain temps.

**Principe général** : Enquête sur le rôle de l'attention visuel sur la mémorabilité de l'image via :

1. Expérience d'eye-tracking sur un ensemble d'images ayant des scores différents de mémorabilité.
2. Prédiction de la mémorabilité, les caractéristiques liées à l'attention peuvent remplacer des caractéristiques bas-niveau pour prédire la mémorabilité.

**Expérience d'eye-tracking sur des images ayant des scores différents de mémorabilité** : L'attention a été mesuré avec des données d'eye-tracking (images et zones de fixations du regard) récoltées sur 17 volontaires ayant regardé 135 images extraites du jeu de données [IXTO11] composé de 2222 images annotées avec la mémorabilité. Deux mesures de l'attention sont observées :

- *La durée de fixation* reflète la profondeur du traitement visuel du cerveau. Les 20 images les plus mémorables sont plus longuement fixées que les 20 les moins mémorables (différence significative).

- *La congruence inter-observateurs* donne le degré de similarité des fixations entre plusieurs observateurs. Forte au début et baisse avec le temps. Les images des deux classes les plus mémorables ont une différence significative avec la dernière classe d'images (les moins mémorables) dans leur congruence inter-observateurs.

Ces résultats montrent que la mémorabilité et l'attention sont liés. Il serait donc raisonnable d'utiliser les caractéristiques liées à l'attention pour prédire la mémorabilité d'images.

**Prédiction de la mémorabilité** : Le modèle utilise deux caractéristiques liées à l'attention (la couverture et la visibilité) pour prédire la mémorabilité, puis utilise un classifieur de type SVR (Support Vector Regression).

- *La couverture des cartes de saillance* : Une valeur basse indique au moins une région saillante, tandis qu'une valeur haute indique qu'il n'y a rien d'important dans l'image ou bien plusieurs régions d'intérêt localisées à différents endroits de l'image. Utilise l'algorithme RARE pour obtenir la couverture moyenne des cartes de saillance.
- *La visibilité selon le contraste des structures de l'image* : Simule le processus humain d'oubli avec des filtres low-pass appliqués sur les images en mesurant la corrélation entre l'image d'origine et l'image filtrée. Les grandes structures contrastées seront plus résistantes au filtrage tandis que les petits détails et les structures avec arrière-plan encombré seront moins résistants.

**Résultats principaux** Ces caractéristiques sont exploitées seules et donnent qu'une corrélation assez faible avec la mémorabilité (0.1 pour la couverture et 0.274 pour la visibilité). Si l'on ajoute à ces deux caractéristiques celles des méthodes SIFT, HOG, SSIM et Pixel on obtient une corrélation de 0.479, ce qui est mieux que [IXTO11] ayant une corrélation de 0.462 avec GIST, SIFT, HOG, SSIM, Pixel.

Cette recherche a donc permis de remplacer quelques caractéristiques de bas-niveau (GIST) de [IXTO11] tout en réduisant la dimensionnalité de l'ensemble des caractéristiques.

### A.3.2 Relative Spatial Features for Image Memorability [KYP13]

Publication : 21 octobre 2013

**Problème abordé** Les caractéristiques intrinsèques de la mémorabilité ne sont pas pleinement comprises, le caractère insolite ou la beauté d'une image pourrait ne pas être corrélée avec la mémorabilité de l'image.

L'article enquête donc sur deux nouvelles caractéristiques spatiales, corrélées à la mémorabilité : **la zone d'objet pondérée (WOA, Weighted Objet Area)** prenant en compte la position et la taille des objets, ainsi que **le rang relatif de la zone (RAR, Relative Area Rank)** capturant le caractère inhabituel relatif de la taille des objets.

Une analyse qualitative des images indique que ce qui corrèle avec la mémorabilité n'est pas le caractère inha-

bituel de l'image mais le caractère inhabituel de chaque objet relativement à leur taille et position attendue.

**Principe général** Les 2 caractéristiques spatiales suivantes sont utilisées ainsi que d'autres caractéristiques mentionnées ci-après. Ils utilisent un classifieur SVR pour prédire la mémorabilité.

**Weighted Objet Area** : Caractéristique considérant la taille et la position des objets d'une image qui sont des facteurs déterminant leur importance. La taille et position des objets ont une forte corrélation avec la mémorabilité d'une image. Plus un objet est large ou plus il est proche du centre de l'image, la probabilité qu'il soit annoté augmente (ainsi que la valeur du WOA).

Concrètement, on donne davantage de poids aux pixels autour du centre et on réduit les poids de ceux proches de la bordure. Une fonction Gaussienne bivariée est utilisée sur les positions des pixels et permet d'évaluer cette caractéristique.

**Relative Area Rank** : Caractéristique du rang relatif capturant l'importance d'un objet selon son ordre d'annotation dans la liste des étiquettes d'une image. Défini par le centile du rang de l'étiquette par rapport à tous les rangs de la même étiquette dans les images d'entraînement.

**Résultats principaux** Les caractéristiques proposées (WOA et RAR) sont comparées à des caractéristiques uti-

lisées dans des précédents travaux de recherche (Object, Scene et Attributes).

Ces caractéristiques sont testées pour leur prédiction de la mémorabilité des images. Les résultats montrent que WOA et RAR peuvent améliorer la prédiction.

WOA + RAR surpassent Object + Scene. RAR à lui seul atteint presque la même performance que Object + Scene. Mais les 127 attributs de Attributes donnent une meilleure performance.

D'autres tests sont réalisés en combinant ces caractéristiques (WOA et RAR) aux caractéristiques visuelles (VIS) provenant des histogrammes Pixel, GIST, SIFT, HOG2  $\times$  2 et SSIM.

WOA remplace Object dans l'ensemble : Object + Scene + VIS + Attributes et améliore l'approche de 0.56 à 0.57 et de 0.57 à 0.58 pour respectivement 6 et 127 attributs de Attributes. De plus, RAR est testé de la même manière et permet d'augmenter la prédiction sur les images les moins mémorables (car capture le caractère inhabituel des tailles d'objets).

Sans Attributes, WOA + RAR remplacent Object dans l'ensemble Object + Scene et améliorent les prédictions de 0.50 à 0.52 sans VIS et de 0.54 à 0.57 avec VIS.

### A.3.3 Understanding and Predicting Image Memorability at a Large Scale [KRTO15]

Publication : Décembre 2015 à l'IEEE International Conference on Computer Vision (ICCV)

**Problème abordé** Le progrès dans l'estimation de la mémorabilité visuelle est limité par les données de référence disponibles : une trop petite quantité ainsi qu'un manque de variété.

L'article propose donc un large jeu de données ainsi qu'une mesure objective pour mesurer la mémorabilité.

Le modèle et les données sont disponibles à l'adresse : <http://memorability.csail.mit.edu/>

### Principe général

**Jeu de données : LaMem** : Ils ont construit un jeu de données nommé **LaMem**, qui est le jeu de données annoté de mémorabilité le plus large (au moment de la présentation de l'article à la conférence internationale sur la Vision par ordinateur de l'IEEE en 2015). Il contient **60 000 images de sources diverses**.

Grâce aux caractéristiques haut-niveau des couches des réseaux de neurones convolutifs, ils ont montré que les objets et régions sont positivement ou négativement corrélés à la mémorabilité. Cela a permis de créer des cartes de mémorabilité pour chaque image et donc une méthode pour manipuler la mémorabilité d'une image.

Le **score humain de mémorabilité** (vérité terrain) a été défini grâce à la plateforme Amazon's Mechanical Turk, où des suites d'images du jeu sont montrées à des participants. Ceux qui échouaient à détecter la même image répétée moins de 7 images après celle-ci ont été bloqués et leurs résultats non pris en compte. Avec cette expérience, ils obtiennent environ **80 scores par image**.

**Comprendre la mémorabilité** : Grâce à différents attributs des images, propres à chaque source du jeu de données, ils peuvent tester la corrélation entre la mémorabilité et les attributs suivants :

- **Les émotions** : les images qui provoquent des émotions négatives (peur, colère, dégoût) sont les plus mémorables (un cas particulier est l'amusement qui implique aussi une bonne mémorabilité). Tandis que les émotions positives (admiration et satisfaction) donnent des mémorabilités basses.
- **La saillance** : les fixations du regard et la mémorabilité ont une corrélation de 0.24
- **La popularité** : score de popularité des images les plus mémorables (1er quartile) statistiquement plus élevé que pour les autres quartiles.
- **L'esthétique** : peu voire pas du tout de corrélation

**Modèle développé** : MemNet , Initialise ce réseau avec le **CNN-Hybride** d'une étude précédente [ZLX<sup>+</sup>14]. Il s'agit d'un réseau pré-entraîné avec les jeux de données *ILSVRC 2012* (reconnaissance d'objets) et *Places* (reconnaissance de scènes). La mémorabilité est une seule valeur réelle donc ils utilisent une loss Euclidienne pour faire du fine-tuning sur le réseau.

**Résultats** Ils ont utilisé les réseaux de neurones convolutifs (CNN) et ont montré que des caractéristiques "fine-tuned" surpassent toutes les autres par une large avance, atteignant une corrélation de 0.64, proche de celle humaine de 0.68.

**Test avec le jeu de données SUN** Surpasse Hog2x2 d'environ 0.15 (0.63 vs 0.48 avec fausses alarmes, et 0.60 vs 0.45 sans fausse alarme). Le fine-tuning baisse la performance à cause du nombre limité de données et de l'overfitting. Il est aussi montré que gérer les fausses alarmes permet d'augmenter significativement la corrélation.

**Test avec le jeu de données développé LaMem** Corrélation de 0.64 sur le jeu de test LaMem, proche de la corrélation humaine sur le même jeu. Le fine-tuning fonctionne bien sur ce large jeu de données et permet de bien généraliser car obtient 0.61 sur le jeu de données de test SUN, ce qui est comparable (vs 0.63) au modèle entraîné avec les données SUN.

### A.3.4 Deep Learning for Image Memorability Prediction : the Emotional Bias [SHGKD18]

Publication : 1 octobre 2016

**Problème abordé** Les jeux de données d'images annotées avec score de mémorabilité ne prennent pas en compte le biais émotionnel. La littérature de psychologique donne la preuve que les images générant une émotion sont associées à une meilleure mémorabilité que les images neutres. L'émotion générée par une image est donc un élément clé pour prédire à quel point l'image sera mémorisée.

Ce travail propose une méthode de prédiction de mémorabilité d'image et analyse ce biais.

**Principe général** 1. Les approches utilisant des caractéristiques intrinsèques seulement performent modérément pour la prédiction de la mémorabilité. Ces approches peuvent être améliorées en utilisant des informations de haut-niveau (comme la sémantique). L'article propose donc un modèle *MemoNet* basé sur l'apprentissage profond pour prédire la mémorabilité et surpassé la performance des modèles précédents. Ils utilisent le modèle *GoogleNet* [Szegedy et al.] (état de l'art de 2014 pour la classification d'objets et de scènes sur le jeu de données ImageNet). Dans le cas de données insuffisantes pour fournir au CNN qui requiert une large quantité de données annotées, on fait du fine-tuning : cela consiste en le fait de pré-entraîner un CNN sur un large jeu de données externe (comme ImageNet) et de peaufiner (=fine-tune) le réseau pré-entraîné en poursuivant la back-propagation sur les données cibles pour s'adapter à une tâche spécifique de classification (prédire la mémorabilité). Le fine-tune est réalisé sur le modèle GoogleNet pour obtenir MemoNet.

