# Visual exploration of web pages : understanding its dynamic for a better modelling

Alexandre Milisavljevic

**Université de Paris**
Ecole doctorale 261 : Cognition,
Comportements, Conduites Humaines
Laboratoire Vision Action Cognition

**Université de Mons**
Faculté Polytechnique de Mons
Laboratoire Information, Signal et
Intelligence Artificielle

THÈSE

En vue d'obtenir le grade de **Docteur en Sciences Cognitives** d'Université de Paris et le grade de **Docteur en Sciences de l'Ingénieur et Technologie** d'Université de Mons

# L'exploration visuelle de pages web : comprendre sa dynamique pour une meilleure modélisation

Visual exploration of web pages: understanding its dynamic for a better modelling

Présentée par
**ALEXANDRE MILISAVLJEVIC**

Sous la direction des Professeurs
**Karine Doré-Mazars & Bernard Gosselin**

Soutenance prévue le 9 Décembre 2020

## Composition du Jury

| | | | |
|---|---|---|---|
| Olivier LE MEUR | Maître de conférence | Université de Rennes | Rapporteur |
| François MAQUESTIAUX | Professeur | Université de Franche-Comté | Rapporteur |
| Véronique MOEYAERT | Professeure | Université de Mons | Examinatrice |
| Thérèse COLLINS | Professeure | Université de Paris | Examinatrice |
| Olivier DEBEIR | Professeur | Université Libre de Bruxelles | Examinateur |
| Alan CHAUVIN | Maître de conférence | Université Grenoble-Alpes | Examinateur |
| Matei MANCAS | Maître de conférence | Université de Mons | Membre Invité |
| Bernard GOSSELIN | Professeur | Université de Mons | Directeur |
| Karine DORÉ-MAZARS | Professeure | Université de Paris | Directrice |

# REMERCIEMENTS

Cette thèse CIFRE est issue de la collaboration entre l'Université de Paris (ex-Paris Descartes), l'Université de Mons et l'entreprise Sublime Skinz.

J'aimerais dans un premier temps remercier les rapporteurs de cette thèse, **Olivier Le Meur** et **François Maquestiaux**, ainsi que les examinateurs(trices) **Véronique Moeyaert**, **Thérèse Collins**, **Alan Chauvin** et **Olivier Debeir** de me faire l'honneur d'évaluer mon travail de thèse. Merci pour votre présence et votre temps.

Je souhaiterais remercier **Sublime Skinz** sans qui cette thèse n'aurait pas pu être possible. Merci à **Marc** d'avoir eu cette brillante idée qui constitue aujourd'hui un chapitre entier de ma vie. Merci à **Romain** de m'avoir permis de présenter mes travaux tout au long de ma thèse. Merci à **Coralie** de m'avoir donné cette opportunité de rejoindre cette fabuleuse famille qu'est Sublime Skinz et d'avoir tout fait pour me permettre de réaliser ce travail dans les meilleures conditions possibles. Merci également à **Damien** et **Fouad** de m'avoir accueilli jeune innocent que j'étais. Le temps passé avec vous à été un des moment fort de ma thèse que je ne suis pas prêt d'oublier. Merci à **Kévin**, mon *crazy man* préféré, tous ces moments passés à réfléchir, échanger, philosopher et à boire ! Ta bonne humeur aussi constante que Pi, m'aura plus d'une fois réchauffé le coeur.

Je tiens à remercier très sincèrement mes deux directeurs de thèse **Karine Doré-Mazars** et **Bernard Gosselin** ainsi que mon co-encadrant **Matei Mancas** qui ont cru en moi et qui m'ont permis de réaliser ma thèse dans les meilleures conditions. Tout d'abord merci **Karine** d'avoir accueilli ce projet fou de thèse cotutelle/cifre d'un inconnu au bataillon avec la même élégance et clairvoyance que les sujets que tu encadres habituellement. Dire que tout est parti d'une bouteille à la mer qui était vouée à disparaitre dans le flot légendaire de tes mails et qui par le coup du sort est passée entre les mailles du filet. Merci d'avoir eu la patience de me former à la recherche, à la rigueur scientifique et surtout à ce domaine passionnant qu'est la psychologie cognitive ! Je souhaiterais également remercier **Matei** de m'avoir fait confiance dès le début de ce projet. En plus de la patience dont tu as fais également preuve pour mon initiation au monde de la saillance, merci de m'avoir accueilli les bras ouverts et permis de m'intégrer au pays des bières. Je remercie également **Bernard** de m'avoir accompagné pendant ces quatres années. Les péripéties n'auront pas manquées et ce jusqu'au bout mais tu as toujours répondu présent. Merci de m'avoir patiemment aiguillé et aidé dans cette

