A perceptual approach to film editing : exploring cinematography through visual attention and computational saliency

Alexandre Bruckert

DOCTORAT BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *AST - Signal, Image, Vision*

Par
Alexandre BRUCKERT

A perceptual approach to film editing

Exploring cinematography through visual attention and computational saliency

Thèse présentée et soutenue à Rennes, le 24/03/2022
Unité de recherche : IRISA

Rapporteurs avant soutenance :

Riccardo LEONARDI — Professor, University of Brescia
Neil D.B. BRUCE — Associate Professor, University of Guelph

Composition du Jury :

Président :	Patrick LE CALLET	Professeur, Université de Nantes
Examinateurs :	Riccardo LEONARDI	Professor, University of Brescia
	Neil D.B. BRUCE	Associate Professor, University of Guelph
	Tim J. SMITH	Professor, Birkbeck University of London
	Hui-Yin WU	Research Scientist, Université Côte d'Azur, INRIA
Dir. de thèse :	Kadi BOUATOUCH	Professeur émérite, Université de Rennes 1
Co-dir. de thèse :	Marc CHRISTIE	Maitre de conférence, Université de Rennes 1

Invité(s) :

Olivier LE MEUR — Maitre de conférence, Xiaomi Technology

Acknowledgements

First and foremost, I would like to thank both of my advisors, Olivier Le Meur and Marc Christie. Their help and guidance have made this PhD journey an incredible experience, during which I learned so much, in research and life, and I will always be grateful to them for that. A big shout-out also to Kadi Bouatouch, for agreeing to take the direction of this thesis during these last few months, and for all the wisdom he shared with me during these three (and some) years.

I would like to thank Dr. Neil Bruce and Pr. Riccardo Leonardi for agreeing to review this thesis, and Pr. Patrick Le Callet, Pr. Tim Smith, Dr. Hui-Yin Wu and Dr. Lu Zhang to be part of this jury.

I would like to thank all the members of the late Percept team, current Mimetic team and future Virtus team at IRISA. It was always a pleasure working with you all, with a special mention to Ific, Anne-Flore, Maxime, Matthieu and Pierre-Adrien.

In this wild place that is the IRISA lab, I would like to particularly thank Maryse, Laurence and Nathalie, for their immeasurable help with all the administrative tasks.

I wish to thank my family for their help and support during these three years, and for always believing in me, even when I did not.

Finally, I would like to thank you, Solène, with all my heart. Having you by my side during this challenge made these three years feel wonderful, as are all the times we are together.

I dedicate this thesis to Jean and Jean-Pierre, whose love for science and humanity made me choose this path in life.

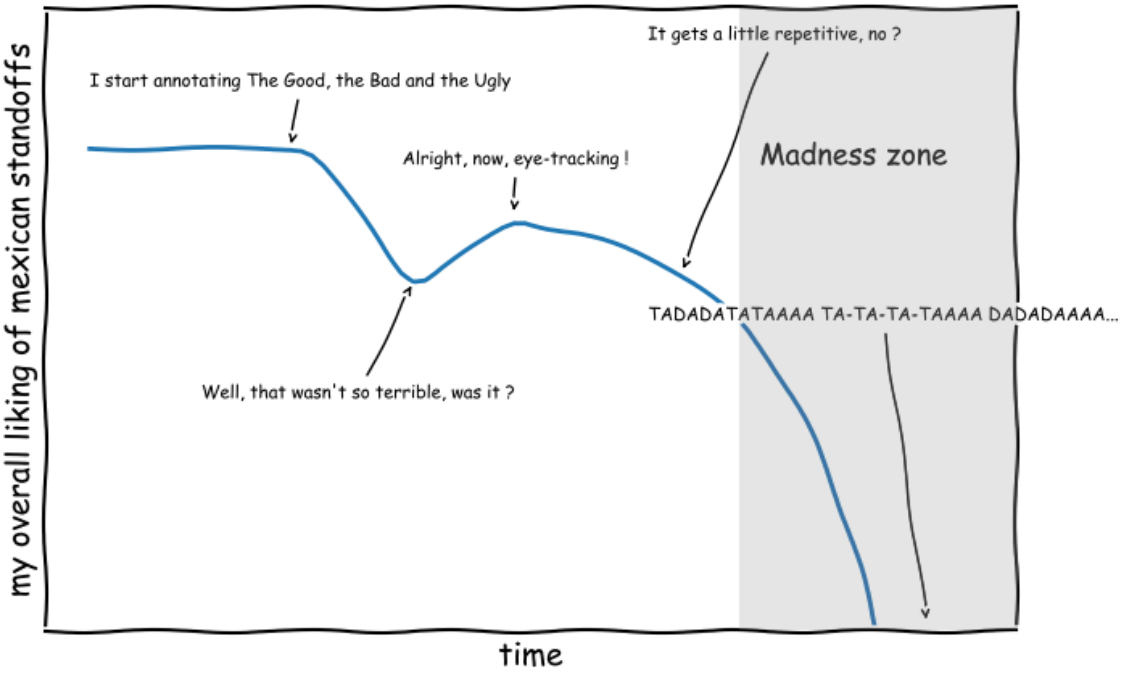

my overall liking of mexican standoffs
time
I start annotating The Good, the Bad and the Ugly
Alright, now, eye-tracking !
Well, that wasn't so terrible, was it ?
It gets a little repetitive, no ?
Madness zone
TADADATATAAAA TA-TA-TA-TAAAA DADADAAAA...

Table of Contents

Résumé en français

Que regardons-nous quand nous regardons des films ? Au delà de l'aspect naïf de cette question – quand nous regardons un film, nous regardons l'écran – il est important de s'intéresser à la façon dont nous percevons notre environnement visuel et les nombreux stimuli qui le composent, particulièrement dans le contexte cinématographique.

En effet, pour faire face à l'écrasante masse d'information, nous avons développé un ensemble d'outils cognitifs et biologiques destinés à réduire la quantité de données visuelles à traiter par notre cerveau. L'ensemble de ces dispositifs est regroupé sous le terme d'**attention visuelle**.

L'attention visuelle est intimement liée aux **mouvements oculaires** : en effet, consciemment ou non, nous dirigeons notre regard vers les zones de notre champ visuel qui nous semblent être les plus pertinentes, dans le but de réduire la charge cognitive due au traitement des zones moins importantes. Ces mécanismes se révèlent être tellement efficaces que nous construisons une image mentale de notre environnement à la fois précise et riche en détails, alors même que la surface de notre rétine contenant la plus grande densité de cellules photoréceptrices nous permettant de distinguer les couleurs ne représente qu'environ 3° de notre champ visuel, c'est à dire l'équivalent de l'aire couverte par un pouce à bout de bras !

Ce lien étroit entre mouvements oculaires et attention visuelle suscite beaucoup d'intérêt dans de nombreux champs de recherche : en effet, l'utilisation de technologies de suivi de l'œil et d'oculométrie a permis de récolter, puis d'expoiter, une quantité importante de données concernant ce que nous regardons dans une image ou non, et par extension ce que nous considérons comme visuellement pertinent ou non. Forts de ces connaissances, de nombreux systèmes de traitement d'image peuvent être améliorés, comme les algorithmes de compression d'images et de vidéo par exemple.

Lorsqu'un réalisateur conçoit un film, il joue en permanence, volontairement ou non, avec l'attention visuelle de ses spectateurs. En effet, l'ensemble du processus de création d'un film est d'une certaine manière consacrée à trouver la meilleure façon de raconter ce que l'auteur veut raconter, et ce visuellement. Tout d'abord, le spectateur est contraint

L'oculométrie : une façon d'observer le spectateur quand il observe Lee Van Cleef observer Clint Eastwood dans *Le Bon, la Brute et le Truand* (Sergio Leone, 1966)

dans ses choix visuels par l'œil de la caméra : l'ensemble de la scène lui est invisible, et il ne peut la découvrir qu'à travers ce que le réalisateur décide de lui montrer. Ensuite, par le montage et le choix des plans dans la séquence, le réalisateur décide également de l'ordre dans lequel l'audience doit parcourir visuellement l'image, et contrôle de ce fait la dimension temporelle de l'attention visuelle. Enfin, une multitude de techniques de réalisation peuvent être utilisées pour inciter encore plus les spectateurs à porter leur attention sur certaines zones de l'écran : le mouvement de la caméra, le choix de la valeur de plan, la composition du plan, les déplacements des acteurs dans la scène, la profondeur de champ, et bien d'autres encore.

Problématiques

Durant un peu plus d'un siècle, les réalisateurs ont créé, codifié et développé tout un ensemble de règles, ainsi qu'une forme de langage visuel destinés à communiquer au mieux leurs intentions narratives et artistiques. Dès lors, les chercheurs s'intéressant à la cinématographie ont tenté de comprendre **comment ces techniques et ces conventions de cinéma influencent l'attention visuelle du spectateur**. Plusieurs relations entre les propriétés cinématographiques d'une séquence et les mouvement des yeux des personnes la regardant ont ainsi pu être mises en évidence.

En parallèle, la **modélisation de l'attention visuelle humaine** a suscité un intérêt particulier dans le domaine du traitement d'image et de la vision par ordinateur. Les

approches orientées données récentes (machine learning, deep learning) ont permis d'importantes hausses de performances dans la prédiction des fixations oculaires, à la fois sur des stimuli statiques (i.e. images) et dynamiques (i.e. vidéos). Cependant, ces méthodes se concentrent principalement sur des paramètres dits *bottom-up*, c'est à dire des caractéristiques intrinsèques des images (ou des vidéos), telles que la couleur, l'éclairage, le contraste ou le mouvement. Même si l'utilisation de ces propriétés aboutit à de très bons résultats, on peut légitimement faire l'hypothèse que **de l'information de haut niveau supplémentaire à propos des techniques de réalisation propres au cinéma devrait améliorer significativement les performances de modèles d'attention sur ce type particulier de stimuli.**

De plus, la plupart des modèles de l'état de l'art actuel s'inscrivent dans le cadre de la **saillance visuelle**, c'est à dire qu'ils prédisent une probabilité de distribution représentant *où* des observateurs seront susceptibles de regarder lorsqu'ils seront confrontés à un certain stimulus visuel. Cependant, une question tout aussi importante, en particulier quand on considère des films, devrait être : **quelles sont les conditions pour que des observateurs manifestent ou non des comportements oculaires similaires ?** Répondre à cette question nous ouvre ensuite de nombreuses possibilités : par exemple, la **congruence visuelle inter-observateurs** (i.e. une mesure quantifiant à quel point des observateurs regardent la même zone de l'image, ou encore à quel point le chemin visuel d'un individu permet de prédire les chemins visuels d'un groupe entier) peut être utilisée comme une borne supérieure aux performances des modèles de saillance visuelle. Ce type de mesure est également très utile pour des réalisateurs, qui peuvent volontairement adapter leur scène afin que leurs spectateurs concentrent leur attention sur un point unique du cadre, ou au contraire les laissent explorer l'ensemble de la scène de manière dispersée.

En psychologie cognitive, dans le domaine de la cinématographie, le terme **attentional synchrony** [SH08] est utilisé pour décrire le phénomène où l'attention des observateurs converge en un seul point, dans l'espace et dans le temps. De nombreuses études ont montré que cette convergence est particulièrement importante lorsque l'on regarde des films, comparativement à des images statiques ou à d'autre type de vidéos. Cela implique donc que les séquences extraites de films possèdent des caractéristiques propres qui tendent à uniformiser les comportements visuels. Du point de vue de la modélisation, cela signifie qu'afin de prédire efficacement la congruence inter-observateurs, et par extension

l'attention visuelle, il est nécessaire de prendre en compte ces propriétés.

Contributions

Dans cette thèse, nous proposons une exploration des caractéristiques cinématographiques et de leur impact sur l'attention visuelle, du point de vue de la vision par ordinateur. Nous proposons les contributions suivantes :

1. Une nouvelle base de données oculométrique, destinée à étudier les fixations de l'œil sur des séquences cinématographiques. En plus des données oculaires et des stimuli, nous fournissons également des annotations décrivant les caractéristiques cinématographiques des séquences. Nous évaluons également leur influence sur l'attention visuelle, et leur intérêt à des fins de modélisation. Enfin, nous montrons les failles des modèles d'attention actuels, et les situations dans lesquelles ils se révelent insuffisants.
2. Un nouveau modèle de saillance visuelle, prédisant la distribution des fixations oculaires sur des séquences cinématographiques. Pour ce faire, nous avons conçu une manière d'intégrer l'information cinématographique dans un modèle orienté données.
3. Une nouvelle métrique permettant de mesurer la congruence visuelle inter-observateurs pour les stimuli dynamiques, conçue particulièrement pour prendre en compte la dimension temporelle des chemins oculaires.
4. Deux nouveaux modèles d'apprentissage profond destinés à prédire la congruence visuelle inter-observateurs : le premier pour des stimuli statiques, le second pour des extraits cinématographiques. Similairement au modèle de saillance visuelle, nous incluons de l'information cinématographique afin d'améliorer les performances.

Pour chacun de ces modèles, nous donnons quelques idées d'applications très simples, dans l'objectif de montrer l'intérêt de ce genre d'approche perceptuelle.

Organisation

Ce manuscrit est organisé en six chapitres principaux, en plus d'une introduction et d'une conclusion générale. Le chapitre 1 est destiné à donner le contexte général concernant l'attention visuelle et ses applications. Le chapitre 2 est une revue des travaux existants concernant la modélisation de l'attention visuelle. Le chapitre 3 consiste en une

explication des spécificités des stimuli cinématographiques, ainsi que d'une revue des travaux existants faisant spécifiquement le lien entre attention visuelle et cinématographie. Le chapitre 4 présente la base de données oculaires que nous avons récoltées, ainsi que les conclusions que nous pouvons tirer de son étude. Le chapitre 5 présente une méthode pour prédire l'attention visuelle sur des extraits de films, utilisant des caractéristiques cinématographiques de haut niveau. Enfin, le chapitre 6 se concentre sur la problématique de la congruence visuelle inter-observateur, sa quantification et sa modélisation.

Ci dessous, nous donnons un bref résumé de ces différents chapitres et des résultats principaux associés.

Le **chapitre 1** présente le contexte général lié à l'attention visuelle, et se concentre sur le rôle des mouvements oculaires. Nous y décrivons les différents usages des données oculométriques dans les systèmes de vision par ordinateur, et présentons quelques applications.

Dans le **chapitre 2**, nous proposons un état de l'art des différents modèles d'attention visuelle, en se concentrant sur les modèles de saillance visuelle. Nous y décrivons les différentes bases de données, méthodes d'évaluation et métriques, puis nous passons en revue les différents modèles de saillance, statiques et dynamiques. Ce tour d'horizon nous permet d'identifier les différents problèmes et lacunes des approches actuelles.

Le **chapitre 3** se concentre sur les caractéristiques propres aux stimuli cinématographiques. Nous décrivons différentes conventions et règles de cinéma, dans le but de décrire et de formaliser le langage visuel utilisé par les réalisateurs pour transmettre leurs intentions narratives. Nous passons également en revue les différents systèmes de formalisations de ce langage cinématographique. Finalement, nous proposons un bref tour d'horizon des travaux reliant attention visuelle et cinématographie. Ces travaux se situant très majoritairement dans le domaine de la psychologie cognitive, nous nous proposons d'apporter dans cette thèse une perspective différente, en abordant le problème du point de vue du traitement de l'image et des approches de modélisation orientées données.

Dans le **chapitre 4**, nous présentons une base de données oculométrique destinée à étudier l'influence des décisions cinématographiques sur l'attention visuelle. Nous avons ainsi récolté les données de fixations oculaires de 24 participants, regardant 20 séquences extraites de films de différentes époques et genres. En parallèle, nous avons annoté ces séquences afin de caractériser certaines leurs propriétés cinématographiques : la valeur des plans, les mouvements et angles des caméras, et leurs transitions entre les différents plans.

Nous identifions certains biais dépendants de ces caractéristiques, comme par exemple l'existence d'un biais centré sur la ligne de tiers supérieure, dont la dispersion dépend de la valeur de plan. Les mouvements panoramiques de caméra tendent également à attirer l'attention du spectateur, dans la direction du mouvement. Nous évaluons également les performances d'un échantillon de modèles de saillance représentatifs de l'état de l'art sur cette base, et nous identifions différents cas d'échecs, également dépendants des caractéristiques cinématographiques.

Dans le **chapitre 5**, nous proposons un nouveau modèle de saillance visuelle, conçu de telle sorte à intégrer de l'information cinématographique de haut niveau. Pour ce faire, nous proposons un modèle d'apprentissage profond, utilisant une architecture *two-stream* : une branche dédiée à l'extraction de caractéristiques visuelles liées aux séquences d'images, et l'autre branche dédiée à l'extraction de caractéristiques temporelles, prennant en entrée des séquences de flux optique. Les cartes de caractéristiques ainsi extraites sont ensuite fusionnées ensembles, mais également avec des cartes de caractéristiques de haut niveau extraites à partir des annotations cinématographiques : biais spécifiques aux valeurs de plans et mouvement de caméra, carte de flicker, carte d'anticipation du mouvement, etc. Après avoir entrainé notre modèle sur différentes bases de données, nous montrons qu'il est effectivement plus performant que l'état de l'art quand il s'agit de prédire l'attention visuelle sur des extraits de films.

Le **chapitre 6** se concentre sur la congruence visuelle inter-observateur. Cette mesure représente la diversité des comportements visuels de différents observateurs regardant un stimulus identique. En effet, de nombreuses études (voir par exemple Smith *et. al.* [SM13]) tendent à indiquer des comportements de fixations oculaires particulièrement similaires entre les observateurs lorsque ceux-ci sont confrontés à des stimuli cinématographiques. Nous proposons tout d'abord une métrique permettant de mesurer ce phénomène, tout d'abord dans le contexte statique, puis dynamique. Pour ce faire, nous utilisons une approche *leave-one-out*, en comparant les fixations de chaque individu aux fixations de tous les autres observateurs à l'aide de métriques utilisées dans le contexte de la saillance visuelle. Nous proposons ensuite deux modèles permettant de prédire cette valeur de congruence, pour des images, puis pour des séquences de films. De façon similaire au modèle de saillance du chapitre 5, nous utilisons pour le modèle dynamique des caractéristiques de haut niveau liées aux annotations cinématographiques, incluses dans une architecture *two-stream* fusionnant des caractéristiques images et des caractéristiques temporelles de mouvement.

Enfin, nous proposons pour conclure plusieurs pistes de recherche, afin d'approfondir les travaux présentés dans ce manuscrit. L'évaluation des modèles de saillance dynamique nous semble être l'un des problèmes persistants : il n'existe en effet pas de méthode permettant d'évaluer un modèle de saillance visuel dynamique autrement qu'en évaluant la qualité des cartes prédites frame par frame. Il parait dès lors important de définir de nouvelles méthodes prenant en compte l'aspect intrinsèquement temporel du problème.

Les travaux présentés dans ce manuscrit nous permettent également d'envisager des applications dans le domaine de la cinématographie virtuelle : la prévision de mouvement de caméras, ou encore le placement automatique de scènes pourrait grandement bénéficier des différentes perspectives perceptuelles, en particulier des caractéristiques de saillance. Enfin, les caractéristiques d'IOC paraissent être particulièrement prometteuses pour des systèmes de montage automatiques, permettant de prendre en compte le phénomène d'*attentional continuity* décrit notamment par Smith [Smi12].

General introduction

Where do we look when watching movies? This question might seem simple, or even irrelevant: when watching movies, we look at the screen. However, it becomes highly important when considering how humans perceive their environment: in order to cope with the overwhelming amount of visual information, we developed a vast array of biological and cognitive tools dedicated to process only a very small fraction of this data. These mechanisms are gathered together under the name of **visual attention**.

Visual attention is intrinsically linked to **eye movements**: indeed, we direct our gaze, consciously or not, so that we can spend more cognitive resources on the parts of our visual field that we deem relevant, while reducing the processing load on less important areas. This almost magic trick works so well that we build a mental image of the world that is full of details and colors, while the area containing the highest density of color-sensitive photoreceptor cells in our retina only accounts for around 3° of our visual field, which is roughly the area covered by a thumb when viewed at arm's length!

This relationship between eye movements and attention is tremendously interesting, as it allows researchers in very diverse fields to use eye-tracking techniques to gather data about what we look at and why we look at it, and by extension how we select and decide what is relevant or not. With this knowledge, one can improve image and video compression algorithms for instance, or learn about the reactions of an individual to a specific stimulus.

When creating movies, **filmmakers are always, voluntarily or not, playing with the visual attention of their audience**. The whole process of making a movie is, in fact, dedicated to tell the story that the author wants to tell, and to do so visually. First, the director forces the spectator to see only what the camera sees: the whole scene is now restricted, and the only elements to focus on are the elements captured by the eye of the camera. Then, by editing and choosing in which order to show the shots, the director also restricts the order in which the viewer will look at things, and controls the timing of when an element becomes relevant. Finally, a multitude of filmmaking techniques can be used to direct the gaze of the viewers on specific areas of the frame, from the movement of the

Figure 1 – Watching the audience watching Lee Van Cleef watching Clint Eastwood in *The Good, the Bad and the Ugly* (Sergio Leone, 1966)

camera to the choice of the size of the shot, the staging of the objects or the depth of field.

Problems

Over a little more than a century, filmmakers have developed a whole array of rules and stereotypical ways of filming dedicated to convey at best their artistic and storytelling intentions. Since then, cognitive film theorists have tried to understand **how these conventions and techniques influence the viewer**, and showed several relationships between the cinematographic properties of a sequence and the gaze patterns of people watching it.

In the meantime, the computer vision community has taken a particular interest in finding the ways of **modeling human visual attention**, on both images and videos. Data-driven approaches have recently exhibited impressive performances in predicting where humans will look. However, these methods are mostly focusing on *bottom-up* factors, i.e. the intrinsic features of the image (or the video), like color, lighting or motion. While these characteristics allow for very decent results, we can legitimately make the hypothesis that additional **high-level information regarding the filmmaking techniques should significantly improve the performances of data-driven visual attention models on this very specific type of stimuli**.

Moreover, in the current state-of-the-art, most visual attention models from the computer science part of the field are **visual saliency models**, i.e. models dedicated to predict a probability distribution of *where* people look when watching something. However, we argue that an equally important question, especially when considering movies, should be: **when do people exhibit similar or different gaze behaviors?** The answer to this question is useful in many ways: for instance, when evaluating visual saliency models, the **inter-observer visual congruency** (i.e. a measure of how much people look at the same place, or how the gaze of a single individual is predictive of the gaze patterns of a whole group) can be used as an upper-bound of the performances. Filmmakers can also make a great use of knowing when people will focus on a single point, and when they will explore the frame in a more dispersed way.

In the field of cognitive film theory , the phenomenon of people looking at the same place at the same time is referred as **attentional synchrony** [SH08]. Numerous studies have shown that this synchrony is particularly high when viewers watch movies, compared to static images or other type of videos. This would imply that movie clips possess certain proper characteristics that tend to uniformize gaze patterns. From a modeling point of view, it means that, in order to properly predict inter-observer congruency (IOC), and by extension visual attention, we must take into account such features.

Contributions

In this thesis, we propose an exploration of filmmaking features and their impact on gaze patterns, from a computer vision modeling point of view. The main contributions are the following:

1. A new eye-tracking database, dedicated to study eye fixation patterns on cinematic sequences. Alongside the eye-tracking data and the movie clips, we use hand-crafted annotations about the cinematographic characteristics of the sequences, in order to evaluate their influence on visual attention and to be used for modeling purposes.
2. A novel visual saliency model, dedicated to predict fixation distributions on movie sequences. To achieve this, we designed a way to incorporate the cinematic information previously mentioned.
3. A new inter-observer visual congruency metric for dynamic stimuli, specifically designed to take into account the temporal dimension of the gaze tracks.
4. Two new models dedicated to predict inter-observer visual congruency: the first for

static stimuli, and the second for movie clips. Similarly to the visual saliency model, we also designed it to allow filmmaking information to improve the predictions.

For each of those models, we provide a few simple applicative examples, in order to showcase the interest of these kinds of data-oriented approaches. Indeed, these can provide a large amount of perceptual data, even though imperfect, that can be used for quantitative studies regarding filmmaking.

Outline

This manuscript is organized as follows:

Chapter 1 provides a brief background on visual attention, and more specifically on the role of eye movements. We describe the general ways of using eye-tracking data in computer vision systems, and discuss a few applications.

Chapter 2 is a review of visual attention models, focusing on visual saliency. We describe the main databases available, the evaluation metrics, and review the state-of-the-art in static and dynamic visual saliency models.

Chapter 3 is dedicated to explaining what makes cinematographic videos different from other kinds of dynamic visual stimuli. We review a few cinematic features, conventions and rules, and different ways of formalizing them to be used by automated systems. Finally, we give a quick review of previous work, mostly in the domain of cognitive psychology, regarding visual attention and movies.

Chapter 4 introduces our dataset. We analyze how the cinematic features that we gathered influenced the location of eye fixations, and provide an evaluation of visual saliency models on it.

Chapter 5 proposes our visual saliency model, designed to improve fixation distribution predictions on movie clips. We show how our approach can give reliable predictions in some of the situations where the other models failed.

Chapter 6 is dedicated to inter-observer visual congruency. We describe the way that we compute visual agreement between observers, analyze the influences of cinematic features on these scores, and propose our two models.

We conclude by discussing future research perspectives and possible applications.

Chapter 1

Visual attention : How do we look at things ?

In this chapter, we define and describe the various mechanisms on which visual attention is built. We focus on overt visual attention, and the role of eye movements, and more specifically eye fixations, as a marker of relevant information within a visual stimulus. Finally, we introduce visual saliency, how it relates to visual attention, and its impact on the field of computer vision through a few applications.

1.1 Introduction

In our visual environment, we are confronted to an overwhelmingly large amount of information, as natural scenes are cluttered with various objects, textures, movements, colors, lighting, and so on. As the human visual system is ultimately limited by its biology, for instance the amount of photoreceptor cells in the retina, this huge quantity of information far exceeds its processing capacities. However, we seem to build an internal model of the world that is both rich in details and coherent, to such an extent that we actually believe that this mental depiction where everything from our environment is simultaneously present, describes in a stable and detailed way the reality of the world around us [Gom72].

Indeed, in our everyday life, such a representation is tremendously helpful as it is almost never challenged, except for some rare cases like optical illusions. This amazing feature has been made possible by several mechanisms and strategies to reduce and process the flow of information received by our eyes. *Visual attention* consists in this series of mechanisms, designed to extract what is relevant in a visual stimulus, and to focus our cognitive resources on the most important parts of it. This way, only a very small subset of the visual information arriving to our eyes is transmitted to the visual cortex, but this subset is enough for our brains to reconstruct a detailed representation of the world, even

if not perfect [Ren00].

1.2 Visual attention

1.2.1 Passive attention mechanisms

These mechanisms include both passive characteristics of the eye, and active focalisations of the visual system. Passive attention first includes the photoelectric transduction of the eye. Indeed, only a very small subset of the electromagnetic radiations, called the *visual spectrum*, is transformed into code for the brain to process. A typical human visual spectrum spans the wavelengths between 380 and 750 nanometers (see Fig. 1.1). The term *light* is used to refer to any elecromagnetic waves in this range of wavelengths.

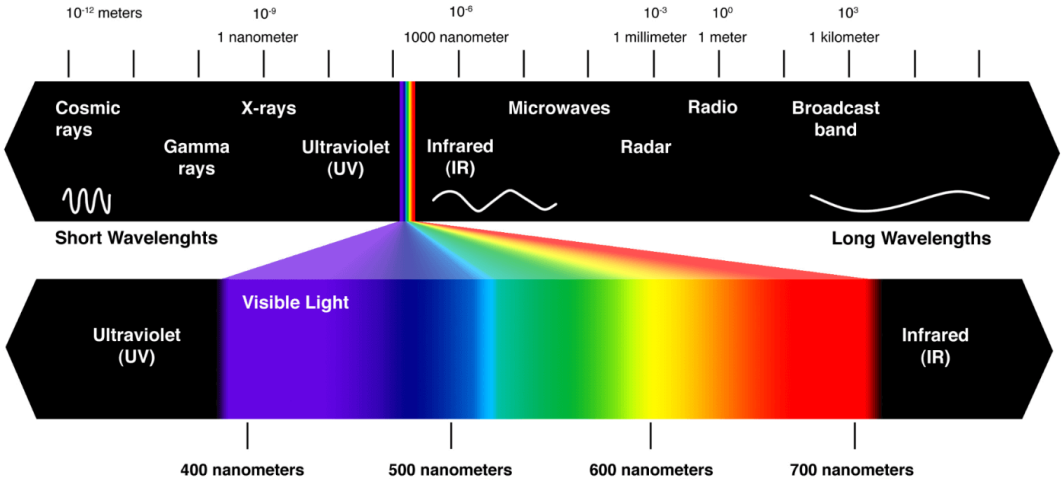


Figure 1.1 – Human visual spectrum.

A second passive way to compress the visual information of an environment is through the uneven distribution of photoreceptive cells on the retina. Indeed, the central part of the retina, called the *foveal area*, contains a high density of *cone cells*, which are color-sensitive photoreceptor cells. This high density allows for a very high spatial resolution in the fovea, while other parts of the retina contain a much lower amount of cone cells, and thus a low spatial resolution. On the other hand, peripheral areas of the retina contain a higher density of *rod cells*, which are more sensitive to dim lights than cone cells, but limited in terms of color vision. Fig. 1.2 shows this particular distribution of photoreceptive

cells on the retina.

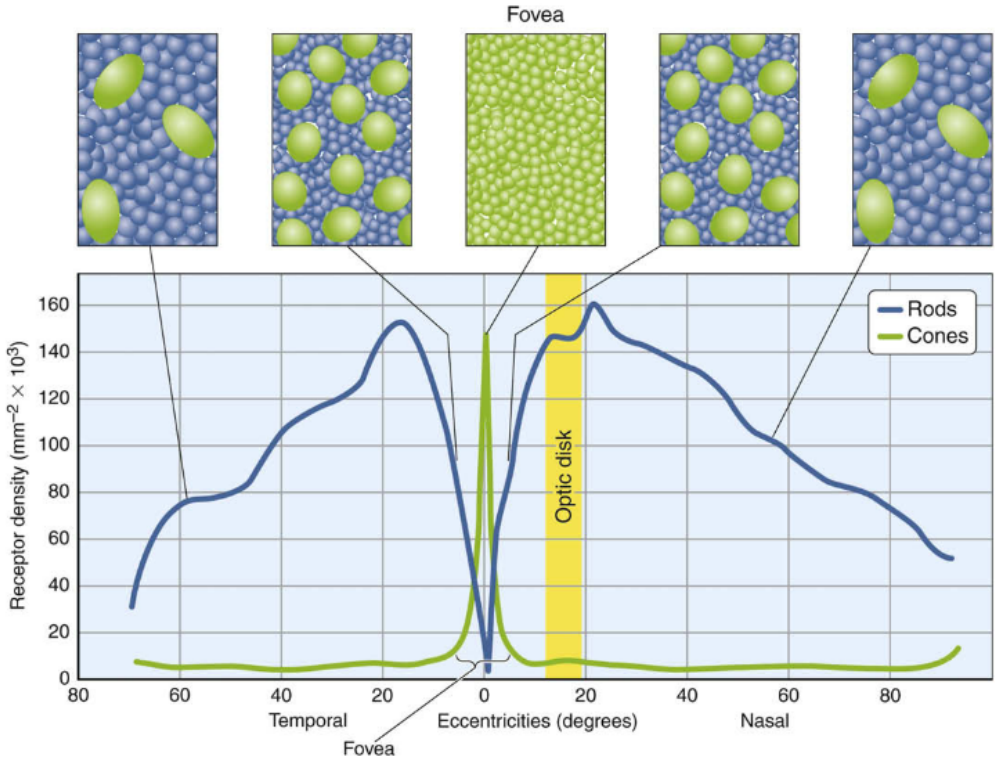


Figure 1.2 – Cone and rod cells distribution on a typical human retina, from [MEP09]

Finally, when passing from the photoreceptor cells to the retinal ganglions, the spatial information is compressed to eliminate redundancy, *i.e.* it carries out contrast information. This process is needed, as there are around a hundred times more photoreceptive cells than retinal ganglion cells. It is carried out by center-surround structures implemented by the bipolar and ganglion cells, and an easy equivalent to understand the way it works is edge detection algorithms, using decorrelation.

1.2.2 Overt and covert visual attention

Active mechanisms refer to the focalisation of attention. Instinctively, we tend to think about visual attention as to where we look and how our eyes are oriented, thus intuitively linking attention to eye movements. However, we do not necessarily attend to objects in the center of our gaze. Indeed, very early discoveries by James [Jam90] showed that humans are able to focus their attention on peripheral areas of their vision field, without

moving their eyes. This led to the distinction between *overt* visual attention and *covert* visual attention.

Overt focalisations involve an eye movement, usually to put the object of attention into the foveal zone, while covert focalisations refers to the action of focusing on areas in the peripheral vision. While eye movements, and thus overt attention, are necessarily sequential (*i.e.* one focalisation after the other), covert attention can be deployed simultaneously on multiple targets.

Covert visual attention is obviously more difficult to investigate, as there are no obvious measures to detect covert visual shifts. However, the easiest and most convenient way to determine covert focalisations seems to be the measure of overt focalisations, or in other words, eye-tracking. Indeed, the exact relationships between overt and covert attention are still debated and studied, but a general consensus seems to be that covert attention shifts precede eye movements [SFH86; SD95; Car11; Kow11; NM11]. This way, overt focalisations act as a good substitute for covert focalisations.

1.2.3 Endogenous and exogenous visual attention

In 1890, William James [Jam90] described two kinds of attention mechanisms : an involuntary and reflexive mechanism, called *exogenous*, and a voluntary and conscious mechanism, referred as *endogenous*. Endogenous attention, sometimes also called *sustained* attention, is the act of wilfully focusing on the information at a certain location. This mechanism takes around 300 milliseconds to be deployed, and is driven by the task at hand for the observers. The deployment of this kind of attention strongly depends on the observers, and is highly task-dependent, relying on *top-down* characteristics (i.e. observer-dependent). Fig. 1.3 shows several visual search strategies depending on different tasks that the observers had to fulfill. Personal emotional states, cultural backgrounds, or histories strongly influence top-down attention, and makes its modeling a difficult task. However, certain recurrent behaviors can be exposed, such as center-bias [BBD14] (people tend to look more at the center of an image), or lefward-bias [Fou+13] (people tend to make their first saccades on the left side of an image when discovering it).

Meanwhile, exogenous attention, or *transient* attention takes around 100 to 120 milliseconds to deploy, and is used, for instance, to spot new interesting locations in the peripheral vision field. This mechanism is fundamentally signal-driven, by *bottom-up* factors (i.e. stimulus-dependent). This implies that this kind of visual attention does not rely on prior knowledge of the stimulus, but rather on spatio-temporal features of the stimu-

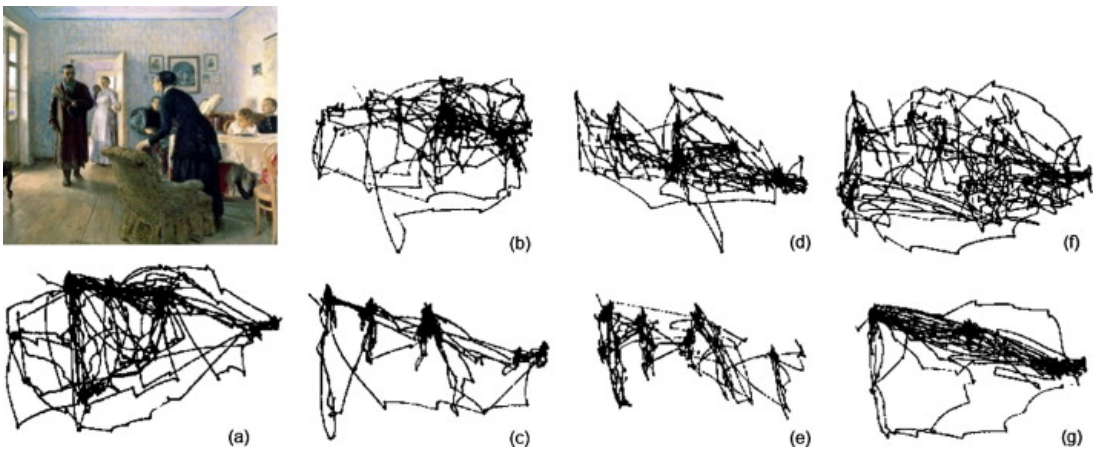


Figure 1.3 – Influence of the task on visual search on a painting of I.E. Repin, *Unexpected Visitors*. (a) no task is asked (free viewing); (b) observer is asked to estimate the social class of the characters; (c) observer is asked to estimate the age of the characters; (d) observer is asked to guess what the family was doing before the visitor entrance; (e) observer is asked to remember the clothes of the family members; (f) observer is asked to remember the positions of objects and people; (g) observer is asked to estimate how long the visitor had been away from the family. Adapted from Yarbus [Yar67].

lus. For instance, high-contrast areas (in terms of luminance, colors, texture, or motion) tend to draw the attention of the observer [MRW97; RZ99; PN04]. Sudden appearance of objects also has a strong impact on overt visual attention [YJ96; RJY92], as well as contextual discrepancies (i.e. when an object seems to be out of place in a scene) [GS14].

Neurophysiological studies seem to find that bottom-up and top-down attention are associated with two separated but communicating areas of the brain [DD95; CS02]. These mechanisms are indeed intertwined, work jointly and interact when dealing with a natural scene. Moreover, even during a top-down attention phase, where the visual focus is consciously controlled by the observer, bottom-up factors influence the attention. For example, Theeuwes [The04] showed that when asked to find an a specific shape on an image containing various objects, people tend to focus on a colour outlier if there is one.

1.3 Eye movements and overt visual attention

As mentioned earlier, one of the mechanisms of visual attention includes moving the eye in order to put the image of an object of interest on the foveal area. Thus, eye movements and overt visual attention are intrinsically linked, and a better understanding of the way eyes move leads to a better understanding of the way we process visual information.

1.3.1 Saccades

Saccades are quick movements of both eyes, which speed is typically between 100 and 700 degrees per second, but can sometime reach up to 900°/s. Saccades are intended to shift the foveal area to another part of the visual stimulus. Thus, they are a crucial piece of the mechanical processes that selects relevant visual information. Saccades do not necessarily follow the shortest path from one fixation point to another, as they can sometimes follow a incurved path. They can be classified into several categories, like visually guided saccades (the eyes move towards a part of the stimulus, either because of the appearance or disappearance of a salient object, or because of an endogenous decision to scan the environment), antisaccades (the eyes move away from a visual cue), memory guided saccades (the eyes move towards an area that was remembered) or predictive saccades (the eyes anticipate a movement of an object).

1.3.2 Fixations

Eye fixations are not eye movements per se, but rather the moment between saccades where the gaze is maintained on a single location. The alternance of saccade phases and fixation phases is a common trait of almost all animals with good vision, including vertebrates and most arthropods and cephalopods [Lan19]. For humans, fixations often occur when the foveal area is located on a part of the visual stimulus deemed relevant, consciously or not. However, even though the eye seems to be still during a fixation phase, it can actually exhibit several eye movements, called *fixational eye movements*, namely microsaccades, ocular drift and ocular microtremors.