2. Ils testent la performance du modèle sur le jeu de données composé de 2 222 images labellisées collectées de [IXTO11] provenant du jeu de données SUN [XHE<sup>+</sup>10].

Et testent la généralité de leur modèle en développant un nouveau jeu de données de 150 images annotées avec la mémorabilité et les émotions procurées. Ces images proviennent du jeu de données International Af-

fective Picture System. Les scores d'émotions sont disponibles en terme de valence, arousal, et dominance. Valence = d'une émotion négative à positive, Arousal (stimulant) = d'inactif à actif, Dominance = de "dominé" à "en contrôle".

Protocole expérimental sur 50 participants : le premier jour on passe des images à suivre aux participants pour obtenir un score de mémorabilité sur ces images (comme [IXTO11]), le second jour on passe des images pour obtenir un score d'émotion basé sur une échelle de mesure pour l'arousal et la valence (avec un système de pictogramme).

**Résultats** Avec MemoNet 30k (30 000 itérations d'entraînement), ils obtiennent Spearman  $p = 0.636$  sur le jeu de données de [IXTO11] comprenant 2222 images avec score de mémorabilité. Ils surpassent l'état de l'art en obtenant une augmentation de 32.78% sur ce jeu de données. Comme prévu, MemoNet 30k est moins performant sur le nouveau jeu de données avec mémorabilité et émotions, ils obtiennent  $p = 0.251$ .

La valence and arousal expliquent la majeure partie de la variance indépendante donc la dominance n'est pas prise en compte pour l'analyse du biais émotionnel.

Biais émotionnel : MemoNet 30k a la plus haute performance de prédiction pour les images suscitant des émotions négatives stimulantes (valence et arousal élevés), tandis que les images procurant des émotions neutres ou positives et peu stimulantes rendent la prédiction moins fiable (valence moyenne à élevée et arousal moyen à bas).

Cela montre l'importance d'avoir un jeu de données d'images distribuées de manière appropriée dans l'espace émotionnelle.

### A.3.5 Deep Learning for Predicting Image Memorability [[SHDGD18](#)]

Publication : Avril 2018 à l'IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

**Problème abordé** La quantité de données disponible augmente exponentiellement (réseaux sociaux par exemple). Comprendre le contenu joue un rôle clé dans ces systèmes pour optimiser leur traitement. La prédiction de la mémorabilité est donc importante pour une compréhension de haut-niveau des images.

L'article propose une méthode de prédiction de la mémorabilité avec réseaux profonds.

**Principe général** Prédit la mémorabilité en utilisant des caractéristiques visuelles et caractéristiques sémantiques (relatives à la légende de l'image, image caption) basées sur des CNN (réseaux neuronaux convolutifs).

Traite le problème comme un problème de classification (et non de régression comme [[KRT015](#)]).

Se sert du large jeu de données de mémorabilité *LaMem* en le divisant en 4 classes de 15 000 images.

Pour extraire des caractéristiques visuelles, utilise le réseau VGG16 pré-entraîné sur ImageNet. De plus, il utilise des caractéristiques sémantiques avec un modèle

d'*image captioning*, IC (légende d'image). Ce modèle est basé sur un CNN et un LSTM (long short-term memory recurrent network).

Le classifieur contient une branche pour les caractéristiques CNN et une pour les caractéristiques IC. Une branche est composée d'une ou plusieurs couches de MLP. Une couche merge les deux branches, puis des couches de MLP et enfin un softmax permet d'obtenir les probabilités de chaque classe.

**Résultats principaux** Surpasse l'état de l'art (MemNet [[KRT015](#)]) lors de la publication de l'article en obtenant une corrélation de Spearman  $p = 0.72$  sur le jeu de données LaMem.

Les caractéristiques de sémantique basées sur l'image captioning avec MLP ou SVR offrent une meilleure performance de prédiction que les caractéristiques CNN malgré leur dimension 4 fois plus petite.

La générativité du modèle entraîné sur LaMem est testé sur d'autres bases de données. Fonctionne bien sur les jeux de données d'images avec des catégories de scènes (proche du domaine de LaMem). Mais fonctionne mal sur des images de visages humains ou sur des visualisations (graphiques).

Ils confirment comme la littérature l'avait indiqué, la non corrélation entre la mémorabilité et le caractère intéressant des images.

### A.3.6 Embracing New Techniques in Deep Learning for Estimating Image Memorability [NB21]

Publication : 8 Janvier 2022

**Introduction** Présentation d'une nouvelle architecture de modèle d'apprentissage profond réalisant la tâche d'estimation de la mémorabilité d'une image. Avec l'avancée notable de la recherche en intelligence artificielle, de l'apprentissage profond, et les dernières recherches dans le domaine de la mémorabilité sur ces dernières années, les auteurs cherchent à re-visiter cette question avec de nouveaux types d'algorithmes pour améliorer les résultats obtenus.

Code source : [github.com/Brain-Bridge-Lab/resmem](https://github.com/Brain-Bridge-Lab/resmem)

**Approche générale** Pour l'entraînement et le test des modèles, les jeux de données utilisés sont : - LaMem [KRT015] contenant 58471 images qui est une compilation de jeux de données similaire (notamment MIR Flickr [HL08], AVA [MMP12], et Affective Image Set [MH10]) - MemCat [GW19] contenant 10000 images de meilleure qualité que dans LaMem. Le score de corrélation de reconnaissance et de mémorabilité des images d'un humain sur le jeu de données LaMem est de 0.67 et de 0.78 sur le jeu de données MemCat.

L'avantage du jeu de données LaMem est sa généralité et sa diversité de l'ensemble des données disponibles. L'inconvénient est que celui-ci contient beaucoup d'images qui sont plus orientées sur des scènes que sur

des objets précis. C'est pourquoi MemCat se focalise sur les objets pour que la combinaison des deux jeux de données forme un ensemble divers et varié.

**MemNet** [KRT015] Pour comparer les données avec le modèle proposé dans l'article de recherche [KRT015], les auteurs ont ré-implémenté deux versions du modèle *MemNet* en *PyTorch*, provenant de [KRT015], car celui-ci fut développé en *Caffe*, technologie désormais obsolète. La première version fut entraînée comme dans l'article [KRT015], et la seconde version repose sur la même architecture mais avec l'entraînement réalisé sur la combinaison des jeux de données et des dernières techniques pour améliorer les performances d'un modèle lors de l'apprentissage du réseau.

**ResMem** S'inspirant des connexions résiduelles introduites avec les modèles ResNet [HZRS15] (destinés à la classification d'image), *ResMem* réutilise les poids du réseau *ResNet-152* pré-entraîné pour le transfert des connaissances de segmentation d'image. Une autre branche du réseau ayant en entrée l'image, suit la typologie d'extraction des caractéristiques d'image du réseau *AlexNet* [KSH17]. Après concaténation des vecteurs de caractéristiques, le modèle utilise plusieurs couches entièrement connectées pour obtenir le résultat. Lors de l'entraînement de ce modèle, la partie d'extraction de caractéristiques *ResNet* a ses poids fixés pour que cette partie reste entraînée sur le jeu de données utilisé dans l'article de *ResNet* [HZRS15]. Un troisième modèle est introduit, nommé *ResMemRetrain*. Dans cette version, les

poids de la partie *ResNet* ne sont pas fixés et seront ré-entraînés lors de l’entraînement global du réseau.

**M3M** Un autre réseau utilisant la segmentation d’image est testé dans l’article. Celui-ci ajoute une branche de segmentation avec poids fixes en suivant l’architecture de *ResMem*. L’extraction des caractéristiques est réalisé par un réseau pré-entraîné nommé *fcn\_ResNet-50*. Pour pouvoir concaténer ces caractéristiques avec le vecteur de caractéristiques du modèle, une couche de convolution est appliquée.

Lors de l’entraînement des réseaux, la fonction de perte utilisée est "MSE" *Mean Squared Error* qui calcule la moyenne des distances aux carrés entre la prédiction et la vérité-terrain. La corrélation de Spearman est analysée en tant que métrique et cherche à montrer dans quelle mesure la relation entre deux variables peut être représentée à l’aide d’une fonction monotone entre les prédictions et les performances de la mémoire humaine (vérité-terrain).

**Résultats** Le tableau A.1 résume les scores obtenues des différents modèles testés. Via une analyse d’un histogramme des scores de mémorabilité obtenus sur le jeu de test ainsi que le tableau A.1, les auteurs indiquent que les résultats de la version classique de *ResMem* améliore grandement les résultats comparé à *MemNet*. La version avec ré-entraînement *ResMemRetrain* améliore sensiblement les résultats au coût d’un entraînement plus long. Et la dernière méthode *M3M* améliore sensiblement les résultats au coût de la lenteur du modèle.

Nom	MSE ↓	Corrélation de Spearman ↑
MemNet	0.012	0.55
ResMem	0.009	0.66
ResMemRetrain	0.008	0.67
M3M	0.009	0.68

TABLE A.1 – ResMem - Résultats des modèles testés  
La version de *MemNet* est la version ré-entraînée par les auteurs de cet article avec PyTorch.

Tous les résultats, listés dans le tableau A.1, des réseaux testés sont entraînés avec la combinaison des deux jeux de données.

## A.4 Intensité émotionnelle

Nous allons voir dans cette partie des articles concernant la caractéristique de la reconnaissance de l’expression faciale et de l’intensité émotionnelle.

### A.4.1 Facial Expression Recognition in Videos Using Dynamic Kernels [PRC20]

Publication : juillet 2020

**Introduction** Dans le papier, un nouveau modèle permettant la détection d’émotion utilisant l’émotion faciale est présenté. Utilisant des séquences d’image en entrée, le modèle utilise un modèle de mélange Gaussien avec différent types de kernel. Pour cela, il ont utilisés des

kernels dynamiques pour permettre l'utilisation de séquence d'image. Pour récupérer les caractéristiques du visage deux techniques sont testés utilisant le marquage des points du visage.

**Approche générale** La première étape est l'alignement du visage. Dans une vidéo, l'utilisateur bouge et donc il faut rester focaliser sur son visage. Pour cela, l'algorithme "Discriminative Response Map Fitting" est utilisé. Il permet de réaliser le marquage du visage de manière efficace. Il permet d'obtenir le marquage du visage, et donc la possibilité de supprimer le fond de la vidéo.

Pour chaque point du marquage du visage, deux histogrammes sont réalisés durant toute la durée de la vidéo. Le premier est l'histogramme du flow optique, et le second est l'histogramme "motion boundary". Le premier permet d'étudier la variation des pixels autour d'un point de marquage, et le second comment il se déplace.