configuration compliquée. Au-delà de tout ce que vous m'avez appris, je vous remercie tous les trois pour vos qualités humaines. J'ai conscience de la chance que j'ai d'avoir pu bénéficier de votre encadrement et j'espère pouvoir continuer à collaborer avec vous dans le futur.

En plus de la recherche, j'ai grandement apprécié enseigner et cela n'aurait pas été possible sans toi **Dorine**. Je te remercie de ta confiance et la gentillesse dont tu as fait preuve envers moi. Merci également à **Alma** de m'avoir permis d'enseigner auprès des Master 2.

J'aimerais remercier tous les membres du VAC de m'avoir accueilli avec autant d'enthousiasme et de bonne humeur au sein du laboratoire. J'ai beaucoup apprécié participer à la vie du laboratoire : les congrès, les pauses cafés, les raclettes de Noël, les zooms apéro ainsi que les différentes activités qu'on a pu partager tous ensemble. Merci à **Céline**, **Anne**, **Alma** et **Hélène** de m'avoir laissé les 7 derniers kilomètres de l'Ekiden 2018, je n'oublierai pas cette arrivée place du Trocadéro sous vos encouragements ! **Anne**, un grand merci pour t'être si gentiment proposée de relire ma thèse, tes corrections d'une grande qualité m'ont beaucoup aidé. **Christelle**, je te remercie pour nos échanges passionnants, que ce soit nos (tentatives) conversations matinales en espagnol, la narration de nos voyages ainsi que tes références illimitées en terme de jeux. Merci **Céline** pour nos échanges Matlab ainsi que pour ta disponibilité et réactivité concernant les ITER/APR. **Patrice**, merci pour tes anecdotes toujours ludiques et passionnantes qui animent avec beaucoup d'entrain les déjeuner au RU. Merci également pour tes conseils avisés et pour ce fameux poster où tu as apposé mon nom. Merci à **Karima** pour tes questions toujours à la pointe sur l'éthique. Nos échanges constructifs m'ont beaucoup aidé dans mon apprentissage du milieu de la recherche. Je remercie également **Priscilla** et **Marine** pour leur bonne humeur au quotidien. Merci de leur aide substantielle dans les déboires administratives (qui jusqu'au bout auront été traumatisantes) ainsi que pour leur patience et leur investissement dans la vie du laboratoire.

Merci aux membres de l'école doctorale 261 de l'Université de Paris de fournir un bon environnement afin de préparer et soutenir sa thèse dans les meilleures conditions. Un grand merci à **Franck**, **Coralie**, **Emma** et **Karolina** de m'avoir aidé à me sortir des méandres administratifs de la cotutelle à d'innombrables reprises.

Merci à tous mes voisins de bureaux sans qui ces années auraient été bien fades. Merci **Martin** pour ta gentillesse et ta bonne humeur. Tes passages au bureau et tes déconvenues avec les transports en communs parisiens me manquent. **Jérôme** merci de m'avoir supporté toutes ces années, tu es d'une patience rare. Merci également pour ton soutien, tes conseils avisés et ta relecture attentive de mon manuscrit. **Alexandra** merci

pour ces discussions passionnantes sur les voyages et le sport mais aussi ces moments de rire qui ont rythmés nos années de doctorants. Un grand merci **Gabriella** pour la relecture de mes travaux, tes corrections ont été précieuses. Je te souhaite bon courage pour la suite de ta thèse ! Enfin, un grand merci à tous les membres du laboratoire VAC : **Louisa**, **Nicole**, **Philippe**, **Agnès**, **Henri**, **Nadia**, **Alain**, **Katarina** et **Agathe**, de votre gentillesse et votre bonne humeur.