Microsaccades

Microsaccades are a kind of fixational eye movement that occur involuntarily between saccades. They usually have an amplitude of 2 to 12 min-arc of visual angle (i.e. bewteen 0.03 to 0.2 degrees) [CK08], and occur once or twice per second during a fixation. However, recent studies seem to highlight the existence of a saccade-microsaccade continuum [Ote+08], with an asymptote in the distribution of the magnitudes of microsaccades around 1 degree. Fig. 1.4 shows an example of microsaccades happening during the fixation phases of the visual exploration of a scene. The purpose of these microsaccades still remains unclear, but several hypotheses are being explored, like the control of the fixation position, or the prevention of perceptual fading [Rol09]. Indeed, in the early 1950s,

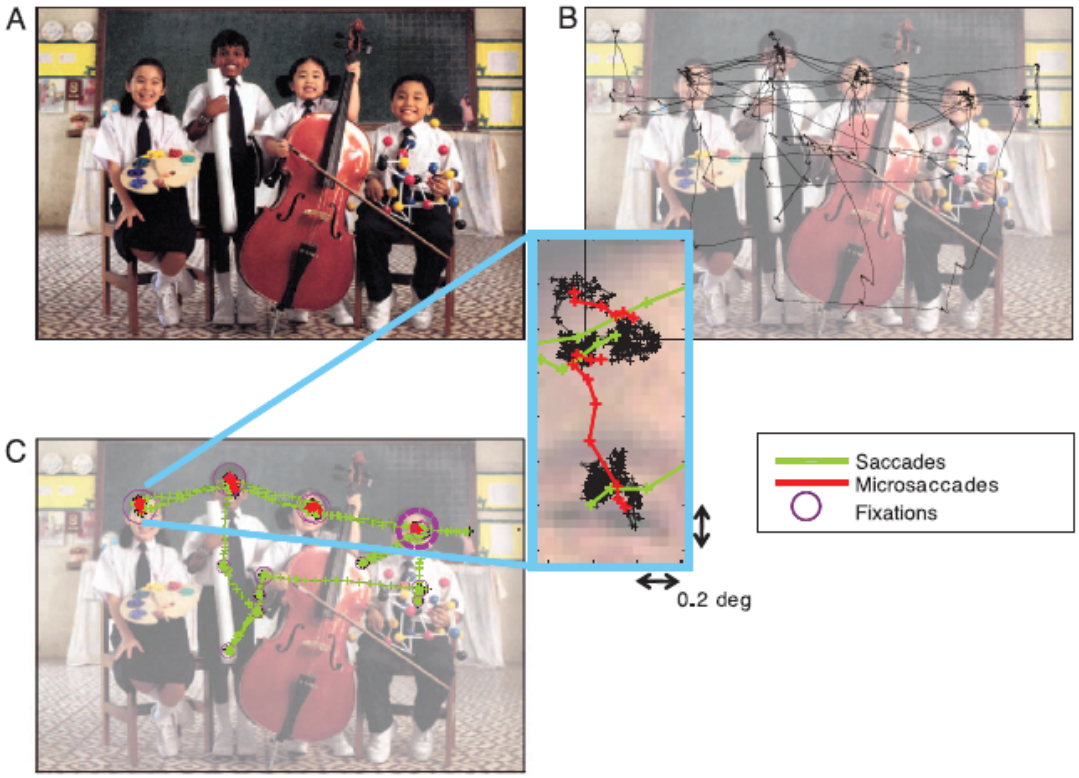


Figure 1.4 – Example of saccades and microsaccades occuring during the free viewing of a scene. B shows the ocular scanpath on a 45 seconds time period, and C shows the saccades (green) and microsaccades (red) happening during a 10 seconds time period. Reproduced from Martinez-Conde *et. al.* [Ote+08].

several studies showed that, when an observed image is set in motion to counteract eye movements, the observer sees the image fading away within a few seconds [Yar57]. Finally, neurological studies showed that microsaccades usually trigger an increase in neural activities, especially in the V1 area of the visual cortex, which could give microsaccades a rather important role in visual attention [MOM13].

Ocular drift

During a fixation, the eye follows an erratic, low velocity path. This slow motion is called occular drift, and can be described as a random walk [Fin74]. The speed of these drifts is usually below 30 min-arc/s. It also seems that these drifts are coordinated between the two eyes. Recent studies showed that these movements are useful to process spatio-temporal information on a scene, as well as the details on stationary objects [AA12; RV15].

Ocular microtremors

Ocular microtremors are irregular wave-like movements, with a high frequency and a very small amplitude (just around a few arcseconds), that happen during the drift movements. Like the ocular drifts, the function of theses fixational eye movements is still debated, but they seem to be linked to the process of perceiving high-grain details, and correlates with the activity of several areas of the brainstem [RV15; MOM13].

The two following eye movements that we describe are secondary for the topic of this thesis; we will just give a quick overview of their specificities and purposes.

1.3.3 Smooth pursuit

Smooth pursuit refers to the movement of the eye following a moving object, in order to stabilize it in the visual field. Contrarily to saccades, the eye moves in a continuous way during smooth pursuit, with a velocity usually under 30°/s. If the object moves faster than that, follow-up saccades are necessary to keep tracking. Humans are ordinarily unable to initiate a smooth pursuit movement without a moving visual stimulus. The purpose of this movement is to stabilize the moving target in the foveal area, in order to examine it with an important power of resolution.

1.3.4 Vergence

Vergence is a movement where both eyes move in different horizontal directions. The reason for that is to keep a binocular vision into an object that moves towards or away from the observer. The eyes then have to rotate horizontally, towards each other when the object is close, and away from each other when the object is farther away.

1.4 Studying eye fixations to inform on visual attention

As mentioned earlier, eye fixations are strongly linked to visual attention, whether they are endogenous or exogenous. Moreover, extensive literature used this knowledge and eye-tracking measurements to show that some visual features can draw and drive attention [Wol98]. However, these studies mostly take place in the setting of the lab, and use low-level features of the stimuli. A natural extension of this work was then to study the allocation of visual attention on complex scenes, closer to the real world. One of the computational approaches dedicated to model the deployment of visual attention on a stimulus is the concept of *saliency map*.

1.4.1 Visual saliency maps

In 1985, Koch and Ullman [KU85] proposed a topographic representation of an image, where the scalar values at each location represent the *saliency* of this area, i.e. the the likelihood of an eye fixation to occur at said location. This notion of saliency is computed using an array of visual features, that could influence overt or covert attention. The main hypotheses of this approach are, that overt attention is driven by bottom-up characteristics that can be extracted from the stimulus, and that overt visual shifts can serve as a proxy for covert attention, as mentioned earlier. This idea was soon implemented in several computational models by Itti, Koch and Niebur [IKN98; IK00]. In these original models, saliency is inferred from low-level features extracted in parallel from a scene, at different scales (see Fig.1.5).

It is worth noting that this approach is fundamentally bottom-up, and thus cannot take into account top-down visual discrepancies; this is why these models are much more accurate at explaining human attention allocation in free-viewing conditions, and when the stimulus is still unknown to the viewer [PLN02].

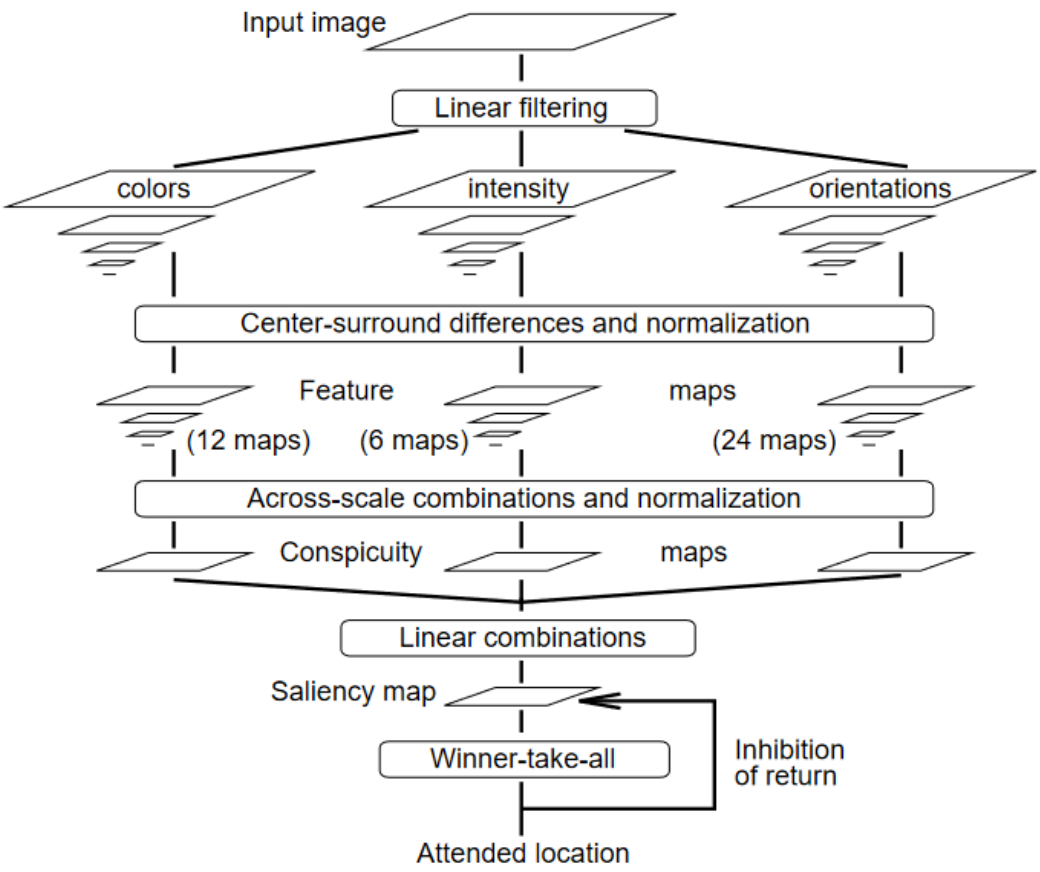


Figure 1.5 – Pipeline of Itti's saliency model, from Itti *et. al.* citeItti98 The image is passed through three channels of feature extraction (color, intensity and orientation), resulting into three separate conspicuity maps, which are then merged together linearly.

1.4.2 Visual saliency and eye fixations

In the work of Itti and Koch [IKN98], making the feature map was the first step to compute visual attention. Once all of the features were combined, the overall distribution was used to predict eye fixations, which were derived to be at the location of the local maxima of the saliency map, in decreasing peak value order. To evaluate this model, and other approaches to saliency, one must rely on ground-truth data, i.e. ground-truth fixations obtained through eye-tracking. However, due to the top-down component of visual attention, it is difficult to just rely on a fixation map from a single or a few observers. To solve the issue, most models rely on *fixation density maps* as a ground-truth representation. The traditional way to create such maps is by extracting the location of fixation points, and then convolving this binary map with a gaussian kernel, whose size is determined by the projected size of the fovea on the viewing device, as well as the precision and accuracy of the eye-tracking device. Fig. 1.6 illustrates this classical pipeline, from eye-tracking experiments to fixation density maps.

Maps created that way are fixation density maps, but are often referred to as ground-truth *saliency maps*, thus allowing a certain confusion between the low-level set of features described by Itti and Koch and these fixation densities. Moreover, there is not a single way to build fixation density maps: these maps are highly dependent on the experiment conditions, such as the viewing time, the number of observers, or the task at hand. Engelke *et. al.* [Eng+13] showed that even though fixation density maps are usually very similar and their differences have a very low impact on any kind of application, they strongly depend on the experimental conditions of the laboratories. This should call for a cautious approach when analyzing results of visual attention models, and more especially when these models are data-driven.

1.5 Applications of eye fixations and visual attention in image processing

Eye-tracking research has now reached an era of new innovations and applications. Indeed, now that eye-tracking technologies become less and less invasive and expensive, thus the knowledge about relationship between eye movements and visual attention gets more and more extensive, so that a whole variety of applications are being investigated. In this section, we will only focus on computer vision related applications, but eye-tracking

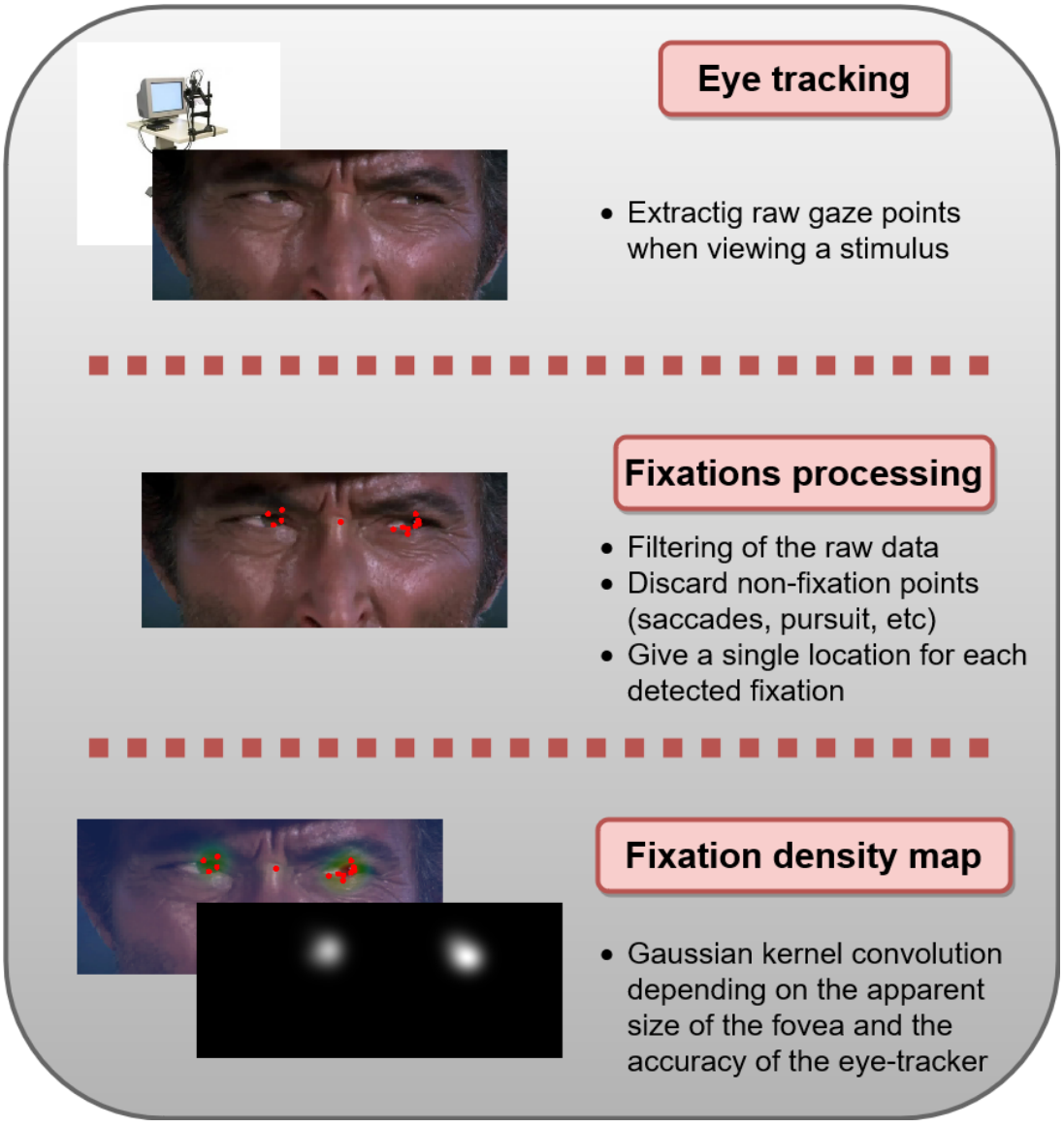


Figure 1.6 – From eye-tracking to fixation density maps. Adapted from Le Callet and Niebur [LN14]

has become a tool used in many fields, like psychology, neurosciences, robotics or advertising (see Duchowski [Duc02] for an extensive review, and Mancas and Le Meur [ML16] for saliency specific applications). The aim here is not to give a comprehensive review of visual attention applications, but rather to give a very brief overview of the diversity of uses of eye-tracking data in the field of image processing. More complete and detailed literature on each of the topics that we mention can be found in the various surveys and reviews cited in the next subsections.

1.5.1 Attention-driven compression

Image and video content production and distribution has recently seen a rather radical increase, due to the advances in communication technologies, and the always increasing bandwidth of multimedia devices. This results in the very fast development of new video and image compression methods, i.e. converting a visual signal in a way that reduces the storage space, while maintaining a good visual quality. The first methods based on visual attention were introduced in the late 1990's [MDN96; KG96].

The main idea was either to find the less interesting areas in a frame and primarily compress them, or transmitting the most salient areas first during a data transfer. By treating the different regions of the image differently, depending on their visual appeal, one can achieve compression without degrading too much the perceived quality. Since then, these techniques have improved, following the advances in visual attention modeling [LE12; ZX18; Itt04; LQI11; HB14]. The generalization of low-cost eye-tracking devices also lead to new compression techniques for network video streaming [Fen+11].

Recently, 360° image and video processing has gained a lot of attention, as the important amount of data created by the spherical viewing range creates a great need for efficient compression methods. While the main approach was to project the 360° images into the 2D plan, and use more traditional compression techniques [SD18], several novel approaches proposed models specific to this kind of content (see the review of Xu *et. al.* [Xu+20]).

1.5.2 Perceptual image quality assessment

As the transmission pipelines for images and videos get more and more complex, and involve more and more steps susceptible to degrade the quality of a visual stimulus, quality assessment plays an increasingly important role at every stage, from the acquisition of the signal, to its compression, transmission and display. These metrics are used to ensure

that the content delivered to the user aligns with the intentions of the transmitting entity. Quality assessment methods are usually separated into two categories : objective and subjective. Subjective assessment involves showing the visual stimulus to a set of human observers, and thus is the most reliable and accurate technique. However, these subjective tests are time and resources-consuming, and cannot be included in an automated pipeline. This is why objective image quality assessment algorithms have been developed, in order to approximate the quality evaluation of a human observer.

Visual attention was introduced in this field, as studies showed that artifacts are usually more annoying when located in salient areas [Eng+11]. Several approaches were proposed incorporating visual attention data, by giving more weight to salient areas when penalizing image distortions during the pooling stage [Nin+07] (see Zhai and Min [ZM20] for a detail review of such metrics). However, even though many studies show that visual attention does indeed affect image quality assessment, there are still a number of issues remaining, as this weighting is not always efficient [Nin+09; LH11].

Similarly to attention-driven compression, attention-driven quality assessment for 360° images [Xu+20] and 3D stereoscopic content [ZM20] has gained a lot of attraction, as saliency maps were found very useful for evaluating quality in these particular cases.

1.5.3 Medical imaging

Recent medical imaging techniques, such as computed tomography scanners or magnetic resonance imaging, opened the field to significant improvement in medical decisions. Analysis of medical images has become a central part of many diagnostic processes, either to detect and locate anomalies, like tumors or lesions, or to interpret the image itself. However, the amount of data and the low number of experts makes it difficult to properly get the full amount of information (and in some cases, the right medical decisions) from these images.

By studying eye movements of medical imaging experts, Krupinsky [Kru12] showed that gaze patterns vary depending on the amount of experience of the viewer. This opened the idea that visual attention models could provide an interesting approach to partially automate and eventually improve medical images analysis. These approaches were successfully used to detect slices of CT scans containing tumors [Man+07], on MRI images [AK14], or for other various purposes (see Lévêque *et. al.* [Lév+18] for a recent survey of the field).

1.5.4 Other attentive systems in computer vision

Most data-oriented applications in computer vision, like for instance image captioning, action recognition, segmentation or classification tasks, are based on human perception. Indeed, annotations and labels are usually provided by humans, and thus heavily rely on human visual attention. This led computer vision researchers to include visual saliency and human attention in a whole variety of models, applied to a large number of tasks and problems. These applications are often referred as *attentive systems*.

For instance, Karessli *et. al.* [Kar+17] showed that gaze patterns are class-discriminative, and thus can be used to perform efficient image classification. Object recognition tasks also benefit from visual attention knowledge: indeed, knowing the spatial distribution of eye fixations provides information about the likelihood that an object is in fact in that area, before even knowing what this object is [WK06; ADF10]. Segmentation [Che+15; Qin+14], scene classification [BI11], caption generation[BA18] or object tracking [Bor+12] can also be improved by using human attention characteristics.

More applications in computer vision and artificial intelligence can be found in the reviews of Zhang *et. al.* [Zha+20] and Nguyen *et. al.* [NZY17].

1.6 Conclusion

In this chapter, we have seem various mechanisms that humans use to reduce the cognitive load of processing their visual environment. We explored how eye movements can inform about the visual attention of the viewer, and how computer vision systems make use of this kind of data.

Visual attention as a research topic is particularly rich, in the sense that it benefits from a whole variety of approaches and disciplines, from computer vision to cognitive psychology, robotics, and many more. The numerous resulting applications have shown how useful eye-tracking technology can be. For instance, we have not mentioned the vast number of interactive systems that rely on quantitative measures of overt visual attention: foveated rendering in virtual reality headsets, communication devices for disabled users, and so on.

During the last decades, significant progresses have also been made in understanding the various neurobiological mechanisms of visual attention. These insights have allowed computer scientists to propose many models of visual attention inspired by the way we process images, for example by simulating the operations taking place in the different

areas of the visual cortex.

Finally, recent techniques in the field of image processing, like machine learning approaches, have considerably increased the sheer amount of data that is used by computer vision systems. In this perspective, using eye-tracking and overt attention knowledge essentially helps reducing this mass of information, by indicating what is -or is not- relevant to the human visual system.

Chapter 2

Modeling visual attention on images and videos

In this chapter, we present a review of the most influential and relevant models of visual saliency. We first present a variety of eye-tracking databases commonly used for these tasks, and the methods used to evaluate visual saliency models. We then review static, and more importantly dynamic saliency models. Finally, we give an overview of other kinds of models, outside the visual saliency paradigm.

2.1 Introduction

Since the first theoretical method to compute visual attention using feature integration outlined by Koch and Ullman [KU85], and implemented later by Itti *et. al.* [IKN98], visual saliency models have seen significant improvements, alongside with our comprehension of how visual attention is deployed on all kind of stimuli. In the early 2000's, a lot of work have been dedicated to finding better hand-crafted features and learning methods to compute saliency maps closer and closer to ground-truth fixation density maps. These improvements lead the quality of the saliency predictions to grow at a relatively stable rate. However, the growing availability of large quantity of eye-tracking data, alongside with the recent resurgence of neural networks and the application of deep learning approaches created a sharp difference in performances, and accelerated significantly the amount of models and new methods. These improvements go to such extent that it is, in some cases, almost impossible to differentiate ground-truth fixation densities from computed saliency maps (see Fig.2.1).

Nevertheless, most of this work focuses on images, i.e. *static stimuli*. In comparison, the same task applied to video, i.e. *dynamic stimuli*, still remains a little bit less explored, although recent studies seem to indicate a growing interest in the challenges presented by it. Multimedia approaches are also more and more investigated, with the addition for

Figure 2.1 – Images with ground-truth fixation densities and computed saliency predictions from the SALICON model [Hua+15]. Maps in the second and third rows belong to either Model (M) or Humans (H) (i.e. are complementary). Try to guess which one is which. Zoom on the bottom-right text to see the answer for the second row. Reproduced from Borji [Bor18]

instance of audio cues to dynamic saliency models.

It is important to distinguish visual saliency models from other kinds of approaches related to visual attention, as the term "*saliency*" is sometimes used with different meanings. In this thesis, it will only refer to bottom-up approaches dedicated to model human fixation densities. Other kinds of frameworks will be shortly discussed in the last section of this chapter.

2.2 Eye-tracking datasets for dynamic visual saliency

In order to evaluate dynamic attention models, or even to build them in the case of data-driven methods, there is an important need for eye-tracking databases. Those bases are usually built by showing visual stimuli to observers, and recording their gaze patterns. However, as mentioned earlier, the experimental conditions in which the fixation points are gathered can influence the collection of the data, and consequently have an impact on the models built out of it. It is then necessary to have a good understanding of the existing eye fixations databases, what kind of stimuli they contain, and in which conditions the data was gathered. In the following, we list the most influential and relevant eye-tracking databases, as well as their characteristics and limitations.

2.2.1 Static stimuli

Eye-fixation datasets on static images play an important role on the development of dynamic saliency models. Indeed, it is common to see a video saliency computation relying at least partially on static features, extracted from such databases. The other advantage is the time observers get to look at the images: for each image, one can follow a relatively long gaze track, while for videos, a single fixation is often carried over several frames, meaning that more observers are usually needed to obtain a reliable fixation density on a single frame.

The **MIT dataset** [Jud+09] was the first large-scale database of eye-tracking experiments dedicated to study visual saliency. It consists of eye-tracking data for 15 free-viewing observers, aged 18 to 35, over 779 landscape images and 228 portrait images. Later, the authors proposed a follow-up dataset, composed of 300 natural scenes, with held-out gaze records from 39 observers, in order to use it as a benchmark for saliency models.

CAT2000 [BI15] includes eye-fixations records on 4000 images, separated uniformly into 20 different categories. Gaze data was collected from 29 observers per image, free-viewing each stimulus for 5 seconds. Half of the resulting fixation maps and saliency maps are held-out, again for benchmark purposes. Alongside with the MIT set, they are used to rank saliency models on the MIT/Tübingen benchmark [Küm+].

OSIE [Xu+14] contains 700 natural images, with eye fixation data from 15 free-viewing observers, each image being seen during 3 seconds. The particularity of this database is the annotations the authors provide, including over 5000 segmented objects, as well as semantic annotations.

SALICON [Jia+15] is a rather particular database: instead of eye fixations, it is composed of mouse-tracking data, over more than 10000 images from the MS COCO image database [Lin+14]. Using the OSIE dataset as an eye-tracking baseline, the authors show that mouse-tracking data, with proper preprocessing, can be used as a good ground truth to train visual saliency models. The large scale of this dataset makes it an important contribution especially in the context of data-demanding methods, like deep learning. However, Tavakoli *et. al.* [Tav+17] looked into the correlation between mouse tracking and eye tracking at finer details, showing the data from the two modalities are not exactly the same. They demonstrated that, while mouse tracking is useful for training a deep model, it is less reliable for model selection and evaluation in particular when the evaluation standards are based on eye tracking.

This list of databases is far from exhaustive; for a more complete view of the topic,

we refer the interested reader to the MIT/Tübingen list of saliency datasets [Küm+]. To this day, they list 31 saliency datasets, with various properties and purposes.

2.2.2 Dynamic stimuli

As for the static case, eye-tracking databases on videos are important to both understand and model the way we watch dynamic stimuli. Several datasets have been proposed over the last few years; in the following, we highlight some of the bases relevant to this thesis, as well as their interest and shortcomings.

DHF1K [Wan+19] is considered as the main standard for evaluating dynamic saliency models. It consists of eye fixations data over 1000 videos, from 17 observers. The authors used videos from Youtube, based on searches using key words, in order to maintain a variety of content and objects. The total duration of the set is around 5h20 (19420 seconds), making it one of the largest dynamic saliency dataset. Out of the 1000 gaze-tracking records, 300 are held-out to evaluate saliency models on the benchmark proposed by the authors [Wan+].

Hollywood-2 [MS15] includes 1707 movie clips, from 69 Hollywood movies, as well as fixation data on those clips from 19 observers. Observers were split into three groups, each with a different task (3 observers free-viewing, 12 observers with an action recognition task, and 4 observers with a context recognition task). Each group being relatively small, the common way to use this data for visual attention modeling is by merging those groups, thus introducing potential biases. The large scale of this dataset (around 20 hours of video) is well fit for training deep saliency models, however few conclusions regarding gaze patterns on movies can be drawn from the data itself, since it mainly focuses on task-driven viewing mode, and that each clip is only around 15 seconds long.

SAVAM [Git+14] includes 41 high-definition videos, 28 of which are movie sequences (or use movie-like realisation, like commercials for instance). Eye fixations are recorded from 50 observers, in a free viewing situation. As for Hollywood-2, the each clip is quite short, only 20 seconds on average.

The **DIEM** project [Mit+11] is an investigation of gaze patterns on videos. The authors first released a dataset composed of eye-tracking records of 42 observers, on 26 movie sequences, for a total of 2605 seconds of content. In their study, the authors showed that temporal features were the most predictive of eye fixations, compared to spatial and static features. Since then, the dataset has grown, and now includes data from over 250 observers, on 85 videos. These videos cover a large range of genres, including advertisements,

movie trailers, music clips, or sports videos.

Breeden and Hanrahan [BH17] proposed eye-tracking data from 21 observers, on 15 clips from 13 films, for a total of 38 minutes of content. Each clip is between 1 and 4 minutes. Alongside this data, they also provide high-level feature annotations, such as the camera movements in shots, the temporal location and types of edits, the presence or absence of faces on screen, and whether or not the characters are speaking. However, the main limitations of this dataset are the relatively low precision of the eye-tracking device used, and the duration of the total content of the base itself.

Study Forrest [Pro14] is a large-scale project centered on the movie *Forrest Gump*, and dedicated to understanding a large spectrum of the sensory impact of the movie. It includes a huge amount of data, including extensive neurological imagery, movie-related annotations and gaze-tracking data [Han+16]. The gaze pattern dataset includes eye-tracking data of 30 observers watching the movie, 15 of the participants being in a fMRI scanner, and the other 15 in a lab setting.

In this list, we focused mainly on movie-related databases. A more complete list, dedicated to the visual saliency paradigm, can be found on the DHF1K benchmark page [Wan+].

2.3 Evaluation of saliency models

Measuring the differences or similarities between two gaze behaviors can be a challenging task, but is fundamental for our understanding of human visual attention. Obviously, it plays a key role in evaluating the performances of saliency models. However, there is not a single unified metric that is used as a consensus reference when it comes to evaluating saliency maps. Le Meur and Baccino [LB13], and later Bylinskii *et. al.* [Byl+19] offered reviews of the way scanpaths, fixations density maps and saliency maps could be compared, and how those metrics behave and what they actually measure. Li *et. al.* [Li+15] used human evaluations of saliency maps to define a subjective ranking, against which they evaluated other metrics. However, human perception is not a very good discriminator for visual saliency, as humans tend to favor certain features of the saliency maps over other. For instance, small variations of the saliency values in low-saliency areas tend to go unnoticed. Finally, Emami and Hoberock [EH13] proposed to rank the metrics based on how well they discriminate between human fixations density maps and randomly-generated saliency maps.

In the following, we list the main metrics used to evaluate saliency models, as we will extensively use them in the remainder of this thesis.

2.3.1 Distribution-based metrics

Pearson's correlation coefficient (CC)

The Pearson's correlation coefficient (CC) evaluates the linear relationship between two variables, and thus can be used to interpret differences between saliency maps and fixations density maps. For a predicted saliency map $\hat{S}$ and a fixation density D, the value of the correlation coefficient is:

$$CC(D, \hat{S}) = \frac{\sigma(D, \hat{S})}{\sigma(D)\sigma(\hat{S})} \tag{2.1}$$

where $\sigma(D, \hat{S})$ is the covariance of D and $\hat{S}$. The correlation coefficient takes values between -1 and 1, 1 indicating perfect correlation, -1 perfect correlation in the other direction, and 0 no correlation. It is symmetric, and thus does not distinguish between false positives (i.e. a predicted salient area where no fixations occur experimentally) and false negatives (i.e. a predicted non-salient area where fixations occur).

Similarity (SIM)

The similarity metric (SIM), or histogram intersection, measures the similarity between two distributions represented as histograms. For a predicted saliency map $\hat{S}$ and a fixation density D, both normalized (i.e. $\sum_i \hat{S}_i = \sum_i D_i = 1$) the value of the metric is:

$$SIM(\hat{S}, D) = \sum_i \min(D_i, \hat{S}_i) \tag{2.2}$$

where i are the pixel locations. A value of 0 indicates no histogram overlap at all, while a value of 1 indicates perfect overlap. The similarity metric is highly sensitive to false negatives, and penalizes them significantly more than false positives.

Kullback-Lieber divergence (KL)

The Kullback-Lieber divergence (KL), also called relative entropy, comes from the field of information theory. It is used to measure how a probability distribution differs from an

other. In the context of visual saliency maps, there exists several ways to compute this metric. However, the most common is the following, for a predicted saliency map $\hat{S}$ and a fixation density D:

$$KL(\hat{S}, D) = \sum_i D_i \log\left(\varepsilon + \frac{D_i}{\varepsilon + \hat{S}_i}\right) \tag{2.3}$$

where i iterates over the pixels of the map and ε is a regularization constant. The value of ε will affect how pixels with a prediction of zero will be penalized; it is usually set to built-in epsilon value of the language used (usually $2^{-52} \approx 2.22\text{e}{-16}$ for 64-bit systems). KL divergence is very sensitive to zero-values, and thus penalizes a lot sparse predictions. Identical maps will score very close to zero, and the score gets higher as the compared distributions differ. The upper-bound for the metric depends on the size of the maps and the chosen value of ε.

Earth mover's distance (EMD)

The earth mover's distance (EMD) measures the minimal cost needed to transform one histogram into another (or in our case, a saliency map into another). It incorporates a ground distance, as to include the notion of space into the computation of the metric. For a saliency maps $\hat{S}$ and a fixation density D, the EMD is defined as

$$\begin{aligned} &EMD(\hat{S}, D) = \min_{f_{i,j}} \frac{\sum_{i,j} f_{ij} d_{ij}}{\sum_{i,j} f_{ij}} \quad s.t. \\ &(1)\ f_{ij} \geq 0,\ (2)\ \sum_j f_{ij} \leq \hat{S}_i\ (3)\ \sum_i f_{ij} \leq D_j\ (4)\ \sum_{i,j} f_{ij} = \min\left(\sum_i \hat{S}_i, \sum_j D_j\right) \end{aligned} \tag{2.4}$$

where f_{ij} represents the flow (i.e. the amount of value transported from pixel i to pixel j), and d_{ij} is the spatial distance between pixel i and pixel j. In our case, the euclidean norm is commonly used for the ground distance. However, this computation is very costly; we then use the following variant, proposed by Pele and Werman [PW08], for which there exists a linear-time algorithm:

$$\begin{aligned} &\widehat{EMD}(\hat{S}, D) = \min_{\{f_{ij}\}} \left(\sum_{i,j} f_{ij} d_{ij}\right) + \left|\sum_i \hat{S}_i - \sum_j D_j\right| \times \max_{i,j} d_{ij} \quad s.t. \\ &(1)\ f_{ij} \geq 0,\ (2)\ \sum_j f_{ij} \leq \hat{S}_i\ (3)\ \sum_i f_{ij} \leq D_j\ (4)\ \sum_{i,j} f_{ij} = \min\left(\sum_i \hat{S}_i, \sum_j D_j\right) \end{aligned} \tag{2.5}$$

An EMD of zero indicates that the distributions are the same, while a larger value indicates more differences. EMD penalizes false positives, depending on the spatial distance between them and the ground truth. As this metric requires to solve an optimal transportation problem (and so a global optimization over the whole saliency map), it still remains computationally costly, which is why it is sometimes not reported when evaluating saliency models.

All of those distribution-based metrics can of course be used to compare a predicted saliency map to a ground-truth fixation density, but also to compare two predicted saliency maps, and so are useful to evaluate saliency models relatively to each others.

2.3.2 Location-based metrics

Area under ROC curve (AUC)

One way of interpreting a visual saliency map is to consider it as a classifier of which areas are fixated or not. This advocates for the use of signal detection metrics to evaluate saliency maps performances. The Receiver Operating Characteristic (ROC) curve represents the rate of false positives (FPR) as a function of the rate of true positives (TPR), when treating the saliency map as several binary classifiers, based on a set of thresholds. The area under the ROC curve (AUC) then provides a measure indicating the performances of the overall classification. Several implementations of the AUC metric exist, in the context of visual saliency, depending on the way the true and false positive rates are calculated. A value of 1 indicates perfect classification, while 0.5 is the chance level.

Judd *et. al.* [Jud+09] proposed a first AUC variant (AUC-J), by computing the TPR as the ratio of true positives (i.e. fixation pixels where the predicted saliency value is above the considered threshold) to the total number of fixations, and the FPR as the ratio of false positives (i.e. unfixated pixels where the predicted saliency value is above the considered threshold) to the total number of pixels of the saliency map above the threshold.

Another variant (AUC-B) was proposed by Borji *et. al.* [BSI13], by using a random uniform sampling of the pixels as negatives, and so defining false positives as the pixels in this set where the saliency map values are higher than the threshold. This is a discrete approximation of the FPR made by AUC-J, and thus is less computationally costly.

Finally, a common shortcomming of visual saliency models is the way they include center bias. Indeed, eye fixation densities on an image will often exhibit higher values

towards the center of the image [Tat07]. Because of that, a model including such bias will be able to predict well at least a part of the fixations, independently from the considered stimulus, especially if the considered dataset has a strong center-bias. In order to penalize this behavior, Tatler *et. al.* [TBG05] introduced the shuffled AUC metric (sAUC), where the negatives are sampled among fixation locations from other images rather than uniformly random. This results in sampling negatives mostly from the center, and thus penalizes models incorporating center biases.

AUC metrics are then computed by varying the threshold, and doing so, measure different aspects of the saliency map compared to ground truth. Indeed, lower thresholds will measure the coverage similarities, while higher thresholds will measure peak similarities [Byl+19].

Normalized scanpath saliency (NSS)

The normalized scanpath saliency (NSS) is a metric comparing a predicted saliency map to ground truth fixations [Pet+05]. The saliency map is first normalized, such that the mean is zero, and unit standard deviation. Then, the normalized saliency values are evaluated at the fixation locations. For a predicted saliency map $\hat{S}$ and a ground-truth binary fixation map F (i.e. a matrix where the value of the fixated pixels is 1, and has 0 on all of its other coordinates), the value of the NSS is:

$$\begin{aligned} NSS(\hat{S}, F) &= \frac{1}{N} \sum_i \bar{S}_i F_i \\ \text{where} \quad \frac{1}{N} &= \sum_i F_i \quad \text{and} \quad \bar{S} = \frac{\hat{S} - \mu(\hat{S})}{\sigma(\hat{S})} \end{aligned} \tag{2.6}$$

iterating over the pixels i, and N is the number of fixated pixels. The chance level of the NSS metric is 0, negative values indicate anti-prediction, and the higher the value, the better the prediction. This measure is particularly sensitive to false positives.

Information gain (IG)

Information gain (IG) is a metric inspired by information theory, proposed by Kümmerer *et. al.* [KWB15] to measure the amount of information predicted by a saliency model beyond a given baseline, usually a centered-bias. It assumes that the saliency map output by the model is a fixation probability density, properly regularized, and that the model

includes a center prior. In this case, given a predicted saliency map $\hat{S}$, a ground-truth binary fixation map F and baseline map B, the information gain is:

$$IG(\hat{S}, F, B) = \frac{1}{N} \sum_i F_i \left(\log(\varepsilon + \hat{S}_i) - \log(\varepsilon + B_i) \right) \tag{2.7}$$

iterating over the pixels i, with N the number of fixations and ε a regularization parameter, similarly to the KL case. An IG score above zero will indicate that the model predicts fixation locations better than the considered baseline. Another interesting property of this measure is that it allows model comparison: the baseline map can be a saliency prediction from another model; the metric will then quantify how much improvement is brought by the new model.

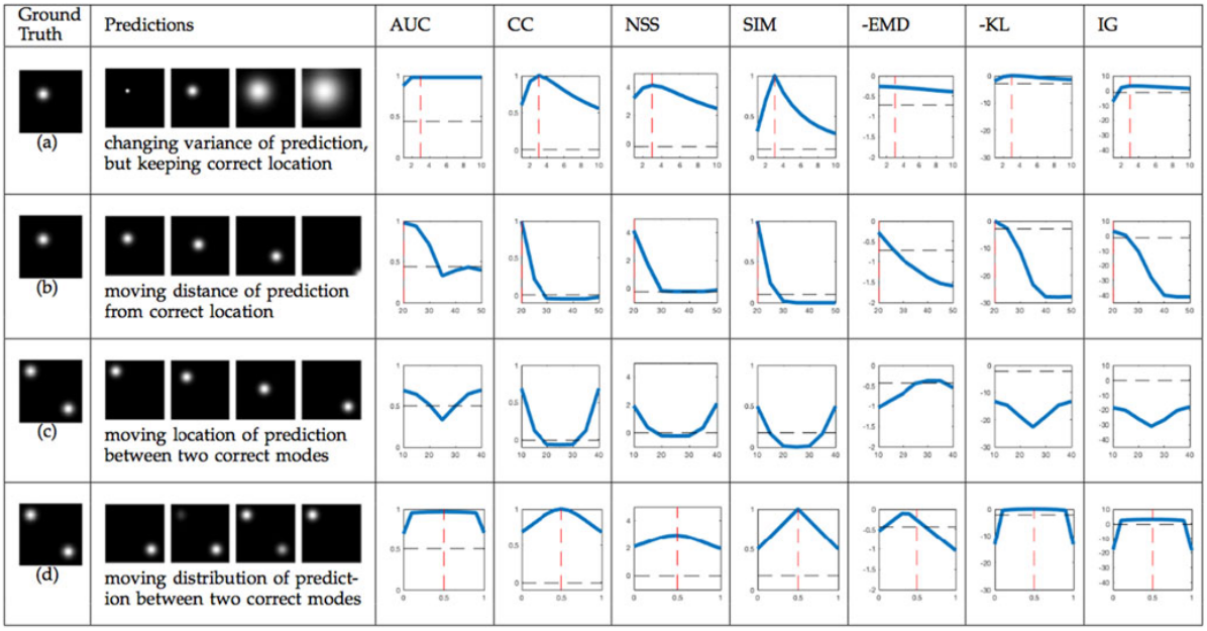


Figure 2.2 – Variation of parameters of a saliency map in order to quantify effects on metric scores. Each row corresponds to varying a single parameter value of the prediction: (a) variance, (b-c) location, and (d) relative weight. The x-axis of each subplot spans the parameter range, with the dotted red line corresponding to the ground truth parameter setting (if applicable). The y-axis is different across metrics but constant for a given metric. The dotted black line is chance performance. EMD and KL y-axes have been flipped so a higher y-value indicates better performance across all subplots. Reproduced from Bylinskii *et. al.* [Byl+19]

2.3.3 The probabilistic framework

Recently, Kümmerer *et. al.* [KWB18] proposed a new framework to evaluate visual saliency methods. They argue that the variety of the metrics, and the fact that they each evaluate different properties of the saliency predictions makes it difficult to consistently rank and benchmark models. Indeed, they show that a saliency map usually cannot perform well according to every metric.

Instead, they propose to differentiate the notions of saliency model and saliency map. A *saliency model* should be intrinsically probabilistic and output a *fixation probability density*, while a *saliency map* should be a metric-related computation, derived from the density prediction. For each metric, they propose a method to derive the optimal saliency map from the density prediction, maximizing the expected performance on that metric. This way, accurate models should perform well according to all measures.