Par la suite, pour chaque point, un uGMM "universal Gaussian Mixture Model" est entraîné à partir d'un des deux histogramme.

Les trois techniques d'analyse de noyau qui sont testées : Explicit Mapping Based Dynamic Kernel, Probability Based Dynamic Kernel, Matching Based Dynamic Kernel.

Les trois techniques utilisent des noyaux différents pour identifier une émotion à partir des caractéristiques extraites. En fonction de la technique utilisée, la distance entre les caractéristiques sera différente.

Pour finir la classification s'effectue grâce un SVM "Support Vector Machine".

Pour calculer un score de similarité entre deux résultats, le score de Fisher est utilisé.

Pour l'entraînement, trois jeux de données sont utilisés : MMI [lien](#), BP4D [lien](#), et AFEW [lien](#)

**Résultats** En utilisant la technique "Probability Based Dynamic Kernel", obtient un résultat bien supérieur comparés aux autres méthodes en utilisant 256 composants pour le Gaussian Mixture Modèle. Pour chaque jeu de données, le modèle dépasse l'état de l'art actuel.

Pour le jeu de données MMI, les caractéristiques extraites avec l'histogramme "motion boundary" obtiennent un meilleur résultat qu'avec l'histogramme du flow optique. Résultat de 73.2%.

Pour le jeu de données BP4D c'est l'inverse, l'histogramme du flow optique obtient des meilleurs résultats. Résultat de 74.5%.

Enfin, pour le dernier jeu de données AFEW, les meilleures caractéristiques sont extraites via l'histogramme "motion boundary". Résultat de 56.9%.

L'avantage premier de cette méthode est la rapidité d'exécution de l'algorithme.

#### A.4.2 Facial Expression Recognition Using Residual Masking Network [[PVT21](#)]

Publication : 10 Janvier 2021

**Introduction** Partant du constat que la détection du marquage du visage via une représentation géométrique ne fonctionne bien que dans des conditions contrôlées.

Pour améliorer la classification de l'expression, ils proposent un réseau de neurones convolutionnel utilisant des sous-réseaux UNET pour segmenter les zones les plus importantes pour la détection du visage (comme les yeux, les sourcils, la bouche) et qui est beaucoup moins sensibles aux changements de conditions lumineuse, ou d'angle de prise de vue.

Ce modèle fonctionne à partir d'une seule image.

Le code et les poids du réseau sont disponible ici :  
[github.com/phamquiluan/ResidualMaskingNetwork](https://github.com/phamquiluan/ResidualMaskingNetwork)

**Jeux de données** Pour l'entraînement et les tests du modèle, deux jeux de données sont utilisés.

**FER2013** [\[BZCFZ16\]](#) Ce jeu de données ([lien](#)) contient 32298 images d'expressions faciales de 48\*48 pixels en noir et blanc et sont classées en 7 catégories obtenues par crowdsourcing.

**VEMO** Jeu de données privé introduit avec ce papier, contient deux parties. Une première partie contenant 36470 images d'expressions faciales classées en 7 catégories extraites depuis Youtube, Google Image, et Flickr. La classification des visages fut réalisée par un jury de 10 personnes qui ont votés sur chaque images pour définir la catégorie de l'émotion. La deuxième partie contient 30000 images classifiées par un algorithme de classification automatique [\[BQSM16\]](#).

**Architecture** Le modèle se découpe en 4 blocs nommés *Residual Masking Block*. C'est blocs se suivent en

sont composé d'un sous ensemble de modules similaire à un Unet. Chaque *Residual Masking Block* réduit la taille spatiale des features extraites. Un module d'attention à la fin de chaque *Residual Masking Block* contrôle que les données en sorties contiennent les zones importantes qu'il y avait en entrée pour ne pas perdre d'information. La classification finale est réalisée avec une couche densément connectée contenant 7 sorties pour les classes de sortie.

**Résultats** Le modèle obtient des performances de l'ordre de 74.12% sur le jeu de données FER2013. Comparé à d'autres techniques plus classique, le réseau obtient de meilleur résultat en augmentant énormément la taille du réseau pour obtenir ces performances. Avec 142.9 millions de paramètre, le modèle obtient seulement 0.75% de précision en plus que l'avant dernier modèle qui contient seulement 28.5 millions de paramètres. Sur le jeux de données proposés (VEMO), le modèle surpassé différentes méthodes testées et obtient une précision de 65.94%.

#### A.4.3 Frame Attention Networks for Facial Expression Recognition in Videos [\[MPWQ19\]](#)

Publication : Septembre 2019

**Introduction** Dans le papier, un nouveau modèle classifiant les émotions à partir de vidéo est présenté. Nommé FAN pour Facial Attention Networks, le modèle permet

de prendre en entrée une vidéo de taille déterminée, et est capable de sélectionner automatiquement les meilleures caractéristiques pour classifier l'émotion.

Code : [github.com/Open-Debin/Emotion-FAN](https://github.com/Open-Debin/Emotion-FAN)

**Approche générale** Le modèle présenté se découpe en deux parties. La taille des vidéos est fixé à 48 images. Les vidéos sont dans un premier temps modifié pour extraire que la zone du visage de la personne.

**L'encodeur (Feature Embedding Module)** La première partie est un encodeur d'image pour obtenir les caractéristiques focalisé sur chaque image de la vidéo. Il s'agit du Convolutional Neural Network "ResNet-18" pré-entraîné sur les jeux de données MSCeleb-1M et FER+.

D'autre encodeur furent testés : "VGGFace Classifier" et "VGGFace End2End", mais obtiennent des résultats moins intéressants.

**Le décodeur (Frame Attention Module)** Le module est partagé en deux sous parties : Un module de self-attention et un module de relation-attention. Le premier module est appliqué sur chaque vecteur de features extraites pour chaque image, puis est concaténé avec chaque résultat pour toute les images pour créer un contexte global. Le deuxième module vient calculer l'attention des features d'une image par rapport à l'ensemble des images.

La classification est réalisée par un Fully Connected

Layer de deux couches (512 puis 1024 puis le nombre de classe).

**Jeux de données** Deux jeux de données sont utilisés pour entraîner et tester le modèle : [CK+](#) et [AWEF](#).

**Résultats** Le modèle obtient un résultat de 99.69% sur le jeu de données CK+ et 51.18% sur le jeu de donnée AWEF. Plusieurs modèles d'encodeur furent testés dans le papier, mais l'encodeur ResNet-18 offre les meilleurs résultats.

#### A.4.4 POSTER - A Pyramid Cross-Fusion Transformer Network for Facial Expression Recognition [[ZMC22](#)]

Publication : 8 avril 2022

**Introduction** Introduction d'un réseau de neurones convolutionnel utilisant un cross-fusion transformers réalisant le marquage du visage en plus d'extraire des features des images. Le tout est utilisé dans un transformers pour améliorer la détection de l'expression du visage.

Les objectifs visés sont :

- Similitude interclasse : Améliorer la précision du réseau en utilisant le marquage du visage.
- Divergence intra-classe : Améliorer la détection d'une même classe pour tout type de personne (âge, sexe, etc...).

- Sensibilité à l'échelle : Améliorer la détection de l'émotion avec des qualités d'images différentes et des résolutions différentes.

Ce modèle fonctionne à partir d'une seule image.

Le code du réseau est disponible ici : [gi-thub.com/zczcwh/POSTER](https://github.com/zczcwh/POSTER)

## Jeux de données

**RAF-DB [LDD17]** Ce jeu de données contient 29672 images d'expressions faciales avec le marquage du visage. Il s'agit d'images simples et non de séquences d'images. Le jeu de données fut extrait de réseaux sociaux pour obtenir des portraits puis les données furent constituées à partir de crowdsourcing et de modèle prédictif existant. L'émotion est enregistrée sous forme de vecteur avec des probabilités pour 7 classes différentes.

**FER+ [BZCFZ16]** Ce jeu de données est basé sur le jeu de données **FER** et contient 32298 images d'expressions faciales classées en 7 catégories obtenues par crowdsourcing et les émotions sont classées sous forme de probabilité.

**AffectNet [MHM17]** Ce jeu de données est composé de 1 million de portraits avec une émotion associée récupérés sur internet avec des mots clés. Une sous partie du jeu de données composée de 420299 portraits fut annotée manuellement.

**Architecture** Le modèle se découpe en deux parties : une partie sur la détection du marquage facial, et une partie sur l'analyse de l'image. Leur intuition est que la détection du marquage facial permettrait une meilleure similitude interclasse car la position des points sera similaire comparée à des caractéristiques d'image selon l'âge ou le sexe. Par la suite, les caractéristiques de deux parties passent à travers un module d'attention où les clés du module d'attention sont les caractéristiques de la branche différente. Leur intuition est que cela permet aux réseaux de mieux reconnaître certaines zones et donc d'améliorer la reconnaissance du contexte et donc de l'émotion. *Specifically, by performing this operation, we enable the image features to be guided by some prior knowledge of salient regions from the landmarks. Likewise, the representations of the landmark stream are provided with global context from the image features while moving through the block operations.* Ce module est réalisé à plusieurs échelles/résolutions. La sortie finale est un MLP avec le nombre d'émotions à reconnaître.

**Résultats** Le modèle obtient des résultats sur la moyenne de chaque catégorie, une précision qui dépasse l'état de l'art actuel avec : 86.03% sur le jeu de données RAF-DB, 63.34% sur le jeu de données AffectNet et 91.62% sur le jeu de données FER+. Le modèle montre une réelle évolution car les résultats sont de l'ordre  $\pm 3\%$  en fonction du jeu de données.

Bien que le modèle montre une réelle amélioration des résultats, celui-ci contient 71.8 millions de paramètres ce qui rend son exécution assez lente.

## A.4.5 Recognition of Emotion Intensities Using Machine Learning Algorithms : A Comparative Study [MSJ19]

Publication : avril 2019

**Introduction** Dans la détection d'émotion, il existe deux types de données sources. Le type "posé" et le type "spontané". Dans le premier cas, la personne joue le type d'émotion demandé et est prise en photo. Dans le type spontané, la personne réagit en direct à un évènement et c'est cette image associée à la bonne émotion qui sera retenue.

La détection d'émotion peut être utilisée dans de nombreux contextes : sécurité avec identification via reconnaissance d'émotion, détection d'autisme, système de recommandation, et retour automatique de feedback utilisateur sur un produit ou service en temps réel.

**Approche générale** Il existe deux grandes familles de méthode. La première fonctionne sur une image seule, et la deuxième sur des séquence d'images. Dans le papier, l'approche est d'utiliser des séquences d'image et de réaliser le processus en temps réel.

Un large nombre de jeux de données existe. Spécialisé avec différentes ethnicités, sexes, et outils de capture différent. Certains jeux proposent des visages avec expressions spontanées, des visages avec expressions posées, ou avec les deux.