Un grand merci aux étudiants qui ont travaillé de près ou de loin sur ce projet de thèse et plus particulièrement **Andria**, **Thomas** et **Fabrice**.

Merci à **Thierry Dutoit** de m'avoir accueilli au sein de l'institut Numediart. Merci à **Ambroise**, **Michael**, **Kevin** et tout l'institut Numediart de m'avoir fait une place dans ce pays qui bien que voisin m'était inconnu. Votre gentillesse m'a permis de me sentir comme chez moi dans ce fabuleux pays qu'est la Belgique.

Merci **Josiane** de m'avoir si gentiment accueilli chez toi pendant la grande partie de ma période d'écriture. Tu as généreusement sacrifié tes épisodes des feux de l'amour et je t'en remercie.

Je tiens également à remercier mes amis **Tom**, **Yann**, **Yoann** et **Chris** de m'avoir soutenu tout au long de cette aventure. Merci Tom pour tes idées "think outside the box" qui m'ont beaucoup inspiré durant toute ma thèse. Merci Yann de ta sagesse que tu n'as jamais hésité à partager avec moi ainsi que tes conseils toujours aussi avisés. Merci Yoann pour tes triggers toujours aussi... rafraîchissantes qui m'ont remonté le moral plus d'une fois. Merci Chris de ton oreille attentive et tes conseils qui m'ont permis de surmonter certaines épreuves que je pensais insurmontables. Merci également à Flora et toi de votre bienveillance et votre inestimable aide.

Merci à mes **parents**, mes **frères** et toute ma famille pour m'avoir particulièrement supporté pendant ces 4 ans. Votre présence et votre soutien m'aura été plus que jamais nécessaire pour mener ce projet à terme.

Pour finir, je tiens à remercier ma si chère **Angélique**, ces dernières lignes ne sont qu'une maigre expression de la gratitude et de la reconnaissance que j'ai pour toi. Toi qui m'a apporté un soutien indéfectible tout au long de ces deux dernières années. Toi qui m'a permis de me transcender dans les moments où je pensais avoir atteint mes limites. Toi qui a toujours cru en moi alors que le doute me submergeait. C'est donc affectueusement que je te remercie de m'avoir soutenu dans cette aventure qui, j'en suis sûr, n'est que la première d'une longue liste.

# Abstract

Since its creation, the web has contributed to the transformation of our society through the massive diffusion of knowledge. Understanding how we access and select information on this new medium has become crucial. Contrary to static images, a web page needs to be scrolled in order to fully explore its content. Thus, the understanding of ocular behaviour on web pages requires analyses taking into account the dynamic from the visual exploration and the scroll. Therefore, we set up an experimental study to demonstrate the importance of the understanding of these dynamics in the prediction of eye movements. In this study, we asked 150 participants to browse 18 web pages of variable lengths, and perform either a free viewing task or a visual search task. The aim of the first axis of this work was to better describe the dynamic of eye movements on web pages through the use of a composite indicator. Recent research has shown a link between the fixation duration, the saccade amplitude and the two main visual pathways involved in vision. Short fixations followed by long saccades (ambient visual mode) would be related to the dorsal stream involved in objects localisation and visually guided actions, while long fixations followed by short saccades (focal visual mode) would be related to the ventral stream involved in object recognition. Thus, the ambient mode would dominate the exploration at the beginning, and the focal mode at the end. We used the definition of visual modes to study to which extent it could explain eye movement temporal evolution. To this end, we investigated existing ratios describing ambient and focal visual modes. We showed that these ratios only evaluated visual modes' intensity rather than their dynamics. Hence, we proposed new measures describing the number of switches between modes and the average time spent in each mode. The second axis of this thesis was to investigate eye movement behaviours on web pages through their relationships with mouse cursor movements and scrolling. We specifically focused on the relationships between their parameters, and the influence of scroll on eye movements. We provided a detailed statistical description of eye movements on web pages along with the mouse movements and scroll statistics. Moreover, we studied how eye movements were influenced before and during the scroll through the study of their parameters, including eyes position, scroll amplitude and scroll speed. Based on these findings, we introduced a more precise definition and segmentation of scrolling events. The third axis goal was to integrate findings from previous axes in web pages scanpath modelling in order to improve prediction accuracy. Existing scanpath models rarely address web pages, but when they do, they consider web pages as static screenshots without the need to scroll. To tackle this problematic, we proposed the first saccadic model including scrolling. Furthermore, scanpath modelling usually include some oculomotor biases, which are mostly considered as stable through visual exploration. In our approach we addressed these biases through their evolution over time. Thus, this work highlights the importance of dynamics in the prediction of eye movements when exploring web pages.