This approach has become a standard, as it has become the new way of evaluation of the MIT/Tübingen benchmark [Küm+]. However, this implies radical changes to saliency models, and makes the comparison with older saliency models harder, as the desired saliency output is fundamentally different. Moreover, the process of designing a saliency map for each metric from a fixation density prediction makes it computationally costly. This also explains why this approach is not yet used to evaluate dynamic saliency models, which usually involve way more data (i.e. one prediction per frame). This is why, while acknowledging the interest of the probabilistic framework, we will use the more traditional metrics in the rest of this thesis.

2.4 Static models of attention

In the two following sections (Section 2.4 and 2.5), we give an overview of the methods used to predict bottom-up human visual attention. The objective is not to give an exhaustive review of all the existing models and their performance (we refer the interested reader to the following reviews: Riche *et. al.* [Ric+13], Borji and Itti [BI13], Borji [Bor19], Wang *et. al.* [Wan+19]), but rather to give a short exploration of the main frameworks and models used to create saliency maps.

2.4.1 Traditional methods

During the pre-deep learning period, a significant number of saliency models was introduced, and numerous survey papers looked into these models and their properties. Most of those models usually used a three-stage approach : (1) **feature extraction**, where feature vectors are extracted from the image at various locations, (2) **activation**, where one or multiple activation maps are computed based on the feature vectors, and (3) **normalization** (and/or combination), where the activation maps are unified into a single saliency prediction.

In their seminal 1998 paper, Itti *et. al.* [IKN98] used cognitive and neuro-physiological concepts to extract features from the images. Inspired by the *feature integration theory* in the study of human visual system, they create three feature channels, for color, intensity and orientations. The image is subsampled by a Gaussian pyramid, and these features are extracted at each level of the pyramid, before being normalized and linearly summed into "conspicuity maps", once more linearly combined into a single saliency map. A lot of cognitive-influenced models were proposed, in the idea of modeling the structure of the human visual system, or a subset of it. Le Meur *et. al.* proposed one [Le +06], where they implemented contrast sensitivity functions, early visual features extraction, masking, perceptual grouping and centered-surround interactions. Murray *et. al.* [Mur+11] also introduced an interesting model following this idea, where the image is processed based on early human pathway (color and luminance channels, with a multiscale decomposition), followed by a normalization inspired by the inhibition mechanisms performed by the visual cortex cells, and the integration of the resulting maps with an inverse wavelet transform, using biologically-justified weights.

Several other approaches have been proposed, for instance based on information maximisation [BT05]: they use Shannon's self-information measure on RGB patches of images, which dimension is reduced using independent component analysis, to compute the information a region conveys relatively to its surroundings, and infer the saliency map from there. Other techniques involve for instance graph-based methods to normalize and combine feature vectors [HKP06b; AL10], or Bayesian modeling to combine feature vectors to contextual priors [Zha+08].

2.4.2 Deep-learning era

Thanks to the deep learning revolution in the early 2010's, and alongside with the growing availability of large-scale eye-fixation databases, the field of saliency models has seen a renewed interest, while the performance of saliency models drastically improved. The characteristics of most of the models shifted towards data-oriented models based on deep convolutional neural networks (CNNs). The deep saliency models fall into two categories, (1) those using CNNs as fixed feature extractors and learn a regression from feature space into saliency space using a non-neural technique, and (2) those that train a deep saliency model end-to-end.

The number of models belonging to the first category is limited, as it quickly appeared that end-to-end approaches lead to significantly better performance. For instance, Vig *et. al.* [VDC14] use a hyperparameters search to optimize the blending of features learned by several deep neural networks on image classification tasks. They then use the resulting feature vector to learn a linear support vector machine (SVM) to perform fixation prediction. Similarly, Tavakoli *et. al.* [R T+17] extracts deep CNN features, and then uses a set of extreme learning machines trained on an image similar to the input image.

Within end-to-end deep learning techniques, the main research has been on architecture design. Many of the models borrow the pre-trained weights from an image recognition network and experiment combining different layers in various ways. In other words, they engineer an encoder-decoder network that combines a selected set of features from different layers of a recognition network. In the following we discuss some of the most well-known models.

Huang *et. al.* [Hua+15] proposed a multi-scale encoder based on VGG networks [SZ14b] and learns a linear combination from responses of two scales (fine and coarse). Kümmerer *et. al.* [KTB15] use a single scale model using features from multiple layers of AlexNet. Similarly, Kümmerer *et. al.* [Küm+17] and Cornia *et. al.* [Cor+16] employed single scale models with features from multiple layers of a VGG architecture.

There has been also a wave of models incorporating recurrent neural architectures. Han and Liu [LH18] proposed a multi-scale architecture using convolutional long-short-term memory (ConvLSTM). It is followed by [Cor+18b] using a slight modified architecture using multiple layers in the encoder and a different loss function. Recurrent models of saliency prediction are more complex than feed-forward models and more difficult to train. Moreover, their performance is not always significantly better than some recent feed-forward networks such as EML-NET [Jia18].

Generative adversarial networks (GAN) have also been investigated by Pan *et. al.* [Pan+17], where they train a traditional encoder (VGG16)-decoder backbone, with a trained adversarial loss function, discriminating between the generated saliency map and the ground-truth fixation density. Che *et. al.* [Che+20] used the same idea, with a modified U-Net as the generator and a "centered-surround connection" module to increase the model non-linearity.

More recently, Kroner *et. al.* [Kro+20] proposed an architecture where multi-level activation maps from a VGG16 backbone to capture information at different scales, before using an atrous spatial pyramidal pooling module, and a decoder composed of convolution and upsampling layers. Finally, Droste *et. al.* [DJN20] proposed a light-weighted structure incorporating new domain adaptation techniques (domain-adaptive priors, fusion and smoothing, and bypass RNN), in order to unify saliency prediction for both static and dynamic stimuli.

2.5 Dynamic models of attention

In recent years, the prediction of eye fixations on dynamic stimuli has received a significant gain of research interest. As the number of applications including video content (video compression, captioning, action recognition, etc) increases, video saliency detection has become a more and more important part of the visual attention research field. Similarly to the static case, the methods to predict visual saliency on videos can be separated into recent deep-learning based methods, and older more traditional approaches. We refer the reader to Wang *et. al.* [Wan+19] for a review and quantitative benchmark of those models.

2.5.1 Traditional methods

In 2005, Le Meur *et. al.* [Le +05] proposed, and later refined [LLB07] a dynamic visual saliency model, where they include temporal features, inspired by certain areas of the visual cortex, and justified by the assumption that motion contrast is a strong attentional attractor. They compute a temporal feature map using a hierarchical block matching to infer the local motion at each point, and create a motion-contrast map by removing the local motion to the dominant motion, computed using M-estimators. Finally, similarly to the static traditional approaches, they normalize and combine linearly the motion-saliency map and a traditional spatial saliency map, including inter-map competition to

detect redundancy.

These types of approaches, consisting of extending traditional static models using temporal features, before combining the resulting maps, was widely used. In an extension of the seminal model [IKN98], Peters and Itti [PI08] included to the previous color, orientation and intensity channels, a motion and a flicker channel, dedicated to grasp the temporal content. The overall architecture of the model remains the same, with extracting center-surround maps from multiscale feature pyramids, creating conspicuity maps, and finally combining all of it into a single dynamic saliency map. Similar approaches were used by Marat *et. al.* [Mar+08] and Gao *et. al.* [GMV07].

Guo and Zhang [GMZ08] later proposed a new method: instead of combining conspicuity maps, they consider each pixel of a frame as a quaternion consisting of color, intensity and motion features. Then, they use the phase spectrum of the quaternion Fourier transform (PQFT), and convolve a Gaussian kernel to this representation to create the final saliency map.

While, in compression algorithms, visual saliency maps are commonly used as inputs, Khatoonabadi *et. al.* [Kha+15] use the inverse assumption. They use a score of compressibility, the operational block description length, to measure the saliency of an area in a video. The idea is that video compressors usually process spatio-temporal blocks differentially, predicting a block from its neighbours (spatially or temporally). If this prediction is ineffective, the block will require more bits to compress, as the residuals of the prediction are higher. By measuring this number for each block, and then smoothing with a spatial and a temporal Gaussian kernel, this results into a prediction of the saliency.

Finally, Leborán *et. al.* [Leb+17] proposed an approach based on whitening of the spatio-temporal features to remove the correlations and variances of the data, only to use high-order statistical information. Blocks of seven frames are separated into three color channels, chromatically whitened, and then passed through a temporal and a spatial frequency decomposition, to compute the spatial and the temporal components of the final saliency map. These representations are then normalized and combined using a competitive weighted sum, where the weights are proportional to the relative significance of the corresponding maps.

2.5.2 Deep-learning models

Similarly to static saliency models, deep learning techniques allowed the performance of the saliency prediction to drastically improve, and created a significant gap, both in

accuracy and run-time, between deep dynamic models and other approaches. Most of these new models fall into two general frameworks : (1) a two-stream approach, where the temporal and static information are extracted separately and then fused together, and (2) a sequential approach, where the spatial saliency features are extracted on each frame, and fed to a LSTM network to incorporate the temporal content. Recently, multimedia approaches have also been investigated, with the addition of audio streams to the dynamic saliency models.

Bak *et. al.* [Bak+18] proposed the first deep dynamic modeling in 2018. It consists in two encoding CNN streams, one for the spatial and one for the temporal features, and a fusion CNN to predict the saliency map. In order to only consider motion in the temporal stream, they use the optical flow of the sequence as an input. Zhang *et. al.* [ZC19] introduced a similar architecture, where they use a VGG16 features extractor backbone to get features from successive frames, and process them using 3D convolutions in the motion stream to get temporal cues.

Jiang *et. al.* [Jia+18] proposed a mixed architecture: two CNN streams are used to extract objectness (using a pruned YOLO [Red+16] network) and motion (using a pruned FlowNet [Dos+15] network) features from consecutive frames, before concatenating the resulting spatio-temporal features. They are then passed through a convolutional LSTM network, generating inter-frames saliency maps. This idea of combining extracted features with a convolutional LSTM is also used by Wang *et. al.*[Wan+18], where a deep CNN based on a VGG16 network is used to learn intra-frames static features, which are then used by the convolutional LSTM to learn sequential saliency representations. A similar two-streams recurrent approach is also successfully used by Zhang *et. al.* [ZCL21] and Lai *et. al.* [Lai+20].

TASEDNet [MC19] relies on another different type of approach: the authors propose a 3D fully-convolutional network, with an encoder part extracting spatio-temporal features, and a decoder part, creating the spatio-temporal saliency map. This choice of architecture is motivated by the recent good performance of action-recognition 3D convolutional networks, proving that these kinds of models are successfully extracting relevant motion features. For the encoder network, they use pre-trained weights from an action-recognition model. The decoding network takes inputs from different pooling layers in the encoder, in order to treat features at different levels. Finally, Bellitto *et. al.* [Bel+21] take a similar approach, using a 3D convolutional encoder, from which they extract feature at each level. The features are then passed through a multi-branch convolutional decoder to create mul-

tiple layers of conspicuity maps, which are then fused using a point-wise convolution into the final saliency prediction.

Audio-visual approaches

Recently, audio-visual approaches have been investigated, with the idea that directed sound also affects visual attention. While performance can marginally be improved by including this information, the interest of the approach is still very much linked to the experimental conditions in which the data was gathered (headsets or speakers, mono oro stereo, number of sources, etc.).

Tavakoli *et. al.* [Tav+20] proposed a two-stream network based on 3D ResNets [HKS18]. The frames are fed to a 3D ResNet pretrained on action-recognition tasks, and the audio channel is converted into a sequence of Mel spectrograms, also fed to a 3D ResNet, retrained on audio-classification task. The visual and auditory features are then concatenated and fed to a 2D CNN to create the final saliency map. The overall model is trained end-to-end on both static (for the visual branch) and dynamic stimuli. Similar approaches (i.e. a two-stream method, separating the visual branch and the audio branch) are also used by Chang *et. al.* [CZZ21] and Jain *et. al.*, using 3D CNNs to extract spatio-temporal information.

2.5.3 Static saliency for dynamic stimuli

Even though previously mentioned models are specifically designed to extract and make use of temporal information, their overall performance can be a little bit disappointing when compared to performance of static models. Indeed, on the DHF1K benchmark [Wan+], for instance, static models (like SALGAN, or SALICON) applied to frames one-by-one sometimes rank higher than some dynamic models (DeepVS for example). Tangemann *et. al.* [Tan+20] explored this phenomenon, and showed that on the LEDOV dataset, more than 75% of the information defined by a gold standard model (i.e. humans fixations predicting other humans fixations, or *inter-observer visual congruency*) can be explained by static features, completely ignoring temporal information. They conclude that this is probably due to a representation bias in the existing video datasets: while temporal effects affecting visual deployment exist, and can have a tremendous influence on attention, they occur relatively rarely in the datasets usually used to benchmark dynamic saliency models. This highlights the shortcomings of the fully data-driven ap-

proaches, where a lack of relevant examples in the data can lead to bad performance. Indeed, they show that the recent dynamic saliency models all fail in the same situations and on the same sequences.

2.6 Conclusion

In the last few years, thanks to the advances in deep supervised learning, the performance of visual saliency models have significantly improved. The growing availability of large datasets of eye fixations, allowed by the technological advances in eye-tracking, makes this kind of architectures and designs particularly efficient, from an application point of view. However, there still remain important shortcomings that need to be addressed by future efforts.

Datasets remain challenging to create, and the largest ones, like SALICON [Jia+15], often rely on heuristics, like using mouse-tracking instead of eye-tracking. The lack of unified framework to conduct eye-tracking experiments for visual saliency prediction make for somewhat inconsistent datasets that can prove difficult to merge together for training a model. It also induces various biases caused by the variations in experiment conditions: the task at hand, the resolution of the displayed stimuli, the lighting settings of the screens, and many more factors can have an impact on the gathered eye fixations.

In his review of deep visual saliency models, Borji [Bor19] pointed out that one of the most important failure case of recent models is their lack of understanding of high-level information. The *semantic gap* is still an important issue: while models are very good at detecting faces or text, and labeling it salient, it becomes way more difficult when it comes to prioritizing which face or what pieces of text is the most important, for instance. It is unlikely that the solution will solely come from larger and larger datasets, and deeper and deeper models, but cognitive psychology studies on attention can probably be very useful to both detect these shortcomings and help solving them.

The main drawback of these approaches is obviously the lack of explainability: while deep learning models can inform us on the statistical patterns embedded in the eye-tracking data, it remains very challenging to understand what is actually learned.

Finally, dynamic saliency prediction has gained a great deal of interest very recently. However, it appears that very little work has been done to adapt the evaluation and the analysis procedures from static stimuli to dynamic ones: the main metrics are simply applied on a frame-by-frame basis, discarding entirely the temporal dimension. The

probabilistic framework proposed by Kümmerrer *et. al.* [KWB18], applied to a spatio-temporal space, could be for instance an interesting first step in finding appropriate ways of comparing deep visual saliency models on dynamic content.

Chapter 3

Cinematography : Giving meaning to the moving image

In this chapter, we highlight the differences between cinematography and other types of video content. We explore the way cinematic images are constructed and arranged together through the process of editing, and how to formalize these characteristics and rules. Finally, we give a literature survey of eye-tracking studies dedicated to understand visual attention when it comes to movies.

3.1 Introduction

What makes movies such a specific type of stimuli ? In the early twentieth century, one could argue that cinematography was simply putting images into motion. However, the appearance of television, and later low-cost cameras and digital videos has somewhat blurred the lines. While a movie is a video, all sorts of videos are not necessarily movies: newscasts, surveillance footage, sports broadcasts, YouTube tutorials, the list goes on and on. It is then more accurate to describe cinematography as a set of narration techniques, or at least a set of ways for conveying meaning, that, once put all together, creates the film object. Similarly to the way comic books creators use paneling, framing, coloring, dialogues (and so on) to tell a story, filmmakers use cinematography.

However, these features and techniques mostly rely on visual perception (with the notable exception of sound characteristics, i.e. music, sound design and dialogues), and therefore can have an impact on visual attention. This will be the core problematic of this thesis : **how to understand and model the relation between filmmaking techniques and visual attention ?**

We can easily understand that traditional ways of considering visual attention and dynamic stimuli, as described in Chapter 2, might not be a suitable approach, as the focus is on the video-object, i.e. a collection of frames and its low-level characteristics

(color, contrast, motion, etc), rather than the high-level contextual information, such as scene composition, that is carried out by a movie. In the following sections, we will describe what are some of these cinematic-specific high-level features, and how to formalize them in order to use quantitative approaches. Finally, we provide a survey of previous work regarding movies and visual attention.

3.2 Cinematic stimuli and their specific features

The following section is highly inspired by the books of Thompson and Bowen [TB09], and Brown [Bro16]. For more information and details about what cinematography consists of, we refer the interested reader to these works.

Filmmakers have a tremendously large array of tools that they can use to convey information and meaning, and to build a narrative. This collection of techniques creates a form of visual language, that, like words combining into phrases, can combine to create meaning. A good comprehension of these features is therefore needed, in order to understand the director's intentions, and how they influence the viewers.

3.2.1 The frame: a unit of space

The basic unit of space in a movie sequence is the frame. The picture composition, the staging and blocking of the actors (i.e. their spatial position relatively to the camera, and the way they move), the position of objects in the frame, all of those characteristics are consciously (and conscientiously) chosen by the filmmakers to transmit a message, an idea, or a storytelling point.

The size of the frame itself is of importance: the width to height ratio is called the *aspect ratio*, and is often written as the ratio to the standard frame height of 1. For instance, classical Hollywood movies have used the 1.33:1 ratio for a long time, and progressively switched to the wider 1.85:1, and even 2.40:1 nowadays (also called the anamorphic format), as the recording and projection equipment, as well as cinematography techniques and preferences evolved.

The *size of the shot* is also an important feature. It refers to the area that is covered by the main object or objects -by object, we include actors, text, or any actual object that might be of interest- relatively to the size of the frame. The larger the object area, the *closer* the shot. Shot sizes range from the extreme closeups, where, in the case of an

Figure 3.1 – Example of nine framing sizes, all appearing in the same sequence of *The Good, the Bad and the Ugly* (Segio Leone, 1966). Adapted from Wu *et. al.* [Wu+18].

actor, for instance, only a fraction of his face or body parts will take the full frame, to very long and establishing shots, where the actors or objects are barely visible. Figure 3.1 shows an example of several shot sizes, for illustration. Closeup shots are often used to give a detailed and specific view of an object, for instance to highlight the emotions of a character. Medium shots are usually used to approximate the way we experience our environment: characters not really close, but not too far away either, for example. Finally, long shots often convey information about the way objects are included in their environments, and how they might interact with one another: it gives an information not only about an object, but also its surroundings.

The position of the objects within the frame, and relatively to each other, is also important, and is referred as *framing*. The depth of field, the geometry of the environment, the chosen lens, the staging of the actors, all of those factors can be manipulated by the filmmaker to build his or her frame. Several rules, born from the decades of practice in cinematography and photography, also apply to framing. For example, the *rule of thirds*: interesting visual elements are often positioned alongside the thirds lines, i.e. the imaginary lines obtained when dividing the frame into thirds, both vertically and horizontally, and at their intersections (see for instance Figure 3.2).

3.2.2 Following the eye of the camera

In a movie sequence, the camera acts like a virtual eye, forcing the viewer to adopt its perspective and its motion. Playing with this forced point of view is obviously a very important tool for filmmakers to direct the attention and to force the audience to focus

Figure 3.2 – An example of the rule of thirds in *The Lord of The Rings: The Fellowship of the Ring* (Peter Jackson, 2001). Note how both characters are positioned alongside the vertical thirds lines, and how their heads appear at the intersection of vertical and horizontal thirds lines, even though they do not have the same apparent size.

on what they want to show and tell.

Camera angles are a useful mean to this end. By varying the angle under which a character or an object is filmed, the director can inform about the emotional state of the character, or the balance of power between several objects. For instance, when a character is filmed with a high angle (i.e. seen from above), the implied meaning is usually that something or someone is looking down on this character, either figuratively or literally. It can be used to suggest that the character is overwhelmed by its environment, or dominated by other persons. In the opposite, a low camera angle will suggest a more powerful, dominating persona. Horizontal angles can also be used to convey meaning : a front angle, with a good view on the actor's face is useful to understand the feelings of the character, while a side or a back angle, by hiding parts of the face, can imply mystery. Finally, rolled shots often convey a sense of discomfort, implying that something is wrong, or abnormal.

As the viewer has to look at what the eye of the camera shows, he or she also has to move when the camera moves. The most basic setup in this regard is the static shot : the camera is stabilized on a support, and does not move during the shot. Cameras can also be handheld, in order to track a moving object, for instance. This kind of camera motion will often create small jerky movement of the frame. When placed on tripods, cameras can also smoothly pan (i.e. rotate along the vertical axis) or tilt (i.e. rotate along the horizontal axis). Finally, heavier equipment allows the filmmakers to translate the camera, or even

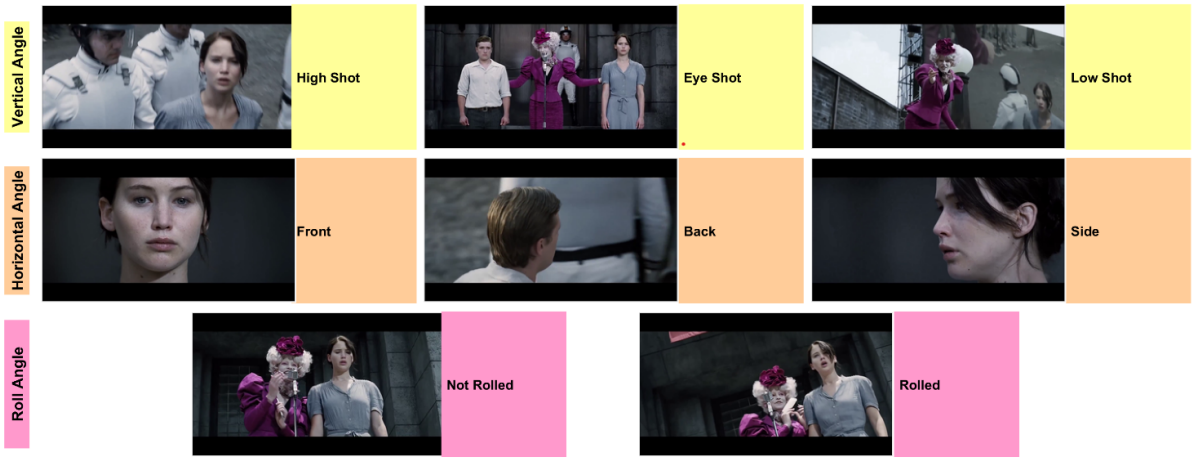


Figure 3.3 – Example of various camera angles, all appearing in the same sequence of *Hunger Games* (Gary Ross, 2012). Adapted from Wu *et. al.* [Wu+18].

to create very complex kind of motion, using cranes (Figure 3.4 illustrates some of the types of camera motion). Camera movements are a particularly effective tool for directors to direct the attention of viewers: often, a moving camera will be justified by a moving object on screen, or a switch in areas of interest, like an object appearing on the border of the frame or in the background. To this extent, zooms (i.e. a dynamic variation of the focal length) and rack focus (i.e. switching the focus from one object to another by changing the depth of field) are two other very useful techniques.

3.2.3 Editing, or how to put the shots together

Finally, when all the shots are captured, often with several cameras under several angles and points of view, the editor cuts and puts together the shots. The same way words are put together into sentences, which are then put together to form a text, shots are combined into scenes, which are often combined into acts, to create the film. The choice of when and how putting two shots together is important: it must be visually coherent and consistent, maintaining a rhythm to keep control of the viewer's attention, and convey information and meaning.

The way shots are put together also obeys certain rules, which were developed (and transgressed) over time, in order to optimize the flow of information while maintaining a continuity and a coherence in the visual stimulus. For example, the 180° rule states that, when editing a dialogue scene between two characters, the editor must establish an

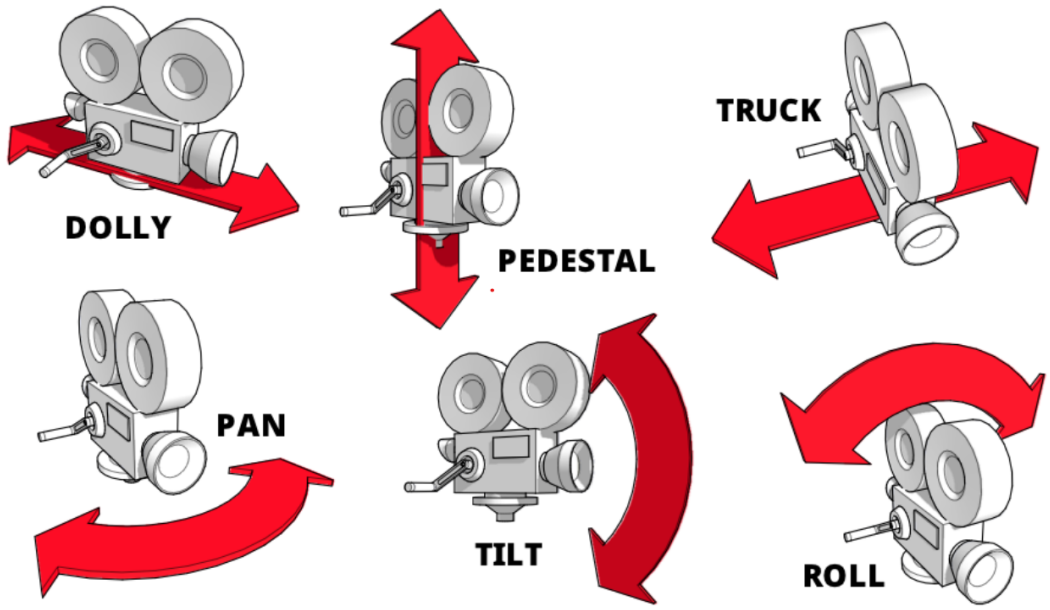


Figure 3.4 – Different types of camera motions

imaginary line between the actors, which must never be crossed by the camera in order to maintain a smooth and continuous feeling of the scene. Deviations from this rule can lead (intentionally or not) to confusing situations in which the viewer is somewhat lost, and have a harder time understanding what is happening on screen.

3.3 Virtual cinematography and formalization of cinematic rules

In order to perform quantitative studies relying on movies-specific features, various techniques had to be developed to describe a frame, a shot or a scene. These descriptive systems usually encompass a rather small set of features, like the frame composition, camera motion, or film idioms (i.e. stereotypical ways to arrange a series of shots, in order to capture at best a situation). Ultimately, these languages and features are used either for film analysis, or for the automation of a part of the film making process.

Some of the first to propose such a descriptive system were Drucker and Zeltzer [DZ94]: they proposed an idiom-based language to describe camera motion constraints, in order to generate smooth camera paths in a virtual environment. Shortly later, He *et. al.* [HCS96]

proposed to encode cinematic idioms using finite state machines, and used it alongside with camera modules implementing the placement of the cameras in the scene. Christianson *et. al.* [Chr+96] also used such idioms to design the Declarative Camera Control Language (DCCL), for camera placement planning. Other constraint-based systems for optimizing the placement of the camera include information about the occurring events in the scene [BGL98], or geometrical constraints set up by the user [BTM00].

More recently, Lino *et. al.* [Lin+11] proposed an automatic camera placement tool, included in a motion-tracked device, relying on an annotated script, which describes the objects on the set and the actions taking place. Ronfard *et. al.* [RGB13] introduced the Prose Storyboard Language, a descriptive language for cinematographic stimuli introducing elements describing the size of the objects, their placement in the frame, and their movements.

In another application, Galvane *et. al.* [Gal+15] proposed an automated editing method based on continuity editing rules. They define the importance of characters and objects, and use their apparent size an position in the frame to select which rush to select, while also considering editing rhythm and continuity rules for the edits. Leake *et. al.* [Lea+17] proposed another automated editing system based on a set of rushes and the script of a dialogue, and using idioms to propose the edits. Finally, Wu *et. al.* [Wu+18] proposed a language called Film Editing Patterns (FEP), that includes many properties of framing, shot relations within a sequence, or objects positions. They show that this language is both useful for movie analysis, but also for automated camera placement and editing of 3D animated scenes.

However, the main drawback of these approaches is that they rely on hand-crafted inputs: the various characteristics must be defined by the user itself, therefore limiting the amount of data that can be efficiently processed. In order to automate this part, several computational techniques to analyze movies have been proposed. For instance, Corridoni and Del Bimbo [CD98] proposed an automated way to detect cuts and camera motions, and used film making rules to derive high-level sequence information, such as the detection of shot-reverse shot sequences. Rasheed *et. al.* [RSS05] used several computational features, such as shot length, color variance, lighting and motion to classify movie clips into genres. This type of automated analysis can also focus on the type shot [Sva+15], or the extraction of meaningful scenes [TZ04].

Finally, Huang *et. al.* [Hua+20] recently released a very large dataset, containing 1100 movies with various annotations: scripts, subtitles, scene segmentations, boxes tagging

the characters on screen, actions annotations, shot scale and movement of the camera. The huge scale of this database allows numerous approaches of quantitative analysis and data-oriented models: for instance, Rao *et. al.* proposed a way of automatically predict the size of the shot and the motion of the camera [Rao+20b], or a segmentation model dedicated to divide a movie sequence into scenes and shots [Rao+20a].

3.4 Visual attention and cinema

Studying film perception and comprehension is still an emerging field, relying on broader studies on scene perception [SLC12; Smi13]. While the effects of low-level features have been studied in great detail, in part thanks to the progress of saliency models, the effects of higher-level film characteristics are far less well understood. Loschky *et. al.* [Los+14] showed that the context of a sequence, meaning the prior knowledge of who the characters are or what the action is about, is particularly relevant to understand the way humans are viewing a particular shot, thus underlying the need for a better comprehension of the high-level features. Valuch and Ansorge [VA15] studied the influence of colors during editorial cuts, showing that continuity editing techniques result in faster re-orientations of gaze after a cut, and that color contributes to directing attention during edits. Other studies showed strong relationships between eye movement patterns and the number and the size of faces in a scene [RPH14; CA16].

A few studies focused on gaze congruency, or attentional synchrony. Goldstein *et. al.* [GWP07] showed that observers tend to exhibit very similar gaze patterns while watching films, and that the inter-observer agreement would be sufficient for effective attention based applications, like magnification around the most important points of the scene. In subsequent studies, Mital *et. al.* and Smith [Mit+11; SM13] showed that attentional synchrony was positively correlated with low-level features, like contrast, motion and flicker. Breathnach [Bre16] also studied the effect of repetitive viewing on gaze agreement, showing a diminution of the inter-observer congruency when movie clips were watched several times.

More generally, it appears that understanding human visual attention while watching movies ultimately requires a framework combining both low- and high-level features. From a cognitive point of view, Loschky *et. al.* [Los+20] recently proposed a perception and comprehension theory, distinguishing between the front-end processes, occurring during a single fixation, and back-end processes, occurring across multiple fixations and allowing

a global understanding of the scene.

3.5 Conclusion

In this chapter, we exposed some of the proper characteristics of cinematographic stimuli, and the way these characteristics are formalized and used to automate parts of the filmmaking process or for virtual cinematography.

While significant progress has been made in understanding people understanding movies, there are still a number of complex challenges to address.

Similarly to the visual saliency problematic, the semantic gap in data-oriented models on cinematographic content remains an important issue: this type of stimulus is indeed very rich in subtext or contextual cues, and relies heavily on style and other high-level information. While the recent efforts in building large-scale datasets of annotated movies aim to address this issue, the complexity of the task and the many outliers breaking conventional cinematographic rules – how to consider for instance the use of jumps cuts in Jean-Luc Godard's *À bout de souffle* (1960)? – will probably require a multidisciplinary approach, combining knowledge from the computer vision field and cognitive film theorists.

From the point of view of visual attention, it appears that the problem have mostly been tackled from the cognitive psychology part of the field. While we now understand many effects of filmmaking techniques on the perception of the viewer, getting automated and reliable predictions of the gaze behaviors of observers watching movies is still a challenging task, for which the new computer vision deep learning models can be helpful.

Finally, we argue that a perceptual approach of film editing can be of great interest for several filmmaking applications: automated editing, virtual cinematography, camera placement, and many more problems could benefit from the cues given by visual attention modeling. In the following chapters, we will give simple examples of how this might be included and taken into account.

Chapter 4

An eye-tracking database to understand visual attention on movies

In this chapter, we introduce a new eye-tracking database dedicated to study the influence of cinematographic features on visual attention. To this extent, we propose a set of 20 movie clips, from different genres and epochs, alongside with hand-crafted annotations regarding cinematographic characteristics, such as camera motion or shot size. We then evaluate how visual attention models, and more specifically visual saliency models perform on this kind of stimuli. Finally, we show that some of the considered features tend to direct attention in a way that is not taken into account by these models. This dataset and the results that we describe here are also detailed in the paper *Where to look at the movies: Analyzing visual attention to understand movie editing* [BCM21], submitted for publication.

4.1 Introduction

Over the last century -and a few years-, directors have come to develop an instinctive knowledge of the different ways that they could direct the attention of their audience. Filmmaking provided them with a very large array of tools to use, from camera motions to stitching shots together through editing, all in the goal of conveying the message and emotions they intent to convey. In this regard, they need to be particularly aware of what draws or repels attention: the final scene of *Citizen Kane* (Orson Welles, 1941) would not be as powerful, had Orson Welles not directed the attention of the spectator on the sleigh -nor Christopher Nolan on the spinning top in the ending scene of *Inception* (Christopher Nolan, 2010)-.

In this regard, studying visual attention in movies is of great interest, whether it is

for better understanding the way human perceive movies, or for filmmakers to improve and polish their craft. As developed in Chapter 3, a lot of progress has been made on this topic, especially in the fields of cognitive psychology, leading to a deeper knowledge of the relationship between visual attention and movie making. In order to offer another point of view, this time from a computer science perspective, we want to be able to apply data-driven approaches such as machine learning and deep learning, as these types of methods offer a singular quantitative outlook.

In the case of cinematographic stimuli, high-level features directed by the filmmaker play a great role in our understanding of what is on screen, and we can hypothesise that such features should also be important for visual attention. However, while visual attention on videos is a hot topic in the field of computer vision, most of the recent work focuses on relatively low-level characteristics of the dynamic stimuli. It follows that there is a need for data specifically related to cinematographic videos, in a relatively large scale to allow data-driven approaches. As explained in Section 2.2, this kind of data is somewhat scarce, or suffer from biases, like the Hollywood-2 eye tracking dataset [MS15], where only 3 observers watched the considered movie clips without any task to do.

In this chapter, we propose a new eye-tracking database dedicated to study visual attention in movies, extending the work of Breeden and Hanrahan [BH17]. Alongside the eye fixations, we provide hand-crafted annotations regarding high-level cinematographic features. Finally, we explore how such features can create visual biases, and how visual saliency models perform on this kind of stimuli.

4.2 Dataset overview

In this section, we describe the movie clips that we considered, and their associated cinematic annotations.

4.2.1 Films and clips selection

In their effort to formalize the visual grammar of cinematography, Wu *et. al.* [Wu+18] proposed a language called *Film Editing Patterns* (FEP) to annotate the production and edition style of a film sequence. Alongside this way of describing cinematographic rules, they present an open database of annotations on several film sequences, for pattern analysis purposes. In order to simplify the annotation process of our dataset, we decided

to use the same clips.

We selected 20 clips, extracted from 17 different movies. The movies span different times (from 1966 to 2012) and genres, and are from different directors and editors, in order to eliminate bias coming from individual style. Table 4.1 gives an overview of the selected clips. The sequences were selected as they were the most memorable or famous sequences from each movie, based on scenes that users uploaded to YouTube, indicating popularity and interest to the general public.

Title	Director	Genre (IMDb)	Nb. Frames	Aspect ratio	Year
American History X	Tony Kaye	Drama	5702	1.85	1998
Armageddon	Michael Bay	Action, Adventure, Sci-Fi	4598	2.39	1998
The Curious Case of Benjamin Button	David Fincher	Drama, Fantasy, Romance	4666	2.40	2008
Big Fish	Tim Burton	Adventure, Drama, Fantasy	3166	1.37	2003
The Constant Gardener	Fernando Meirelles	Drama, Mystery, Romance	5417	1.85	2005
Departures	Yôjirô Takita	Drama, Music	10117	1.85	2008
Forrest Gump	Robert Zemeckis	Drama, Romance	2689	2.39	1994
Gattaca (1)	Andrew Niccol	Drama, Sci-Fi, Thriller	3086	2.39	1997
Gattaca (2)	Andrew Niccol	Drama, Sci-Fi, Thriller	3068	2.39	1997
The Godfather	Francis Ford Coppola	Crime, Drama	1918	1.37	1972
The Good, The Bad & The Ugly	Sergio Leone	Western	9101	2.35	1966
The Hunger Games	Gary Ross	Action, Adventure, Sci-Fi	5771	2.35	2012
Invictus	Clint Eastwood	Biography, Drama, History	2203	2.39	2009
LOTR : The Fellowship of the Ring	Peter Jackson	Action, Adventure, Drama	5109	2.40	2001
Pulp Fiction	Quentin Tarantino	Crime, Drama	3211	2.39	1994
The Shawshank Redemption (1)	Frank Darabont	Drama	5374	1.85	1994
The Shawshank Redemption (2)	Frank Darabont	Drama	4821	1.85	1994
The Shining	Stanley Kubrick	Drama, Horror	4781	1.33	1980
The Help (1)	Tate Taylor	Drama	4151	1.85	2011
The Help (2)	Tate Taylor	Drama	5244	1.85	2011

Table 4.1 – Overview of the selected clips

Here we give a small description of each scene, and its most remarkable characteristics:

— **American History X**: Flashback scene, dialogue between characters seated at a table. Mostly static shots on the faces of the characters. This scene is in black and white.

— **Armageddon**: Action scene, high frequency of edits. The shot size varies a lot, from extreme closeups to large establishing shots. A lot of camera movements.

— **Benjamin Button**: Flashback scene. A lot of camera movements tracking the characters. A narrator comments the whole sequence. Some of the shots are replicated, with variations, in order to indicate alternative possibilities in the unfolding of the narrated story.

— **Big Fish**: Crowd scene, with two main characters walking through the crowd. A few shots take place in a whole different location, with only the two characters conversing.

— **The Constant Gardener**: Dramatic scene, the camera is handheld, and follows a single character throughout the sequence.
— **Departures** : Closing scene, alternation of static camera shots. Three characters are present, but no dialogue.
— **Forrest Gump**: Flashback scene, narrated by a character. Camera movements are used to reveal actors in the scene.
— **Gattaca (1)**: Dialogue scene between two characters. A lot of play on camera angles, since one of the characters is in a wheelchair, and the other one is standing.
— **Gattaca (2)**: Dialogue scene between three characters.
— **The Godfather** : Dramatic sequence, where the edits alternate back and forth from one central quiet scene to several simultaneous dramatic situations.
— **The Good, The Bad and The Ugly**: Mexican standoff scene, with three characters, where the frequency of the edits accelerate and the shot sizes go from larger to closer as the tension builds up.
— **The Hunger Games**: Dramatic scene, alternating a lot of different camera movements, angles and shot sizes. A crowd is present, but several tricks (colored clothing, focus) are used to distinguish the main characters.
— **Invictus**: Contemplative scene, starting in a cell and ending in outdoors. Characters appear and disappear as ghosts. A narrator reads a poem.
— **Lord of The Rings**: Dialogue scene between two characters, alternating with flashbacks, mostly of action scenes. Different camera movements, angles and shot sizes.
— **Pulp Fiction**: Dialogue scene between two characters seated face to face. The exact same camera angle is used throughout the scene.
— **Shawshank Redemption (1)**: Dialogue between several characters, various camera movements, angles and shot sizes.
— **Shawshank Redemption (2)**: Flashback scene, following a single character, explaining a prison escape. A narrator comments a part of the sequence. Various camera movements, angles and shot sizes.
— **The Shining**: Dialogue scene between two characters. Very low frequency of edits, and abundant presence of the color red in the scene.
— **The Help (1)**: Flashback scene, dialogue between two characters.
— **The Help (2)**: Flashback scene, in between a dialogue scene between two characters. A lot of faces and colored clothing.

The length of the clips varies from 1 minute 30 to 7 minutes. This length is voluntarily higher than in the other datasets presented in Section 2.3, in order to allow the observer to feel immersed in the sequence, and thus exhibiting more natural gaze patterns. In total, the dataset contains roughly one hour of content. Table 4.2 show the lengths of the average shots for each sequence, and Figure 4.1 shows the overall distribution of shot lengths in the database. The high diversity in terms of shot lengths underlines the diversity in terms of editing styles. However, there is a clear tendency for relatively short shots: 70% of the shots have a length of less than 100 frames, i.e. around 4 seconds.