Dans la plupart des travaux de recherche, 5 expressions ont reçu beaucoup d'attention car plus simples à détec-

ter : la joie, la tristesse, la peur, le danger, et le neutre. Reconnaître l'expression du visage d'un individu revient à décrire l'émotion humaine, c'est pourquoi cette tâche est compliquée.

Les expressions du visage sont classées en Action Unit (AU). Chaque Action Unit est classifiée d'un rang de 0 pour Neutre à 5 pour une expression extrêmement marquée. Par exemple, un sourire est une Action Unit de niveau extrême qui permet de renforcer la classification pour la joie par exemple. Les Action Units permettent de décrire un visage à un instant T. Que ce soit un sourire, les yeux plissés, ou des joues relevées, tout cela permet d'effectuer une meilleure classification.

**Résultats** Pour l'extraction des caractéristiques, trois algorithmes sont testés : Gabor Features, Histogram of Oriented Gradients, et Local Binary Pattern. Pour la classification des caractéristiques, trois algorithmes sont testés : Support Vector Machine, Random Forest, et k-Nearest Neighbors.

Chaque image est transformée, et cadrée sur le visage de la personne, ensuite un des trois algorithmes d'extraction de caractéristiques est appliqué. Avec chaque caractéristique d'une suite d'images, un algorithme (ici Laplacian eigenmap algorithm) vient réduire les dimensions à travers la suite d'images pour permettre une classification en temps réel. Avec les caractéristiques finales, un algorithme de classification est appliqué et le résultat obtenu. Les résultats des différents tests sont disponibles dans le tableau A.2.

Les meilleurs résultats obtenus utilisent la méthode Lo-

#	Feature Type	KNN	RF	SVM
1.	HOG Feature	68.64	71.95	88.62
2.	Gabor Wavelets	82.13	85.6	87.20
3.	LBP	92.11	96.33	<b>97.16</b>

TABLE A.2 – Résultats des méthodes testées

cal Binary Pattern + Support Vector Machine. Les résultats surpassent l'état-de-l'art actuel du papier mais il faut prendre en compte que lors d'inférence avec le modèle, le capteur, les conditions lumineuses, les rotations et déplacements du visage peuvent dégrader les performances du réseau.

#### A.4.6 Spatio-Temporal Facial Expression Recognition Using Convolutional Neural Networks and Conditional Random Fields [HM17]

Publication : Juin 2017

**Introduction** Présentation d'une nouvelle architecture de reconnaissance d'émotion faciale à partir d'une vidéo d'un visage. Les techniques existantes reposent sur des SVM ou des classifiants Bayesien ne sont pas adaptés pour les vidéos alors que les réseaux neuronaux montrent des résultats intéressants dans d'autres domaines.

**Approche générale** Pour l'entraînement et le test des modèles, les jeux de données utilisés sont : Ck+ [LCK<sup>+</sup>10], MMI [PVRM05], FERA [BMS11].

Ils découpent le problème en deux : Une première partie sur l'analyse de chaque image, puis dans une seconde partie ils analysent les relations entre les caractéristiques extraites au cours du temps entre les images voisines.

Pour la première partie, ils utilisent un réseau neuronal convolutif résiduel (Inception-ResNet). Pour la seconde partie, ils utilisent un Conditional Random Fields (ou CRFs), à l'inverse d'un LSTM, celui-ci permet de connaître le passé et le futur car il prend en compte toute la durée de la vidéo et donc d'améliorer la classification.

Les deux parties sont entraînées séparément due au Conditional Random Fields. Au vu de la petite taille des jeux de données, les réseaux sont entraînés utilisant la technique de validation croisée K-Fold (ici k=5).

**Résultats** Pour tester la précision du réseau, ils ont entraînés une version avec une couche softmax en fin de la première partie pour une analyse uniquement sur une image, et ils ont entraînés le réseau avec le Conditional Random Fields pour montrer les performances que celui-ci permet d'apporter.

De plus, ils ont entraînés les modèles sur chaque jeu de données pour évaluer les modèles puis ils ont mélangés les données des 3 jeux de données et évalués les performances sur chaque dataset. Il faut prendre en compte que dans l'entraînement de jeux de données croisés, il est noté que l'entraînement des méthodes "state-of-the-art" n'utilise pas exactement les mêmes données pour l'entraînement que leur modèle.

On peut voir dans le tableau A.3 que le module Conditional Random Fields permet d'améliorer grandement les

	InceptionResNet With CRF	InceptionResNet Without CRF	State of the art
CK+	93.04%	85.77%	93.6% [ <a href="#">ZMM15</a> ]
MMI	78.68%	55.83%	86.7% [ <a href="#">SGM09</a> ]
FERA	66.66%	49.64%	56.1% [ <a href="#">LLS<sup>+</sup>15</a> ]

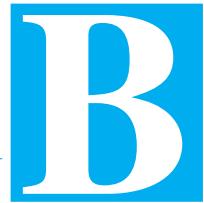
TABLE A.3 – Résultats des entraînements par jeux de données

résultats du réseaux (par rapport à la version sans ce module) cependant on peut voir que les résultats ne dépasse pas l'état de l'art. La raison expliqué dans le papier est que le réseau développé est trop profond et contient trop de paramètre pour le peu de données disponible.

On peut voir dans le table A.4, les résultats de la méthode d'entraînement avec les jeux de données croisés. Les résultats sont bien inférieur comparés à chaque modèle sur chaque jeux de données, cependant on peut noter que le modèle avec le module CRF à le meilleur taux de reconnaissance des expressions du visage sur chaque jeux de données.

	InceptionResNet Without CRF	InceptionResNet With CRF	State of the art
CK+	73.91%	64.81%	64.2% [ <a href="#">MCM16</a> ]
MMI	68.51%	52.83%	66.9% [ <a href="#">ZMM15</a> ]
FERA	53.33%	47.05%	39.4% [ <a href="#">MCM16</a> ]

TABLE A.4 – Résultats des entraînements par jeux de données croisés



---

# Planification

Dans cette partie sont détaillés les différentes planifications ainsi que les diagramme de Gant associés. Les diagrammes sont réalisés et générés grâce au logiciel open-source Gant Project.

## B.1 Phase 1

La figure B.1 présente le planning prévisionnel sur la phase 1. Tandis que la figure B.2 présente le planning réel de l'avancement du travail sur la phase 1.

La principale différence entre la planification prévisionnelle et réelle réside en le temps passé sur l'étude bibliographique des caractéristiques perceptuelles qui est intervenue plus rapidement que prévu. Notre projet consistant en l'intégration de ces caractéristiques dans la génération de résumé vidéo, il nous a paru essentiel de nous familiariser rapidement avec ces notions, en lisant des articles de recherche sur les émotions, la mémorabilité, ainsi que sur la congruence visuelle inter-observateurs. La partie de test sur des algorithmes exis-

tant vient prolonger l'étude bibliographique mais est très impactée par le fait d'avoir les algorithmes disponibles et leurs facilités d'utilisation (notamment si l'environnement de développement n'est pas indiqué et que le projet date de plusieurs années). En conséquence, nous n'avons pas pu tester des modèles de congruence visuelle inter-observateurs et de modèle de résumé vidéo par manque de temps et de modèle disponible.

Lors de l'étude bibliographique, nous nous sommes rendus compte du temps important dédié à la lecture d'articles de recherche pour bien comprendre l'article mais aussi pour s'intéresser aux références afin d'élargir le sujet.

## B.2 Phase 2

La figure B.3 présente le planning prévisionnel sur la phase 2. Tandis que la figure B.2 présente le planning réel de l'avancement du travail sur la phase 2.

Nous avons pris du retard sur les tests de modèle

puisque nous n'avions pas à notre disposition de modèle d'inférence de la congruence visuelle inter-observateurs. Pour le modèle de Résumé vidéo, nous avons privilégié dans un premier temps le test et l'inférence sur les caractéristiques que nous cherchons à extraire avant de tester ce modèle ci. De plus les auteurs du modèle AC-SUM-GAN n'ont pas mis à disposition leur modèle pré-entraîné. Nous avons donc laisser de côté cette partie puisque nous avions déjà la vérité-terrain de chaque base de données (scores d'importances de chaque image de chaque vidéo, annotés par des utilisateurs)

Néanmoins, nous avons pu travailler davantage et plus tôt sur le clustering dans le but de trouver des profils-types de vidéos, que ce soit avec la réduction de dimensions via divers méthodes ou bien en testant différentes solutions pour comparer les vidéos de tailles différentes.

Suite à divers problème rencontrés avec le cluster GPU proposé par notre commanditaire, nous avons pris la décision d'exécuter les différents modèles sur nos ordinateurs. Le plus gros problème fut que la machine virtuelle s'arrête automatiquement au bout d'un certain nombre de minutes même si un programme est en cours d'exécution sur le serveur.

Suite à ce délai concernant l'inférence des caractéristiques, nous avons commencé le travail des différents tests uniquement quand les données furent disponible, c'est-à-dire au fur et à mesure des inférences réalisées. A savoir : les tests de corrélation, d'indépendance et de similarité.

Néanmoins, nous sommes rentrés dans les temps pour pouvoir réaliser les modèles de prédiction de score d'im-

portance et de prédiction de frames sélectionnées.