# Résumé

La ruée vers l'information engendrée par l'invention du web a totalement transformé notre société. Il est aujourd'hui devenu crucial de comprendre comment nous sélectionnons et accédons à certaines informations. Contrairement aux images statiques, une page web est dynamique en raison de la nécessité de faire défiler le contenu pour le voir en intégralité. Par conséquent, la compréhension du comportement oculaire sur les pages web nécessite des analyses qui prennent en compte la dynamique de l'exploration visuelle et celle du défilement. Afin de démontrer l'importance de la dynamique dans la compréhension et la prédiction des mouvements oculaires, nous avons mis en place une étude comportementale. Lors de celle-ci, nous avons demandé à 150 participants de parcourir 18 pages webs de longueur variable soit en exploration libre, soit en recherchant une cible. Le premier axe de ce travail visait à contribuer à une meilleure compréhension de la dynamique des mouvements oculaires sur les pages web grâce à l'utilisation d'un indicateur unique. Des recherches récentes ont permis de trouver un lien entre la durée de fixation, l'amplitude de saccade et les deux principales voies du traitement visuel. Des fixations courtes suivies de longues saccades (mode visuel ambient) seraient ainsi liées à la voie dorsale impliquée dans la localisation d'objets et les actions guidées visuellement, alors que des fixations longues suivies de saccades courtes (mode visuel focal) seraient liées à la voie ventrale impliquée dans la reconnaissance d'objets. Ainsi, le mode ambient serait plus présent au début de l'exploration et le mode focal à la fin. Nous avons utilisé cette définition pour étudier dans quelle mesure celle-ci pouvait décrire la dynamique des mouvements oculaires. À cette fin, nous avons étudié les ratios existants décrivant les modes visuels ambient et focal. Nous avons montré que ces ratios n'évaluaient que leur intensité et non leur dynamique. Nous proposons de nouvelles mesures décrivant le nombre de changements de modes et le temps moyen passé dans un mode. Le deuxième axe de cette thèse consistait à étudier le comportement oculaire lors de l'exploration de pages web en tenant compte de la relation entre les mouvements des yeux, le déplacement du pointeur et le défilement de la page. Nous nous sommes spécifiquement concentrés sur la relation entre leurs paramètres et l'influence du défilement sur les mouvements des yeux. Ainsi, nous avons proposé une description statistique détaillée des mouvements des yeux sur les pages web ainsi que des mouvements de la souris et du défilement. De plus, nous avons étudié comment les mouvements des yeux étaient influencés par le défilement. Nous avons montré différents comportements avant et pendant le défilement, comprenant la position des yeux, l'amplitude du défilement et la vitesse de défilement. Ces travaux nous ont permis de proposer une définition et une segmentation plus précise des événements de défilement. L'objectif du troisième axe était d'inclure les résultats des axes précédents dans la modélisation du chemin oculaire lors de l'exploration de pages web afin d'améliorer la précision des prédictions. Les modèles existants utilisent généralement des images de scènes naturelles et moins des stimuli tels que les pages web. Dans ce dernier cas, ils utilisent la plupart du temps des captures d'écran plutôt que des pages webs dont l'exploration requiert un défilement du contenu. Pour résoudre cette problématique, nous avons proposé le premier modèle saccadique incluant le défilement. De plus, la modélisation des mouvements oculaires inclut également la prise en compte de certains

biais la plupart du temps considérés comme stables tout au long de l'exploration visuelle. Sur la base de nos résultats, notre approche prend en compte ces différents biais et leur évolution au cours du temps. Ainsi, l'ensemble de ce travail permet de mettre en avant l'importance de la dynamique dans la prédiction des mouvements oculaires lors de l'exploration de pages web.