Sequence	**Sequence Length (s)**	**Longest shot (s)**	**Shortest shot (s)**	**Average shot (s)**
Armageddon	191.8	12.1	0.0	1.6
The Hunger Games	240.8	16.7	0.6	2.4
The Curious Case of Benjamin Button	194.7	11.8	0.3	2.5
The Godfather	80.0	6.8	0.5	2.7
Big Fish	132.1	7.6	0.7	2.8
The Constant Gardener	226.0	13.8	0.4	3.5
LOTR : The Fellowship of the Ring	213.1	8.4	0.5	3.6
The Good, The Bad & The Ugly	379,7	36.5	0.2	3.8
The Help (2)	218.8	14.0	1.0	4.0
Invictus	91.9	8.6	1.8	4.2
American History X	237.9	14.7	1.0	4.2
Pulp Fiction	134.0	12.2	1.4	4.6
The Shawshank Redemption (1)	224.2	19.2	0.8	4.7
The Help (1)	173.2	17.7	1.8	6.0
Gattaca (1)	128.7	23.7	0.2	6.1
Departures	422.0	21.8	1.8	6.6
Forrest Gump	112.2	16.6	1.8	6.7
Gattaca (2)	128.0	17.1	1.8	6.7
The Shawshank Redemption (2)	201.1	18.0	1.8	7.7
The Shining	199.5	107.1	8.8	39.9

Table 4.2 – Lengths of the sequences, and of the longest, shortest and average shots of each sequence.

4.2.2 Handcrafted high-level features annotations

Films typically contain many high-level features aiming to attract or to divert the observers' visual attention [SLC12]. These features can be of different sorts : the presence of faces or text, the framing properties, the scene composition, or the camera motion and angle, for instance. The timing of the shots, the selection of the shots from rushes by the editor and the narrative it creates are also high-level features specific to films. Audio cues, like the presence of music or dialogue can also be considered as a form of high-level movie features, and have been increasingly studied as a way to improve visual attention models [Tav+20]. However, all of those features can prove very challenging to extract

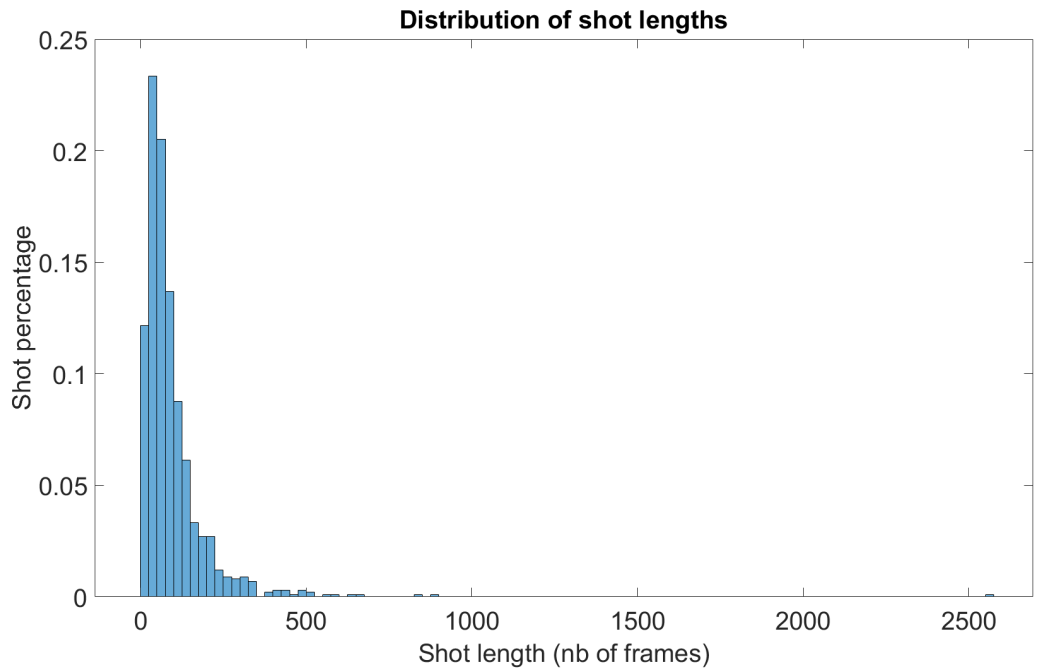


Figure 4.1 – Distribution of the shot lengths across the dataset. Note that 70 percent of the shots have a length of less than 100 frames, i.e. around less than 4 seconds.

automatically, which can explain why saliency models seem to only learn non-temporal image characteristics, at the scale of the frame, like contrast- or texture-like information. We then used the database of Film Editing Patterns described in Wu *et. al.* [Wu+18] to select a hand-crafted set of high-level annotations (described thereafter) that can help in the study of visual attention and gaze patterns on films. More particularly, such annotations enable us to conduct quantitative analysis on the influence of these cinematographic features over visual attention.

Camera motion

Camera motion is an efficient tool used on set by the filmmaker to direct attention. For each shot of the database, we differentiate several possible camera motions:

— *Static*: The camera is mounted on a stand and does not move.

— *Track*: The camera moves in order to keep an object or a character in the same region of the image

— *Zoom*: The camera operator is zooming in or out

— *Pan*: The camera rotates on the horizontal plan

— *Tilt*: The camera rotates on the vertical plan
— *Dolly*: The camera is being moved using a dolly
— *Crane*: Complex camera motion, where both the camera base and the mount are in motion
— *Handheld*: The camera operator holds the camera by hand, creating a jerky motion
— *Rack focus*: The focus of the lens shifts from one point of the scene to an other

Those features are binary for each shot, and a single shot can include different camera motions.

Camera angle

In order to convey the emotional states of the characters, or power relationships, filmmakers often use camera angles [TB09]. For instance, a rolled plan will often indicates that the characters are lost, or in an unstable state of mind, while filming actors with a low angle will give them an impression of power over the other characters, as they tower over the scene. We relied on six different degrees of camera angles [Wu+17]:

— *Eye*: The camera is at the same level as the eyes of the actors
— *Low*: The camera is lower than the eyes of the actors, pointing up
— *High*: The camera is higher than the eyes of the actors, pointing down
— *Worm*: The camera is on the ground, or very low, pointing up with a sharp angle
— *Bird*: The camera is very high, pointing down with a sharp angle
— *Top*: The camera is at the vertical of the actors, pointing straight down

Shot size

The size of a shot represents how close to the camera, for a given lens, the main characters or objects are, and thus how much of their body area is displayed on the screen. Shot size is a way for filmmakers to convey meaning about the importance of a character, for instance, or the tension in a scene. Very large shots can also be used to establish the environment in which the characters will progress. To annotate the shot sizes, we use the 9-size scale defined by [TB09]: extreme closeup (XCU), big closeup (BCU), closeup (CU), medium closeup (MCU), medium shot (MS), medium long shot (MLS), long shot (LS), very long shot (VLS) and establishing shot (EST).

Faces

As shown by Cerf *et. al.* [Cer+08], the presence of faces in images is a very important high-level information to take into account when studying visual attention. We then provide bounding boxes delimiting each face on each frame. Recent state of the art face detection models show that deep learning models extract this information very well. It is then probable that deep visual attention models are also great at extracting faces features, making this hand-crafted feature redundant. However, we include it as it permits an easier analysis of the editing style: for instance, continuity edits will often display faces on the same area of the image, while shot/reverse shots often display faces on opposite sides of the image [TB09].

4.3 Eye-tracking data collection

In this section, we describe the way that we gathered and preprocessed the eye-tracking data that we will use in the rest of this thesis.

4.3.1 Participants and experimental conduct

We have collected eye-tracking data from 24 volunteers (11 female and 13 male), aged 19 to 56 (average 28.8). Participants were split into two groups, each group watching half of the videos. Four observers were part of both groups, and viewed the whole dataset. In total, we acquired exploitable eye fixation data for 14 participants for each video.

Viewers were asked to fill an explicit consent form, and to perform a pre-test form. The objective of the pre-test form was to detect any kind of visual impairment that could interfere with the conduct of the experiment (colourblindess, or strabism, for instance). In order to ensure that they could understand English language well enough, as sequences were extracted from the English version of the movies, all of the interactions between the viewers and the operator (welcoming, description of the experiment, consent and pre-test forms) were conducted in English. Participants were informed that they could end the experiment at any moment.

During a session, subjects viewed the 10 movie sequences assigned to their group, in a random order. Sound was delivered by a headset, and volume was set before the first sequence. They could also adjust the volume at will during the experiment. After each sequence, a 15 seconds dark gray screen was displayed. After a series of five clips (around

15 to 20 minutes of video), participants were asked to make a break, as long as they needed, and fill a form, recording whether or not they could recall the scenes they saw, whether or not they had seen the movies previously, or if they recognized any actors in the scenes. After the second series of five clips, at the end of the experiment, they were asked to fill the same form. The total duration of the experiment for a participant was between fifty minutes and one hour.

4.3.2 Recording environment and calibration

Eye movements were recorded using a Tobii X3-120 eye tracker, sampling at 120 Hz. The device was placed at the bottom of a 24,1" screen with a display resolution of 1920×1200 pixels. All stimuli had the same resolution of 96 dpi, and were displayed respecting the original aspect ratio, using letterboxing. The participants were asked to sit at a distance of 65cm from the screen. They were asked to sit as comfortably as possible, in order to minimize head movements. In order to replicate natural viewing conditions, we did not use chin rests.

Stereo sound, with a sampling frequency of 44100Hz, was delivered to the participant, using a headset. Calibration was performed using the 9-points Tobii calibration process. In the case of errors of more than one degree, the participant was asked to reposition, and recalibrate. After the break, before viewing the five last clips, participants were asked to validate the previous calibration, and to recalibrate if necessary.

After recording the data for all participants, we used the following cleaning procedure. First, we ensured that every participant had a gaze sampling rate of more than 90% (i.e. more than 90% of the sampled points were considered as valid). We then kept only points that were flagged as fixations, eliminating tracking errors due to blinks or other factors, as well as points recorded during saccades. This choice was motivated by the relatively low frequency rate of the eye-tracker, making the analysis of saccadic data impossible. Then, we discarded all points that fell in the letterboxing or outside the screen. Finally, we used the position of the remaining raw points to construct binary fixation maps : for each frame, we create an image the same size of the frame, where we give the value 1 to each pixel where a fixation point was flagged during the time the frame was on screen (i.e. 1/24th of a second), and 0 to each pixel where no fixation occurred.

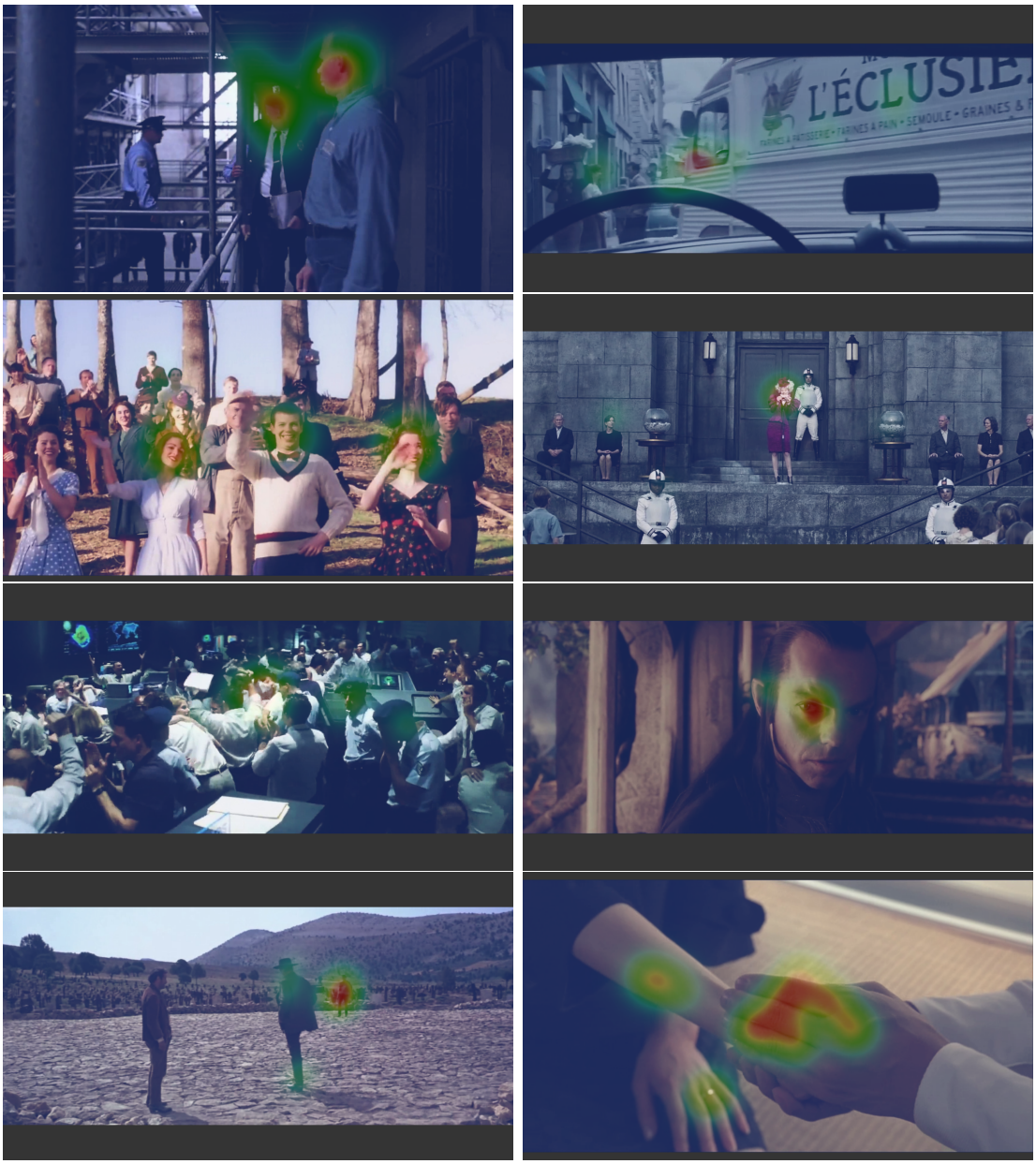


Figure 4.2 – Examples of saliency heatmaps created from the collected fixation points

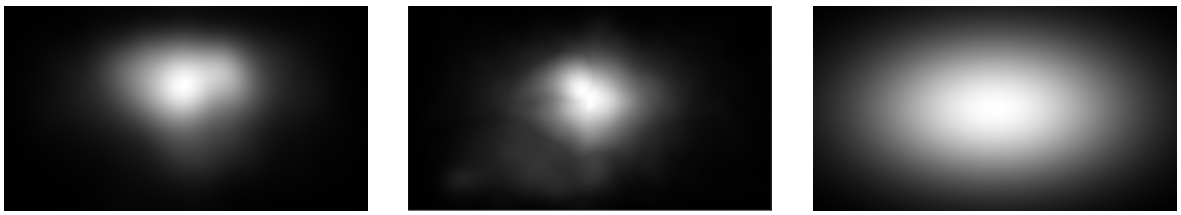
4.4 Exploring the effects of film making patterns on gaze

In this section, we explore several characteristics throughout our database, and analyze underlying relationships between editing patterns and eye fixations patterns. In the following, we will often refer to *fixation maps* and *saliency maps*. For each frame, the fixation map is the binary matrix where each pixel value is 1 if a fixation occurred at the pixel location during the frame, and 0 if not, as described previously. Saliency maps, as explained in Chapter 1, are obtained by convolving the fixation maps with a 2-D Gaussian kernel, which variance is set to one degree of visual angle (in our case, one degree of visual angle equals to roughly 45 pixels), in order to approximate the size of the fovea.

4.4.1 Editing-induced visual biases

Studying the average of the saliency maps usually reveals strong attentional biases. For instance, on static images, Tatler *et. al.* [Tat07] showed that humans tend to look at the center of the frame. That center bias is also commonly used as a lower baseline for saliency models. In order to avoid recording this bias too much, we did not take into account for our analysis the first 10 frames of each clip, as people tend to look in the middle of the screen before each stimulus. This center bias is also strong on video stimuli: for instance, Fig. 4.3 (a) and (b) shows the average saliency map on our dataset and on the DHF1K dataset [Wan+19] respectively. However, the latter is composed of Youtube videos, with a great diversity in the content, and no cinematographic scenes, which might cause a different viewing bias. Fig. 4.3 (a) shows a peak density slightly above the center of the frame, which would indicate that filmmakers use a different composition rule. Fig. 4.3 (c) shows a centered Gaussian map, often used as a baseline for centered bias. Correlation between the average saliency map on our dataset and this centered Gaussian is 0.81, whereas the correlation between the average map on DHF1K and the centered Gaussian is 0.84, which highlights this position discrepancy between the two average saliency maps. This is consistent with the findings of [BH17], and is most likely due to the rule of thirds [Bro16] stating that in cinematography, important elements of the scene should be placed on thirds lines, *i.e.* lines dividing the frame in thirds horizontally and vertically.

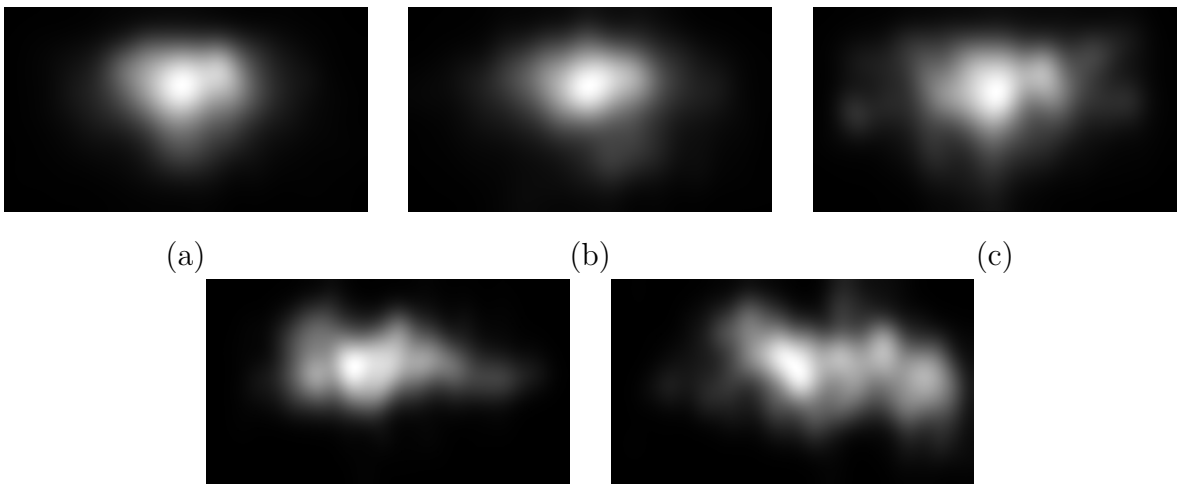
We also observe disparities in this bias depending on the size of the shot: the wider the shot, the more diffuse that bias is, indicating that directors tend to use a bigger part of



(a) (b) (c)

Figure 4.3 – Average saliency map of our dataset (a) compared to DHF1K [Wan+19] dataset (b) and to a centered gaussian map (c). Both average maps exclude the first 10 frames of each clip.

the screen area when shooting long shots, while using mostly the center of the frames for important elements during closeups and medium shots (Fig. 4.4, (a,b,c). We also observe a leftward (resp. rightward) bias during pans and dolly shots, where the camera moves towards the left (resp. right), as exposed in Fig. 4.4 (d,e). This confirms that camera movements are an important tool for filmmakers to guide the attention of the spectators.



(a) (b) (c)

(d) (e)

Figure 4.4 – Average saliency maps for closeup shots (XCU-BCU-CU) (a), medium shots (MCU-MS-MLS) (b) and long shots (LS-VLS-EST) (b). Subfigure (d) is the average saliency map during pans and dolly shots moving to the left, and (e) is the average saliency map during pans and dolly shots moving to the right.

4.5 Visual attention modeling

In this section, we evaluate several visual saliency models on our database, and highlight certain limitations of current dynamic saliency models. We also discuss how editing patterns can explain some of the failure cases of the models.

4.5.1 Performance results

In Table 4.3, we show the performance of state-of-the-art static and dynamic saliency models. In order to evaluate the models, we used the following six classic saliency metrics, described in [LB13]:

— Pearson's correlation coefficient (CC $\in [-1, 1]$) evaluates the degree of linear correlation between the predicted saliency map and the ground truth map.

— SIM (SIM $\in [0, 1]$) evaluates the similarity between two saliency maps through the intersection between their histograms.

— AUC (AUC-J, AUC-B $\in [0, 1]$) is the area under the Receiver Operator Curve (ROC). Differences between AUC-J and AUC-B relies on the way true and false positive are computed (see [LB13] for more details).

— Normalised Scanpath Saliency (NSS $\in [0, +\infty[$) is computed between the predicted saliency map and the ground truth fixation map by measuring the saliency values at the locations of the fixations.

— Kullback-Lieber Divergence (KLD $\in [0, +\infty[$) between the two probability distributions represented by the saliency maps.

In general, those results are quite low, compared to performance on non-cinematic video datasets (see for instance [Wan+19]).

This would indicate, in the case of deep-learning models, that either the training sets do not contain enough of videos with features specific to cinematic stimuli, or the deep neural networks cannot grasp the information from some of those features. Even though the best performing model is a dynamic one [ZC19], we observe that static models (DeepGaze II and MSINet) performance are quite close to those of dynamic models. This might support the latter hypothesis, that dynamic models fail to extract important temporal features.

Recent work from [Tan+20] on the failure cases of saliency models in the context of dynamic stimuli also highlight this point, refering to appearing objects, movements or interactions between objects as some of the temporal causes of failure. Figure 4.5 shows an example from our database of such a failure. It should be noted that all the deep

	Model	CC ↑	SIM ↑	AUC-J ↑	AUC-B ↑	NSS ↑	KLD ↓
Baseline	Center Prior*	0.398	0.302	0.859	0.771	1.762	2.490
Dynamic models	PQFT* [GZ10]	0.146	0.189	0.702	0.621	0.783	2.948
	Two-stream [Bak+18]	0.404	0.329	0.873	0.830	1.738	1.410
	DeepVS [Jia+18]	0.457	0.361	0.880	0.829	2.270	1.245
	ACLNet [Wan+19]	0.544	0.429	0.892	0.858	2.54	1.387
	ACLNet (retrained)†	0.550	0.423	0.890	0.858	2.592	1.408
	Zhang *et. al.*[ZC19]	**0.608**	**0.454**	**0.903**	0.881	2.847	**1.154**
Static models	Itti* [IKN98]	0.208	0.195	0.756	0.640	1.005	2.573
	SalGAN [Pan+17]	0.533	0.390	0.897	0.781	2.622	1.372
	DeepGaze II [Küm+17]	0.584	0.362	0.846	0.774	**3.188**	2.307
	MSINet [Kro+20]	0.597	0.417	0.901	**0.893**	2.893	1.226

Table 4.3 – Scores of several saliency models on the database. Non-deep models are marked with *. Best performances are bolded. †Note that the testing dataset for the retrained ACLNet model is not exactly the same as the other models, as it is a subset of half of our dataset.

learning models are trained on non-cinematic databases, with the exception of ACLNet, which include the Hollywood 2 dataset in its training base. Indeed, this base is not well-fit to learn meaningful cinematographic features, as explained in Chapter 2.

In order to confirm this hypothesis, we retrained the ACLNet model using the same training procedure described in [Wan+18]. For the static batches, we used the same dataset (SALICON [Hua+15]), and for the dynamic batches, we created a set composed of half of our videos, randomly selected, leaving the other videos out for testing (roughly 490000 frames for training, and 450000 frames for test). We only obtained marginally better results on some of the metrics (0.550 instead of 0.544 on the correlation coefficient metric, 2.592 instead of 2.54 on the NSS metric), and did not outperform the original model settings on the other. All of this would tend to indicate that some features, specific to cinematographic images, could not be extracted by the model.

4.5.2 Editing annotation and model performance

We also studied how the two best dynamic models, [Zha+20] and ACLNet [Wan+19], performed on our database, depending on shot, camera motion and camera angle characteristics. Table 4.4 shows the average results of the models depending on the annotation characteristics. We performed one-way ANOVAs to ensure that results within each table would yield significant differences. In all cases, p-values were under 10^{-5}.

As shown in Table 4.4 (a), it appears that saliency models perform relatively well

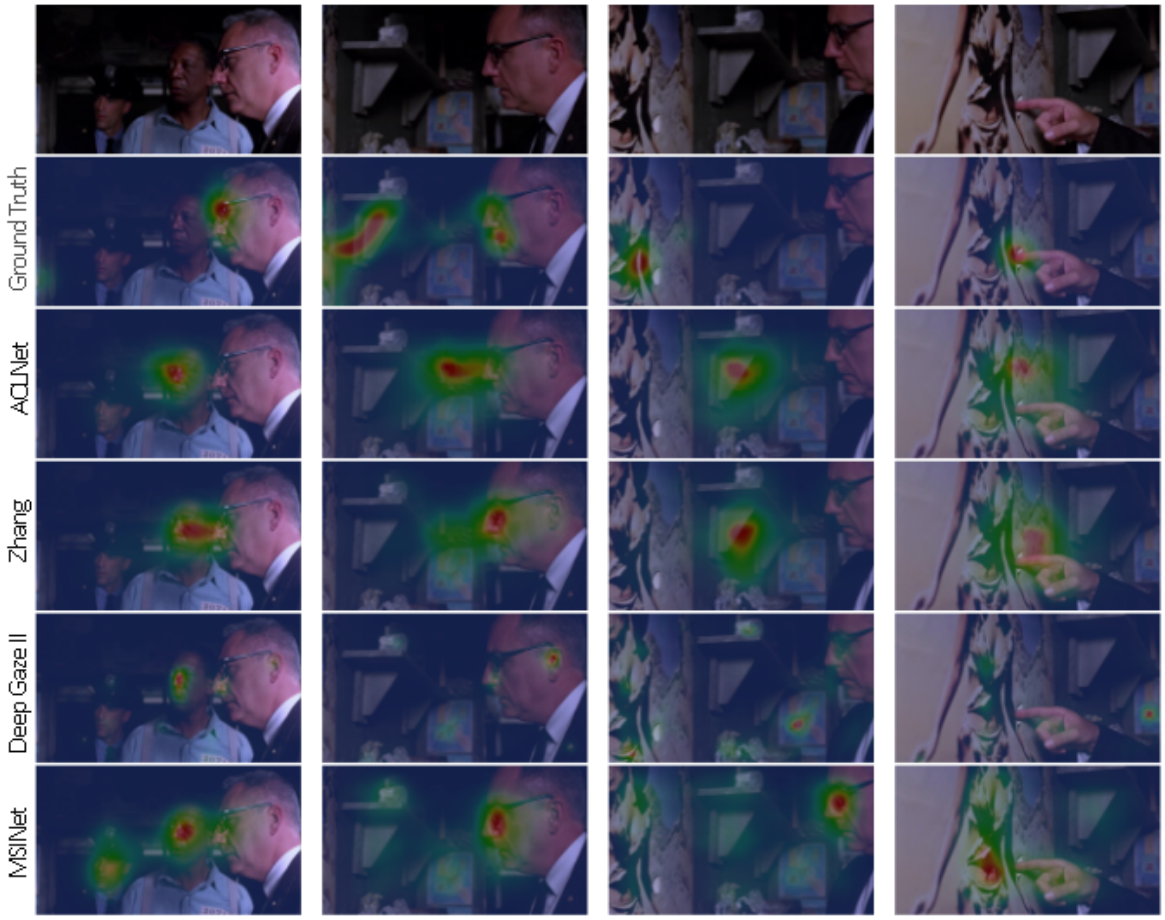


Figure 4.5 – An example of failure case in *ShawshankRedemption*. Here, the camera pans from the face of the prison director to the poster on the wall. While observers quickly shift their attention towards the poster, as suggested by the camera movement, even though it is not yet on screen, models tend to predict areas of interest on the faces.

Model	Metric	Static	Track	Zoom	Pan	Tilt	Dolly	Rack Focus
ACLNet	CC	**0.561**	0.545	0.538	*0.466*	0.488	0.517	0.545
	NSS	**2.631**	2.610	2.523	*2.138*	2.269	2.481	2.610
Zhang *et al.*	CC	0.637	0.608	0.643	*0.556*	0.584	0.615	**0.675**
	NSS	3.014	2.908	3.118	*2.615*	2.797	3.022	**3.338**

(a) Scores depending on camera motion

Model	Metric	Eye	High	Low	Bird	Worm	Top
ACLNet	CC	0.552	*0.500*	0.525	**0.544**	0.532	0.540
	NSS	2.602	*2.343*	2.465	**2.699**	2.679	2.628
Zhang *et al.*	CC	0.621	*0.582*	0.605	0.648	**0.679**	0.672
	NSS	2.932	*2.777*	2.918	3.286	**3.513**	3.375

(b) Scores depending on camera angles

Model	Metric	XCU	BCU	CU	MCU	MS	MLS	LS	VLS	EST
ACLNet	CC	0.526	0.532	**0.586**	0.549	0.497	0.510	*0.473*	0.520	0.512
	NSS	2.596	2.271	**2.689**	2.677	2.497	2.481	*2.255*	2.478	2.543
Zhang *et al.*	CC	0.656	0.607	**0.663**	0.645	0.580	0.615	*0.567*	0.628	0.636
	NSS	**3.320**	*2.679*	3.099	3.186	2.889	3.027	2.733	3.089	3.221

(c) Scores depending on shot size

Table 4.4 – Scores of two saliency models on the database, depending on hand-crafted editing features. Highest score for each metric and each model is bolded, lowest score is italicized.

on static scenes, or when the camera movement tracks an actor, or an object on screen. Performance are also quite good on shots including rack focuses, which was expected, as this is a very strong tool for the filmmaker to direct attention, and deep feature extractors distinguish very well blurry background from clear objects. However, when a more complex camera motion appears, like pans or tilts, models seem to fail more often; this might indicate that saliency models are unable to anticipate that an object is likely to appear in the direction of the motion, which humans usually do.

With Table 4.4 (b), we observe that camera angles show little variations in the performances of the models. However, it seems that scenes with high amplitude angles (Bird or Worm) are easier for a model to predict. This is probably due to the fact that those camera angles are often used when filming characters and faces, in order to convey a dominant or a submissive feeling from the characters [TB09]; since deep learning models are very efficient at recognizing faces, and faces tend to attract gaze, saliency models naturally

perform better on those shots.

Finally, looking at Table 4.4 (c), saliency models seem to exhibit great performance on closeups scenes, which could be, again, because closeup scenes often display faces. Medium to long shots are however harder to predict, maybe because a wider shot allows the director to add more objects or actors on screen, and as shown by [Tan+20], interactions between objects is often a failure case for deep saliency models. Close-up shots also display one of the lowest mean IOC, which could also explain why they are easier to predict.

4.6 Conclusion

In this chapter, we introduced a new eye-tracking dataset dedicated to study visual attention deployment and eye fixations patterns on cinematographic stimuli. Alongside with the gaze points and saliency data, we provide annotations on several film-specific characteristics, such as camera motion, camera angles or shot size. These annotations allow us to explain a part of the causes of discrepancies between shots in terms of the performance of visual saliency models.

In particular, we highlight the conclusions of Tangemann *et. al.* [Tan+20] regarding failure cases of state-of-the-art visual attention models. Video stimuli sometimes contain a lot of non-static information, that, in some cases, is more important for directing attention than image-related spatial cues. As directors and editors include consciously a lot of meaning with their choices of cinematographic parameters (camera motion, choice of the shots within a sequence, shot sizes, etc.), we would advocate researchers in the field of dynamic saliency to take a closer look at movie sequences, in order to develop different sets of features to explain visual attention.

Looking forward, we can investigate whether or not the high-level cinematic features that we provided would be of help to predict visual deployment, by building a model that includes this kind of metadata at the shot level. Another crucial point that we did not pursue is the context of the shot : the order of the shots within the sequence has been proven to influence gaze patterns [Los+14; Los+20]. As these questions have been tackled from a psychological or cognitive point of view, they remain to be studied in computer science, and to be included in visual attention models. This would greatly benefit multiple areas in the image processing field, like video compression for streaming, or automated video description.

Furthermore, we hope that this data would help cinema scholars to quantify potential

perceptual reasons to filmmaking conventions, assess continuity editing on sequences and hopefully improve models of automated edition [Gal+15].

Finally, developing automated tools to extract similar high-level cinematic information could be particularly of interest, both for the design of such tools, as it would give cues on the way to design better visual attention models on cinematographic data, but also with its outcome, as it would allow the provision of large-scale annotated cinematic databases, which would help giving a – quantitative – dimension to research on movie contents by film scholars.

Chapter 5

A visual saliency model for studying movies

In this chapter, we propose a visual saliency model, designed to be particularly efficient with cinematic stimuli. We explore how high-level information about the considered movie clip can be useful to infer where observers will look, and we propose a way of including this kind of information. Finally, we show how our model can overcome some of the difficult situations detailed on the previous chapter.

5.1 Introduction

As shown in the previous chapter, current visual saliency models suffer from serious limitations when it comes to processing videos. More specifically, they seem to fail when it comes to motion, and high-level and long-term features, such as contextual information for instance. This causes these models to perform quite poorly on movie clips, which are very rich in this type of features. While Tangemann *et. al.* [Tan+20] advocate for the development of larger databases, containing more of those difficult cases, we believe this is probably just a partial solution to a more complex problem: deep saliency models seem to be inherently unable to extract dynamic high-level characteristics which seem to play an important role directing our gaze.

Recently, the field of action recognition has brought up interesting solutions, relying on 3D convolutional networks. Deep feature extractors, like I3D [CZ17], have proven very efficient to discriminate between different kinds of motion, and to detect interactions between objects. Using transfer learning, recent dynamic saliency models have relied on such extracted features, which lead to significant improvements on the traditional benchmarks [Wan+].

Simultaneously, alternative ways of including information are being explored. Architectures including audio cues have gained a lot of attraction, even if the performance gain

remains yet unclear [Jai+21; Tav+20].

Similarly, we tried to find a way to include high-level movie information in a visual saliency model. We propose here a new saliency model, inspired by the hierarchical architecture of the ViNet model [Jai+21], and relying on two streams, one for motion and dynamic features, and the other for the spatial cues. We designed a simple, yet efficient way to include cinematic-induced bias, and manage to achieve state-of-the-art results on cinematographic stimuli.

5.2 Proposed architecture

Considering the weaknesses of dynamic saliency model when it comes to temporal features [Tan+20], we decided to use a two-stream approach: one taking as an input a stack of successive frames, and the other a stack of successive optical flow frames. Both stacks are then passed through an S3D encoder, from which features at different hierarchical levels are extracted. These feature maps are integrated at different levels by a decoding module, consisting in a sequence of 3D convolutions and upsampling, using trilinear interpolation. At the output of both streams, a saliency map is predicted. The two predictions are then merged together by a shallow 2D convolution network, which also integrates cinematic prior maps, computed upon available annotation and the frames themselves. The overall architecture is shown in Figure 5.1.

The choice of using 3D convolutional feature extractor (or more precisely *separated* 3D convolutions; more on that in subsection 5.2.1) is justified by the excellent performance of such architectures in the field of action recognition. Indeed, we can argue that video classification based on actions and visual saliency prediction are in some ways very similar, as it is strongly dependant on motion and temporal features. It is then not surprising that recent dynamic visual saliency models and action recognition models share similar traits. For instance, Carreira and Zisserman [CZ17] considered four types of architectures for feature extraction: (1) successive frames with 2D CNNs, fused with LSTM; (2) 3D CNNs, taking a stack of successive frames as inputs; (3) two-stream with 2D CNNs, with a single frame and successive optical flow maps as inputs; and (4) two-stream with 3D CNNs, using both a stack of frames and a stack of optical flow maps as inputs.

Interestingly, most dynamic saliency models fall into these categories: ACLNet [Wan+18], Linardos *et. al.* [Lin+19] and UNISAL [DJN20] use 2D CNNs alongside with recurrent units (1); TASEDNet [MC19], ViNet [Jai+21] and HD2S [Bel+21] use 3D CNNs with a

stack of successive frames (2); OM-CNN [Jia+18] use a two-stream approach, with a 2D CNN to extract spatial information from a single frame, and a flow estimation CNN, and similarly, STRA-Net [Lai+20] use 2D CNNs to extract features from a stack of optical flow maps and a single frame (3). However, the best-suited framework to deal with motion features in the work of Carreira and Zisserman was the fourth option, using two streams of 3D CNNs based on optical flow and RGB frames. We then decided to use this approach, hoping that this would remain true for saliency prediction.

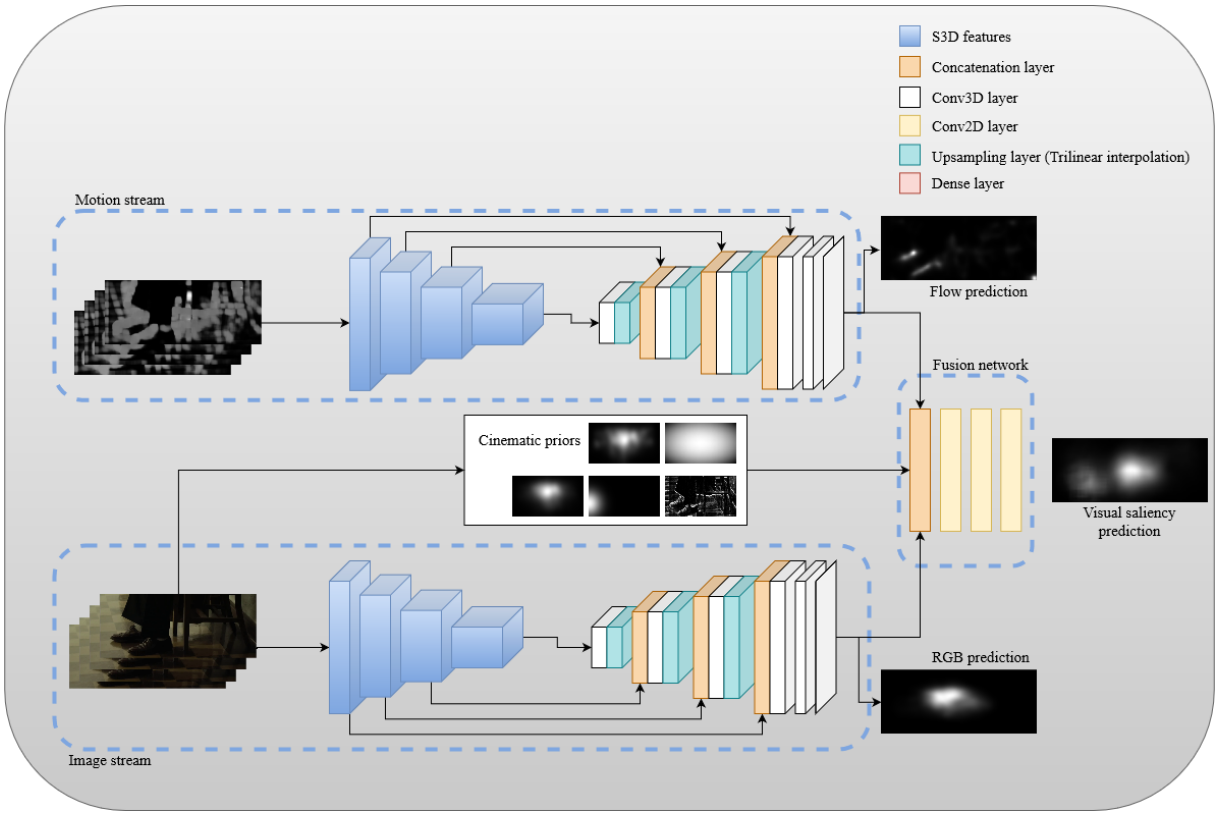


Figure 5.1 – Overview of the architecture of the proposed model.