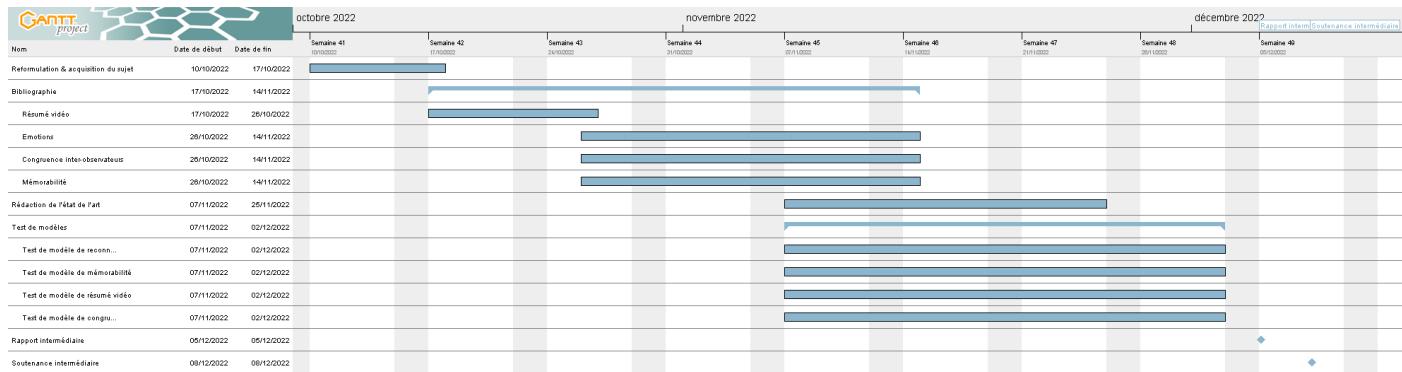


FIGURE B.1 – Phase 1 - Planification prévisionnelle

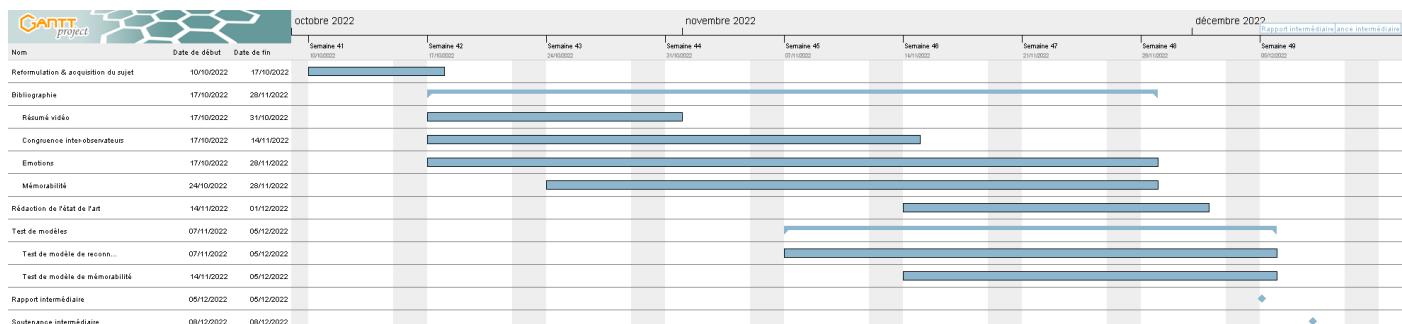


FIGURE B.2 – Phase 1 - Planning effectif

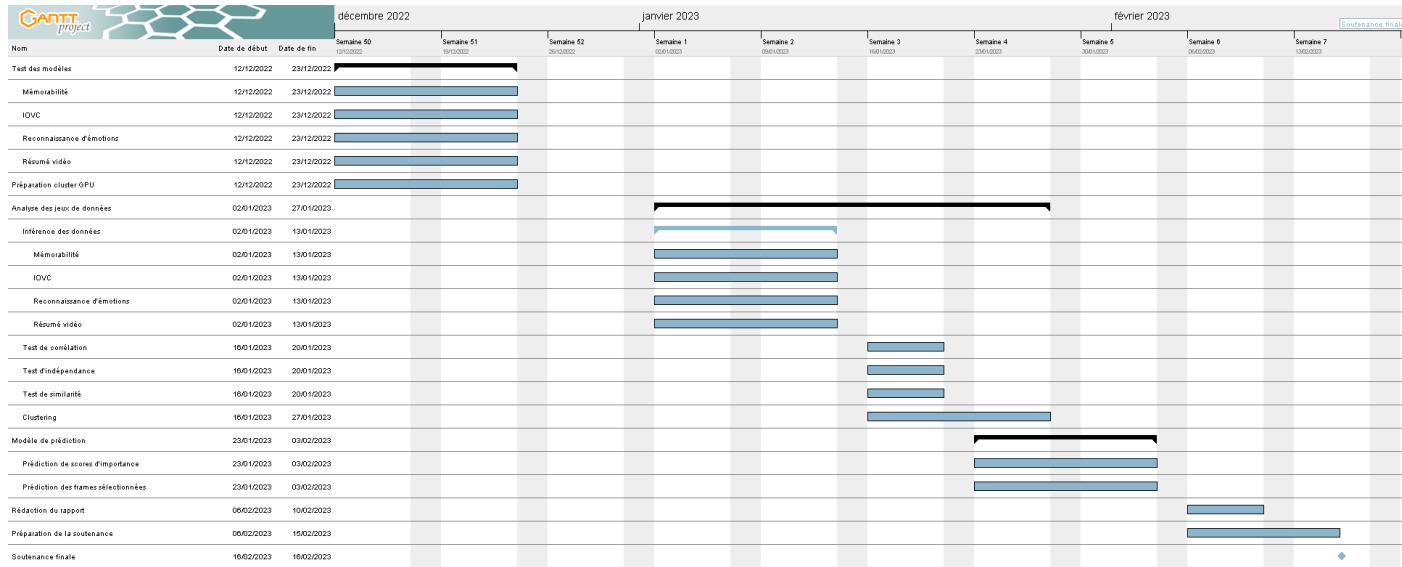


FIGURE B.3 – Phase 2 - Planification prévisionnelle

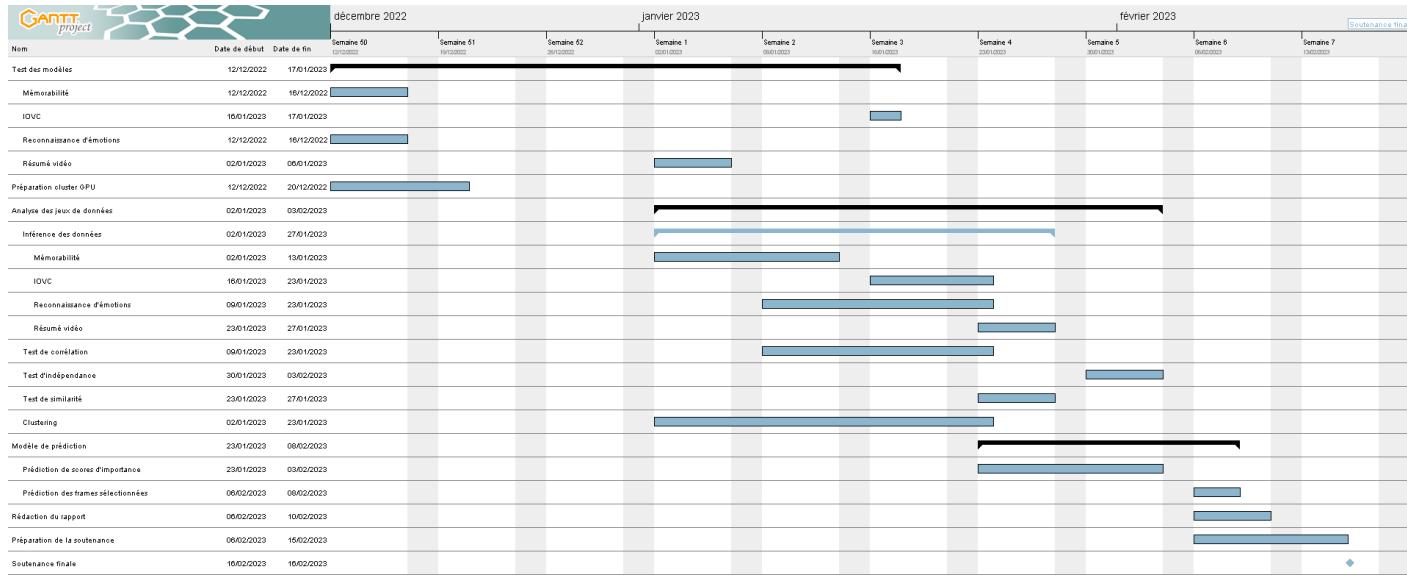


FIGURE B.4 – Phase 2 - Planning effectif



# Fiches de suivi

---

## Fiche de suivi de la semaine 1 du 10 octobre 2022 au 15 octobre 2022

---

Temps de travail de Josik SALLAUD: 10 h 0 m

Temps de travail de Nathan ROCHER: 9 h 0 m

### Travail effectué.

- Lecture des travaux de recherche fournis par le commanditaire : 1 / 6

*Video Summarization Using Deep Neural Networks : A Survey* [[AAM<sup>+</sup>21b](#)]

- Rédaction d'un premier rapport lié au contexte et à la problématique

### Échanges avec le commanditaire.

- Réunion le 11/10/2022 à 10h

Question : En quoi consiste le test subjectif qui sera à concevoir ?

Réponse :

C'est une façon d'évaluer une méthode ou un algorithme qui fait appel à la subjectivité donc à l'humain dans un certain contexte car il est compliqué

d'avoir une métrique pour évaluer, par exemple, la qualité d'une vidéo. Dans notre cas il s'agira de juger de la qualité et de l'attractivité du résumé vidéo. Ce test sera défini en fonction de l'idée que l'on se fait d'un résumé vidéo.

Deux possibilités :

1. Un résumé vidéo doit être informatif et doit reprendre les informations importantes de l'ensemble de la vidéo.
2. Un bon résumé vidéo donne envie de voir la vidéo complète et doit être attractif.

Dans le premier cas, on pourrait imaginer un test impliquant que l'utilisateur aura vu et pris connaissance de la vidéo avant la réalisation du test. Cet utilisateur devra juger parmi un ensemble de résumés vidéo celui qui est le plus informatif avec un classement.

Dans le second cas, on pourrait demander aux utilisateurs de regarder des résumés vidéo puis de donner une note ou un classement des résumés pour

savoir lesquels sont les plus attractifs.

### Planification pour la semaine prochaine.

- Continuer la lecture des travaux de recherche fournis par le commanditaire
- Écrire des fiches de lectures sur les différents travaux de recherche
- Trouver des travaux de recherche en lien avec notre sujet

---

### Fiche de suivi de la semaine 2 du 17 octobre 2022 au 21 octobre 2022

---

Temps de travail de Josik SALLAUD: 10 h 30 m

Temps de travail de Nathan ROCHER: 10 h 0 m

#### Travail effectué.

Lecture et rédaction de fiche pour les articles :

- Video Summarization Using Deep Neural Networks
- Inter-Observer Visual Congruency in Video-Viewing
- Prediction of the Inter-Observer Visual Congruency (IOVC) and Application to Image Ranking
- Deep Learning For Inter-Observer Conguency Prediction

#### Échanges avec le commanditaire.

— Réunion le 18/10/2022 à 11h

— Sélectionner les caractéristiques (haut-niveau) à garder : - non négociable : congruence inter-observateurs - à faire aussi : mémorabilité - pourquoi pas : expressions faciales et reconnaissances d'émotions - on abandonne pour le moment : multi-modale avec de l'audio (piste audio au lieu des frames)

— L'idée n'est pas ajouter les caractéristiques sur les modèles. Mais évolution des caractéristiques des modèles et les corréler avec les caractéristiques perceptuelles extraites. Construire un modèle naïf avec caractéristiques perceptuelles qu'on a. **Preuve de concept** : preuve que l'on fait mieux que le hasard. Ouverture sur l'amélioration des modèles d'évaluation de résumés vidéo avec caractéristiques perceptuelles

— L'évaluation des résumés n'est pas bien définie et est difficile et subjective.

— Articles sur les caractéristiques perceptuelles auxquelles il faut s'intéresser pour lier avec la génération de résumé vidéo :

— Expression faciale : [[PRC20](#), [MPWQ19](#), [HM17](#), [ZX18](#)]

— Mémorabilité : [[KRT015](#), [KYP13](#), [SHGKD18](#), [DHKC<sup>+</sup>20](#)]

— Congruence inter-utilisateurs : [[LMBR11](#), [RB16](#), [BLCLM19](#), [YLZ<sup>+</sup>21](#)]

— 1ère lecture rapide sur les articles : problème + comment il le résout + globalement comment il

fait.