Cette thèse est le fruit d'une cotutelle de doctorat associant l'Université de Mons en Belgique, et l'Université de Paris. La thèse étant rédigée en langue anglaise, un résumé substantiel de la thèse est présenté ci-dessous.

# Introduction

L'attention sélective permet de sélectionner les informations les plus pertinentes, en inhibant les éléments distracteurs tout en rehaussant la cible à viser. L'information visuelle qui parvient à notre rétine est extrêmement riche. Après une succession d'étapes, l'information visuelle est acheminée dans le cortex visuel primaire. Par la suite elle est transmise à différentes aires corticales afin d'extraire les propriétés de l'objet ou bien d'effectuer des traitements plus complexes tels que l'identification (voie ventrale) ou la localisation spatiale (voie dorsale). Ces traitements sont essentiellement possibles lorsque l'objet se situe sur la partie de l'oeil ayant la plus haute acuité visuelle : la fovéa. Pour ce faire, l'oeil effectue un mouvement appelé saccades permettant d'amener l'objet d'interêt sur la fovéa. Lorsque nous réalisons une saccade vers un objet, notre attention se déplace également sur ce dernier : c'est l'attention overt. Ces saccades peuvent être déclenchée de manière réactive ou volontaires et se distinguent dans leurs paramètres de latence, de direction et d'amplitude. Après l'exécution d'une saccade, une période de stabilisation s'ensuit, appelée fixation. La fixation se caractérise par sa durée et est entrecoupée de mouvements involontaires tels que les microsaccades, les dérives et les tremblements. Les différents paramètres de la saccade et de la fixation évoqués ci-dessus évoluent à travers le temps ce qui rend complexe leur étude. Ainsi, une approche dynamique semble nécessaire afin de mieux appréhender l'exploration visuelle. C'est dans ce but qu'un ensemble d'études a mis en lien les voies ventrales et dorsales

avec les mouvements des yeux en y associant deux modes visuels : ambiant et focal. Afin d'analyser ces modes lors de l'exploration nous avons utilisé deux ratios, le premier est présenté dans le **Poster 1** et le second dans le **Poster 2**, l'**Article 1** et l'**Article 3**.

Comme l'attention visuelle, l'exploration oculaire peut-être influencée par différents types de facteurs que l'on peut classer en trois catégories. Les biais oculomoteurs constituent le premier type de facteurs influençant cette exploration. Ces derniers sont une combinaison de contraintes physiologiques et de comportements appris au cours de l'évolution. Le second facteur, dit ascendant, regroupe les éléments propre au stimulus. Le dernier et troisième facteur, appelé descendant, désigne l'influence des processus cognitifs associés au contexte. Ces facteurs descendants peuvent également dépendre du type de stimulus présenté. Par exemple, les pages webs suivent une organisation similaire à travers le web. Cette organisation présente sur la majorité des sites webs a mené les utilisateurs à développer des stratégies afin d'optimiser leur exploration visuelle. On parle alors de biais spécifiques au web. Toutefois, les stimuli web possèdent une particularité qui les distingue d'une image classique : ils sont interactifs au travers de la souris. Cette interactivité a mené les recherches à tenter de mieux comprendre le comportement visuel lors d'évènements tels que le défilement de la page ou le mouvement de la souris. Ces aspects sont abordés dans les **Articles 2 et 3** ainsi que dans le **Poster 3**. En ce sens, l'étude du mouvement de la souris s'est rapidement concentrée sur la coordination entre l'oeil et la souris afin d'essayer de prédire les mouvements des yeux en fonction de la position du curseur. L'étude de cette coordination sera abordée de manière détaillée dans l'**Article 2**.

Les yeux sont au centre de la perception visuelle, et à travers l'enchainement des fixations et des saccades, ils permettent d'acquérir des informations visuelle de manière efficiente. C'est pourquoi de nombreux modèles ont tenté de reproduire cette mécanique. Deux types de modèles ont alors émergés: les modèles de saillances et les modèles saccadiques. Les premiers ont pour objectif de fournir un résumé des endroits visité par un participant lors de l'exploration d'une scène visuelle. On catégorise souvent ces modèles selon les types de facteurs qu'ils prennent en compte. D'un côté, nous