5.2.1 S3D encoder

For the feature extraction, we relied on the popular S3D network [Xie+18]. It has been used as a backbone in many recent visual saliency models, thanks to its excellent generalization properties and its computational efficiency. Indeed, instead of relying on 3D convolutions, which are computationally heavy and prone to overfit, they disentangle the spatial and the temporal operation by replacing each 3D convolution by a separable

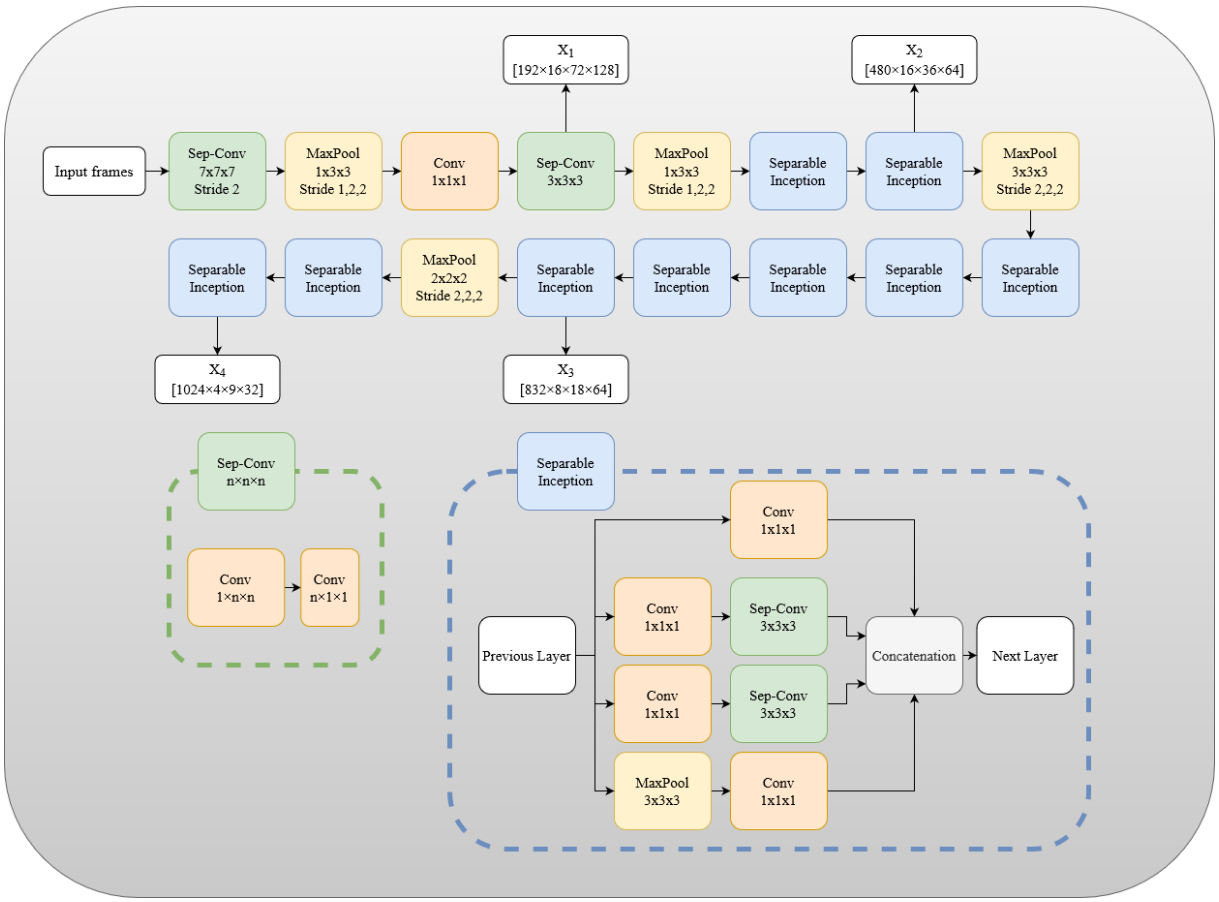


Figure 5.2 – Architecture of the S3D encoder. The green doted box shows the separable convolution operation, and the blue doted box details the separable inception module.

convolution, i.e. a 2D spatial convolution, which learns spatial features, followed by a 1D temporal convolution. The S3D network is then build by applying this to the whole I3D network [CZ17].

Figure 5.2 shows the architecture of the model. A stack of successive 3-channels frames of dimension $[T, H, W]$ is passed to the network, and passed through several separable convolutions, pooling layers and inception blocs. Inception modules are an efficient way to deepen the network while remaining efficient and avoiding overfitting: instead of stacking multiple kernel filters sequentially, they are operating on the same level, in parallel with max pooling. The resulting feature maps are then concatenated and passed to the next layer. $[1 \times 1 \times 1]$ convolutions are also added before convolutions with higher kernel sizes, in order to make the operation less costly, by reducing the dimension.

In our model, we extract features from this network at four different levels, at the

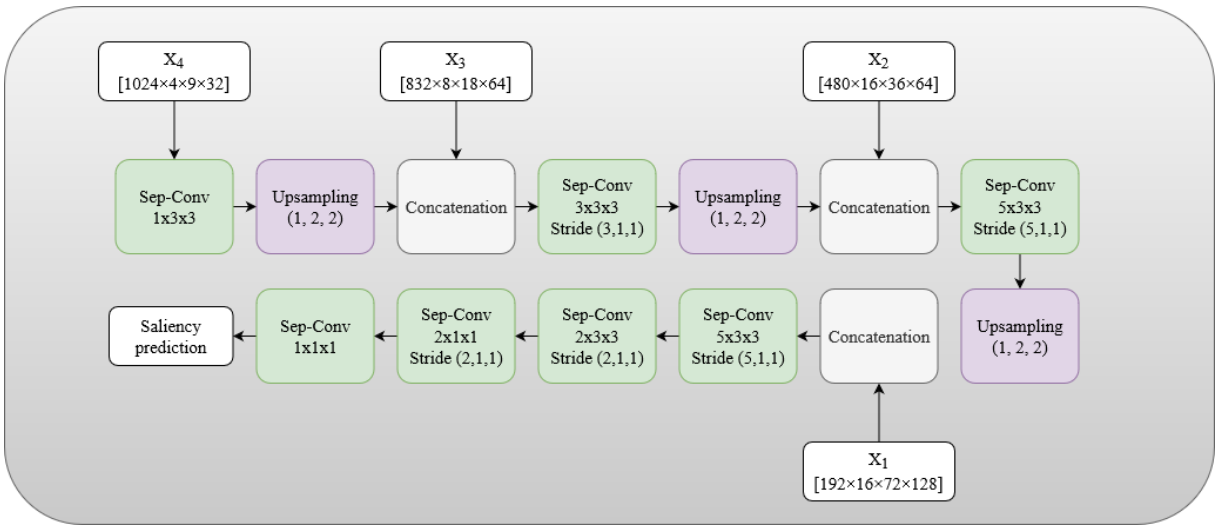


Figure 5.3 – Architecture of the decoder module. It is based on the ViNet architecture [Jai+21], replacing the 3D convolution layers by separable convolutions.

end of each convolution (or inception) block. Doing so, we extract four feature tensors, X_1, X_2, X_3 *and* X_4, of respective dimensions $[192 \times \frac{T}{2} \times \frac{H}{4} \times \frac{W}{4}]$, $[480 \times \frac{T}{2} \times \frac{H}{8} \times \frac{W}{8}]$, $[832 \times \frac{T}{4} \times \frac{H}{16} \times \frac{W}{16}]$ and $[1024 \times \frac{T}{8} \times \frac{H}{32} \times \frac{W}{32}]$. These features are extracted for both the sequence of RGB frames and the sequence of optical flow maps; indeed, Xie *et. al.* [Xie+18] showed that this architecture is also efficient to extract features from optical flow inputs.

5.2.2 Decoder module

The decoder module is based on the architecture of the ViNet model [Jai+21], and is shown in Figure 5.3. We chose to use this design due to its simplicity, efficiency and robustness, allowing us to build more components on top of it. The extracted X_1, X_2, X_3 and X_4 tensors are injected at different levels of the decoding process, by concatenation with the previous upsampled feature maps.

Inputs pass through several separable convolution blocks (as for the encoders, using separable convolution instead of 3D convolutions is performed to reduce the possibilities of overfitting, and to improve efficiency), and are regularly upsampled in the spatial dimensions, using trilinear interpolation. Finally, a saliency prediction is built, of size $[\frac{H}{4}, \frac{W}{4}]$. This prediction is done by two of these decoder modules, for both RGB inputs and optical flow inputs.

5.2.3 Cinematic feature maps

We then designed a way to include cinematographic information, by computing priors and features maps based on the available high-level editing annotations. Five different maps are included:

— a global centered-bias prior;
— a shot size specific bias;
— a camera motion bias map;
— a cut-dependent prior;
— a flicker map;

These different maps are then concatenated with the outputs of the motion and the image streams, and passed through the fusion network.

Centered-bias prior

It is a common practice for saliency models to include a prior in order to take into account the center-bias in eye fixations. In our case, for cinematic stimuli, as discussed in Chapter 4, this bias is not centered, but rather located on the upper horizontal third line. We then replace the traditionally used centered Gaussian by the average fixation distribution map of our entire dataset.

Shot size bias

Similarly, we showed in Chapter 4 that the shot size also induces a specific bias. If the information about the shot size is available, we then include the average fixation distribution map on all the shots having the same size of our dataset; if it is not, an empty map is used.

Camera motion bias map

In the shots where the camera is moving from one object to another, an interesting phenomenon illustrated by Figure 4.5 is that observers sometimes tend to anticipate the appearance of the new object by fixating the border of the frame in the direction of the camera motion, and that even though there might still be salient objects elsewhere. It is obviously a very hard task for a visual saliency model to predict such behaviors, as it needs to anticipate the possibility of something salient appearing.

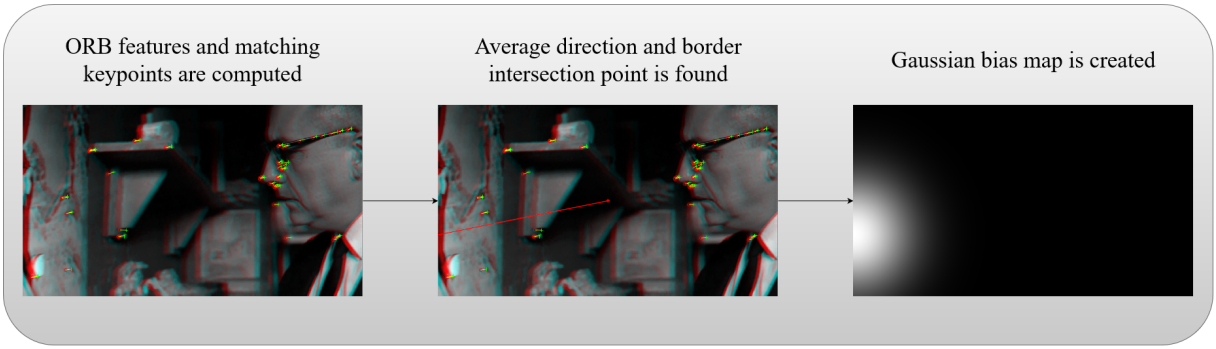


Figure 5.4 – Computation of the camera motion bias map

In order to account for this event, we compute a bias map based on the estimated camera motion. First, we use editing annotations to assess whether or not the camera is moving, and if it is, we ensure it is not a tracking motion. Indeed, the anticipation of the appearance of an object on screen will not take place if the camera is just following a salient object in order to keep it rather centered in the frame.

Then, if we consider the motion at time t, we extract keypoints using the ORB features (Oriented FAST and Rotated BRIEF [Rub+11]) on the frames $F_{t-5}, ..., F_t$. ORB descriptors are an efficient and robust way to match spatial points in two scenes with viewpoint changes, and in our case, will be used to interpolate the direction of the camera motion in the plane of the frame. Keypoints are then matched between each pair of frames $\{F_{t-i}, F_{t-i+1}\}$, and for each match, we compute the direction and the amplitude (as a L_2 norm) of the shift. We then discard the 10% of matching pairs with the highest shift amplitude, in order to account for eventual mismatches, and compute the average direction. Finally, we average one more time the direction between the five pairs of frames, and get a resulting vector representing the camera motion in the plane of the frame.

We then compute the intersection point between the border of the frame and the ray, starting from the center of the frame and following the computed direction. The final step is building the bias map, by creating a 2D Gaussian centered on the intersection border point.

Cut-dependent prior

A common phenomenon, observed by Dorr *et. al.* [Dor+10], Mital *et. al.* [Mit+11] and many other works, is the tendency, immediately after a cut, to make their first fixations towards the center of the screen. To take that into account, we include a cut-dependent

prior: if the current frame falls within 500ms following a cut (i.e. in the 12 frames following a cut, for a 24 fps clip), we include a centered Gaussian map as a prior; in all the other cases, we pass an empty map.

Flicker map

As pointed out by Tangemann *et. al.* [Tan+20], a difficult scenario to predict fixations is the sudden appearance of objects on the screen. In order to do so, we include the flicker, i.e. the change of luminance over time, as it might be an interesting feature to detect such sudden appearances.

The flicker is integrated by computing a flicker map: at time t, we consider frames $F_{t-4}, ..., F_t$, and transfer them from RGB to the CIE-LAB color space. We then compute the absolute difference of the frames luminance values $(L_{t-4}, ..., L_t)$, and average it:

$$Fl_t = \frac{1}{N} \sum_{i=1}^{N} |L_{t-i} - L_{t-i+1}| \tag{5.1}$$

Where Fl_t is the flicker map at time t and N is the number of successive frames considered. In our case, we use $N = 5$, similarly to Smith and Mital [SM13], in order to reduce the influence of noise due to compression artifacts.

5.2.4 Fusion network

The final stage of the model is a simple 2D CNN dedicated to fuse the outputs of the image stream, the optical flow stream, and the cinematic feature maps all together. It simply consists in two successive convolution layers with respective output channel sizes 64 and 128 and kernel size $[3 \times 3]$, and a final convolution to output the saliency prediction. The resulting map can be later upsampled if needed, as it is of dimensions $[\frac{H}{4}, \frac{W}{4}]$.

5.3 Training

To train this model, we use sequences of 32 consecutive frames and optical flow maps. To predict the saliency map for a frame F_t, we use the previous frames $F_{t-31}, ..., F_t$. If the current frame is among the 31 first frames of a clip, we just repeat the first frame of this clip the adequate amount of times. When training the model, random sequences of

32 frames (and their associated flow) are selected, and passed forward. Frames are resized and padded to keep the original aspect ratio, to dimensions $[288 \times 512]$

The optical flow is computed the same way as the original S3D model: the TV-L1 algorithm [ZPB07] is used to extract optical flow from the consecutive frames. The flow is then truncated into the range $[-20, 20]$, and encoded as JPEG files.

Pre-trained weights on the Kinetics dataset [CZ17] are used, for both the image and the motion streams.

5.3.1 Training phases

We trained this model in three phases. First, each stream is trained separately on the DHF1k dataset. For this, we use the annotated part of the dataset, composed of 700 clips. We used the original split of 600 clips for training, and 100 clips for validation.

Then, the whole model is trained end-to-end using the Hollywood2 dataset [MS15]; the 823 train sequences are split into a training set of 700 training clips and 123 validation clips. While, as discussed in Chapter 2, this dataset is not properly a free-viewing base, the sheer amount of sequences proves very useful to train a deep architecture like ours.

Finally, we freeze the weights of the motion and image stream, and we use 15 movie clips from our dataset (12 for training, 3 for validation) to fine-tune the fusion network, making full use of the cinematic feature maps.

5.3.2 Loss function

As mentioned in Chapter 2, there are many ways to evaluate the quality of a saliency prediction, and, as shown by Kümmerer *et. al.* [KWB18], none of them is fully satisfactory. It follows that no single loss function can capture every quality factor. In order to evaluate this, we proposed a comparative study, dedicated to find the best strategies when designing loss functions for deep visual saliency models [Bru+21].

In this work, we show that the best way to get an efficient loss is by combining several metrics, especially those which measure fundamentally different properties: for instance, a pixel-based metric with a distribution-based one. Bearing that in mind, we used the following combination of losses to train our model:

$$\mathcal{L}(D, F, \hat{S}) = \alpha\mathcal{L}_{w-MSE}(D, \hat{S}) + \beta\mathcal{L}_{NSS}(F, \hat{S}) + \gamma\mathcal{L}_{CC}(D, \hat{S}) + \delta\mathcal{L}_{KL}(D, \hat{S}) \qquad (5.2)$$

where D is a ground truth fixation density map, F the ground truth fixation map and $\hat{S}$ is the predicted saliency map. The coefficients are set to $\alpha = 2$, $\beta = -1$, $\gamma = -2$, $\delta = 10$.

Weighted MSE loss

The mean squared error is a traditional pixel-based loss. Here, similarly to Cornia *et. al.* [Cor+16], we introduce a weighting term $\frac{1}{k-D_i}$ dedicated to penalize more the errors occurring on salient areas:

$$\mathcal{L}_{w-MSE}(D, \hat{S}) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{k - D_i} \left(D_i - \hat{S}_i\right)^2 \tag{5.3}$$

where D_i represents the value of the i-th pixel of the density fixation map, N is the total number of pixels, and k is a constant, set to 1.1. This way, the error is multiplied by a factor of 10 for a pixel on the ground truth map which value is 1, and by 0.9 for a pixel which value is 0.

NSS loss

We included the normalized scanpath saliency metric into our loss, as it captures several properties that are specific to saliency maps. Using a ground truth fixation map also allows for a better inter-operability between datasets and experimental conditions, as it does not rely on the choice of the smoothing Gaussian kernel used to create the fixation density map.

$$\mathcal{L}_{NSS}(F, \hat{S}) = \frac{1}{N} \sum_{i=1}^{N} \frac{\hat{S}_i - \mu(\hat{S}_i)}{\sigma(\hat{S}_i)} F_i \tag{5.4}$$

CC loss

Similarly, we also use Pearson's correlation coefficient, as it is an important (and relatively robust) metric for visual saliency maps

$$\mathcal{L}_{CC}(D, \hat{S}) = \frac{\sigma(D, \hat{S})}{\sigma(D)\sigma(S)} \tag{5.5}$$

KLD loss

Finally, we consider the ground truth and predicted saliency maps as probability distributions, and thus, we use Kullback-Leibler divergence to measure the dissimilarity

between the two histograms:

$$\mathcal{L}_{KL}(D, \hat{S}) = \sum_{i=1}^{N} D_i \log\left(\varepsilon + \frac{D_i}{\varepsilon + \hat{S}_i}\right) \tag{5.6}$$

where ε is a regularization constant set to 2^{-52}.

5.4 Experiments

5.4.1 Benchmark and state-of-the-art

We evaluated our model on three different datasets, containing cinematographic clips: the 884 clips from the testing set of Hollywood-2, the 5 held-out clips from our dataset, and the 15 clips from Breeden and Hanrahan's whole dataset. We used five metrics to evaluate the quality of the models: AUC-J, CC, SIM, sAUC and NSS (see Chapter 2 for more details about these metrics). We compared our model with four static models, and eight dynamic models: Itti [IKN98], SALICON [Hua+15], Deep Gaze II [Küm+17], MSINet [Kro+20], PQFT [GMZ08], Two-stream [Bak+18], DeepVS [Jia+18], ACLNet [Wan+18], TASED-Net [MC19], UNISAL [DJN20], ViNet [Jai+21] and HD2S [Bel+21]. We selected these models for how well they represent the diversity of the state-of-the-art, and used the publicly available implementation for each model, using the parameters provided by the authors. Figure 5.5 shows a few qualitative examples of saliency map predictions, compared to the ground truth.

Hollywood-2

Table 5.1 shows the comparative results on the Hollywood-2 test dataset. Our proposed model outperforms the other methods in most metrics. It improves the results of the ViNet model, which was to be expected, as we added information using optical flow through a motion stream, as well as the prior flicker map. The model scores well in all metrics, which is probably due to the combination of losses used to train it.

Breeden and Hanrahan's

Table 5.2 shows the comparative results on Breeden and Hanrahan's dataset. In our perspective, these results are particularly interesting, as this database includes cinematographic annotations, and thus allows us to evaluate the full potential of our model. We

		Hollywood-2				
	Model	CC ↑	SIM ↑	AUC-J ↑	sAUC ↑	NSS ↑
Baseline	Center Prior*	0.421	0.331	0.869	0.615	1.808
Dynamic models	PQFT* [GZ10]	0.153	0.201	0.723	0.621	0.755
	Two-stream [Bak+18]	0.382	0.276	0.863	0.710	1.748
	DeepVS [Jia+18]	0.446	0.356	0.887	0.693	2.313
	ACLNet [Wan+19]	0.623	0.542	0.913	0.757	3.086
	TASED-Net [MC19]	0.646	0.507	0.918	0.768	3.302
	UNISAL [DJN20]	0.673	0.542	0.934	0.795	3.901
	ViNet [Jai+21]	0.693	0.550	0.930	0.813	3.730
	HD2S [Bel+21]	0.670	0.551	0.936	0.807	3.352
Static models	Itti* [IKN98]	0.257	0.221	0.788	0.607	1.076
	SALICON [Hua+15]	0.425	0.321	0.856	0.711	2.013
	DeepGaze II [Küm+17]	0.591	0.378	0.855	0.778	3.225
	MSINet [Kro+20]	0.627	0.430	0.916	0.778	2.956
Proposed	Ours	0.703	0.562	0.939	0.816	3.878

Table 5.1 – Scores of several saliency models on the Hollywood-2 database. Non-deep models are marked with *. Best performances are marked in red, second best are marked in blue.

observe that our model outperforms every other one, on every metric, except for the shuffled-AUC, which can be explained by the use of centered priors, which is strongly penalized by this metric. Moreover, it is interesting to note that it is the only model which scores improved compared to the scores on the Hollywood-2 dataset, reinforcing our hypothesis that additional cinematographic information is in fact useful for this task.

Ours

Finally, Table 5.3 reports the results of the models on the testing hold-out of our dataset. As expected, our model still scores higher than all the other ones, as the dataset also includes cinematic annotations. Results are quite similar to those obtained on Breeden and Hanrahan's base, which was to be expected, as both datasets are very similar, in terms of type, content and length of the stimuli.

When looking at the predictions qualitatively, we observe that our model handles some of the difficult cases way better than the rest of the models, especially when the camera is moving. For instance, Figure 5.5 shows the predictions of our model on a sequence of Forrest Gump, where the camera pans down from Forrest's face to Jenny's: while she is

Breeden and Hanrahan

	Model	CC ↑	SIM ↑	AUC-J ↑	sAUC ↑	NSS ↑
Baseline	Center Prior*	0.356	0.297	0.848	0.763	1.679
Dynamic models	PQFT* [GZ10]	0.139	0.170	0.683	0.625	0.748
	Two-stream [Bak+18]	0.376	0.269	0.859	0.704	1.722
	DeepVS [Jia+18]	0.430	0.342	0.880	0.684	2.155
	ACLNet [Wan+19]	0.537	0.420	0.883	0.700	2.461
	TASED-Net [MC19]	0.638	0.493	0.911	0.759	3.146
	UNISAL [DJN20]	0.664	0.522	0.928	0.777	3.780
	ViNet [Jai+21]	0.679	0.536	0.909	0.807	3.591
	HD2S [Bel+21]	0.668	0.550	0.931	0.806	3.287
Static models	Itti* [IKN98]	0.234	0.218	0.773	0.601	0.971
	SALICON [Hua+15]	0.418	0.323	0.852	0.712	2.002
	DeepGaze II [Küm+17]	0.593	0.381	0.858	0.767	3.201
	MSINet [Kro+20]	0.611	0.407	0.901	0.768	2.873
Proposed	Ours	0.708	0.557	0.940	0.803	3.885

Table 5.2 – Scores of several saliency models on Breeden and Hanrahan's database. Non-deep models are marked with *. Best performances are marked in red, second best are marked in blue.

not yet entirely on the frame, the model correctly anticipates the appearance of a salient object, thanks to the vertical camera motion.

5.4.2 Ablation study

In order to validate the architecture choice and the benefits of the biases and feature maps that we include, we tested some variations of our model, removing each component, while keeping the same training procedure. We tested the following settings: (1) Image stream alone; (2) Motion stream alone; (3) Two streams without cinematic biases; (4) Full model without the center-bias prior; (5) Full model without the shot-size bias; (6) Full model without the camera motion bias; (7) Full model without the flicker bias; (8) Full model without the cut bias. Results for each configuration, alongside the results of the full architecture are given in Table 5.4. All the configurations are evaluated on the test hold-out set of our database.

Image stream alone performs very similarly to the ViNet model, as they share the same design. However, the motion stream, while insufficient on its own, manages to improve the results of the model when combined to the image stream, proving that 3D-CNN

Figure 5.5 – Example of predictions on a sequence from *Forrest Gump* (Robert Zemeckis, 1994). Frames: the frames sequence; GT: ground truth fixation density; Ours: our proposed model; ACL: ACLNet [Wan+18] predictions; TASED: TASED-Net [MC19] predictions; DGII: Deep Gaze II [Küm+17] predictions.

Ours

	Model	CC ↑	SIM ↑	AUC-J ↑	sAUC ↑	NSS ↑
Baseline	Center Prior*	0.387	0.290	0.851	0.757	1.688
Dynamic models	PQFT* [GZ10]	0.141	0.178	0.697	0.626	0.758
	Two-stream [Bak+18]	0.382	0.303	0.870	0.720	1.712
	DeepVS [Jia+18]	0.441	0.356	0.874	0.687	2.159
	ACLNet [Wan+19]	0.546	0.431	0.893	0.715	2.557
	TASED-Net [MC19]	0.637	0.489	0.905	0.743	3.108
	UNISAL [DJN20]	0.669	0.533	0.931	0.804	3.824
	ViNet [Jai+21]	0.685	0.542	0.924	0.801	3.585
	HD2S [Bel+21]	0.657	0.546	0.930	0.797	3.291
Static models	Itti* [IKN98]	0.226	0.201	0.762	0.589	1.038
	SALICON [Hua+15]	0.437	0.331	0.860	0.708	2.084
	DeepGaze II [Küm+17]	0.586	0.384	0.862	0.754	3.301
	MSINet [Kro+20]	0.619	0.428	0.909	0.766	2.920
Proposed	Ours	0.711	0.574	0.937	0.800	3.905

Table 5.3 – Scores of several saliency models on our database. Non-deep models are marked with *. Best performances are marked in red, second best are marked in blue.

architectures on their own are unable to grasp every aspect of the temporal dependencies of visual attention.

We observe slight, yet significant improvements of the scores when adding the different cinematic biases. It appears however that the least useful bias map is the general center-bias prior; this is probably because this information is already at least partially taken into account within the shot-size bias, which is somewhat similar. It follows that such high-level information is indeed important, and that context is a crucial information to take into account when designing visual saliency models. We would argue that, now that there are very efficient general saliency models, a particular attention should be paid to their use, and how to optimize their specificity for a specific task.

5.5 Application

In this section, we propose a very simple example of possible application for our saliency model, dedicated to improve systems of automatic editing.

As explained on Section 3.3, automated movie editing is a difficult task, as it requires a knowledge of what is happening on screen for each camera, contextual cues, traditional film editing idioms, and so on. In order to improve such systems, we argue that attention

Ablation study

Configuration	CC ↑	SIM ↑	AUC-J ↑	sAUC ↑	NSS ↑
Image stream (1)	0.678	0.539	0.915	0.787	3.574
Flow stream (2)	0.351	0.291	0.867	0.704	1.701
Two-streams (3)	0.688	0.552	0.928	0.804	3.740
Two-streams (4)	0.707	0.570	0.935	0.802	3.894
Two-streams (5)	0.705	0.571	0.934	0.794	3.890
Two-streams (6)	0.693	0.566	0.933	0.795	3.843
Two-streams (7)	0.697	0.570	0.933	0.798	3.856
Two-streams (8)	0.697	0.571	0.931	0.799	3.892
Full model	0.711	0.574	0.937	0.800	3.905

Table 5.4 – Scores of several configurations of our model. Best performances are marked in red, second best are marked in blue.

features such as visual saliency should be considered, as they allow a form of feedback on what is important on screen or not, and how the audience will perceive the stimuli. For instance, in order to make a cut seem seamless, an editor could decide to position a salient object in the same area of the frame before and after the edit, so that the viewers would not have to shift their attention to another part of the frame after the cut.

For dynamic stimuli, we propose a way to measure the activity of what is happening on screen, by considering the temporal gradient of the visual saliency map. Considering a predicted dynamic saliency map $\hat{S}(x, y, t)$:

$$Activity(t) = \frac{1}{N} \sum_{x,y} \left| \frac{\partial \hat{S}(x, y, t)}{\partial t} \right| \tag{5.7}$$

iterating and averaging over all pixels of the frame, where N is the number of pixels.

Values close to zero indicates that no significant motion is happening on screen. This does not mean that the shot is necessarily uninteresting: for instance, a static closeup shot on the face of a character speaking will exhibit very little activity in the saliency map. However, it is very useful to spot sudden movements, and moments when potentially interesting events are happening. To illustrate this, we compute the gradients of saliency on different camera angles from a virtual scene modeled after a sequence of *Back to the Future* (Robert Zemeckis, 1985). Figure 5.6 shows the evolution of those gradients on several cameras, alongside with keyframes; we observe that spikes in the gradient correspond to events happening on screen: a character entering or leaving the frame, turning his head, or even moving in the background. Similarly, long sequences of frames

in the first camera have a very low gradient value, due to nothing particular happening on screen, and could be easily discarded by an automatic editing system.

5.6 Conclusion

In this chapter, we proposed a novel visual saliency model, dedicated to predict visual attention on movie sequences. By using a two-stream approach, with both the sequence of frames and the optical flow as an input, we are able to have a better –while still imperfect– grasp on the temporal components of visual attention. We also show that additional information regarding high-level features also contributes to the performances, as such characteristics are difficult to handle for deep feature extractors.

Going forward, a few research ideas and directions can be explored to extend this work. First, it seems important to find a way to automate the annotation process, so that we can have a larger dataset of high-level cinematographic features, which will allow us to design other ways of including them into the saliency prediction. In this perspective, the work of Rao *et. al.* [Rao+20b] shows encouraging results, and should be extended.

It is also important to consider the contextual semantic information, which plays an essential role in the context of a movie: as of today for instance, visual saliency models will flag faces as salient, with the emphasis on the faces in the center of the frame, even though they are not necessarily the main character's of the scene. In the previous chapter, Figure 4.5 shows this lack of understanding: the main character of the shot is clearly the warden, in the right side of the frame, but all of the models consider the character played by Morgan Freeman to be the most salient area. It becomes then crucial to find new ways to take these semantic considerations into account, perhaps using data from the written scripts.

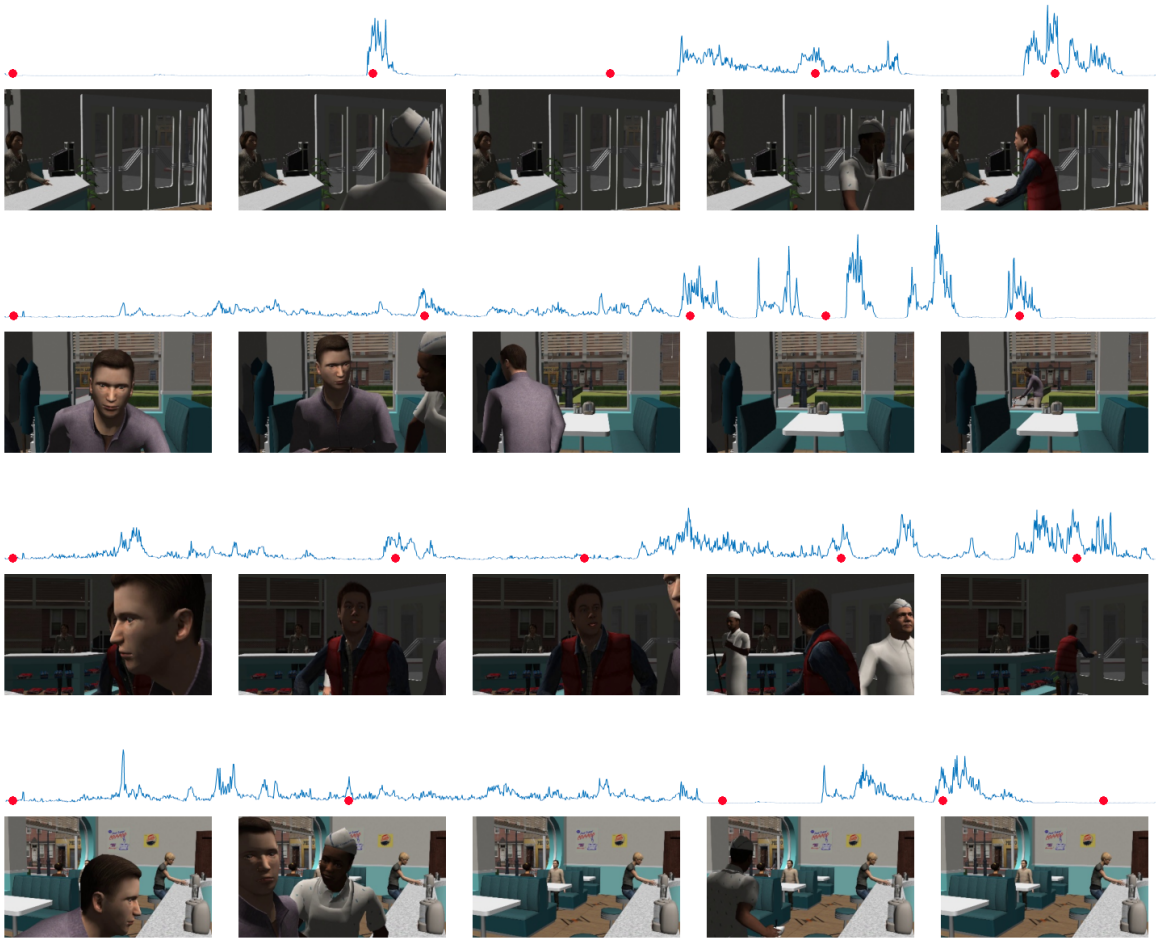


Figure 5.6 – Example of gradients of saliency on rushes of a 3D scene, modeled after *Back to the Future* (Robert Zemeckis, 1985). Red dots indicates where the frames were extracted.

Chapter 6

Inter-observer visual congruency: when will people look at the same place ?

In this chapter, we introduce the concept of inter-observer visual congruency (IOC), which reflects the amount of agreement between the gaze patterns of several observers watching the same visual stimulus. We define a metric, for both the static and dynamic cases, and propose two models, one for each case, to predict such scores, and we analyze how IOC varies during movies, and how those variations relate to specific cinematic features, such as camera motions or edits. Finally, we propose two deep learning models to infer IOC from the stimulus, either static (image) or dynamic (video).

6.1 Introduction

As discussed in the introduction chapters, learning where a human will look when watching a scene is important, and has a multitude of applications in various fields. However, visual behavior is not always consistent between observers, either because of top-down factors (for instance, observers with a previous knowledge of the stimuli will exhibit different gaze patterns [Dor+10]), or bottom-up characteristics. For example, people will tend to exhibit very similar behaviors when viewing a scene containing a single salient object, while cluttered scenes, or scenes lacking strong visual attractors will induce more diversity in eye fixation locations. This similarity, or dissimilarity between visual trajectories among observers is referred as *attentional synchrony*, and metrics quantifying this synchrony are commonly called *inter-observer congruency* (IOC) metrics.

Such metrics have proven very useful in a whole variety of applications, such as image ranking, quality assessment, or even visual saliency: indeed, IOC has been shown to provide an upper-bound on the performances of models predicting the locations of eye fixa-

tions. However, this measure in itself has received way less attention. Le Meur *et. al.* [MBR11] offered a first image-processing approach, where they studied the influence of several image features, such as the depth of field or the image complexity, on IOC scores. Following this work, Rahman and Bruce [RB16] explored more image characteristics, coupled with top-down features. They proposed a predictive model of IOC based on both those feature sets, as well as information yielded by the predictions of visual saliency models.

In the context of movies, attentional synchrony has also been studied, from a more cognitive point of view. Dorr *et. al.* pointed out several differences in the variation of eye fixations and saccade amplitudes when watching the same stimulus several times over two days, and compared the synchrony observed on Hollywood movies and natural scenes. Mital *et. al.* [Mit+11] showed that the most predictive features for gaze clustering when viewing dynamic stimuli were temporal and motion-related, like flicker or contrast in motion. Smith and Mital [SM13] also studied the influence of the viewing task on attentional synchrony, highlighting a significant influence of it, but mostly after the first few fixations, which were usually guided by the exogenous attention mechanisms.

In the following, we first describe how to compute a reliable and adapted IOC score, inspired by visual saliency metrics. We start by defining a metric for still images, and then extend it to dynamic stimuli. We then explore the influence of cinematographic features on the variations of IOC scores, and how these measures can provide valuable knowledge when studying filmmaking. Finaly, we propose two models dedicated to predict IOC scores, for static and dynamic stimuli, focusing on movie clips.

6.2 Measuring inter-observer congruency

6.2.1 Static stimuli

Previous metrics

A lot of methods have been proposed to describe the amount of visual congruency among observers when viewing a stimulus. All of those methods use different hypotheses about the distribution of gaze patterns, but overall, these metrics are highly correlated to one another [Dor+10]. Rajashekar *et. al.* [RCB04] used the average z-score between the individual human fixations and the overall fixation density, using Kullback-Lieber divergence as a metric. Peters *et. al.* [Pet+05] used the normalized scanpath saliency metric (NSS) to compare each individual gaze track to a global inter-observer model,

composed of the aggregation of individual saliency heatmaps.

Sawahata *et. al.* [Saw+08] used a criterion based on information theory, the entropy of the fixation distribution, or more precisely, the entropy of a Gaussian mixture model (GMM) fitted on the the gaze points divided into clusters based on the Bayesian information criterion (BIC). Similarly, Mital *et. al.* [Mit+11] used GMMs, and more specifically the weighted covariance value, to discriminate between "tightly and loosely clustered frames", i.e. frames in which attentional synchrony is higher or lower. Smith and Mital [SM13] also used these GMM clusters and their covariance, expressed as the visual angle enclosing 68% of gaze points. Finally, several area-based methods have been proposed: for instance, Goldstein *et. al.* [GWP07] computed the area of the best-fit bivariate contour ellipse, whereas Breeden and Hanrahan [BH17] used the area of the convex hull of the fixation points.

Finally, more saliency-inspired methods consist in comparing the gaze tracks of a single observer to the joint distribution of all the other observers. This leave-one-out approach was used by Torralba *et. al.* [Tor+06] and Le Meur *et. al.* [MBR11], where they use the rate of fixations falling in a saliency classifier, created from a thresholded fixation distribution map, and Rahman and Bruce [RPH14], where they compute the AUC score between the individuals and the aggregated fixation distribution of all other observers.

Our approach

Since we are interested in the relations between inter-observer visual congruency and other computing features, such as saliency, we used an IOC score inspired by the leave-one-out approach. It is indeed justified as the metric used to quantify IOC is the same that can be used to compare two saliency maps, and thus can be used, for instance, as an upper bound on the performance of saliency models.

Assuming that an image have been seen by N_o observers, we first threshold the scanpaths of each observer to only consider the N_{fp} first fixations. This is done in order to ensure that each observer has the same weight in the final score. Empirically, there is very little variation depending on the chosen length of the scanpath, as long as there are enough observers, as shown by Figure 6.1. A good rule of thumb can be to select N_{fp} as the size of the shortest scanpath among all observers, as long as it is longer than 5, in order to take into account the evolution of the gaze dispersion with the viewing time.

Once the number of fixation points per observer is thresholded, for each observer i, we consider the aggregated binary fixation map of the $(N_o - 1)$ other observers, i.e. a binary

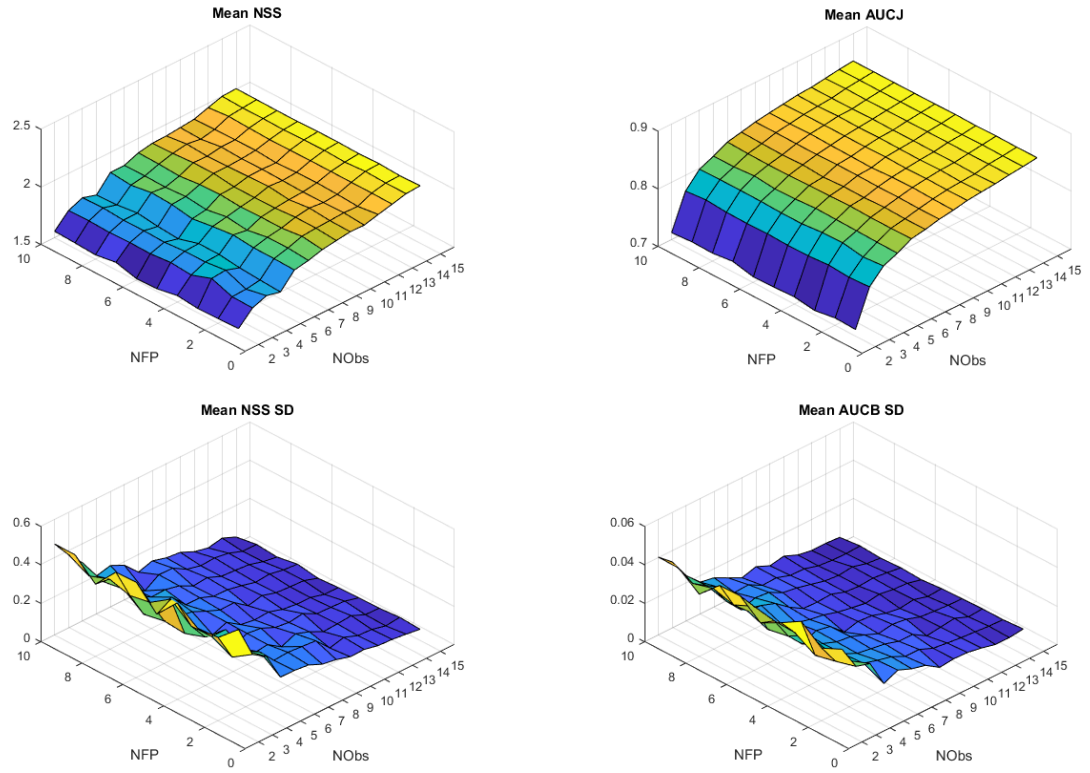


Figure 6.1 – First line shows the average IOC-NSS and IOC-AUC scores on the CAT2000 dataset, depending on the considered number of observers and length of scanpaths. Second line shows the avereage standard deviation of IOC-NSS and IOC-AUC scores on the same dataset.

matrix which pixel values are 1 if the pixel was fixated, and 0 otherwise. Similarly to a fixation density map, we convolve these fixation maps with a 2D Gaussian kernel, which covariance is set to roughly approximate the size of the fovea (around one degree of visual angle). This operation is dependent on the conditions in which the fixation points have been recorded, which allows for a fair comparison between datasets. If such comparison is needed, one should also threshold the number of observers: indeed, the more observers, the more likely it is that the scanpath of a viewer falls into the salient areas of the fixation density map, and thus the higher the IOC score.