- **Modèle de fiche de lectures** proposé par M. Bruckert :

- Contexte
  - Problème abordé
  - Approche / principe général
  - Résultats principaux
  - (critiques / limites du papier optionnellement)
- Quelques lignes (ou des bullet points) dans chaque section suffisent.

#### **Planification pour la semaine prochaine.**

- Continuer la lecture des articles
- Rédiger une fiche de lecture par article

---

#### **Fiche de suivi de la semaine 3 du 24 octobre 2022 au 28 octobre 2022**

---

Temps de travail de Josik SALLAUD: 17 h 15 m

Temps de travail de Nathan ROCHER: 17 h 00 m

#### **Travail effectué.**

- Rédaction d'un premier jet de l'état de l'art dans le rapport
- Recherche de code pour certains articles
- Lecture et rédaction de fiche pour les articles :
  - Recognition of Emotion Intensities Using Machine Learning Algorithms : A Comparative

Study

- POSTER : A Pyramid Cross-Fusion Transformer Network for Facial Expression Recognition
- Factors underlying Inter-observer agreement in gaze patterns : Predictive modelling and analysis
- Understanding and Predicting Image Memorability at a Large Scale
- Relative Spatial Features for Image Memorability
- Memorability of natural scenes : the role of attention
- Facial Expression Recognition in Videos Using Dynamic Kernels
- Frame Attention Networks for Facial Expression Recognition in Videos

#### **Échanges avec le commanditaire.**

- Réunion le 25/10/2022 à 17h
- Mesure de l'IOC (congruence inter-observateurs) : méthode leave one out = mesure pour chaque observateur en se basant sur les zones de fixation de regard de tous les autres observateurs, donne la capacité d'un observateur à prédire le cas général, ou inversement, comment le cas général se distille.
- On devrait avoir des données (user-generated content) à la rentrée pour commencer.
- Pour nous : ne pas hésiter à lire des articles annexes pour plus de papiers.

- Faire un tour sur les papiers de résumé vidéo : se baser sur la review globale : regarder les modèles les plus influents et les plus intéressants et essayer de trouver du code.

### **Planification pour la semaine prochaine.**

- Continuer la lecture des articles
- Rédiger une fiche de lecture par article
- Trouver des sources de code de différents modèles
- Point à la rentrée le mardi 8 novembre

vidéos

- Écriture de la bibliographie

### **Échanges avec le commanditaire.**

- Réunion le 08/11/2022 à 16h30
- Discussion autour de l'extraction des caractéristiques des vidéos
- Deux jeux de données contenant des vidéos :
  - SumMe [link](#)
  - VSUMM [link](#)

### **Planification pour la semaine prochaine.**

- Continuer la lecture des articles
- Rédiger une fiche de lecture par article
- Trouver des sources de code de différents modèles
- Extraire des caractéristiques des vidéos

---

### **Fiche de suivi de la semaine 4 du 07 novembre 2022 au 11 novembre 2022**

---

Temps de travail de Josik SALLAUD: 11 h 00 m

Temps de travail de Nathan ROCHER: 15 h 00 m

#### **Travail effectué.**

- Lecture et rédaction de fiche pour les articles :
  - Spatio-Temporal Facial Expression Recognition Using Convolutional Neural Networks and Conditional Random Fields
  - Deep Learning for Predicting Image Memorability
- Création du Git pour le projet
- Recherche de code pour certains articles
- Utilisation du code de certains papiers pour commencer à extraire des caractéristiques à partir des

---

### **Fiche de suivi de la semaine 5 du 14 novembre 2022 au 18 novembre 2022**

---

Temps de travail de Josik SALLAUD: 12 h 00 m

Temps de travail de Nathan ROCHER: 15 h 00 m

#### **Travail effectué.**

- Lecture et rédaction de fiche pour les articles :
  - Deep Learning for Image Memorability Prediction : the Emotional Bias
  - Test de modèles de prédiction de la mémorabilité d'image

## Échanges avec le commanditaire.

- Réunion le 15/11/2022 à 15h
- Lien vers modèle LaMem <http://memorability.csail.mit.edu/download.html>
- Deux liens vers reconnaissance faciale <https://github.com/minhnhatvt/glamor-net> et <https://github.com/pedrodiamei/ferattention>
- Nous n'avons pas trouvé de sources de modèles de prédiction de l'IOVC. M. Bruckert nous enverra dès que possible ces modèles.
- Base UGC de Ytb, 20 secondes : <https://media.withyoutube.com/ugc-dataset>
- Explication mémorabilité pour compléter la bibliographie : [https://link.springer.com/chapter/10.1007/978-3-030-81465-6\\_8#Sec18](https://link.springer.com/chapter/10.1007/978-3-030-81465-6_8#Sec18)
- Tester les modèles de mémorabilité sur les bases de données de résumés pour voir les résultats
- Soutenance le jeudi 8 décembre à 17h

## Planification pour la semaine prochaine.

- Lecture d'articles sur des méthodes de résumé vidéo
- Rédiger une fiche de lecture par article
- Continuer les tests de modèles de prédiction des caractéristiques (mémorabilité, IOVC, émotion)
- Test de modèles de résumé vidéo
- Finir la bibliographie
- Faire une présentation (diapo) pour résumer l'état de l'art afin de pour préparer la soutenance

- Réunion 14h mardi prochain

---

## Fiche de suivi de la semaine 6 du 22 novembre 2022 au 25 novembre 2022

---

Temps de travail de Josik SALLAUD: 22 h 00 m

Temps de travail de Nathan ROCHER: 15 h 30 m

### Travail effectué.

- Lecture et rédaction de fiche pour les articles :
- Facial Expression Recognition Using Residual Masking Network
- Rédaction du rapport intermédiaire
- Début du travail sur la présentation

### Échanges avec le commanditaire.

- Réunion le 23/11/2022 à 15h
- Réunion concernant la soutenance de mi-projet
- Idée du plan :

- Contexte (à quoi servent les résumés vidéo, pourquoi c'est intéressant pour les UGC...)
- État de l'art du résumé vidéo (il nous manque l'aspect critique des articles)
- Montrer comment ajouter nos différentes caractéristiques pour améliorer la prédiction des résumés
- État de l'art des caractéristiques (il nous manque l'aspect critique des articles)
- Points

- positifs et négatifs
- Expliquer notre façon de faire : Regarder comment les métriques évoluent pour effectuer le travail de résumer les vidéos

Exemple d'aspect critique : - Regarder les données utilisées puis faire des critiques pour les modèles - Résumé vidéo : vidéos sans diversité, vidéo courte, de mauvaise qualité DONC généralisation faible et contenu différent

#### **Planification pour la semaine prochaine.**

- Terminer la rédaction du rapport
- Faire une présentation (diaporama) pour préparer la soutenance
- Réunion à 14h mardi prochain

- Discussion et retours sur une V1 du rapport intermédiaire avec des commentaires pour améliorer les différentes parties développées

#### **Planification pour la semaine prochaine.**

- Terminer la présentation
- Préparer la soutenance

---

#### **Fiche de suivi de la semaine 8**

**du 5 décembre 2022 au 9 décembre 2022**

---

Temps de travail de Josik SALLAUD: 12 h 30 m

Temps de travail de Nathan ROCHER: 10 h 00 m

#### **Travail effectué.**

- Travail sur la présentation
- Répétition de la soutenance
- Soutenance le 08/12/2022 à 17h

#### **Échanges avec le commanditaire.**

- Réunion 06/12/2022 à 14h
- Evocation de SVM pour l'entraînement sur les caractéristiques extraites
- Discussion autour de la soutenance et retours sur notre présentation
- Discussion pour la semaine prochaine : voir si on a besoin de GPU car possibilités d'avoir accès à des GPU pour notre projet.

#### **Planification pour la semaine prochaine.**

---

#### **Fiche de suivi de la semaine 7**

**du 28 novembre 2022 au 2 décembre 2022**

---

Temps de travail de Josik SALLAUD: 17 h 30 m

Temps de travail de Nathan ROCHER: 20 h 00 m

#### **Travail effectué.**

- Rédaction du rapport intermédiaire
- Travail sur la présentation
- Fiche de lecture : Embracing New Techniques in Deep Learning for Estimating Image Memorability

#### **Échanges avec le commanditaire.**

- Voir les différentes bases de données de résumé vidéo et leur qualité (biais émotionnel par ex.)
- Test de modèles
- Construction d'un modèle d'apprentissage

récupérer le jeu de données car les liens ne fonctionnent plus

Réponse : L'auteur nous a répondu qu'il ne peut plus partager le jeu de données car les vidéos ne sont plus publiques et ne sont plus disponibles.

### **Échanges avec le commanditaire.**

- Réunion 15/12/2022 à 9h30
- Dans un premier temps, analyser les données (caractéristiques inférées et vérité-terrain). Cela permettra de voir s'il y a des profils types de vidéos afin d'analyser les différents jeux de données à notre disposition. Une bonne base serait variée en terme de profils. Pour retrouver ces profils, on applique un algorithme d'apprentissage automatique de clustering (K-Means par exemple) sur les données de plusieurs vidéos. Nos données = pour chaque vidéo, on a une courbe d'évolution des caractéristiques inférées et de l'importance des frames de la vidéo.
- Ensuite l'idée serait de voir si les profils types de vérité-terrain sont les mêmes que pour les caractéristiques à inférer (mémorabilité, congruence inter-observateurs, intensité émotionnelle).

### **Planification pour la semaine prochaine.**

- Temps de travail de Josik SALLAUD: 10 h 00 m
- Temps de travail de Nathan ROCHER: 10 h 00 m
- Travail effectué.**
  - Accès au GPU de l'université pour nos développements
  - Sur la base TVSum comprenant des vidéos de quelques minutes avec annotations de 20 observateurs ayant donné une note d'importance de chaque frame des vidéos (<https://gitub.com/yalesong/tvsum>)
  - Développement d'un script Python permettant de tester la corrélation de la base TVSum, entre mémorabilité (inférée via ResMem) et vérité-terrain (importance des frames)
  - Développement d'un script permettant d'appliquer un algorithme de clustering (K-Means)
  - Téléchargement des jeux de données sur le serveur
  - Contact avec l'auteur de "MED Summaries" pour

- Tester différentes bases de données de résumé vidéo

---

## **Fiche de suivi de la semaine 10 du 2 janvier 2023 au 6 janvier 2023**

---

Temps de travail de Josik SALLAUD: 7 h 00 m

Temps de travail de Nathan ROCHER: 6 h 30 m

### **Travail effectué.**

- Inférence des scores de mémorabilité sur les vidéos de TVSum avec ResMem
- Clustering K-Means : sur les scores de mémorabilité des vidéos de TVSum (inférés par ResMem), ainsi que sur les scores d'importance selon des observateurs. Affichage de 3 profils-types pour la mémorabilité et 3 profils-types pour la vérité-terrain.

### **Échanges avec le commanditaire.**

- Nous avons convenu d'une réunion mardi 10 janvier à 14h.