Once the leave-one-out fixation density map is computed, we can evaluate the proximity between the observer's fixations and the density using any of the metrics used to evaluate visual saliency models. In our case, we only consider the AUC and NSS metrics, as they do not require to transform the fixations of the left-out individual into a fixation

Figure 6.2 – Examples of IOC-AUC scores on a few images of the CAT2000 dataset.

density, which would be tricky considering the relatively small number of fixations. The whole process is then iterated and averaged, over all observers, to get a single IOC score (either IOC-AUC or IOC-NSS). In the case of the two considered metrics, a high score means that the observers will tend to exhibit similar fixation patterns, while a low score will indicates more variety in the fixation locations.

Generally, a low IOC score is associated with either images exhibiting no particularly salient content, or cluttered scenes, where there are too much salient locations, especially if the viewing time is short. Examples of such situations can be observed on Figure 6.2.

6.2.2 Dynamic stimuli

Extending the measure of IOC to the spatio-temporal domain is not as straight forward as it may seem. For instance, applying an IOC measure on a frame-by-frame basis can be

problematic, as there might not be enough fixations to avoid a significant amount of noise: indeed, in the case of cinematographic movies, each frame will be displayed for around 42 milliseconds, while the average eye fixation spans around a few hundred milliseconds, implying that each frame will only display one or two fixations per observer.

More generally, designing an IOC measure for dynamic stimuli implies answering questions about what we actually want to measure. For example, let us consider a sequence containing two spatially separate salient locations A and B (a dialogue between two characters, for instance), and two observers. If, during a short time period, the first observer fixates location A first and location B second, and the second observer does the opposite, both observers will exhibit similar spatial gaze patterns, and only differ temporally. However, a frame-by-frame measure will (in the worst scenario) treat the case as if the first observer only fixated location A and the second only location B. We then argue that a well-designed IOC metric should take into account the temporal continuity: two non-simultaneous fixations at the same spatial location should be considered as "close" based on the temporal dimension.

In order to address this issue, we propose a new approach to compute an IOC measure in the spatio-temporal domain.

First, we define the spatio-temporal fixation density map for a stimulus. For each frame, we compute the traditional fixation density map by convolving the binary fixation map with a Gaussian kernel, which covariance is chosen so that it approximates the size of the fovea. Figure 6.3 shows an example of this spatio-temporal representation. Then, we stack those density map into a spatio-temporal volume, and smooth it in the temporal dimension using a Gaussian kernel, which variance is set to approximate 250 ms, i.e. the average duration of a fixation. In the case of a 24 frames per second cinematic stimuli, this amounts to 6 frames. Now, this spatio-temporal map can be compared to ground truth fixations using the NSS metric on the whole volume:

$$\begin{aligned} NSS(S,F) &= \frac{1}{N}\sum_i \bar{S}_i F_i \\ \text{where} \quad \frac{1}{N} &= \sum_i F_i \quad \text{and} \quad \bar{S} = \frac{S-\mu(\hat{S})}{\sigma(S)} \end{aligned} \tag{6.1}$$

where N is the number of fixated voxels, S is the fixation density volume, F is a spatio-temporal binary fixation map, i.e. a volume where each voxel is either 1 if a fixation occurred at its location and time, and 0 otherwise. The choice of the NSS metric in this

case comes straight forward, as it is way less time- and memory-consuming than AUC metrics.

Figure 6.3 – Example of spatio-temporal fixation density map on a sequence of *Big Fish* (Tim Burton, 2004).

From there, we use the exact same leave-one-out approach than the static case. A fixation density is computed for each group of $(N_o - 1)$ observers, and compared using the NSS metric to the fixations of the remaining observer. The scores are then averaged over the observers to get a global IOC value. In order to track the evolution of attentional synchrony over time, we keep the global fixation densities and fixation maps, and compute the NSS values over a sliding time-window, which size can be chosen depending on the context: a shorter time window (e.g. four or five frames) allows for a finer-grained analysis, but is more sensitive to noise, for instance.

However, the main drawback of this method is its memory consumption. Indeed, we need to store a volume of size $H \times W \times T$ (where H is the height of the frame, W the width and T the duration of the whole sequence) for each observer, which can quickly become overwhelming when working with high-resolution stimuli and (relatively) long movie sequences. In order to solve this issue, we designed a simple, yet useful heuristic.

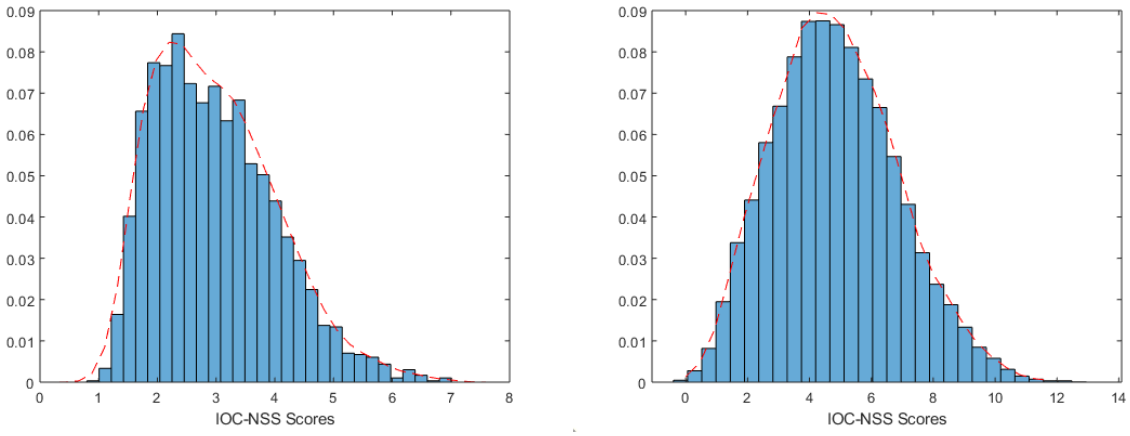
We only consider a sliding time-window of size t; for each group of $(N_o - 1)$ observers, we gather all their fixations during this period, and report it on a 2D binary fixation map, which is then smoothed into a fixation density. This map is then compared to the binary fixation map of the remaining observer using the NSS metric. The process is iterated and averaged over all the observers to get an IOC score over the considered time frame. The duration of the time window can be freely chosen, once again depending on the context. In our analyses, we considered two window sizes: 5 frames for a fine-grain approach and 20 frames for a more general view. On our database, described in Chapter 4, we found a

strong significant correlation between this heuristic and the memory expensive approach (for time windows of 5 frames: $r = 0.7912$, $p < 0.001$; for time windows of 20 frames: $r = 0.8531$, $p < 0.001$). From now on, we will then only refer to this heuristic when we mention spatio-temporal IOC.

6.3 Inter-observer congruency and cinematography

Now that we have a reliable measure of inter-observer congruency for dynamic stimuli, we are interested in studying how cinematographic characteristics may influence it. Since the annotations in our dataset are at the level of the shot, in the following, we use the 20-frames bin measure of the IOC, unless stated otherwise.

When observing the IOC values over the whole dataset, we first note that IOC values are relatively high, especially compared to IOC values in the static case. Figure 6.4 shows the distribution of IOC values for static datasets and for our cinematic dataset. The average IOC value over all the clips is 4.1273, compared to 2.9819 for the CAT2000 [BI15] and MIT [Jud+09] datasets. This discrepancy corroborates the findings of Smith and Mital [SM13], where they find smaller clusters of fixations during the free viewing of dynamic scenes compared to free viewing of similar static scenes, implying that movement on screen will tend to cause attentional synchrony.



(a) (b)

Figure 6.4 – Distribution of the IOC-NSS scores: (a) in the static case (CAT2000 [BI15] and MIT [Jud+09] datasets), and (b) in the dynamic case (on our dataset).

We also observe a disparity in IOC scores across movies; the scene with the highest

score on average is *The Shining* (IOC= 5.7621), while the lowest score is the clip from *Armageddon* (IOC= 3.4056) (see Figure 6.5). This would tend to indicate that inter-observer congruency is influenced by certain editing features. Figure 6.6 shows the variations of IOC on the beginning of several clips, and illustrates significant discrepancies, both between the clips and within it. For instance, the scene from *Invictus* exhibits sudden variations of IOC, while the scene from *Pulp Fiction* exhibits a smoother profile.

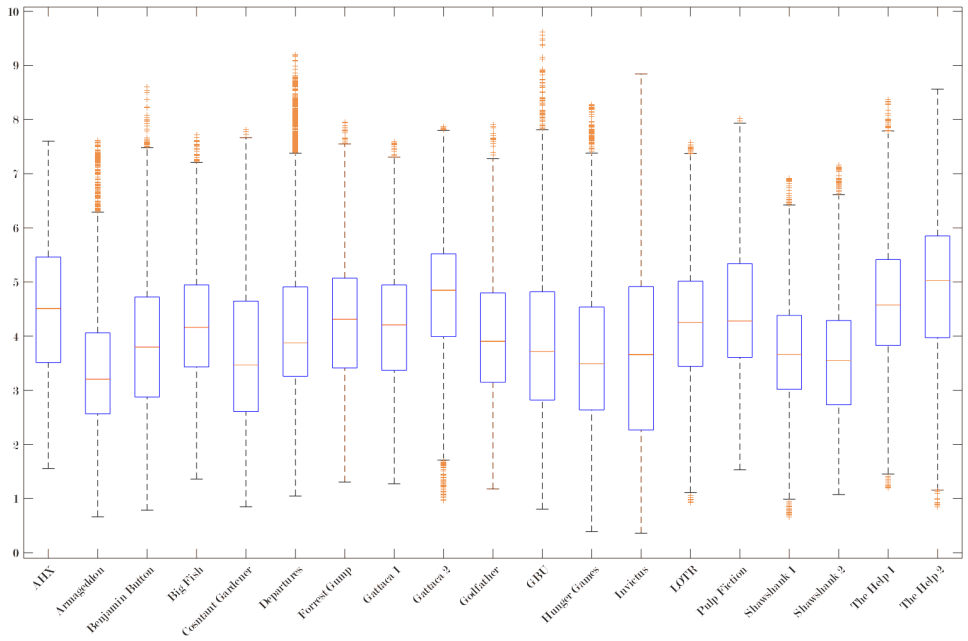


Figure 6.5 – IOC score distribution among the different clips of the dataset.

6.3.1 Camera movements and IOC

In order to evaluate the impact of camera motion on inter-observer congruency, we compared the IOC values on shots that contain at least one camera movement with fully static shots, using one-way ANOVA. On average, static shots show slightly higher IOC values ($M = 4.331$, $SD = 1.60$) than shots exhibiting camera motion ($M = 4.025$, $SD = 2.310$) ($p << 10^{-5}$), but it is worth noting that standard deviation is significantly higher in the shots where the camera moves, indicating that camera motion plays an important role in increasing or decreasing IOC. This is consistant with the findings of Mital *et. al.* [Mit+11], showing that motion-related features (not specifically camera motion) are a good predictor of eye fixations clustering.

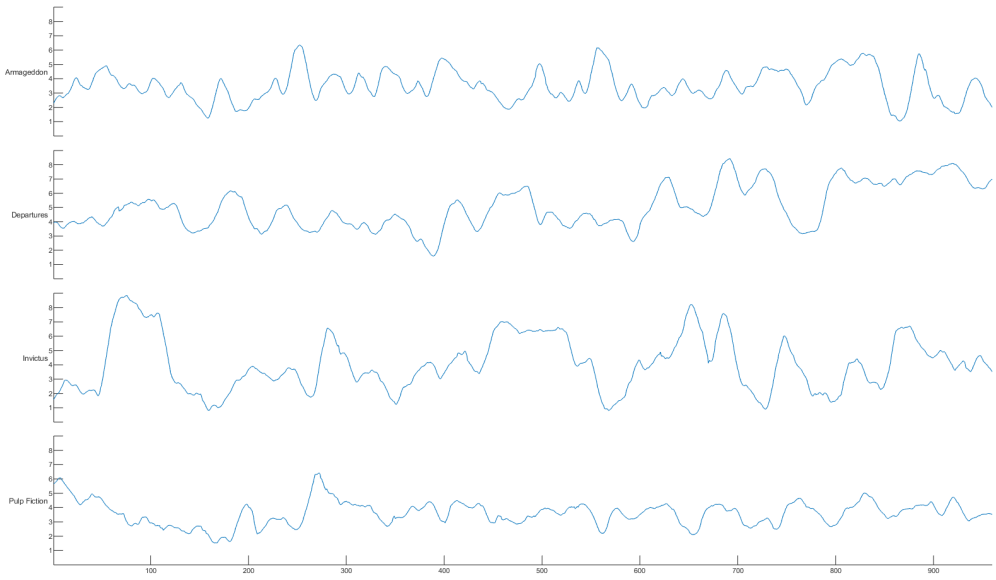


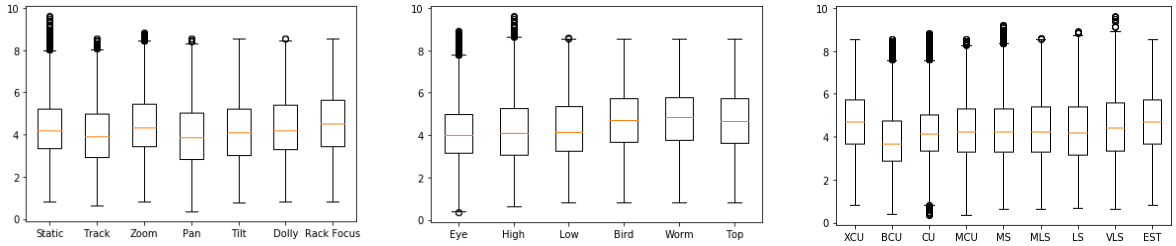
Figure 6.6 – Examples of variations of IOC during the first 30 seconds of several movie clips (*Armageddon*, *Departures*, *Invictus* and *Pulp Fiction*)

Considering this, we performed post-hoc pairwise t-tests (using Bonferonni correction for multiple tests, i.e. multiplying the p-values by the number of comparisons) between the annotation groups for camera movement, showing significant differences ($p << 10^{-5}$) between most shot characteristics, except between static and dolly shots ($p = 0.026$). As expected, the highest average IOC values are in zoom shots and rack focuses, which are camera features specifically designed to direct visual attention; these values are shown in Figure 6.7 (a).

Camera angles (Figure 6.7 (b)) show no significant differences between the choice of camera angle and inter-observer congruency. At first glance, it may seem that extreme camera angles (bird shots, worm shots and top shots) are associated with higher IOC values, but this might just be an artifact due to the relatively low number of such shots in the dataset.

6.3.2 Shot size and IOC

Similarly, we looked at the average scores depending on the size of the shots (Figure 6.7 (c)). Extreme closeup shots are associated with the highest IOC scores ($M = 4.863$,



(a) (b) (c)

Figure 6.7 – IOC scores depending on camera movement features (a), camera angles (b) and shot size (c)

$SD = 1.758$), while interestingly, big closeups have the lowest IOC averages ($M = 3.967$, $SD = 1.840$). This difference is confirmed by a t-test ($p << 10^{-5}$), and might be explained by the way these shots are used in the overall scene: on average, extreme closeup shots are very short ($M = 34.111$, $SD = 26.434$ frames) compared to big closeups ($M = 78.683$, $SD = 67.476$ frames), thus leaving more time for exploration. A perfect example of this is the Mexican standoff scene from *The Good, the Bad and the Ugly* (Sergio Leone, 1966): the shots come closer and closer to the characters as the tension builds up; when it reaches the big closeup shot size, a little bit of time is given to the spectator to read the characters faces. After that, a series of very short shots show extreme closeups on the eyes and guns of the characters, forcing eye fixations on the salient elements (i.e. the eyes and the guns).

Medium shot categories (MCU, MS, MLS and LS) show little to no significant differences of IOC. This might be due to categories sometimes not very well defined, as it can be hard distinguishing between a medium shot and a medium-long shot, for instance.

6.3.3 Cuts and edits

As mentioned just before, the rhythm of the cuts and edits play an important role in directing attentional synchrony. A well-known effect [Dor+10; Mit+11; SM13] is the sudden augmentation in inter-observer agreement immediately following a cut. This tendency is observable in our dataset, when taking into account the binning effect linked to the size of the temporal window used to compute IOC scores. Figure 6.8 shows this effect on a clip from *Armageddon* (Michael Bay, 1998). We observe a significant difference ($p << 10^{-5}$) between the IOC scores of the frames within the first 500ms immediately following a cut ($M = 5.712$, $SD = 1.882$) and the rest of the frames ($M = 3.510$, $SD = 1.974$).

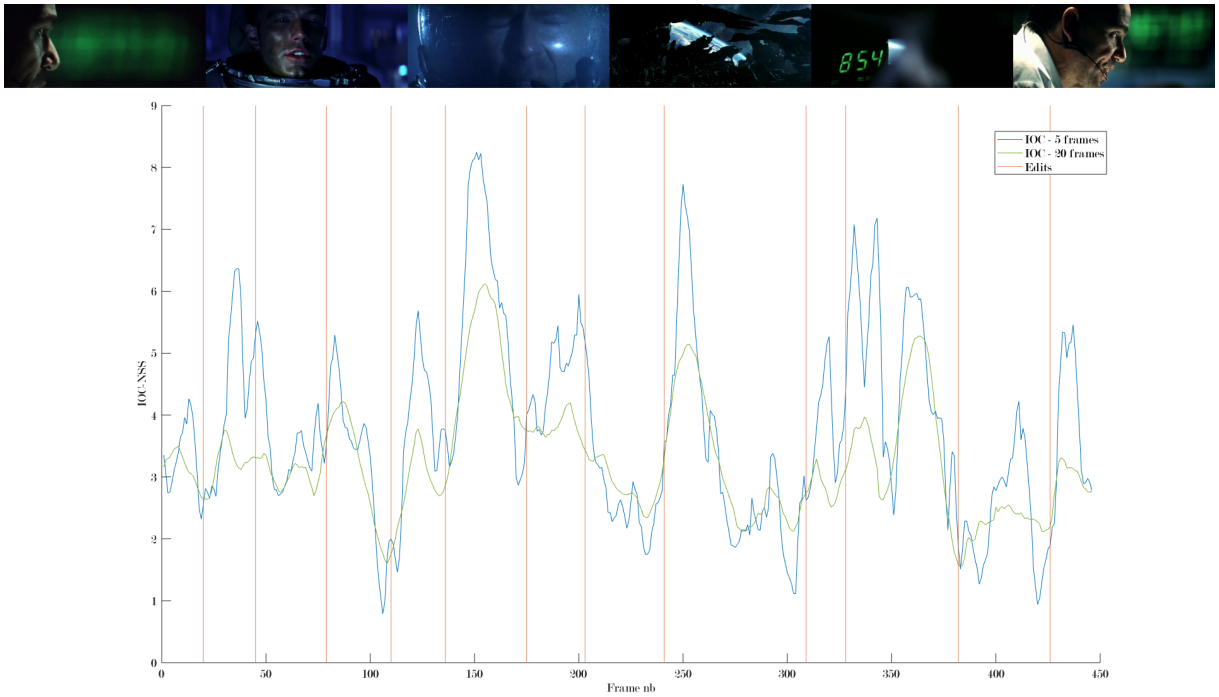


Figure 6.8 – Example of the variations of IOC scores (5-frames window: blue; 20-frames window: green) on a clip from *Armageddon* (Michael Bay, 1998). Note the peak usually following an edit.

6.4 A first model of IOC prediction, for static stimuli

We propose to rely on a learning approach to predict the inter-observer congruency score of an image. We exploit the IOC scores computed from the CAT2000 and Judd/MIT databases to train a network in a two-staged approach: first (i) by designing an encoder based on VGG19 [SZ14b] to extract deep features, and second (ii) by designing a straightforward shallow network as a decoder to perform the regression.

6.4.1 A two-staged model architecture

The lack of images labeled with eye tracking data makes the creation of a reliable and robust model challenging. This is why we used transfer learning, using the pre-trained weights from a feature extractor architecture. Moreover, the shallow regression network seemed a good compromise between that lack of data, that precludes any kind of learning that uses too many parameters, and the capacities of a method that is more complex and efficient than simple linear regression.

The overall architecture is presented in Figure 6.9. The model first uses VGG19 network for extracting a set of deep features. We chose that architecture for its excellent performances in the field of visual attention, especially through popular models such as DeepGazeII [Küm+17] and MLNET [Cor+16]. VGG-based networks are well-known for their very good generalization properties, as well as their simplicity. The particularity of that structure is the use of multiple convolution layers with small kernel size (3×3). The layer stacking is then more discriminative due the multiple non-linear rectification layers. It also decreases the number of parameters, hence easing the training process. We used two different versions of that encoder, one for AUCB-based IOC score, and one for NSS-based score. The first is the full VGG19 feature extractor, containing 5 max-pooling layers. In that case, the output of the encoder is a tensor of size $[37 \times 50 \times 1472]$. For the NSS score, we observed better performances when removing two of the five max-pooling layers, leaving the output dimensions as $[37 \times 50 \times 1280]$.

We then design a simple shallow network as a decoder to perform the regression task. After the input, a dropout layer is applied, followed by three convolution layers with $[3 \times 3]$ kernel sizes, reducing the number of features maps to 320, 64 and 1. Batch normalization is used to normalize the output, followed by a flattening layer, a second dropout and three fully connected layers, reducing the dimension from 1850 to 1024, 256 and 1. The final output is the predicted score of IOC. Best performances were achieved for AUCB

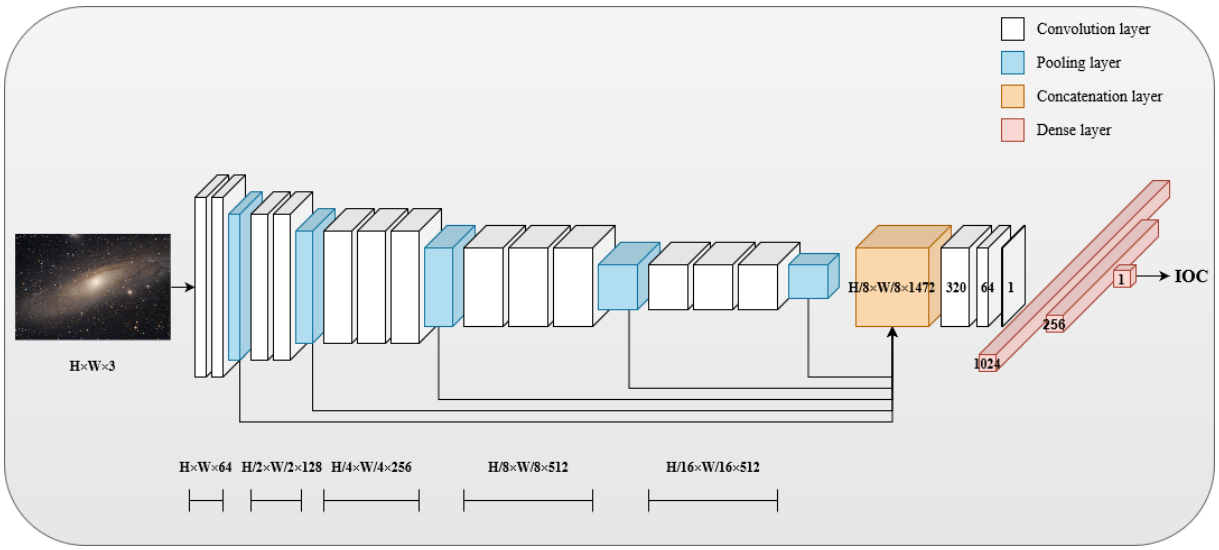


Figure 6.9 – Architecture of the proposed static IOC prediction model.

by setting the dropout rate at 0.5, using sigmoid activation functions within the dense layers, and binary categorical entropy as a loss function. For NSS, the dropout rate is set as 0.2, with ReLU activation functions and mean squared error as a loss function.

6.4.2 Training database

To train our model and evaluate IOC prediction, we used the IOC scores (based on AUCB and NSS) computed over the Judd/MIT database [Jud+09] and the CAT2000 database [BI15]. The Judd/MIT database includes 1000 images with different resolutions, 500 of them being used for training, 200 for validation along the training process, and 300 have been held out to evaluate the performances. For the CAT2000 we use 1200 images for the training process, 400 for the validation and 400 for the test. This database is also split into 20 different categories of images, which allows to compare the precision of the prediction in each category.

Since the Judd/MIT database only contains 500 training images, we performed a first training merging the two databases together. We also performed data augmentation by flipping the images horizontally, making the assumption that such a transformation should not disturb the IOC score. The images are then rescaled into $[400 \times 300 \times 3]$ images. We used the pre-trained ImageNet weights [Den+09] for the encoder network and froze those layers during the training. A fine-tuning phase has also been performed, during which we froze the first convolution layer of the decoder network.

6.4.3 Results

Figure 6.10 shows the distribution of the ground truth and predicted IOC scores on the CAT2000 dataset. For the AUCB prediction, the average is exactly the same as the ground truth mean, but its standard deviation is slightly smaller. There is a tendency that the prediction value is closer to the mean, meaning that the prediction of outliers is more difficult. Similar effects can also be noticed with NSS score. When the global prediction mean is slightly smaller than that of the ground truth, the spreading of prediction NSS is significantly smaller than that of the ground truth. We also observe that when the ground truth AUCB distribution is almost symmetric along the mean value, the ground truth NSS is more right-skewed and has a long tail at high NSS values. It may explain the poorer performance in NSS prediction. This is most probably due to the regularization and dropout we used in the decoder network, in order to prevent overfitting issues. The range of the predictions, for both scores, are also smaller than the ground truth, which is due to the capacity of the model to generalize properties and to perform really well when averaged over a few images.

Overall, Table 6.1 indicates that the correlation coefficient between ground truth and prediction is 0.611 and 0.642 on Judd/MIT and CAT2000 databases, respectively. On Judd/MIT database, the proposed method significantly outperforms Le Meur [6] and Bruce [8] methods. We also applied our model on two other databases, namely the Memorability [23] and Bruce's database [24]. Compared to Bruce method [8], the proposed model is better on Memorability database while Bruce method provides the best results on Bruce database. Note that the proposed method has not been trained over neither Memorability nor Bruce database. Both results suggest that the proposed method has good generalizing properties.

Dataset	Judd/MIT	CAT2000	Memorability [ML13]	Bruce [BT09]
Le Meur [MBR11]	0.340	N/A	N/A	N/A
Bruce [BCJ16]	0.456	N/A	0.519	**0.506**
VGG19 Deep Features	**0.611**	**0.642**	**0.537**	0.473

Table 6.1 – Pearson correlation coefficient between predicted IOC scores and ground truth IOC for several models

To confirm these hypotheses, we performed a study of the IOC per image category based on the CAT2000 database. The mean ground truth IOC scores we computed consolidated our original intuitions about what kind of images should have a higher (or lower)

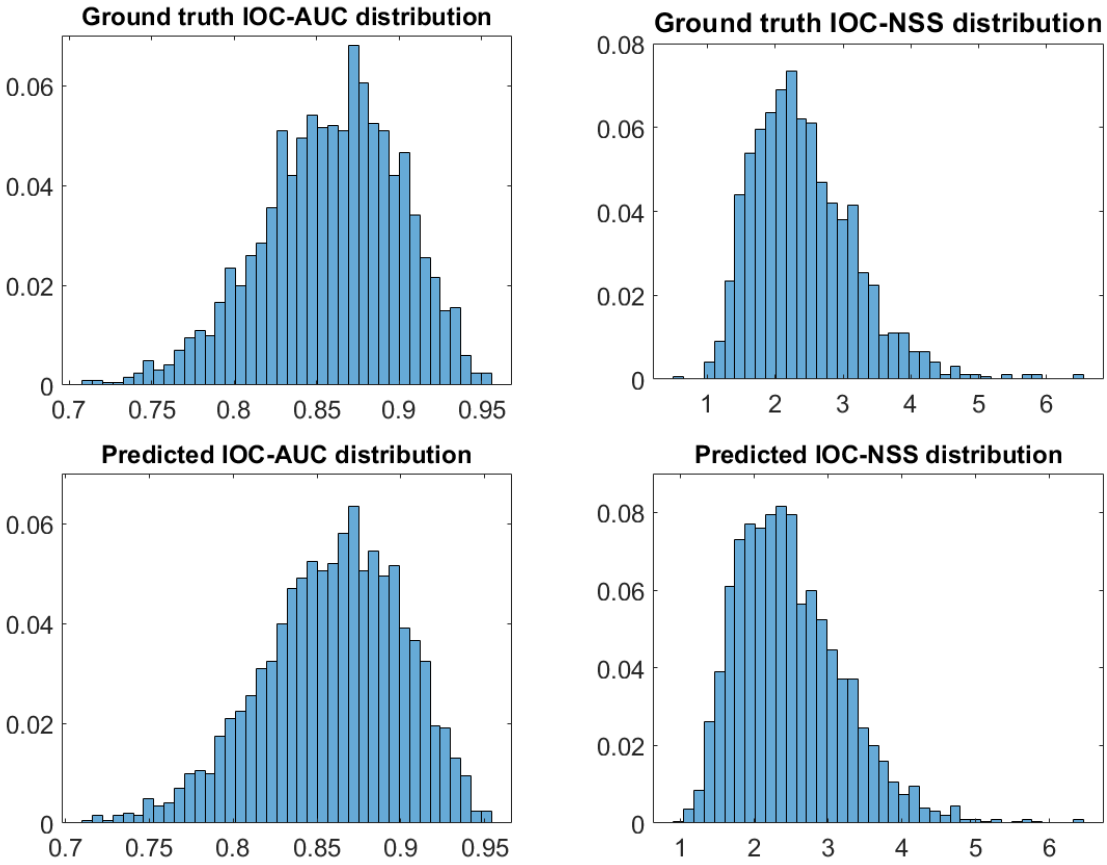


Figure 6.10 – Distributions of the ground truth and predicted IOC scores (both for AUCB and NSS metrics) on the CAT2000 dataset.

IOC score. For example, the category getting the lowest IOC score is composed of satellite images, in which it is hard to distinguish specific objects. On the opposite, the highest IOC scores are performed by sketch drawings, that offer very specific locations to look at, a very high contrast between the drawing and the background, and display familiar objects. It also appears that the predicted mean IOC scores are really close to the ground truth ($r = 0.953$, $p << 10 - 5$ for AUCB, $r = 0.845$, $p << 10 - 5$ for NSS).

IOC values are therefore partly correlated with the high level information of the scene (categories), and partly on the low level information in each individual image. When the effect of individual features is lowered due to the averaging, the categorical visual information becomes more important and leads to an improvement in correlation. This reflects that our model has a capacity to partially understand high-level features common in each category.

6.5 A second model of IOC prediction, for dynamic stimuli

Similarly to the static case, we propose a bottom-up model dedicated to predict inter-observer visual congruency on dynamic stimuli, and more specifically on cinematic stimuli. For this purpose, we designed a two-stream deep neural network, inspired by the architecture of our visual saliency model (see Chapter 5). This model would probably be useful for a wide range of applications; we explore some of them in Section 6.6.

6.5.1 Architecture

The overall architecture of this model is very similar to our visual saliency model. Indeed, we make the assumption that the features that drive attention in videos and that are extracted in deep saliency models should also play an important role into determining whether or not a stimulus will induce high or low visual congruency. This assumption was also made by Rahman and Bruce [RB16], with their Histogram of Predicted Salience features, where they use a stack of feature vectors extracted from several visual saliency models.

Our model is divided into three parts: (i) first, a two-stream encoder extracts features from the optical flow and the frames at different depths; (ii) then, similarly to the ViNet model [Jai+21], these features are passed through 3D convolution layers and upsampling,

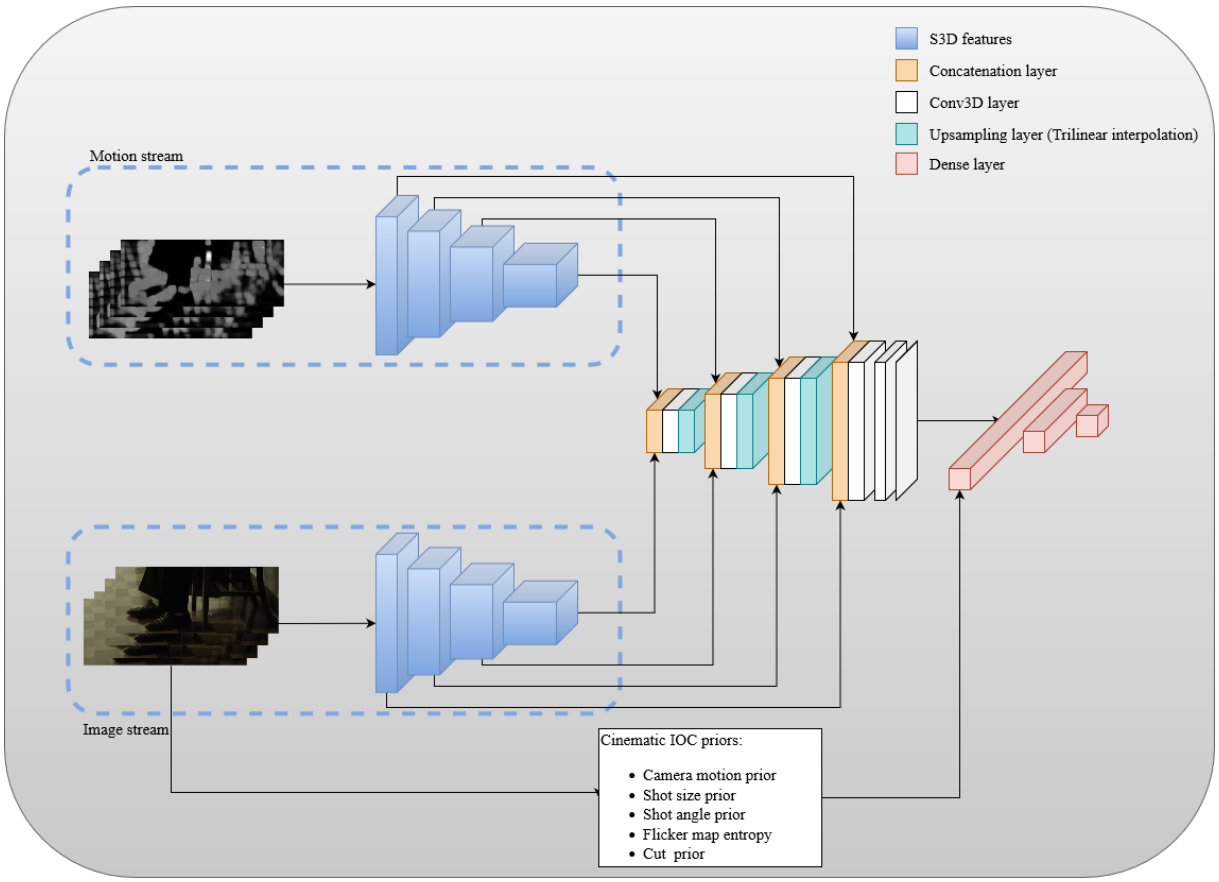


Figure 6.11 – Architecture of the proposed dynamic IOC model

mixing the different hierarchical features using skip connections; (iii) finally, the resulting representation, alongside with IOC priors based on the cinematographic characteristics, is passed through fully connected layers to obtain an IOC value. The overall architecture is shown on Figure 6.11.

Two-stream encoder

Similarly to the visual saliency model (see Chapter 5 for more details and motivations), the encoder part is composed of two S3D networks [Xie+18], one for the spatial features, using a stack of 32 consecutive frames as the input, and the other using the same 32 stack with optical flow. Following the approach of ViNet [Jai+21], for a frame at time t, the input is composed by the frames $F_{t-32+1}, ..., F_t$ and the optical flow maps $O_{t-32+1}, ..., O_t$.

The features are extracted at the end of the four convolution blocks, and passed through skip connections to the decoder module, at different hierarchical levels. For an

input of shape $[T \times C \times H \times W]$, where T is the time window (in our case, 32), C is the number of channels of the input (in our case, 3) and H and W are the height and width of the considered frame, the four features vectors, X_1, X_2, X_3 and X_4 have respective shapes of $[192 \times 16 \times \frac{H}{4} \times \frac{W}{4}]$, $[480 \times 16 \times \frac{H}{8} \times \frac{W}{8}]$, $[832 \times 8 \times \frac{H}{16} \times \frac{W}{16}]$ and $[1024 \times 4 \times \frac{H}{32} \times \frac{W}{32}]$.

Decoder module

The decoder module consists in a succession of concatenations alongside the temporal axis, gathering the hierarchical features from the two stream and the output of the previous upsampling layer, 3D convolution layers, and upsampling using trinlinear interpolation. This integration part is then followed by three 3D convolution layers, to reduce the feature tensor to one in the channel and temporal dimensions. The output features are then flattened, batch-normalized and concatenated with IOC priors, before being passed through three dense layers (similarly to the static IOC model) of size 1024, 256 and 1.

Adding Cinematic IOC priors

As for the visual saliency model, we tried to include high-level cinematic features into the prediction, as we showed it can influence inter-observer visual congruency, and is most likely not taken into account by the feature extractor (as shown in Chapter 4). We include five prior values into the feature vector:

— A camera motion prior, which is the average IOC value for the type of camera movement in the shot of the considered frame,
— A shot size prior, which is the average IOC value for the shot size of the considered frame,
— A shot angle prior, which is the average IOC value for the shot angle of the considered frame,
— The entropy of the flicker map of the considered frame,
— A cut prior, which is the average IOC value of frames within the first 500 milliseconds following a cut if the frame is in this situation, and the average IOC value of the other frames if not.

In their work, Mital *et. al* [Mit+11] showed that flicker, i.e. the change in luminance over time, alongside with motion, is a strong predictor of gaze clustering. Since motion is already taken into account by the optical flow stream, we include flicker by computing the entropy of a flicker map: at time t, we consider frames $F_{t-4}, ..., F_t$, and transfer them

from RGB to the CIELAB color space. We then compute the absolute difference of the frames luminance values $(L_{t-4}, ..., L_t)$, and average it:

$$Fl_t = \frac{1}{N} \sum_{i=1}^{N} |L_{t-i} - L_{t-i+1}| \tag{6.2}$$

Where Fl_t is the flicker map at time t and N is the number of successive frames considered. In our case, we use $N = 5$, similarly to Smith and Mital [SM13], in order to minimize the influence of noise due to compression artifacts.

6.5.2 Training

Implementation details

The frames are first resized to $[288 \times 512]$, using letterboxing if needed to respect the original aspect ratio of the frame. The optical flow frames are processed using the same procedure as Xie *et. al.* [Xie+18]: the optical flow is extracted using the TV-L1 algorithm [ZPB07], the magnitude is truncated into $[-20, 20]$, and the maps are then stored as 3-channels encoded JPEG files.

To process the frame F_t, the sequence $F_{t-32+1}, ..., F_t$ is fed to the model. If any of those frames fall before the first frame of the clip, the first frame is just repeated the adequate amount of times. In order to train the network, we select the 32-frames sequences in a random order among all clips

The priors are computed based on available information; if no editing annotation is provided, we take the average IOC value of the whole dataset for each IOC prior.

The S3D encoder are initialized using weights pre-trained on the Kinetics dataset [Kay+17] on an action-recognition task, using both RGB frames and optical flow. We use the L2 norm as a loss function, with the Adam optimizer, learning rate is initially set at $10e-4$, and the batch size is set at 4.

Training datasets

The model is first trained on the DHF1k dastaset [Wan+]. Ground truth IOC scores are computed based on the supplied scanpaths (using the 20-frames time window). The 500 first clips from the training set are used for training, and the remaining 100 are used for validation, and for early stopping. While the Hollywood2 dataset would have been useful to train on, as it features the type of clips we are interested in, its limitations

prevented us from using it. The low number of free-viewing observers makes it difficult to get a reliable IOC score, and, while adding task-oriented data can be useful for visual saliency, it induces too much of a bias for IOC prediction.

Then, we use 15 clips from our dataset to fine-tune the model (12 for training, 3 for validation), using the IOC priors as we have cinematographic annotations, holding out the 5 remaining clips for testing purposes.

6.5.3 Results

We used three datasets to evaluate the model: the validation set of DHF1k (100 clips), the 5 held out clips from our dataset, and the dataset from Breeden and Hanrahan [BH17].

We observe a Pearson correlation coefficient score between the predicted IOC values and the ground-truth of $r = 0.691$ ($p < 10^{-5}$) for the DHF1k dataset, $r = 0.731$ ($p < 10^{-5}$) for Breeden's dataset and $r = 0.755$ ($p < 10^{-5}$) for ours. These scores are much higher than those we obtained on the static case (see Section 6.4), which can be explained by the prominent role played by motion features on IOC [Mit+11]. DHF1k results also seem to be lower than the other, probably due to the absence of cinematographic priors and annotations, that are used in Breeden's and our dataset. Figure 6.12 shows an example of the predictions from our model on a clip of *The Lord of the Rings*; we observe a pretty good match between the ground truth scores and the predicted scores, especially when peaks are observed.

Ablation study

In order to evaluate how each part of the model contributes to the overall performances, and especially how the cinematic priors play, we performed and ablation study, retraining different settings of the model. First, we tried both branches (RGB and Optical Flow) separated, without any priors. Then, we use the two streams and all of the priors but one each time: the camera motion prior (1), the shot size prior (2), the shot angle prior (3), the flicker map entropy (4) and the cut prior (5). Results for each configuration is shown in Table 6.2 As expected, on the DHF1k set, as there is no significant prior, the correlation scores do not vary when removing priors, except in configuration (4), where the entropy of the flicker map is removed. The camera angle prior does not seem to have any impact on the prediction, which is consistent with what we observed in Section 6.3, and can probably be removed. A small improvement is seen when adding the optical flow stream to the RGB

Dataset	DHF1k [Wan+19]	Breeden [BH17]	Ours
RGB-stream (no prior)	0.631	0.624	0.657
Flow-stream (no prior)	0.471	0.473	0.469
Two-stream+priors (1)	0.690	0.712	0.733
Two-stream+priors (2)	0.689	**0.731**	0.728
Two-stream+priors (3)	0.690	**0.731**	0.754
Two-stream+priors (4)	0.652	0.699	0.718
Two-stream+priors (5)	**0.691**	0.707	0.743
Full model	**0.691**	**0.731**	**0.755**

Table 6.2 – Pearson correlation coefficient between predicted IOC scores and ground truth IOC for several models

stream. The relatively low value for this improvement can be explained by the fact that the RGB-stream already extract at least some motion features, because of its 3D-CNN feature extractor. Finally, overall, adding cinematographic high-level information through these priors seems to be of interest for predicting inter-observer visual congruency.

6.6 Applications

In this section, we describe two simple applications for the IOC metric and the dynamic IOC model.

6.6.1 A tool for style analysis

While studying the results of our IOC model and the ground truth IOC values on our dataset, we made the hypothesis that inter-observer congruency could be a marker of style in movies. We then applied our IOC model to 20 entire movies, from 4 different directors who are known for having specific filmmaking gimmicks and techniques: Roland Emmerich, Stanley Kubrick, Dennis Villeneuve and Robert Zemeckis. The average values of IOC for each movie are summarized in Table 6.3.

The highest values of IOC are observed on Stanley Kubrick's films *2001: A Space Odyssey* (1968) and *The Shining* (1980), which is consistent with his style of very carefully chosen camera movements and well composed shots, where salient elements are easily distinguished from the background, often taking advantage of the symmetry. On the opposite, the cluttered shots of Roland Emmerich, combined with jerky camera motions and the relatively high rhythm of his cuts, dedicated to make the audience feel the gi-

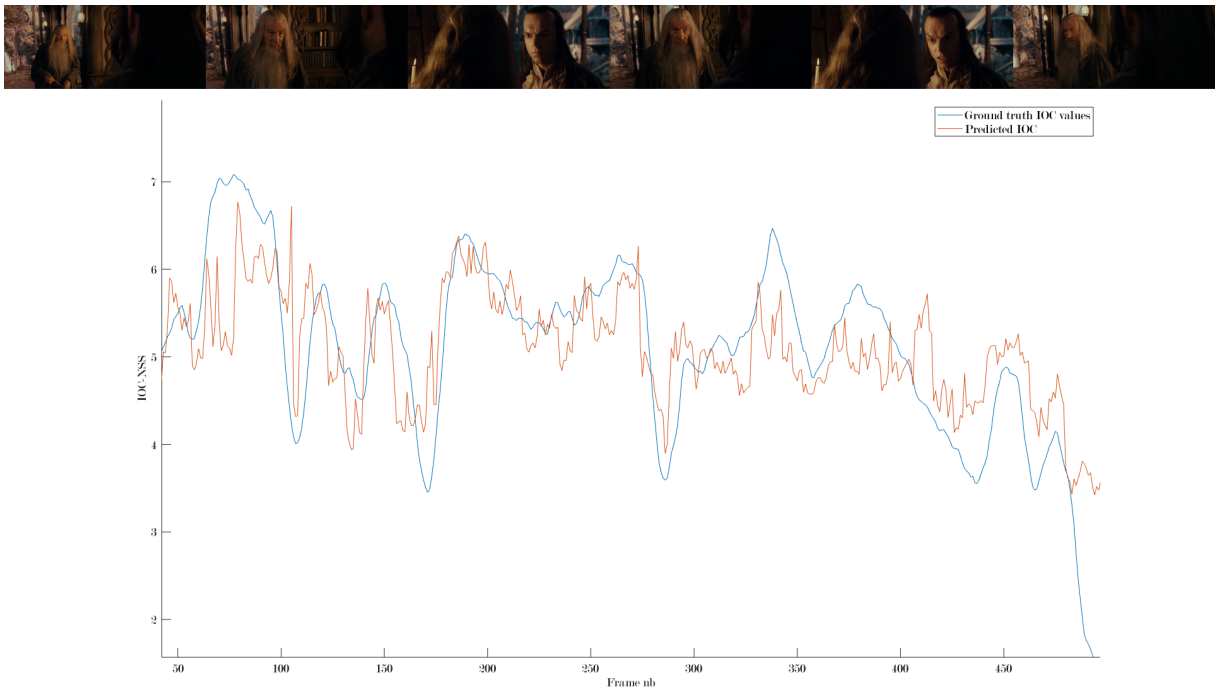


Figure 6.12 – Example of predictions of dynamic IOC on a clip from *The Lord of the Rings: The Fellowship of the Ring* (Peter Jackson, 2001). Ground truth IOC values are displayed in blue, predictions in orange.

Director	Film	Genre (IMDb)	Year	Average IOC
Roland Emmerich	Stargate	Action; Adventure; Sci-Fi	1994	4,020
	Independence Day	Action; Adventure; Sci-Fi	1996	3,974
	The Day After Tomorrow	Action; Adventure; Sci-Fi	2004	3,885
	2012	Action; Adventure; Sci-Fi	2009	3,637
	White House Down	Action; Drama; Thriller	2013	3,711
Stanley Kubrick	2001: A Space Odyssey	Adventure; Sci-Fi	1968	4,876
	A Clockwork Orange	Crime; Drama; Sci-Fi	1971	4,247
	Barry Lyndon	Adventure; Drama; History; War	1975	4,324
	The Shining	Drama; Horror	1980	4,782
	Full Metal Jacket	Drama; War	1987	4,377
Denis Villeneuve	Prisoners	Crime; Drama; Mystery; Thriller	2013	4,109
	Enemy	Drama; Mystery; Thriller	2013	4,360
	Sicario	Action; Crime; Drama; Mystery; Thriller	2015	3,983
	Arrival	Drama; Sci-Fi	2016	4,521
	Blade Runner 2049	Action; Drama; Mystery	2017	4,618
Robert Zemeckis	Back to the Future	Adventure; Comedy; Sci-Fi	1985	4,16
	Who Framed Roger Rabbit	Animation; Adventure; Comedy	1988	3,118
	Forrest Gump	Drama; Romance	1994	4,294
	Contact	Drama; Mystery; Sci-Fi	1997	4,639
	Cast Away	Adventure; Drama; Romance	2000	4,143

Table 6.3 – Average IOC scores on several movies from 4 directors: Denis Villeneuve, Stanley Kubrick, Roland Emmerich and Robert Zemeckis.

gantism and the importance of the events taking place, makes the average IOC scores on his movies the lowest. It is worth noting that the IOC value for *Who Framed Roger Rabbit* (Robert Zemeckis, 1988), which is particularly low, should be discarded as is is certainly an artifact due to the model being trained on live action sequences, while this movie combines live action with 2D animation.

Overall, it appears that there are significant differences of IOC between directors, which supports the hypothesis that IOC can be used as a marker of style. However, more complete studies on selected sequences with identified styles should be conducted to further confirm this.

6.6.2 Attentional continuity

Most movies that are being made today follow more or less the same set of rules when it comes to editing shots together. These rules were developed first in the beginning of the 20th century in the US, and are gathered under the term of *continuity style* [BST85]. The aim of these common editing practices is to provide the audience with a smooth and effortless viewing experience, where the narrative is coherent. While this style has become today a global standard, it is worth noting that this has never been the only way

of considering movie editing: for instance, the *Soviet montage theory* codified by Sergei Eisenstein, or the editing style of the *French new wave* in the 1960s offer significantly different visions of what movie editing conveys, and how it should do so.

The *attentional theory of cinematic continuity* (AToCC) was introduced in 2012 by Smith [Smi12], as a way to explain and formalize this continuity editing style from the perspective of the viewer and its visual attention and perception. A key point of continuity editing is the use of attentional cues to tip the audience that a cut is about to take place, in order to synchronize the cut with an attentional shift. If the viewer shifts his or her attention during a cut, and if the following shot meets his or her expectations (e.g. no sudden change of time, location, or characters if it was not implied before), the cut will feel seamless – or even not be perceived at all.

As a result, if a filmmaker follows the continuity rules, we should expect to see an increase in attentional synchrony right after a cut, as all viewers should shift their gaze at the same time to the new object of interest of the scene. Using our IOC measure, we can quantify this effect, and thus give an idea of how continuous a scene is. To do so, we can simply count, on a sequence, the number of cuts for which the IOC score (computed using the 5-frames window, in order to avoid taking frames before and after the cut for a single measure, due to the binning) 5 frames before the cut is lower than the IOC score 5 frames after the cut. The ratio of such continuous cuts to the total number of cuts in the sequence gives a measure of how continuous the editing is.

Figure 6.13 shows the distribution of the cuts from our dataset based on the value of the IOC before and after the cut. Overall, 74.3% of the edits exhibit a higher IOC value after the cut than before. Interestingly, the highest values are observed on dialogue scenes (*Pulp Fiction*: 89,3%, *Departures*: 87,3%, *Gattaca (2)*: 83,3%, *The Help (1)*: 82,1%, *The Help (2)*: 81,5%), which make an abundant use of classical continuity rules, like the 180° rule, or the traditional structure of shot-reverse shot.

In order to take into account the magnitude of the attentional synchrony peaks, we can also weight each cut by the difference in IOC values after and before the cut:

$$w - Continuity = \frac{\sum_{i \in \mathcal{A}} |IOC_{i-5} - IOC_{i+5}|}{\sum_{j \in \mathcal{C}} |IOC_{j-5} - IOC_{j+5}|} \tag{6.3}$$

where $\mathcal{C}$ is the set of frame numbers where a cut occurs in a sequence, IOC_n is the IOC value of the n-th frame of the sequence, and $\mathcal{A} \subset \mathcal{C}$ is the set of frame numbers where a cut occurs, and for which the IOC value after the cut is higher than before.

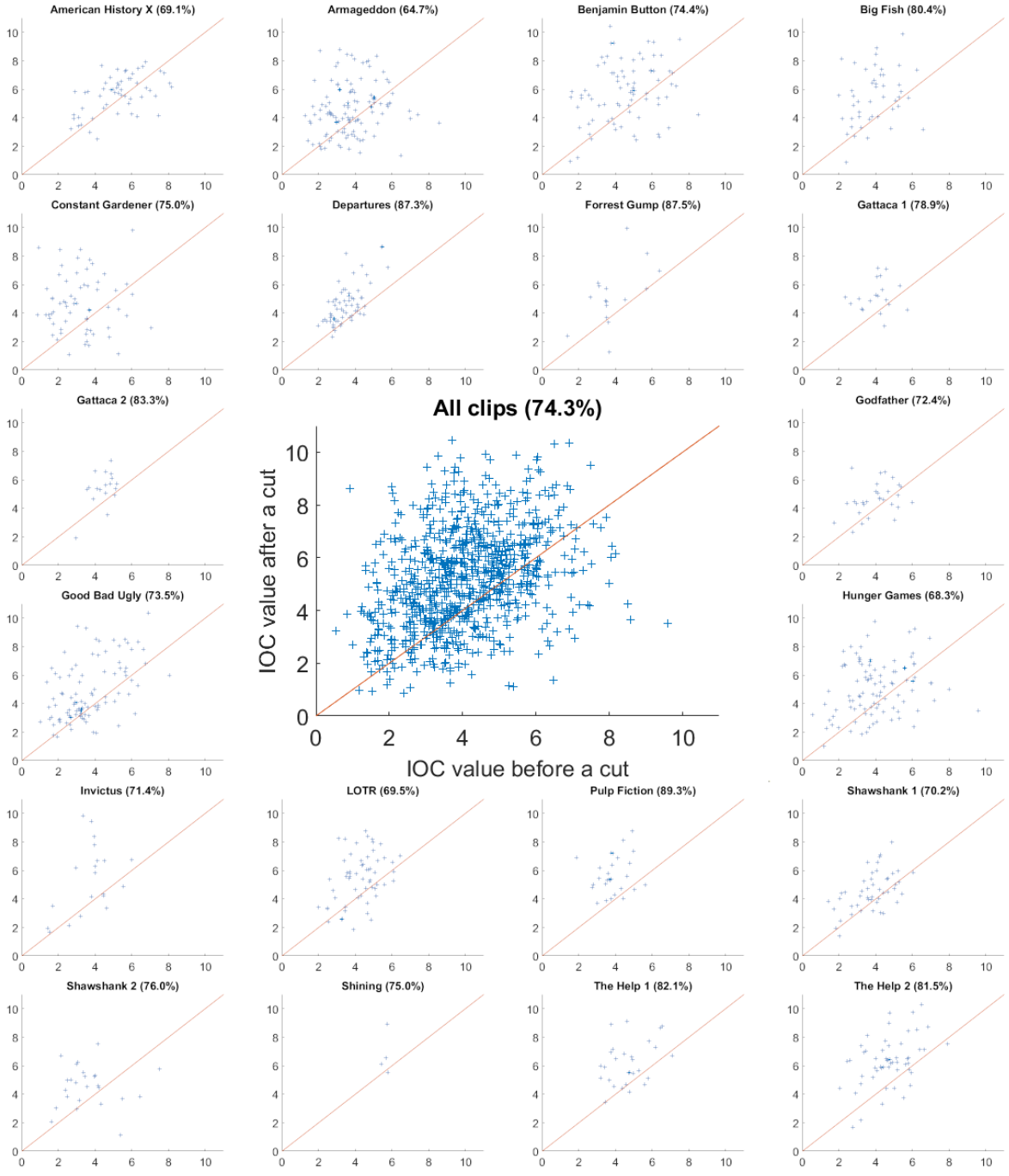


Figure 6.13 – IOC values before and after the cuts, for each movie of our dataset and overall, with the rate of "continuous cuts" (i.e. points over the red $y = x$ line).

We then tried to evaluate whether or not our dynamic IOC model could predict these continuity metrics (continuity or weighted-continuity rates) by computing them using the predictions from our model on the testing part of our dataset. However, it appears that the model is unable to give reliable predictions (correlation coefficients: $r = 0.209$, $p < 10^{-3}$ for the continuity rate, $r = 0.313$, $p < 10^{-3}$ for the w-continuity rate).This is probably caused by the complexity and the variety of attentional cues that are used by continuity editing rules directing the attention during a cut, a lot of which are not taken into account in our model (for instance, off-screen sounds, pointing gestures, gaze directions of the characters, etc).

6.7 Conclusion

In this chapter, we focused our attention on inter-observer visual congruency, a measure of how similar gaze behaviors from different observers are when they are watching the same stimulus. We proposed a way to measure this phenomenon on dynamic stimuli, and introduced two models to predict it, on images and on movie sequences.

While inter-observer congruency (or attentional synchrony) is well known and studied by cognitive psychologists, we argue that more attention should be payed to this measure in computer vision, both from a modeling point of view and for the resulting applications. While its role as an upper bound of the performance of visual attention models is well-known, it can also be used to constraint visual saliency predictions: for instance, a predicted saliency map exhibiting a lot of salient areas will probably be wrong if the IOC is high (meaning that observers tend to look at the same place). In this regard, predicting IOC can be used to give an estimation of how "difficult" a saliency prediction will be, and serve as a likelihood score.

It could also be interesting to evaluate the interest of this measure in the context of image quality assessment: a high degree of visual congruency means that there might be a single strong visual attractor on the image, and thus artifacts on other areas of the frame could be overlooked.

From the perspective of filmmaking, knowing when viewers will focus their attention in the same location is tremendously useful for directors, as it allows them even more control on what the viewer experiences, in order to convey their narrative content and messages at best. For virtual cinematography and automated editing, this can be used to constraint the choice of the cuts, for instance, depending on the desired style.

General conclusion

In this thesis, we aimed to propose an exploration of cinematographic features from the perspective of the viewer, and more specifically their visual attention. We investigated the impact of the director's choices under the lens of visual saliency and inter-observer congruency, asking ourselves what impact these choices had on the viewers' gaze patterns, and how to include those in a modeling context. In this general conclusion, we give a summary of the contributions of the thesis and propose several research perspective, deriving from the outcomes of our work, but also from the experience gathered along the way.

Contributions

An eye-tracking dataset for studying visual attention in movies

First, we proposed to evaluate the effects of editing and directing choices on visual attention by conducting an eye-tracking experiment on movie clips, extending the work of Breeden and Hanrahan [BH17]. Using hand-crafted high-level features regarding several cinematographic aspects, such as the camera motion, the camera angle or the shot size, we were able to quantify the influence of the film grammar. When evaluating visual attention models on these specific kind of stimuli, it became obvious that state-of-the-art models were unable to grasp these high-level semantics, and thus we found significant discrepancies between their prediction and human visual attention. More importantly, a lot of these discrepancies happen when the stimulus contains a lot of non-static information, and is semantically rich. Studying these kind of cinematic stimuli seems then to be of great interest in order to develop richer sets of attention features for a large array of applications.

A visual saliency model for movie sequences

After having discussed the shortcomings of visual saliency models on cinematic stimuli, we addressed the problem of reliably predicting visual attention on movie scenes. To this

end, we have designed a deep saliency model based on two streams using 3D convolution networks. The first stream deals with motion information in the form of optical flow, and the second processes a succession of RGB frames. When each stream have output a prediction, the resulting maps are fused together by a 2D convolutional network, alongside with cinematic feature maps dedicated to include high-level knowledge of the scene. Using ablation analysis, we showed that this approach allows for a significant improvement in the predicting power of the model. We have also shown that our model outperforms other state-of-the-art approaches on movie clips databases.

Inter-observer visual congruency and movies

Third, we considered the problem of inter-observer visual congruency and its entanglements with the cinematic features described earlier. We started by designing a robust way of measuring inter-observer congruency on both static and dynamic stimuli. Using the eye-tracking data that we collected, we were able to highlight several relations between this measure and the cinematic properties of the considered movie clips, implying that director's choices have a high degree of influence on IOC. We then proposed a first IOC prediction model for static images, using an encoder-decoder architecture that relies on 2D convolution networks. By extracting features at different stages of the encoding process, we were able to achieve a 0.642 correlation between our prediction scores and the ground truth values on the CAT2000 dataset [BI15], and 0.611 on the MIT dataset [JDT12]. For the dynamic case, we also proposed a bottom-up IOC model, specifically designed to perform well on cinematic stimuli. To this extent, we use a two-stage structure similar to the visual saliency model of Chapter 5, where optical flow and RGB frames are processed separately, and features are extracted and fused together at different levels. While this model performs very well on cinematic content (respectively 0.731 and 0.755 correlation on Breeden's dataset and ours), we also show that the method is robust enough to deal with non-cinematographic content (0.691 correlation score on the DHF1K [Wan+] dataset.

Perspectives

Here, we briefly discuss a few ideas and research directions that either arose during these last three years working on this topic, or extend the work presented in this thesis.

Automated extraction of cinematic features

During our work, we heavily relied on hand-crafted features to describe the cinematic language of a movie sequence. However, obtaining such features requires tedious work, by annotating each shot, or even each frame of a clip for a finer-grained level of annotations. In this light, automation of the extraction of cinematic features will allow for a better comprehension of cinematographic patterns, which can then be applied to various tasks.

While some annotations can already be extracted automatically, such as bounding boxes for characters, shot boundaries, and to a lesser extent, camera motion, many cinematographic properties still require the human eye to be recognized, like the size of the shot for instance. The recent work of Courant *et. al.* [Cou+21] show promising results being able to accurately detect camera motion, but also frame layering, which could be used to infer a shot size.

Another trail that can be followed with regards to cinematic features is the explainability of deep learning models trained on style-analysis tasks, like movie style classification. By studying the latent representations of clips in such models, we could learn what kind of features are actually extracted and their relative importance. Such knowledge will then allow us to build a robust set of deep cinematic features that could be used in many application, for both filmmakers and cinematic scholars.

Redefining the evaluation of dynamic visual saliency

When performing experiments on deep saliency models, it appeared that the way of evaluating dynamic attention models could somewhat be related to why static saliency models perform surprisingly well on video benchmarks. By only evaluating the generated saliency maps frame-by-frame, the dynamics of the predictions are not taken into account. A small temporal offset in the generated saliency maps for instance could have a dramatic effect on frame-by-frame scores, while still being very close to the ground-truth fixation densities.

An answer to this issue could be the use as ground-truth of 3-dimensional fixation densities smoothed in the temporal domain, such as we did in Chapter 6. The intuition behind this would be that if a fixation is likely to occur at a given pixel of given frame, it is also likely to occur at the same location in temporally close frames.

Another way of evaluating dynamic saliency could also be the extension of the probabilistic framework proposed by Kümmerer *et. al.* [KWB18] to the temporal domain.

Automated editing systems

Editing a video, i.e. gathering a collection of rushes and creating a narrative unit by selecting the best frames, and cutting and joining them together, is a difficult and tedious task. In order to alleviate the work of editors, recent research work have started to introduce automated systems designed to help with the editing process. For instance, Pardo *et. al.* [Par+21] proposed a model that, given two untrimmed shots, returns the plausibility of a cut happening at any given moment, based on a deep representation learned from a high volume of videos.

Considering the importance of perceptual cues in continuity editing, we believe that such a system could greatly benefit from visual attention features, such as inter-observer congruency. Indeed, knowing when the attention of the audience is focused on a specific element or not, or where the attention might be guided to should a cut occur can without a doubt help a model making choices, on whether to make a cut or not, or what clip to cut to.

List of publications

Conference papers

Alexandre Bruckert, Yat Hong Lam, Marc Christie, Olivier Le Meur, «Deep learning for inter-observer congruency prediction», ICIP, 2019, pp.3766–3770

Journal papers

Alexandre Bruckert, Hamed R Tavakoli, Zhi Liu, Marc Christie, Olivier Le Meur, «Deep saliency models: the quest for the loss function», Neurocomputing 453, 2021, pp.693–704

Alexandre Bruckert, Marc Christie, Olivier Le Meur, «Where to look at the movies: Analyzing visual attention to understand movie editing», Behavior Research Methods, 2022

BIBLIOGRAPHY

[AA12] Ehud Ahissar and Amos Arieli, « Seeing via Miniature Eye Movements: A Dynamic Hypothesis for Vision », *in*: *Frontiers in computational neuroscience* 6.*89* (2012), DOI: 10.3389/fncom.2012.00089.

[ADF10] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari, « What is an object? », *in*: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 73–80, DOI: 10.1109/CVPR.2010.5540226.

[AK14] Sharon Alpert and Pavel Kisilev, « Unsupervised detection of abnormalities in medical images using salient features », *in*: *Medical Imaging 2014: Image Processing*, ed. by Sebastien Ourselin and Martin A. Styner, vol. 9034, International Society for Optics and Photonics, SPIE, 2014, pp. 295–300, DOI: 10.1117/12.2043213.

[AL10] Tamar Avraham and Michael Lindenbaum, « Esaliency (Extended Saliency): Meaningful Attention Using Stochastic Image Modeling », *in*: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.*4* (2010), pp. 693–708, DOI: 10.1109/TPAMI.2009.53.

[BA18] Shuang Bai and Shan An, « A survey on automatic image caption generation », *in*: *Neurocomputing* 311 (2018), pp. 291–304, DOI: 10.1016/j.neucom.2018.05.080.

[BAA11] Ali Borji, Majid N. Ahmadabadi, and Babak N. Araabi, « Cost-Sensitive Learning of Top-Down Modulation for Attentional Control », *in*: *Machine Vision and Applications* 22 (2011), pp. 61–76, DOI: 10.1007/s00138-009-0192-0.

[Bak+18] Cagdas Bak et al., « Spatio-Temporal Saliency Networks for Dynamic Saliency Prediction », *in*: *IEEE Transactions on Multimedia* 20.*7* (2018), pp. 1688–1698, DOI: 10.1109/TMM.2017.2777665.

[BBD14] Vincent Buso, Jenny Benois-Pineau, and Jean-Philippe Domenger, « Geometrical Cues in Visual Saliency Models for Active Object Recognition in Egocentric Videos », *in*: *Proceedings of the 1st International Workshop on Perception Inspired Video Processing*, 2014, pp. 9–14, DOI: 10.1145/2662996.2663007.

[BCJ16] N. D. B. Bruce, C. Catton, and S. Janjic, « A Deeper Look at Saliency: Feature Contrast, Semantics, and Beyond », *in*: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 516–524.

[BCM21] Alexandre Bruckert, Marc Christie, and Olivier Le Meur, *Where to look at the movies : Analyzing visual attention to understand movie editing*, 2021, arXiv: 2102.13378 [cs.CV].

[Bel+21] Giovanni Bellitto et al., « Hierarchical Domain-Adapted Feature Learning for Video Saliency Prediction », *in*: (2021), arXiv: 2010.01220.

[BGL98] William H. Bares, Joël P. Grégoire, and James C. Lester, « Realtime Constraint-Based Cinematography for Complex Interactive 3D Worlds », *in*: *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, AAAI '98/IAAI '98, AAAI Press, 1998, pp. 1101–1106, DOI: 10.5555/295240.296260.

[BH17] Katherine Breeden and Pat Hanrahan, « Gaze Data for the Analysis of Attention in Feature Films », *in*: *ACM Transactions on Applied Perception* 14.*4* (2017), DOI: 10.1145/3127588.

[BI11] Ali Borji and Laurent Itti, « Scene classification with a sparse set of salient regions », *in*: *IEEE International Conference on Robotics and Automation*, 2011, pp. 1902–1908, DOI: 10.1109/ICRA.2011.5979815.

[BI13] Ali Borji and Laurent Itti, « State-of-the-Art in Visual Attention Modeling », *in*: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.*1* (2013), pp. 185–207, DOI: 10.1109/TPAMI.2012.89.

[BI15] Ali Borji and Laurent Itti, « Cat2000: A large scale fixation dataset for boosting saliency research », *in*: *arXiv preprint arXiv:1505.03581* (2015).

[Bor+12] Ali Borji et al., « Adaptive object tracking by learning background context », *in*: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 23–30, DOI: 10.1109/CVPRW.2012.6239191.

[Bor+13] Ali Borji et al., « Analysis of Scores, Datasets, and Models in Visual Saliency Prediction », *in*: *2013 IEEE International Conference on Computer Vision*, 2013.

[Bor18] Ali Borji, « Saliency prediction in the deep learning era: An empirical investigation », *in*: *arXiv preprint arXiv:1810.03716* (2018).

[Bor19] Ali Borji, « Saliency Prediction in the Deep Learning Era: Successes and Limitations », *in*: *IEEE transactions on pattern analysis and machine intelligence* 43.*2* (2019), pp. 679–700, DOI: 10.1109/tpami.2019.2935715.

[Bre16] David Breathnach, « Attentional Synchrony and the Affects of Repetitve Movie Viewing », *in*: *AICS*, 2016, URL: http://ceur-ws.org/Vol-1751/AICS_2016_paper_57.pdf.

[Bro16] Blain Brown, *Cinematography: theory and practice: image making for cinematographers and directors (3rd ed.)* Routledge, New York, 2016, ISBN: 978-1-138-21258-9.

[Bru+19] Alexandre Bruckert et al., « Deep Learning For Inter-Observer Congruency Prediction », *in*: *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 3766–3770, DOI: 10.1109/ICIP.2019.8803596.

[Bru+21] Alexandre Bruckert et al., « Deep saliency models : The quest for the loss function », *in*: *Neurocomputing* 453 (2021), pp. 693–704, DOI: 10.1016/j.neucom.2020.06.131.

[BSI12] Ali Borji, Dicky N. Sihite, and Laurent Itti, « Probabilistic learning of task-specific visual attention », *in*: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 470–477, DOI: 10.1109/CVPR.2012.6247710.

[BSI13] Ali Borji, Dicky N. Sihite, and Laurent Itti, « Quantitative Analysis of Human-Model Agreement in Visual Saliency Modeling: A Comparative Study », *in*: *IEEE Transactions on Image Processing* 22.*1* (2013), pp. 55–69, DOI: 10.1109/TIP.2012.2210727.

[BST85] David Bordwell, Janet Staiger, and Kristin Thompson, *The classical Hollywood cinema : film style & mode of production to 1960*, New York: Columbia University Press, 1985, ISBN: 0-231-06054-8.

[BT05] Neil D. B. Bruce and John K. Tsotsos, « Saliency Based on Information Maximization », *in*: *Proceedings of the 18th International Conference on Neural Information Processing Systems*, 2005, pp. 155–162, DOI: 10.5555/2976248.2976268.

[BT09] Neil D. B. Bruce and John K. Tsotsos, « Saliency, attention, and visual search: An information theoretic approach », *in*: *Journal of Vision* 9.*3* (2009), pp. 5–5, ISSN: 1534-7362, DOI: 10.1167/9.3.5.

[BTM00] William H. Bares, Somying Thainimit, and Scott Mcdermott, « A Model for Constraint-Based Camera Planning », *in*: *Proc. AAAI Spring Symp. Smart Graphics*, 2000, pp. 84–91.

[Byl+16] Z. Bylinskii et al., « Where should saliency models look next? », *in*: *European Conference on Computer Vision (ECCV)*, 2016.

[Byl+19] Zoya Bylinskii et al., « What Do Different Evaluation Metrics Tell Us About Saliency Models? », *in*: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.*3* (2019), pp. 740–757, DOI: 10.1109/TPAMI.2018.2815601.

[CA16] James E. Cutting and Kacie L. Armstrong, « Facial expression, size, and clutter: Inferences from movie structure to emotion judgments and back », *in*: *Attention, Perception, & Psychophysics* 78 (2016), pp. 891–901, DOI: 10.3758/s13414-015-1003-5.

[Car11] Marisa Carrasco, « Visual attention: The past 25 years », *in*: *Vision Research* 51.*13* (2011), Vision Research 50th Anniversary Issue: Part 2, pp. 1484–1525, DOI: https://doi.org/10.1016/j.visres.2011.04.012.

[CBN05] Hannah F. Chua, Julie E. Boland, and Richard E. Nisbett, « Cultural variation in eye movements during scene perception », *in*: *Proceedings of the National Academy of Sciences* 102.*35* (2005), pp. 12629–12633, DOI: 10.1073/pnas.0506162102.

[CD98] J. M. Corridoni and A. Del Bimbo, « Structured representation and automatic indexing of movie information content », *in*: *Pattern Recognition* 31.*12* (1998), pp. 2027–2045, DOI: 10.1016/S0031-3203(98)00061-2.

[Cer+08] Moran Cerf et al., « Predicting human gaze using low-level saliency combined with face detection », *in*: *Advances in Neural Information Processing Systems*, vol. 20, 2008, pp. 241–248.

[Cha18] Wang Y. Chang G.J. Zhang Y., « An Element Sensitive Saliency Model with Position Prior Learning for Web Pages », *in*: *ICIAI*, 2018.

[Che+15] Ming-Ming Cheng et al., « Global Contrast Based Salient Region Detection », *in*: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.*3* (2015), pp. 569–582, DOI: 10.1109/TPAMI.2014.2345401.

[Che+17] Liang-Chieh Chen et al., « Rethinking atrous convolution for semantic image segmentation », *in*: *arXiv preprint arXiv:1706.05587* (2017).

[Che+20] Zhaohui Che et al., « How is Gaze Influenced by Image Transformations? Dataset and Model », *in*: *IEEE Transactions on Image Processing* 29 (2020), pp. 2287–2300, DOI: 10.1109/TIP.2019.2945857.

[Chr+96] David B. Christianson et al., « Declarative Camera Control for Automatic Cinematography », *in*: *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 1*, AAAI'96, AAAI Press, 1996, pp. 148–155, DOI: 10.5555/1892875.1892897.

[CK08] Han Collewijn and Eileen Kowler, « The significance of microsaccades for vision and oculomotor control », *in*: *Journal of vision* 8.*14* (2008), pp. 1–21, DOI: 10.1167/8.14.20.

[Cor+16] Marcella Cornia et al., « A Deep Multi-Level Network for Saliency Prediction », *in*: *International Conference on Pattern Recognition (ICPR)*, 2016.

[Cor+18a] Marcella Cornia et al., « Predicting Human Eye Fixations via an LSTM-Based Saliency Attentive Model », *in*: *IEEE Transactions on Image Processing* 27.*10* (2018), pp. 5142–5154, DOI: 10.1109/TIP.2018.2851672.

[Cor+18b] Marcella Cornia et al., « Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model », *in*: *IEEE Transactions on Image Processing* 27.*10* (2018), pp. 5142–5154.

[Cou+21] Robin Courant et al., « High-Level Features for Movie Style Understanding », *in*: *ICCV 2021 Workshop on AI for Creative Video Editing and Understanding*, 2021.

[CS02] Maurizio Corbetta and Gordon L. Shulman, « The Visual Image », *in*: *Nature Reviews Neuroscience 3* (2002), pp. 201–215, DOI: 10.1038/nrn755.

[CZ17] João Carreira and Andrew Zisserman, « Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset », *in*: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 4724–4733.

[CZZ21] Qinyao Chang, Shiping Zhu, and Lanyun Zhu, *Temporal-Spatial Feature Pyramid for Video Saliency Detection*, 2021, arXiv: 2105.04213 [cs.CV].

[DD95] Robert Desimone and John Duncan, « Neural Mechanisms of Selective Visual Attention », *in*: *Annual Review of Neuroscience* 18 (1995), pp. 193–222, DOI: 10.1146/annurev.ne.18.030195.001205.

[Den+09] Jia Deng et al., « ImageNet: A large-scale hierarchical image database », *in*: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255, DOI: 10.1109/CVPR.2009.5206848.

[DJN20] Richard Droste, Jianbo Jiao, and J. Noble, « Unified Image and Video Saliency Modeling », *in*: *ArXiv* abs/2003.05477 (2020).

[Dor+10] Michael Dorr et al., « Variability of eye movements when viewing dynamic natural scenes », *in*: *Journal of Vision* 10.*10* (2010), pp. 28–28, DOI: 10.1167/10.10.28.

[Dos+15] Alexey Dosovitskiy et al., « FlowNet: Learning Optical Flow With Convolutional Networks », *in*: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2758–2766, DOI: 10.1109/ICCV.2015.316.

[Duc02] Andrew T. Duchowski, « A breadth-first survey of eye-tracking applications », *in*: *Behavior Research Methods* 34.*4* (2002), pp. 455–470, DOI: 10.3758/BF03195475.

[DZ94] Steven M. Drucker and David Zeltzer, « Intelligent Camera Control in a Virtual Environment », *in*: *Proceedings of Graphics Interface*, 1994, pp. 190–199, DOI: 10.20380/GI1994.23.

[EH13] Mohsen Emami and Lawrence L. Hoberock, « Selection of a best metric and evaluation of bottom-up visual saliency models », *in*: *Image and Vision Computing* 31.*10* (2013), pp. 796–808, DOI: 10.1016/j.imavis.2013.08.004.

[Eng+11] Ulrich Engelke et al., « Visual Attention in Quality Assessment », *in*: *IEEE Signal Processing Magazine* 28.*6* (2011), pp. 50–59, DOI: 10.1109/MSP.2011.942473.

[Eng+13] Ulrich Engelke et al., « Comparative Study of Fixation Density Maps », *in*: *IEEE Transactions on Image Processing* 22.*3* (2013), pp. 1121–1133, DOI: 10.1109/TIP.2012.2227767.

[Fen+11] Yunlong Feng et al., « Hidden Markov Model for eye gaze prediction in networked video streaming », *in*: *2011 IEEE International Conference on Multimedia and Expo*, 2011, pp. 1–6, DOI: 10.1109/ICME.2011.6011935.

[Fin74] John M. Findlay, « Direction perception and human fixation eye movements », *in*: *Vision Research* 14 (1974), pp. 703–711, DOI: 10.1016/0042-6989(74)90067-4.

[Fin97] John M. Findlay, « Saccade target selection during visual search », *in*: *Vision Research* 37.*5* (1997), pp. 617–631, DOI: 10.1016/S0042-6989(96)00218-0.

[Fou+13] Tom Foulsham et al., « Leftward biases in picture scanning and line bisection: A gaze-contingent window study », *in*: *Vision Research* 78 (2013), pp. 14–25, DOI: https://doi.org/10.1016/j.visres.2012.12.001.

[FU08] Tom Foulsham and Geoffrey Underwood, « What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition », *in*: *Journal of Vision* 8.*2* (2008), DOI: 10.1167/8.2.6.

[Gal+15] Quentin Galvane et al., « Camera-on-Rails: Automated Computation of Constrained Camera Paths », *in*: *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games*, 2015, pp. 151–157, DOI: 10.1145/2822013.2822025.

[Gao+17] Ge Gao et al., « Saliency-guided adaptive seeding for supervoxel segmentation », *in*: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 4938–4943, DOI: 10.1109/IROS.2017.8206374.

[Gar+12] A. Garcia-Diaz et al., « Saliency from hierarchical adaptation through decorrelation and variance normalization », *in*: *Image and Vision Computing* 30.*1* (2012), pp. 51–64, ISSN: 0262-8856, DOI: http://dx.doi.org/10.1016/j.

imavis.2011.11.007, URL: http://www.sciencedirect.com/science/article/pii/S0262885611001235.

[GC18] Siavash Gorji and James J. Clark, « Going from Image to Video Saliency: Augmenting Image Salience with Dynamic Attentional Push », *in*: *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7501–7511, DOI: 10.1109/CVPR.2018.00783.

[GEB15] L.A. Gatys, A.S. Ecker, and M. Bethge, « A neural algorithm of artistic style », *in*: *arXivpreprint*, 2015, arXiv: 1508.06576.

[GHV09] Dashan Gao, Sunhyoung Han, and Nuno Vasconcelos, « Discriminant Saliency, the Detection of Suspicious Coincidences, and Applications to Visual Recognition », *in*: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31.*6* (2009), pp. 989–1005, DOI: 10.1109/TPAMI.2009.27.

[Git+14] Yury Gitman et al., « Semiautomatic Visual-Attention Modeling and Its Application to Video Compression », *in*: *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 1105–1109, DOI: 10.1109/ICIP.2014.7025220.

[GMV07] Dashan Gao, Vijay Mahadevan, and Nuno Vasconcelos, « The Discriminant Center-Surround Hypothesis for Bottom-up Saliency », *in*: *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, 2007, pp. 497–504.

[GMZ08] Chenlei Guo, Qi Ma, and L. Zhang, « Spatio-temporal Saliency detection using phase spectrum of quaternion fourier transform », *in*: *2008 IEEE Conference on Computer Vision and Pattern Recognition* (2008), pp. 1–8.

[Gom72] E. H. Gombrich, « The Visual Image », *in*: *Scientific American* 227.*3* (1972), pp. 82–97, DOI: 10.1038/scientificamerican0972-82.

[GS14] Nurit Gronau and Meytal Shachar, « Contextual integration of visual objects necessitates attention », *in*: *Attention, Perception, & Psychophysics* 76 (2014), pp. 695–714, DOI: 10.3758/s13414-013-0617-8.

[GWP07] Robert B. Goldstein, Russell L. Woods, and Eli Peli, « Where people look when watching movies: do all viewers look at the same place? », *in*: *Computers in Biology and Medicine* 37.*7* (2007), pp. 957–964, DOI: 10.1016/j.compbiomed.2006.08.018.

[GZ10] Chenlei Guo and Liming Zhang, « A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression », *in*: *IEEE Transactions on Image Processing* 19.*1* (2010), pp. 185–198, DOI: 10.1109/TIP.2009.2030969.

[Han+16] Michael Hanke et al., « A studyforrest extension, simultaneous fMRI and eye gaze recordings during prolonged natural stimulation », *in*: *Scientific Data* 3 (2016).

[Har+16] Katarzyna Harezlak et al., « Application of Eye Tracking for Diagnosis and Therapy of Children with Brain Disabilities », *in*: *Intelligent Decision Technologies*, 2016, pp. 323–333, DOI: 10.1007/978-3-319-39627-9_28.