### **Planification pour la semaine prochaine.**

- Réunion mardi 10 janvier à 14h.
- Génération de résumé vidéo pour trouver des profils-types
- Inférence de l'IOVC pour trouver des profils-types
- Test sur d'autres bases de données que TVSum

---

## **Fiche de suivi de la semaine 11 du 9 janvier 2023 au 13 janvier 2023**

---

Temps de travail de Josik SALLAUD: 16 h 30 m

Temps de travail de Nathan ROCHER: 12 h 30 m

### **Travail effectué.**

- Inférence des scores de reconnaissance d'émotions sur les vidéos de TVSum avec ResidualMaskingNetwork
- Travail sur la linéarisation et réflexion sur l'intégration des résultats de la reconnaissance des émotions
- Mémorabilité des vidéos de TVSum : bilan sur les corrélations avec la vérité-terrain & réduction de dimensions (ACP et T-SNE) puis clustering K-Means

### **Échanges avec le commanditaire.**

- Nous avons convenu d'une réunion mardi 17 janvier à 14h.
- M. Bruckert s'occupe de ré-entraîner le modèle d'IOVC pour qu'on puisse inférer les scores des images des vidéos
- Partie reconnaissance faciale : Linéariser les résultats pour éviter les faux positifs
- Partie IOVC : Sur des bases comme VSUMM qui ont seulement les images extraites pour le résumé vidéo (et non pas un score par frame) : on peut binariser un vecteur (1 si image gardée, 0 non) Puis comparer score de mémorabilité sur les images à 1 et sur les images à 0.
- Partie memorabilité :
  - Générer fichier de vérité-terrain agrégée à partir des 20 annotations d'observateurs de TVSum
  - Faire T-SNE et/ou PCA pour réduire le nombre

- de dimensions très élevé de nos données (autant de dimensions que de secondes dans nos vidéos). À faire sur les données brutes pour réduire les dimensions et donner une indication sur le nombre de clusters à utiliser. Sachant qu'il n'y a possiblement pas de cluster.
- Calculer distances au centroïde, si distances très élevées pas de cluster net = pas de profil type.
  - Si on prend l'intégralité des points de la vidéo et qu'il n'y a pas de profil type, est-ce qu'en prenant les pics ou chutes (gradients) on peut obtenir un résultat.
  - A tester : Prendre des fenêtres de taille fixe composée de N frames (30 secondes par exemple) puis faire glisser 1 par 1 ou avec un stride. Comme ça tout les points dans l'espace du clustering ne représentent pas la vidéo entière mais juste une sous-partie. On peut donc comparer des choses qui n'étaient pas comparables auparavant. Notamment des dynamiques de vidéos semblables à une autre au début, et pas à la fin. Taille de la fenêtre en fonction des courbes avec pas mal de variation mais pas trop non plus.
  - Clustering de série temporelle : nombre de pic, min max, beaucoup de méthode classique (voir semaine pro)

### Planification pour la semaine prochaine.

- Réunion mardi 17 janvier à 14h.
- Génération de résumé vidéo pour trouver des

profils-types

- Inférence de l'IOVC pour trouver des profils-types
- Test sur d'autres bases de données que TVSum

---

### Fiche de suivi de la semaine 12 du 16 janvier 2023 au 20 janvier 2023

---

Temps de travail de Josik SALLAUD: 12 h 30 m

Temps de travail de Nathan ROCHER: 12 h 30 m

#### Travail effectué.

- Refactoring du code
- Inférence des scores de mémorabilité sur la base VSumm composée de vidéos de YouTube et d'Open-Video.
- Inférence des scores d'IOVC sur la base TVSum

#### Échanges avec le commanditaire.

- M. Bruckert nous a donné le modèle entraîné de prédiction d'IOVC d'image. Perspective si on avait du temps : entraîner le modèle sur des vidéos.
- Question : Quelles caractéristiques inférer pour les visages ? Pour chaque image de la vidéo : si on détecte un visage on envoi un vecteur de 7 classes, on peut définir à zéro les 7 classes si pas de visages dans le cas de plusieurs visages : on prends la valeur maximales pour chaque classes.

Nombre de visages, Probabilités maximales de

toute les classes pour tout les visages, Moyenne des valeurs maximales (oublie du type d'émotions) On peut aussi voir pour fixer un nombre N de visages et définir un tableau de vecteur pour chaque caractéristique.

- Contrairement à la ressource donnée la semaine dernière (<https://www.kaggle.com/code/izzettunc/introduction-to-time-series-clustering>), il ne faudrait pas utiliser l'algorithme de clustering K-Means sur le nuage de points réduit avec T-SNE (des données de fenêtres de mémorabilité par ex.). D'autres métriques telles que le gradient, les pics etc. seraient préférables à utiliser. A revoir la semaine prochaine.
- Pour le serveur qui s'éteint si on ferme l'onglet de JupyterLab, c'est un paramètre qui a été mis à la création de la machine virtuelle et qui n'est pas changeable, il faut faire avec.

### Planification pour la semaine prochaine.

- Réunion jeudi 26 janvier à 9h30.
- Faire une synthèse des données que l'on a sur chaque base :
  - données inférées : mémorabilité, IOVC, émotions génération de résumé vidéo
  - vérité-terrain
- Sur ces données, faire une analyse de la variance (ANOVA) : distribution des scores inférées en fonction des vidéos, moyenne des scores
- Générer une figure par vidéo avec l'évolution de la caractéristique.

---

### Fiche de suivi de la semaine 13 du 23 janvier 2023 au 27 janvier 2023

---

Temps de travail de Josik SALLAUD: 17 h 30 m

Temps de travail de Nathan ROCHER: 16 h 00 m

#### Travail effectué.

- Fix de la génération du fichier JSON des données inférées
- Sur la base VSumm :
  - Inférence de l'IOVC
  - Inférence des émotions reconnues
- Sur la base SumMe :
  - Lecture de la vérité-terrain (format MatLab)
  - Inférence de la mémorabilité
  - Inférence de l'IOVC
  - Inférence des émotions reconnues
  - Génération de résumé vidéo avec PGL-SUM
- Sur la base TVSum :
  - Test du clustering avec la distance "Dynamic time warping" sur les données lissées d'évolution de caractéristiques (sur la mémorabilité) au cours de fenêtres vidéo (aucun résultat)
  - Génération de résumé vidéo avec PGL-SUM
  - Inférence des émotions reconnues
  - Boxplots des caractéristiques inférées (IOVC et mémorabilité) et de la vérité-terrain
  - Fichier avec les moyennes par vidéo de chaque caractéristique et de la vérité-terrain.

#### Échanges avec le commanditaire.

- Nous avons convenu d'une réunion jeudi 26 janvier

à 9h30 :

- Premier échange autour des données inférées :
  - Problème sur les données VSUMM -> YouTube -> flv (non-utilisation de ces données pour cause de problème de framerate entre la vérité-terrain et les vidéos)
  - Problème de VSUMM -> OpenVideo -> MPG : ne pas tenir compte des erreurs, la vidéo est forcément lue dans l'ordre (en MPG une frame est reconstruite à partir des précédentes car stocke seulement les modifications des pixels par rapport aux précédentes).
- Important : avoir au moins le travail fait sur une des bases
- Faire ces analyses sur TVSum et SumMe car la vérité-terrain = score d'importance de chaque frame estimé par des utilisateurs. Tandis que sur VSumm, la vérité-terrain simplement la sélection ou non des frames par les utilisateurs. Pour SumMe, la vérité-terrain est en format Matlab, pour ouvrir des fichiers Matlab depuis python : `scipy.io.loadmat`.

Analyse des bases de vidéos à faire sur les données inférées / vérité-terrain :

1. Calculer des moyennes par vidéos (voir si des vidéos sortes de l'ordinaire)
2. Boxplots pour voir des distributions pour chaque vidéo (voir si grande variance ou non) -> Important
3. Si différence de distribution : calculer des distances de Wasserstein : Faire une grande ma-

trice (sur iovc par ex ici), contenant les distances entre toutes les paires de vidéos possibles et le faire sur une seule modalité à la fois (IOVC, mémorabilité, vérité-terrain...) distance de Wasserstein permet de calculer la différence entre deux distributions (en calculant combien d'opérations il faut réaliser sur une distribution pour la transformer en une autre). Distance de Wasserstein se calcule à partir de l'histogramme (pour calculer la distance, le nombre de bins est fixé dans la lib).

4. Wasserstein T-SNE : <https://github.com/fsvbach/WassersteinTSNE> -> Pour conclure sur si les vidéos sont biaisées sur une caractéristique ou sur la vérité-terrain.
- Faire une analyse de corrélation entre vidéos, et s'il y a des vidéos qui corrèlent ou anti-corrèlent faire une analyse qualitative (pourquoi visuellement ça fonctionne ou non) et voir aussi les outliers avec une analyse qualitative aussi.
- Normalisation des images : Perte de temps sur une erreur lié à la normalisation des images
- Nos méthodes de clustering ne fonctionnent pas, que faire ? Réponse : ça fait sens, c'est logique, la variance est trop élevée, ça ne sert à rien de continuer à chercher des clusters. Pour le rapport : expliquer pourquoi ça ne donne rien :
  1. Pas assez de vidéos (point dans un espace de dimension 600 avec quelques milliers points c'est trop peu pour arriver à un résultat) même en utilisant la technique des fenêtres glissantes :

il faudrait beaucoup plus de données pour faire cette étude (d'autres bases). Diversité de la base insuffisamment importante pour permettre la réalisation de clustering.

2. Pas de logique donc pas de cluster -> Pas un problème de méthode, mais un problème de sampling.

Bien d'avoir essayé car si on avait eu des clusters reconnaissables, on aurait pu le voir et donc nous indiquer sur la formes des courbes d'une vidéo quelconque à partir des clusters.

- Dans le rapport : quand on présente les caractéristiques, on peut ajouter les graphiques d'évolution de la caractéristique sur certaines vidéos pour faire une analyse qualitative : dire à quoi correspondent les pics et creux du score de la caractéristique.