[HB14] Hadi Hadizadeh and Ivan V. Bajic, « Saliency-Aware Video Compression », *in*: *IEEE Transactions on Image Processing* 23 (2014), pp. 19–33, DOI: 10.1109/TIP.2013.2282897.

[HCS96] Li-wei He, Michael F. Cohen, and David H. Salesin, « The Virtual Cinematographer: A Paradigm for Automatic Real-Time Camera Control and Directing », *in*: *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, New York, NY, USA: Association for Computing Machinery, 1996, pp. 217–224, DOI: 10.1145/237170.237259.

[He+19] Sen He et al., « Understanding and Visualizing Deep Visual Saliency Models », *in*: *arXiv preprint arXiv:1903.02501* (2019).

[HHK12] X. Hou, J. Harel, and C. Koch, « Image Signature: Highlighting Sparse Salient Regions », *in*: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.*1* (2012), pp. 194–201.

[HK03] Amelia R. Hunt and Alan Kingstone, « Covert and overt voluntary attention: linked or independent? », *in*: *Cognitive Brain Research* 18.*1* (2003), pp. 102–105, DOI: https://doi.org/10.1016/j.cogbrainres.2003.08.006.

[HK18] Katarzyna Harezlak and Pawel Kasprowski, « Application of eye tracking in medicine: A survey, research issues and challenges », *in*: *Computerized Medical Imaging and Graphics*, vol. 65, 2018, pp. 176–190, DOI: 10.1016/j.compmedimag.2017.04.006.

[HKP06a] Jonathan Harel, Christof Koch, and Pietro Perona, « Graph-Based Visual Saliency », *in*: *Proceedings of the 19th International Conference on Neural Information Processing Systems*, 2006.

[HKP06b] Jonathan Harel, Christof Koch, and Pietro Perona, « Graph-Based Visual Saliency », *in*: *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, 2006, pp. 545–552, DOI: 10.5555/2976456.2976525.

[HKS18] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh, *Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?*, 2018, arXiv: 1711.09577 [cs.CV].

[Hua+15] Xun Huang et al., « SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks », *in*: *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 262–270, DOI: 10.1109/ICCV.2015.38.

[Hua+20] Qingqiu Huang et al., « MovieNet: A Holistic Dataset for Movie Understanding », *in*: 2020, pp. 709–727, DOI: 10.1007/978-3-030-58548-8_41.

[IK00] Laurent Itti and Christof Koch, « A saliency-based search mechanism for overt and covert shifts of visual attention », *in*: *Vision Research* 40.*10* (2000), pp. 1489–1506, DOI: https://doi.org/10.1016/S0042-6989(99)00163-7.

[IKN98] Laurent Itti, Christof Koch, and Ernst Niebur, « A model of saliency-based visual attention for rapid scene analysis », *in*: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.*11* (1998), pp. 1254–1259, DOI: 10.1109/34.730558.

[Itt04] Laurent Itti, « Automatic foveation for video compression using a neurobiological model of visual attention », *in*: *IEEE Transactions on Image Processing* 13 (2004), pp. 1304–1318, DOI: 10.1109/TIP.2004.834657.

[Jai+21] Samyak Jain et al., *ViNet: Pushing the limits of Visual Modality for Audio-Visual Saliency Prediction*, 2021, arXiv: 2012.06170 [cs.CV].

[Jam90] William James, *The principles of psychology*, New York, Henry Holt, 1890, DOI: doi.org/10.1037/10538-000.

[JDT12] Tilke Judd, Frédo Durand, and Antonio Torralba, « A Benchmark of Computational Models of Saliency to Predict Human Fixations », *in*: *MIT libraries, CSAIL Technical Reports*, 2012, URL: http://hdl.handle.net/1721.1/68590.

[JF16] A. Alahi J. Johnson and L. Fei-Fei, « Perceptual Losses for Real-Time Style Transfer and Super-Resolution », *in*: *2016 European Conferenceon Computer Vision (ECCV)*, 2016, pp. 694–711.

[Jia+15] M. Jiang et al., « SALICON: Saliency in Context », *in*: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[Jia+18] Lai Jiang et al., « DeepVS: A Deep Learning Based Video Saliency Prediction Approach », *in*: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[Jia18] Sen Jia, « EML-NET: An Expandable Multi-Layer NETwork for Saliency Prediction », *in*: *CoRR* abs/1805.01047 (2018), arXiv: 1805.01047, URL: http://arxiv.org/abs/1805.01047.

[JMV16] S. Jetley, N. Murray, and E. Vig, « End-to-End Saliency Mapping via Probability Distribution Prediction », *in*: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[JP07] Sébastien Jodogne and Justus H. Piater, « Closed-Loop Learning of Visual Control Policies », *in*: *Journal of Artificial Intelligence Research* 28.*1* (2007), pp. 349–391, DOI: 10.5555/1622591.1622601.

[Jud+09] Tilke Judd et al., « Learning to predict where humans look », *in*: *12th international conference on Computer Vision*, IEEE, 2009, pp. 2106–2113.

[KAB17] Srinivas S. S. Kruthiventi, Kumar Ayush, and R. Venkatesh Babu, « DeepFix: A Fully Convolutional Neural Network for Predicting Human Eye Fixations », *in*: *IEEE Transactions on Image Processing* 26.*9* (2017), pp. 4446–4456, DOI: 10.1109/TIP.2017.2710620.

[Kan+09] Christopher Kanan et al., « SUN: Top-down saliency using natural statistics », *in*: *Visual cognition* 17.*6-7* (2009), pp. 979–1003, DOI: 10.1080/13506280902771138.

[Kar+17] Nour Karessli et al., « Gaze Embeddings for Zero-Shot Image Classification », *in*: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 6412–6421.

[Kay+17] Will Kay et al., *The Kinetics Human Action Video Dataset*, 2017, arXiv: 1705.06950 [cs.CV].

[KC17] Alex Kendall and Roberto Cipolla, « Geometric loss functions for camera pose regression with deep learning », *in*: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[KG96] Philip Kortum and Wilson S. Geisler, « Implementation of a foveated image coding system for image bandwidth reduction », *in*: *Human Vision and Electronic Imaging*, vol. 2657, SPIE, 1996, pp. 350–360, DOI: 10.1117/12.238732.

[Kha+15] Sayed Hossein Khatoonabadi et al., « How Many Bits Does it Take for a Stimulus to Be Salient? », *in*: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[Koe+14] Kathryn Koehler et al., « What do saliency models predict? », *in*: *Journal of Vision* 14.*3* (2014), DOI: 10.1167/14.3.14.

[Kow11] Eileen Kowler, « Eye movements: The past 25years », *in*: *Vision Research* 51.*13* (2011), pp. 1457–1483, DOI: https://doi.org/10.1016/j.visres.2010.12.014.

[Kro+20] Alexander Kroner et al., « Contextual encoder–decoder network for visual saliency prediction », *in*: *Neural Networks* 129 (2020), pp. 261–270, DOI: 10.1016/j.neunet.2020.05.004.

[Kru12] Elizabeth A. Krupinsky, « On the development of expertise in interpreting medical images », *in*: *Human Vision and Electronic Imaging XVII*, ed. by Bernice E. Rogowitz, Thrasyvoulos N. Pappas, and Huib de Ridder, vol. 8291, International Society for Optics and Photonics, SPIE, 2012, pp. 221–228, DOI: 10.1117/12.916454.

[KTB15] Matthias Kümmerer, Lucas Theis, and Matthias Bethge, « Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet », *in*: *ICLR Workshop*, 2015.

[KU85] C. Koch and S. Ullman, « Shifts in selective visual attention: towards the underlying neural circuitry. », *in*: *Human neurobiology* 4 4 (1985), pp. 219–227, DOI: 10.1007/978-94-009-3833-5_5.

[Küm+] Matthias Kümmerer et al., *MIT/Tübingen Saliency Benchmark*, https://saliency.tuebingen.ai/

[Küm+17] Matthias Kümmerer et al., « Understanding Low- and High-Level Contributions to Fixation Prediction », *in*: *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4789–4798, DOI: 10.1109/ICCV.2017.513.

[KWB15] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge, « Information-theoretic model comparison unifies saliency metrics », *in*: *Proceedings of the National Academy of Sciences* 112.*52* (2015), pp. 16054–16059, DOI: 10.1073/pnas.1510393112.

[KWB16] Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge, « DeepGaze II: Reading fixations from deep features trained on object recognition », *in*: *arXiv preprint arXiv:1610.01563* (2016).

[KWB18] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge, « Saliency Benchmarking Made Easy: Separating Models, Maps and Metrics », *in*: *Computer Vision – ECCV 2018*, 2018, pp. 798–814, DOI: 10.1007/978-3-030-01270-0_47.

[Lai+20] Qiuxia Lai et al., « Video Saliency Prediction Using Spatiotemporal Residual Attentive Networks », *in*: *IEEE Transactions on Image Processing* 29 (2020), pp. 1113–1126, DOI: 10.1109/TIP.2019.2936112.

[Lan19] Michael Land, « Eye movements in man and other animals », *in*: *Vision Research* 162 (2019), pp. 1–7, DOI: 10.1016/j.visres.2019.06.004.

[LB13] Olivier Le Meur and Thierry Baccino, « Methods for comparing scanpaths and saliency maps: strengths and weaknesses », *in*: *Behavior Research Methods* 45.*1* (2013), pp. 251–266, DOI: 10.3758/s13428-012-0226-9.

[Le +05] Olivier Le Meur et al., « A spatio-temporal model of the selective human visual attention », *in*: *IEEE International Conference on Image Processing 2005*, vol. 3, 2005, pp. III–1188, DOI: 10.1109/ICIP.2005.1530610.

[Le +06] Olivier Le Meur et al., « A coherent computational approach to model bottom-up visual attention », *in*: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28.*5* (2006), pp. 802–817, DOI: 10.1109/TPAMI.2006.86.

[Le +17] Olivier Le Meur et al., « Visual Attention Saccadic Models Learn to Emulate Gaze Patterns From Childhood to Adulthood », *in*: *IEEE Transactions on Image Processing* 26.*10* (2017), pp. 4777–4789, DOI: 10.1109/TIP.2017.2722238.

[LE12] Jong-Seok Lee and Touradj Ebrahimi, « Perceptual Video Compression: A Survey », *in*: *IEEE Journal of Selected Topics in Signal Processing* 6.*6* (2012), pp. 684–697, DOI: 10.1109/JSTSP.2012.2215006.

[Lea+17] Mackenzie Leake et al., « Computational Video Editing for Dialogue-Driven Scenes », *in*: *ACM Transactions on Graphics* 36.*4* (2017), pp. 1–14, DOI: 10.1145/3072959.3073653.

[Leb+17] Víctor Leborán et al., « Dynamic Whitening Saliency », *in*: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.*5* (2017), pp. 893–907, DOI: 10.1109/TPAMI.2016.2567391.

[Lév+18] Lucie Lévêque et al., « State of the Art: Eye-Tracking Studies in Medical Imaging », *in*: *IEEE Access* 6 (2018), pp. 37023–37034.

[LH11] Hantao Liu and Ingrid Heynderickx, « Visual Attention in Objective Image Quality Assessment: Based on Eye-Tracking Data », *in*: *IEEE Transactions on Circuits and Systems for Video Technology* 21.*7* (2011), pp. 971–982, DOI: 10.1109/TCSVT.2011.2133770.

[LH18] Nian Liu and Junwei Han, « A Deep Spatial Contextual Long-Term Recurrent Convolutional Network for Saliency Detection », *in*: *IEEE Transactions on Image Processing* 27.*7* (2018), pp. 3264–3274.

[Li+15] Jia Li et al., « A Data-Driven Metric for Comprehensive Evaluation of Saliency Models », *in*: *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 190–198, DOI: 10.1109/ICCV.2015.30.

[Lin+11] Christophe Lino et al., « The Director's Lens: An Intelligent Assistant for Virtual Cinematography », *in*: *Proceedings of the 19th ACM International Conference on Multimedia*, MM '11, New York, NY, USA: Association for Computing Machinery, 2011, pp. 323–332, DOI: 10.1145/2072298.2072341.

[Lin+14] Tsung-Yi Lin et al., « Microsoft COCO: Common Objects in Context », *in*: *Computer Vision – ECCV 2014*, 2014, pp. 740–755.

[Lin+17] Tsung-Yi Lin et al., « Focal loss for dense object detection », *in*: *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[Lin+19] Panagiotis Linardos et al., « Simple vs complex temporal recurrences for video saliency prediction », *in*: *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, 2019, p. 182.

[LLB07] Olivier Le Meur, Patrick Le Callet, and Dominique Barba, « Predicting visual fixations on video based on low-level visual features », *in*: *Vision Research* 47.*19* (2007), pp. 2483–2498, DOI: 10.1016/j.visres.2007.06.015.

[LN14] Patrick Le Callet and Ernst Niebur, « Visual Attention and Applications in Multimedia Technologies », *in*: *Proceedings of the IEEE* 101.*9* (2014), pp. 2058–2067, DOI: 10.1109/JPROC.2013.2265801.

[Los+14] Lester Loschky et al., « What Would Jaws Do? The tyranny of film and the relationship between gaze and higher-level comprehension processes for narrative film. », *in*: *Journal of Vision* 14.*10* (2014), DOI: 10.1167/14.10.761.

[Los+20] Lester C. Loschky et al., « The Scene Perception & Event Comprehension Theory (SPECT) Applied to Visual Narratives », *in*: *Topics in Cognitive Science* 12.*1* (2020), pp. 311–351, DOI: 10.1111/tops.12455.

[LQI11] Zhicheng Li, Shiyin Qin, and Laurent Itti, « Visual attention guided bit allocation in video compression », *in*: *Image and Vision Computing* 29.*1* (2011), pp. 1–14, DOI: 10.1016/j.imavis.2010.07.001.

[Man+07] Matei Mancas et al., « Computational Attention for Event Detection », *in*: *ICVS Workshop on Computational Attention & Applications*, 2007, DOI: 10.2390/biecoll-icvs2007-154.

[Mar+08] Sophie Marat et al., « Spatio-temporal saliency model to predict eye movements in video free viewing », *in*: *2008 16th European Signal Processing Conference* (2008), pp. 1–5.

[MBR11] Olivier Le Meur, Thierry Baccino, and Aline Roumy, « Prediction of the Inter-Observer Visual Congruency (IOVC) and Application to Image Ranking », *in*: *Proceedings of the 19th ACM International Conference on Multimedia*, 2011, pp. 373–382, DOI: 10.1145/2072298.2072347.

[MC19] Kyle Min and Jason J. Corso, « TASED-Net: Temporally-Aggregating Spatial Encoder-Decoder Network for Video Saliency Detection », *in*: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 2394–2403, DOI: 10.1109/ICCV.2019.00248.

[MDN96] Anthony J. Maeder, Joachim Diederich, and Ernst Niebur, « Limiting human perception for image sequences », *in*: *Human Vision and Electronic Imaging*, vol. 2657, SPIE, 1996, pp. 330–337, DOI: 10.1117/12.238729.

[MEP09] Debarshi Mustafia, Andreas H. Engela, and Krzysztof Palczewski, « Structure of Cone Photoreceptors », *in*: *Progress in Retinal and Eye Research* 28 (2009), pp. 289–302, DOI: 10.1016/j.preteyeres.2009.05.003.

[Mit+11] Parag K. Mital et al., « Clustering of Gaze During Dynamic Scene Viewing is Predicted by Motion », *in*: *Cognitive Computation* 3.*1* (2011), pp. 5–24, DOI: 10.1007/s12559-010-9074-z.

[ML13] Matei Mancas and Olivier Le Meur, « Memorability of natural scenes: The role of attention », *in*: *2013 IEEE International Conference on Image Processing*, 2013, pp. 196–200, DOI: 10.1109/ICIP.2013.6738041.

[ML16] Matei Mancas and Olivier Le Meur, « Applications of Saliency Models », *in*: *From Human Attention to Computational Attention: A Multidisciplinary Approach*, ed. by Matei Mancas et al., Springer New York, 2016, pp. 331–377, DOI: 10.1007/978-1-4939-3435-5_18.

[MM14] Jorge Otero-Millan Stephen L. Macknik and Susana Martinez-Conde, « Fixational eye movements and binocular vision », *in*: *Frontiers in integrative neuroscience* 8.*52* (2014), DOI: 10.3389/fnint.2014.00052.

[MOM13] S. Martinez-Conde, J. Otero-Millan, and S. Macknik, « The impact of microsaccades on vision: towards a unified theory of saccadic function », *in*: *Nature Reviews Neuroscience* 14 (2013), pp. 83–96, DOI: 10.1038/nrn3405.

[MR02] Päivi Majaranta and Kari-Jouko Räihä, « Twenty Years of Eye Typing: Systems and Design Issues », *in*: *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications*, ETRA '02, 2002, pp. 15–22, DOI: 10.1145/507072.507076.

[MRW97] S. K. Mannan, K. H. Ruddock, and D. S. Wooding, « Fixation sequences made during visual examination of briefly presented 2D images », *in*: *Spatial vision* 11.*2* (1997), pp. 157–178, DOI: 10.1163/156856897x00177.

[MS15] Stefan Mathe and Cristian Sminchisescu, « Actions in the Eye: Dynamic Gaze Datasets and Learnt Saliency Models for Visual Recognition », *in*: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.*7* (2015), pp. 1408–1424, DOI: 10.1109/TPAMI.2014.2366154.

[Mur+11] Naila Murray et al., « Saliency estimation using a non-parametric low-level vision model », *in*: *CVPR 2011*, 2011, pp. 433–440, DOI: 10.1109/CVPR.2011.5995506.

[MV10] Vijay Mahadevan and Nuno Vasconcelos, « Spatiotemporal Saliency in Dynamic Scenes », *in*: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.*1* (2010), pp. 171–177, DOI: 10.1109/TPAMI.2009.112.

[Nin+07] Alexandre Ninassi et al., « Does where you Gaze on an Image Affect your Perception of Quality? Applying Visual Attention to Image Quality Metric », *in*: *2007 IEEE International Conference on Image Processing*, vol. 2, 2007, pp. 169–172, DOI: 10.1109/ICIP.2007.4379119.

[Nin+09] Alexandre Ninassi et al., « Considering Temporal Variations of Spatial Visual Distortions in Video Quality Assessment », *in*: *IEEE Journal of Selected Topics in Signal Processing* 3.*2* (2009), pp. 253–265, DOI: 10.1109/JSTSP.2009.2014806.

[NM11] Ken Nakayama and Paolo Martini, « Situating visual search », *in*: *Vision Research* 51.*13* (2011), pp. 1526–1537, DOI: https://doi.org/10.1016/j.visres.2010.09.003.

[NZY17] Tam V. Nguyen, Qi Zhao, and Shuicheng Yan, « Attentive Systems: A Survey », *in*: *International Journal of Computer Vision* 126 (2017), pp. 86–110.

[Ote+08] Jorge Otero-Millan et al., « Saccades and microsaccades during visual fixation, exploration, and search: foundations for a common saccadic generator », *in*: *Journal of vision* 8.*14* (2008), pp. 1–18, DOI: 10.1167/8.14.21.

[PAF01] Dale Purves, George J. Augustine, and David Fitzpatrick, *Neuroscience, 2nd. edition*, Sinauer Associates, 2001, ISBN: 0-87893-742-0.

[Pan+16] Junting Pan et al., « Shallow and Deep Convolutional Networks for Saliency Prediction », *in*: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 598–606, DOI: 10.1109/CVPR.2016.71.

[Pan+17] Junting Pan et al., « Salgan: Visual saliency prediction with generative adversarial networks », *in*: *arXiv preprint arXiv:1701.01081* (2017).

[Pet+05] Robert J Peters et al., « Components of bottom-up gaze allocation in natural images », *in*: *Vision research* 45.*18* (2005), pp. 2397–2416.

[PI08] Robert J. Peters and Laurent Itti, « Applying Computational Tools to Predict Gaze Direction in Interactive Visual Environments », *in*: *ACM Transactions on Applied Perception* 5.*2* (2008), DOI: 10.1145/1279920.1279923.

[PLN02] Derrick J. Parkhurst, K. Law, and E. Niebur, « Modeling the role of salience in the allocation of overt visual attention », *in*: *Vision Research* 42 (2002), pp. 107–123, DOI: 10.1016/S0042-6989(01)00250-4.

[PN04] Derrick J. Parkhurst and Ernst Niebur, « Texture contrast attracts overt visual attention in natural scenes », *in*: *The European journal of neuroscience* 19.*3* (2004), pp. 783–789, DOI: 10.1111/j.0953-816x.2003.03183.x.

[Pro14] Study Forrest Project, *Study Forrest*, https://www.studyforrest.org/, 2014.

[PW08] Ofir Pele and Michael Werman, « A Linear Time Histogram Metric for Improved SIFT Matching », *in*: *Computer Vision – ECCV 2008*, Springer Berlin Heidelberg, 2008, pp. 495–508, DOI: 10.1007/978-3-540-88690-7_37.

[Qin+14] Chanchan Qin et al., « Integration of the saliency-based seed extraction and random walks for image segmentation », *in*: *Neurocomputing* 129 (2014), pp. 378–391, DOI: 10.1016/j.neucom.2013.09.021.

[R T+17] Hamed R. Tavakoli et al., « Exploiting Inter-image Similarity and Ensemble of Extreme Learners for Fixation Prediction Using Deep Features », *in*: *Neurocomput.* 244.*C* (June 2017), pp. 10–18, ISSN: 0925-2312.

[Rao+20a] Anyi Rao et al., « A Local-to-Global Approach to Multi-Modal Movie Scene Segmentation », *in*: *CVPR 2020* (2020), pp. 10143–10152, DOI: 10.1109/CVPR42600.2020.01016.

[Rao+20b] Anyi Rao et al., « A Unified Framework for Shot Type Classification Based on Subject Centric Lens », *in*: *ECCV 2020*, 2020, pp. 17–34, DOI: 10.1007/978-3-030-58621-8_2.

[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, 2016, pp. 155–162, DOI: 10.1145/2857491.2857495.

[RCB04] Umesh Rajashekar, Lawrence K. Cormack, and Alan C. Bovik, « Point-of-gaze analysis reveals visual search strategies », *in*: *Human Vision and Electronic Imaging IX*, vol. 5292, SPIE, 2004, pp. 296–306, DOI: 10.1117/12.537118.

[RCY09] Keith Rayner, Monica S. Castelhano, and Jinmian Yang, « Eye movements when looking at unusual/weird scenes: Are there cultural differences? », *in*: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35.*1* (2009), pp. 254–259, DOI: 10.1037/a0013508.

[Red+16] Joseph Redmon et al., *You Only Look Once: Unified, Real-Time Object Detection*, 2016, arXiv: 1506.02640 [cs.CV].

[Ren00] Ronald A. Rensink, « The Dynamic Representation of Scenes », *in*: *Visual Cognition* 7.*1-3* (2000), pp. 17–42, DOI: 10.1080/135062800394667.

[RGB13] Remi Ronfard, Vineet Gandhi, and Laurent Boiron, « The Prose Storyboard Language : A Tool for Annotating and Directing Movies », *in*: *AAAI Workshop on Intelligent Cinematography and Editing* (2013).

[Ric+13] Nicolas Riche et al., « Saliency and Human Fixations: State-of-the-Art and Study of Comparison Metrics », *in*: *The IEEE International Conference on Computer Vision (ICCV)*, 2013.

[RJY92] Roger W. Remington, James C. Johnston, and Steven Yantis, « Involuntary attentional capture by abrupt onsets », *in*: *Perception & Psychophysics* 51 (1992), pp. 279–290, DOI: 10.3758/BF03212254.

[Rol09] Martin Rolfs, « Microsaccades: Small steps on a long way », *in*: *Vision Research* 49.*20* (2009), pp. 2415–2441, ISSN: 0042-6989, DOI: https://doi.org/10.1016/j.visres.2009.08.010.

[RPH14] Anis Rahman, Denis Pellerin, and Dominique Houzet, « Influence of number, location and size of faces on gaze in video », *in*: *Journal of Eye Movement Research* 7.*2* (2014), pp. 891–901, DOI: 10.16910/jemr.7.2.5.

[RSS05] Zeeshan Rasheed, Yaser Sheikh, and Mubarak Shah, « On the Use of Computable Features for Film Classification », *in*: *IEEE Trans. Cir. and Sys. for Video Technol.* 15.*1* (2005), pp. 52–64, DOI: 10.5555/2322561.2323678.

[Rub+11] Ethan Rublee et al., « ORB: An efficient alternative to SIFT or SURF », *in*: *2011 International Conference on Computer Vision*, 2011, pp. 2564–2571, DOI: 10.1109/ICCV.2011.6126544.

[Rud+13] Dmitry Rudoy et al., « Learning Video Saliency from Human Gaze Using Candidate Selection », *in*: *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1147–1154, DOI: 10.1109/CVPR.2013.152.

[RV15] Michele Rucci and Jonathan D. Victor, « The Unsteady Eye: an Information Processing Stage, not a Bug », *in*: *Trends in neurosciences* 38.*4* (2015), pp. 195–206, DOI: 10.3389/fncom.2012.00089.

[RZ99] P. Reinagel and A. M. Zador, « Natural scene statistics at the centre of gaze », *in*: *Network : Computation in Neural Systems* 10.*4* (1999), pp. 341–350, DOI: 10.1088/0954-898X_10_4_304.

[Saw+08] Yasuhito Sawahata et al., « Determining comprehension and quality of TV programs using eye-gaze tracking », *in*: *Pattern Recognition* 41.*5* (2008), pp. 1610–1626, DOI: 10.1016/j.patcog.2007.10.010.

[SB10] Fred Stentiford and Ade Bamidele, « Image recognition using maximal cliques of interest points », *in*: *IEEE International Conference on Image Processing*, 2010, pp. 1121–1124, DOI: 10.1109/ICIP.2010.5649610.

[SD18] Mikhail Startsev and Michael Dorr, « 360-aware Saliency Estimation with Conventional Image Saliency Predictors », *in*: *Signal Processing: Image Communication* 69 (2018), pp. 43–52.

[SD95] Werner X. Schneider and Heiner Deubel, « Visual Attention and Saccadic Eye Movements: Evidence for Obligatory and Selective Spatial Coupling », *in*: *Eye Movement Research*, ed. by John M. Findlay, Robin Walker, and Robert W. Kentridge, vol. 6, Studies in Visual Information Processing, North-Holland, 1995, pp. 317–324, DOI: https://doi.org/10.1016/S0926-907X(05)80027-3.

[SFH86] Martin Shepherd, John M. Findlay, and Robert J. Hockey, « The Relationship between Eye Movements and Spatial Attention », *in*: *The Quarterly Journal of Experimental Psychology Section A* 38.*3* (1986), pp. 475–491, DOI: 10.1080/14640748608401609.

[SH08] Tim J. Smith and John Henderson, « Attentional synchrony in static and dynamic scenes », *in*: *Journal of Vision* 8.*6* (2008), p. 774, DOI: 10.1167/8.6.773.

[SLC12] Tim J. Smith, Daniel Levin, and James E. Cutting, « A Window on Reality: Perceiving Edited Moving Images », *in*: *Current Directions in Psychological Science* 21.*2* (2012), pp. 107–113, DOI: 10.1177/0963721412437407.

[SM13] Tim J. Smith and Parag K. Mital, « Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes », *in*: *Journal of Vision* 13.*8* (2013), pp. 16–16, DOI: 10.1167/13.8.16.

[Smi12] Tim J. Smith, « The Attentional Theory of Cinematic Continuity », *in*: *Projections* 6.*1* (2012), pp. 1–27, DOI: 10.3167/proj.2012.060102.

[Smi13] Tim J. Smith, « Watchin you watch movies : Using eye tracking to inform cognitive film theory », *in*: *Psychocinematics: Exploring cognition at the movies*, 2013, pp. 165–192, DOI: 10.1093/acprof:oso/9780199862139.003.0009.

[Sva+15] M. Svanera et al., « Over-the-shoulder shot detection in art films », *in*: *2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI)* (2015), pp. 1–6.

[SZ14a] Chengyao Shen and Qi Zhao, « Webpage Saliency », *in*: *ECCV*, IEEE, 2014.

[SZ14b] Karen Simonyan and Andrew Zisserman, « Very deep convolutional networks for large-scale image recognition », *in*: *arXiv preprint arXiv:1409.1556* (2014).

[Tan+20] Matthias Tangemann et al., « Measuring the Importance of Temporal Features in Video Saliency », *in*: *2020 European Conference on Computer Vision (ECCV)*, 2020, pp. 667–684, DOI: 10.1007/978-3-030-58604-1_40.

[Tat07] Benjamin W. Tatler, « The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions », *in*: *Journal of Vision* 7.*14* (2007), pp. 4–4, DOI: 10.1167/7.14.4.

[Tav+17] H. R. Tavakoli et al., « Saliency Revisited: Analysis of Mouse Movements Versus Fixations », *in*: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6354–6362.

[Tav+20] Hamed R. Tavakoli et al., *DAVE: A Deep Audio-Visual Embedding for Dynamic Saliency Prediction*, 2020, arXiv: 1905.10693 [cs.CV].

[TB09] Roy Thompson and Christopher Bowen, *Grammar of the Shot (2nd ed.)* Focal Press, 2009, ISBN: 978-0-240-52121-3.

[TBG05] Benjamin W. Tatler, Roland J. Baddeley, and Iain D. Gilchrist, « Visual correlates of fixation selection: effects of scale and time », *in*: *Vision Research* 45.*5* (2005), pp. 643–659, DOI: 10.1016/j.visres.2004.09.017.

[The04] Jan Theeuwes, « Top-down search strategies cannot override attentional capture », *in*: *Psychonomic Bulletin & Review* 11 (2004), pp. 65–70, DOI: 10.3758/BF03206462.

[Tor+06] Antonio Torralba et al., « Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search », *in*: *Psychological Review* 113 (2006), pp. 766–786.

[TZ04] Wallapak Tavanapong and Junyu Zhou, « Shot clustering techniques for story browsing », *in*: *IEEE Transactions on Multimedia* 6.*4* (2004), pp. 517–527, DOI: 10.1109/TMM.2004.830810.

[VA15] Christian Valuch and Ulrich Ansorge, « The influence of color during continuity cuts in edited movies: an eye-tracking study », *in*: *Multimedia Tools and Applications* 74 (2015), pp. 10161–10176, DOI: 10.1007/s11042-015-2806-z.

[VDC14] Eleonora Vig, Michael Dorr, and David Cox, « Large-Scale Optimization of Hierarchical Features for Saliency Prediction in Natural Images », *in*: *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 2798–2805, DOI: 10.1109/CVPR.2014.358.

[Wan+] Wenguan Wang et al., *Revisiting Video Saliency Prediction in the Deep Learning Era*, https://mmcheng.net/videosal/.

[Wan+18] Wenguan Wang et al., « Revisiting Video Saliency: A Large-Scale Benchmark and a New Model », *in*: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4894–4903, DOI: 10.1109/CVPR.2018.00514.

[Wan+19] Wenguan Wang et al., « Revisiting Video Saliency Prediction in the Deep Learning Era », *in*: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019), DOI: 10.1109/TPAMI.2019.2924417.

[WK06] Dirk B. Walther and Christof Koch, « Modeling attention to salient proto-objects », *in*: *Neural networks* 19.*9* (2006), pp. 1395–407, DOI: 10.1016/j.neunet.2006.10.001.

[Wol98] Jeremy M. Wolfe, « What Can 1 Million Trials Tell Us About Visual Search? », *in*: *Psychological Science* 9.*1* (1998), pp. 33–39, DOI: 10.1111/1467-9280.00006.

[Wu+17] Hui-Yin Wu et al., « Analyzing Elements of Style in Annotated Film Clips », *in*: *Eurographics Workshop on Intelligent Cinematography and Editing*, 2017, DOI: 10.2312/wiced.20171068.

[Wu+18] Hui-Yin Wu et al., « Thinking Like a Director: Film Editing Patterns for Virtual Cinematographic Storytelling », *in*: *ACM Transactions on Multimedia Computing, Communications, and Applications* 14.*4* (2018), DOI: 10.1145/3241057.

[Xie+18] Saining Xie et al., « Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification », *in*: *ECCV 2018*, 2018, pp. 318–335, DOI: 10.1007/978-3-030-01267-0_19.

[Xu+14] Juan Xu et al., « Predicting human gaze beyond pixels », *in*: *Journal of Vision* 14.*1* (2014), p. 28, DOI: 10.1167/14.1.28.

[Xu+20] Mai Xu et al., « State-of-the-Art in 360° Video/Image Processing: Perception, Assessment and Compression », *in*: *IEEE Journal of Selected Topics in Signal Processing* 14.*1* (2020), pp. 5–26, DOI: 10.1109/JSTSP.2020.2966864.

[Yar57] Alfred L. Yarbus, « The perception of an image fixed with respect to the retina », *in*: *Biophysics* 2 (1957), pp. 683–690.

[Yar67] Alfred L. Yarbus, *Eye movements and vision*, New York, Plenum Press, 1967, DOI: doi.org/10.1007/978-1-4899-5379-7.

[YJ96] Steven Yantis and John Jonides, « Attentional capture by abrupt onsets: New perceptual objects or visual masking? », *in*: *Journal of Experimental Psychology: Human Perception and Performance* 22.*6* (1996), pp. 1505–1513, DOI: 10.1037/0096-1523.22.6.1505.

[YL09] Stella X. Yu and Dimitri A. Lisin, « Image Compression Based on Visual Saliency at Individual Scales », *in*: *Advances in Visual Computing (ISVC)*, 2009, pp. 157–166, DOI: 10.1007/978-3-642-10331-5_15.

[ZC19] Kao Zhang and Zhenzhong Chen, « Video Saliency Prediction Based on Spatial-Temporal Two-Stream Network », *in*: *IEEE Transactions on Circuits and Systems for Video Technology* 29.*12* (2019), pp. 3544–3557, DOI: 10.1109/TCSVT.2018.2883305.

[ZCL21] Kao Zhang, Zhenzhong Chen, and Shan Liu, « A Spatial-Temporal Recurrent Neural Network for Video Saliency Prediction », *in*: *IEEE Transactions on Image Processing* 30 (2021), pp. 572–587, DOI: 10.1109/TIP.2020.3036749.

[Zha+08] Lingyun Zhang et al., « SUN: A Bayesian framework for saliency using natural statistics », *in*: *Journal of Vision* 8.*7* (2008), p. 32, DOI: 10.1167/8.7.32.

[Zha+20] Ruohan Zhang et al., « Human Gaze Assisted Artificial Intelligence: A Review », *in*: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2020, pp. 4951–4958, DOI: 10.24963/ijcai.2020/689.

[ZK13] Qi Zhao and Christof Koch, « Learning saliency-based visual attention: A review », *in*: *Signal Processing* 93.*6* (2013), pp. 1401–1407.

[ZM20] Guangtao Zhai and Xiongkuo Min, « Perceptual image quality assessment: a survey », *in*: *Science China Information Sciences* 63 (2020), pp. 1–52, DOI: 10.1007/s11432-019-2757-1.

[ZPB07] C. Zach, T. Pock, and H. Bischof, « A Duality Based Approach for Realtime TV-L1 Optical Flow », *in*: *Pattern Recognition*, 2007, pp. 214–223, DOI: 978-3-540-74936-3.

[ZS13] J. Zhang and S. Sclaroff, « Saliency Detection: A Boolean Map Approach », *in*: *2013 IEEE International Conference on Computer Vision*, 2013.

[Zün+13] Fabio Zünd et al., « Content-aware compression using saliency-driven image retargeting », *in*: *2013 IEEE International Conference on Image Processing (ICIP)* (2013), pp. 1845–1849, DOI: 10.1109/ICIP.2013.6738380.

[ZX18] Shiping Zhu and Ziyao Xu, « Spatiotemporal visual saliency guided perceptual high efficiency video coding with neural network », *in*: *Neurocomputing* 275 (2018), pp. 511–522, DOI: https://doi.org/10.1016/j.neucom.2017.08.054.

FILMOGRAPHY

[Bay98] Michael Bay, *Armageddon*, 1998.

[Bur03] Tim Burton, *Big Fish*, 2003.

[Cop72] Francis Ford Coppola, *The Godfather*, 1972.

[Dar94] Frank Darabont, *The Shawshank Redemption*, 1994.

[Eas09] Clint Eastwood, *Invictus*, 2009.

[Emm04] Roland Emmerich, *The Day After Tomorrow*, 2004.

[Emm09] Roland Emmerich, *2012*, 2009.

[Emm13] Roland Emmerich, *White House Down*, 2013.

[Emm94] Roland Emmerich, *Stargate*, 1994.

[Emm96] Roland Emmerich, *Independance Day*, 1996.

[Fin08] David Fincher, *The Curious Case of Benjamin Button*, 2008.

[God60] Jean-Luc Godard, *À bout de souffle*, 1960.

[Jac01] Peter Jackson, *The Lord of the Rings: the Fellowship of the Ring*, 2001.

[Kay98] Tony Kaye, *American History X*, 1998.

[Kub68] Stanley Kubrick, *2001: A Space Odyssey*, 1968.

[Kub71] Stanley Kubrick, *A Clockwork Orange*, 1971.

[Kub75] Stanley Kubrick, *Barry Lyndon*, 1975.

[Kub80] Stanley Kubrick, *The Shining*, 1980.

[Kub87] Stanley Kubrick, *Full Metal Jacket*, 1987.

[Leo66] Sergio Leone, *The Good, the Bad and the Ugly*, 1966.

[Mei05] Fernando Meirelles, *The Constant Gardener*, 2005.

[Nic97] Andrew Niccol, *Gattaca*, 1997.

[Nol10] Christopher Nolan, *Inception*, 2010.

[Ros12] Gary Ross, *The Hunger Games*, 2012.

[Tak08] Yôjirô Takita, *Departures*, 2008.

[Tar94] Quentin Tarantino, *Pulp Fiction*, 1994.

[Tay11] Tate Taylor, *The Help*, 2011.

[Vil13a] Denis Villeneuve, *Enemy*, 2013.

[Vil13b] Denis Villeneuve, *Prisoners*, 2013.

[Vil15] Denis Villeneuve, *Sicario*, 2015.

[Vil16] Denis Villeneuve, *Arrival*, 2016.

[Vil17] Denis Villeneuve, *Blade Runner 2049*, 2017.

[Wel41] Orson Welles, *Citizen Kane*, 1941.

[Zem00] Robert Zemeckis, *Cast Away*, 2000.

[Zem85] Robert Zemeckis, *Back to the Future*, 1985.

[Zem88] Robert Zemeckis, *Who Framed Roger Rabbit*, 1988.

[Zem94] Robert Zemeckis, *Forrest Gump*, 1994.

[Zem97] Robert Zemeckis, *Contact*, 1997.

DOCTORAT BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

Titre : Attention visuelle et cinématographie

Mot clés : Attention visuelle, cinématographie, oculométrie, saillance visuelle, congruence visuelle inter-observateurs.

Résumé : Quand nous regardons un film, nous ne traitons pas toute l'information visuelle émise par l'image tout le temps. À la place, nous dirigeons notre attention sur certaines zones de l'écran que nous considérons comme importantes, que ce soit à cause de leurs propriétés visuelles, ou de leur importance sémantique pour la narration du film. Depuis plus de cent ans, les réalisateurs de films ont appris à jouer avec l'attention visuelle de leur public, en utilisant un ensemble varié d'outils et de techniques. Dans cette thèse, nous nous proposons d'explorer les liens entre ces choix cinématographiques du réalisateur et la perception visuelle qu'en a le public. Bien qu'il existe de nombreux modèles de saillance visuelle, prédisant les zones d'attention visuelle d'observateurs sur des vidéos, nous montrons que les prédictions de ces modèles s'avèrent parfois fausses dans le contexte particulier de stimuli cinématographiques. Nous proposons donc un nouveau modèle de saillance visuelle, incluant des caractéristiques de haut niveau concernant les propriétés cinématographiques de l'extrait de film considéré. Enfin, nous proposons une étude de la congruence visuelle inter-observateurs dans ce contexte, ainsi que deux modèles visant à prédire l'intensité de cette congruence, sur des images et des extraits de films.

Title: A perceptual approach to film editing

Keywords: Visual attention, movie editing, eye-tracking, visual saliency, inter-observer visual congruency.

Abstract: When watching movies, we do not grasp the full image that is displayed at all time. Instead, we focus on several parts of the frame, depending on what we deem relevant, be it for the visual properties of this area or its semantic importance in the narration. With more than a century of cinematographic experience, filmmakers have developed a whole array of tools and techniques to direct the attention of their audience, using cuts, camera motion, staging, and so on. In this work, we propose to explore the links between film editing and the visual perception an audience has of it, using a data-driven approach. While there exists a lot of efficient models predicting where people will look on a video, we found that these models could often be wrong on cinematographic stimuli. We then propose a visual saliency model dedicated to include the high-level information created by the director's editing choices, and we show a significant improvement on cinematic stimuli compared to the state-of-the-art. Finally, we propose two models dedicated to predict the inter-observer visual congruency on both static and dynamic stimuli, with particular care to the case of cinematographic stimuli.