### Planification pour la semaine prochaine.

- Réunion jeudi 2 février à 9h30.
- Soutenance jeudi 16 février à 11h00.
- Continuer l'analyse des bases (distributions des caractéristiques des vidéos, distance de Wasserstein, Wasserstein T-SNE)

---

### Fiche de suivi de la semaine 14 du 30 janvier 2023 au 3 février 2023

---

Temps de travail de Josik SALLAUD: 12 h 00 m

Temps de travail de Nathan ROCHER: 12 h 30 m

### Travail effectué.

- Entraînement SVR pour prédire le score d'importance d'une frame à partir des features extraits
- Entraînement d'un réseau de neurones pour prédire le score d'importance des frames à partir des features extraits (IOVC, memorabilité, et reconnaissance d'émotions)
- Entraînement d'un réseau de neurones pour prédire le score d'importance des frames à partir de 20/50 premières features extraites des 20/50 premières frames
- Rendu des heatmaps pour les matrices de Wasserstein pour l'IOVC et la mémorabilité
- Test de Student (t-test) entre la distribution des caractéristiques sur les frames sélectionnées ou non (selon la vérité-terrain de résumé)

### Échanges avec le commanditaire.

- Plot matrice avec couleur pour voir si des lignes ou des valeurs ressortent, si une valeur ressort alors la vidéo est proche d'une autre, pour voir si des vidéos sont proches d'une autre
- Mettre en valeur les similarités entre les vidéos via les graphiques générées, assembler par écart-type
- Analyser le contenu des vidéos par rapport à leur graphique pour les vidéos qui se démarquent et pour les vidéos plus proches de la moyenne
- Faire matrice Wasserstein sur toutes les caractéristiques
- Faire calcul (corrélation, ou calcul présent dans

- le papier <https://arxiv.org/pdf/2101.06072.pdf> sur ground truth avec les résultats de PGL-SUM (déjà fait par les chercheurs sur PGL-SUM), à utiliser pour évaluer notre modèle.
- Pour le SVR : 15% ce n'est pas super mais il y a quand même une corrélation, ce n'est pas un résultat anormal,  $\text{sqrt}(0.15) = 0.38$  de corrélation. Donc c'est quand même un peu intéressant.
- Garder 5 vidéos entières pour faire du test sur les modèles
- Utiliser le svr pour extraire les images de la vidéo

### Planification pour la semaine prochaine.

- Réunion mercredi 8 février à 14h.
- Rédaction du rapport

---

### Fiche de suivi de la semaine 15 du 6 février 2023 au 12 février 2023

---

Temps de travail de Josik SALLAUD: 17 h 30 m

Temps de travail de Nathan ROCHER: 17 h 00 m

### Travail effectué.

- Rédaction du rapport
- Travail sur la présentation
- Entraînement de réseau de neurone avec LSTM
- Entraînement de réseau de neurone
- Entraînement d'un arbre de décision

- Rendu de graphique

### Échanges avec le commanditaire.

- Réunion mercredi 8 février à 14h.
- Pour la classification binaire des frames :
  - Data Inbalance : Toujours prédire la classe dominante pour la classification binaire
  - Supprimer des lignes où le true label est à 0 pour balancer les classes (pour avoir 50/50)
  - Question 1 : pourquoi c'est linéaire sur figure 4.17 Réponse : c'est étonnant, vérifier si les données sont bien séparées dans jeux de données entraînement / validation PAR VIDEO
  - Question 2 : Peut-on utiliser un lstm sur du multi-variable ? Oui a tester On aussi essayer de normaliser les données Mettre du ReLU après le LSTM Vérifier si les splits entre modèles sont les mêmes
  - Voir pour faire un NN pour classification binaire
- Question sur le test de Student, normal d'avoir autant de vidéos dont la distribution est différente entre le score sur les frames sélectionnées et non sélectionnées ? regarder les vidéos  $pvalue > 0.5$ , peut être plan fixe du coup frame intéressante au hasard.
- Les scores extraits depuis nos modèles utilisent des features qui eux même utilisent des features inférées sur des images. Les features apprises seraient-elles pas déjà les mêmes que celles dans les modèles de résumé vidéo ? Si c'est vrai on a rajouté de l'explicabilité. Sinon ça donne de nouvelles fea-

tures et c'est intéressant.

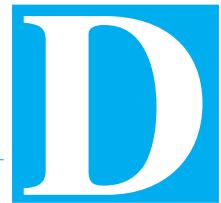
### **Planification pour la semaine prochaine.**

— Soutenance le 16 février à 11h en D005

Le tableau C.1 récapitule le taux d'avancement du projet. Rappelons que le temps de travail théorique *minimal* correspond au temps indiqué sur la maquette pédagogique auquel on ajoute un strict minimum de 20 % correspondant au travail personnel hors emploi du temps. La partie « haute » de la fourchette correspond à 50 % de temps supplémentaire au titre du travail personnel.

Semaine	Temps prévu		Josik SALLAUD				Nathan ROCHER		
	bas	haut	hebdo.	$\Sigma$	%	hebdo.	$\Sigma$	%	
	h : m	h : m	h : m	h : m		h : m	h : m		
1	10 : 00	12 : 30	10 : 0	10 : 00	100 (80)	9 : 0	9 : 00	90 (72)	
2	20 : 00	25 : 00	10 : 30	20 : 30	102 (82)	10 : 0	19 : 00	95 (76)	
3	30 : 00	37 : 30	17 : 15	37 : 45	125 (100)	17 : 00	36 : 00	120 (96)	
4	40 : 00	50 : 00	11 : 00	48 : 45	121 (97)	15 : 00	51 : 00	127 (102)	
5	50 : 00	62 : 30	12 : 00	60 : 45	121 (97)	15 : 00	66 : 00	132 (105)	
6	60 : 00	75 : 00	22 : 00	82 : 45	137 (110)	15 : 30	81 : 30	135 (108)	
7	70 : 00	87 : 30	17 : 30	100 : 15	143 (114)	20 : 00	101 : 30	145 (116)	
8	80 : 00	100 : 00	12 : 30	112 : 45	140 (112)	10 : 00	111 : 30	139 (111)	
9	90 : 00	112 : 30	10 : 00	122 : 45	136 (109)	10 : 00	121 : 30	135 (108)	
10	100 : 00	125 : 00	7 : 00	129 : 45	129 (103)	6 : 30	128 : 00	128 (102)	
11	110 : 00	137 : 30	16 : 30	146 : 15	132 (106)	12 : 30	140 : 30	127 (102)	
12	120 : 00	150 : 00	12 : 30	158 : 45	132 (105)	12 : 30	153 : 00	127 (102)	
13	130 : 00	162 : 30	17 : 30	176 : 15	135 (108)	16 : 00	169 : 00	130 (104)	
14	140 : 00	175 : 00	12 : 00	188 : 15	134 (107)	12 : 30	181 : 30	129 (103)	
15	150 : 00	187 : 30	17 : 30	205 : 45	137 (109)	17 : 00	198 : 30	132 (105)	

TABLE C.1 – Avancement du projet par rapport au temps de travail théorique minimal (respectivement haut)



---

## Auto-contrôle et auto-évaluation

La figure D.1 permet d'énumérer un certain nombre de points importants dans les trois composantes du travail :

1. rapport;
2. présentation orale;
3. travail de fond;

ainsi que d'évaluer notre niveau de satisfaction à l'issue de la phase I, composée de trois étapes :

1. étude préalable;
2. étude bibliographique;
3. conception générale.

La figure D.2 permet d'énumérer un certain nombre de points importants dans les trois composantes du travail ainsi que d'évaluer notre niveau de satisfaction à l'issue de la phase II, constituée de :

1. la conception détaillée;
2. la réalisation;
3. la recette.

Rapport	Organisation	Plan	Équilibre	
		Fluidité	Coherence Introductions (partielles)	
Rédaction	Orthographe	Tableaux, figures	Transitions Conclusions (partielles)	
		Numérotés	X	
Rédaction	Orthographe	Légendes	X	
		Références (non "en ligne")	X	
Rédaction	Orthographe	Coquilles	X	
		Fautes évitables	X	
Rédaction	Orthographe	Franglais, jargon	X	
		Alisées	X	
Bibliographie	Références	Absence de plagiat !	X	
		Suffisantes (nombre, intérêt)	X	
Bibliographie	Références	Pérennes	X	
		Complètes (auteurs, pages...)	X	
Bibliographie	Références	Conséquentes (volume)	X	
		Références dans le texte	X	
Proposition de note du jury		18,28		
Proposition de note haute		13,28		
Proposition de note basse				
Projection	Organisation	Plan		
		Liaisons		
Contenu	Organisation	Numérotation		
		Informatif		
Contenu	Organisation	Concise		
		Clair		
Contenu	Organisation	Orthographe		
		Illustrations		
Oral	Présentation	Assurance		
		Tenue		
Oral	Présentation	Articulation, compréhension		
		Respect		
Durée	Présentation	Temps de parole équilibré		
		Pertinence		
Réponses	Présentation	Argumentation		
Proposition de note du jury		17,41		
Proposition de note haute		12,59		
Proposition de note basse				
Travail	Etude	Bibliographie		
		Cahier des charges		
Travail	Etude	Hypothèses envisagées		
		Validation		
Complexité	Etude	Tableau comparatif		
		Choix argumenté(s)	X	
Complexité	Etude	Faisabilité	X	
Annexes	Etude	Temps consacré		
		Résultats obtenus		
Annexes	Etude	Difficulté		
		Fiches d'avancement		
Annexes	Etude	Gantt		
Proposition de note du jury		?		
Proposition note haute		16,67		
Proposition note basse		11,97		
Proposition de note du jury				

FIGURE D.1 – Points à contrôler à l'issue de la phase I

## Phase II : Conception détaillée, réalisation et recette

Rapport	Organisation	Plan	Équilibre			
		Fluidité	Cohérence	x	x	
Rédaction	Orthographe	Tableaux, figures	Introductions (partielles)	x	x	
			Transitions	x	x	
Rédaction	Rédaction	Numérotés	Conclusions (partielles)	x	x	
		Légendés	Référencés (non "en ligne")	x	x	
Projection	Organisation	Coquilles		x	x	
		Fautes évitables		x	x	
Projection	Contenu	Franglais, jargon		x	x	
		Alisee		x	x	
Projection	Présentation	Absence de plagiat !		x	x	
Proposition de note haute			17,95			
Proposition de note basse			12,95			
Proposition de note du jury						
Projection	Organisation	Plan		x	x	
		Liaisons		x	x	
Oral	Contenu	Numérotation		x	x	
		Informatif		x	x	
Oral	Présentation	Concise		x	x	
		Clair		x	x	
Oral	Durée	Orthographe		x	x	
		Illustrations		x	x	
Oral	Réponses	Alisance		x	x	
		Tenue		x	x	
Oral	Recette	Articulation, compréhension		x	x	
		Respect		x	x	
Oral	Complexité	Temps de parole équilibré		x	x	
		Pertinence		x	x	
Oral	Annexes	Argumentation		x	x	
Proposition de note haute			18,52			
Proposition de note basse			13,52			
Proposition de note du jury						
Travail	Conception	Générale	Clarté	x	x	
			Niveau de détail adéquat	x	x	
Travail	Développement		Argumentation	x	x	
			Validation	x	x	
Travail	Réalisation	Détallée	Formalisation adéquate (algorithme,..)	x	x	
			Validation	x	x	
Travail	Recette	Temps consacré	"Fonctionnalités"	x	x	
		Résultats obtenus	Volume, environnement...	x	x	
Travail	Complexité	Difficulté	Détailles	x	x	
			Détailles	x	x	
Travail	Annexes	Fiches d'avancement	Prévisionnel et justifications	x	x	
		Gant	Efectif, erreurs, corrections	x	x	
Proposition de note haute			18,11			
Proposition de note basse			13,11			
Proposition de note du jury						

FIGURE D.2 – Points à contrôler à l’issue de la phase II