avons les modèles dits ascendants (ou bottom-up), qui ont donné lieux à une littérature fleurissante, prédisent les régions visitées en se basant sur les caractéristiques du stimulus. De l'autre côté, nous avons les modèles dits descendants (ou top-down). La majorité du temps, ces modèles partent d'un modèle ascendant qu'ils "augmentent" avec des paramètres liés à la tâche, les visages présents, les objets, etc. Bien que moins étendu que la littérature des modèles ascendants, les modèles descendants sont aujourd'hui au coeur de l'intérêt des chercheurs. Quant aux seconds types de modèles, leur but est, non plus de fournir un résumé des régions fixées, mais la séquence détaillée et ordonnée des zones vues. Ainsi, les paramètres oculomoteurs ont une bien plus grande importance car ils peuvent influencer le choix de la fixation suivante. C'est pourquoi ce type de modèle, en plus de modéliser le mécanisme de sélection de la prochaine fixation, modélise également ces biais. Les biais les plus souvent modélisés étant l'amplitude de saccade, l'orientation des saccades, la fovéa, et l'inhibition de retour. Toutefois, ces biais sont la plus part du temps modélisés de la même manière. A la fois les modèles de saillance et les modèles saccadiques ont été développés pour prédire les mouvements des yeux lors de l'exploration de scènes naturelles. Lorsque ne nous intéressons aux pages webs, les modèles se font beaucoup plus rares. Cependant, la modélisation sur page web a donné lieu à un autre type d'approche : la modélisation de la position de l'oeil en fonction de la position de la souris. Plusieurs modèles ont été proposés suivant cette approche mais ils se sont focalisés sur les pages de recherches de moteur de recherche plutôt que sur les pages webs classiques. A la lumière de tous ces éléments, nous proposons dans le **Chapitre 8** un modèle saccadique sur des pages webs dynamiques. De plus, nous appliquons les analyses temporelles des paramètres oculaires réalisées dans l'**Article 3** afin de proposer une modélisation des biais oculomoteurs basée sur leur évolution dans le temps.

# Objectifs

Les modèles saccadiques ont pour but de reproduire la dynamique des mouvements des yeux ainsi que les facteurs l'influençant. La motivation principale derrière ce travail de thèse est la démonstration de comment l'évolution temporelle des paramètres des mouvements des yeux ainsi que la dynamique induite par le scroll sont primordiaux dans la modélisation saccadique sur pages web.

Le premier axe de cette thèse se concentre sur comment résumé la dynamique des mouvements des yeux en un seul indicateur. Basé sur la définition des modes de traitements visuels ambient et focal (Pannasch et al., 2008; Unema et al., 2005; Velichkovsky et al., 2005), nous avons évalué la pertinance des indicateurs existants (Dehais et al., 2015; Goldberg & Kotval, 1999; Krejtz et al., 2016) lors de l'exploration d'images naturelles et de pages web.

De nombreuses recherches se sont concentrées sur la coordination entre les yeux et le curseur de la souris, mais ces études ont porté principalement sur les pages de résultats des moteurs de recherche (SERP) et ont négligé le défilement (Guo & Agichtein, 2010; Huang et al., 2012; Rodden et al., 2008). Pour ces raisons, le second axe de cette thèse vise à analyser le comportement des mouvements oculaires lors de la navigation sur des pages web de notre quotidien. Pour ce faire, nous avons mis en place deux études expérimentales et nous avons examiné la relation entre les yeux, le curseur de la souris et le défilement.

Les deux premiers axes ont permis de mieux comprendre la dynamique des mouvements oculaires et la relation entre les yeux, le curseur de la souris et le défilement. Nous avons utilisé certaines de nos découvertes dans le troisième axe afin d'améliorer la modélisation saccadique. Ainsi, nous avons proposé un modèle saccadique incluant un mécanisme de défilement et modélisant l'évolution temporelle des paramètres des mouvements oculaires. Ce modèle a été évalué sur la base de données de haute qualité issues des études expérimentales présentées dans le chapitre suivant.

# Contributions expérimentales

## Ambient et focal comme indicateurs de la dynamique des mouvements oculaires

Le premier axe de cette thèse consiste à déterminer comment le traitement visuel ambient et focal peut décrire la dynamique des mouvements oculaires. Ces modes visuels trouvent leur origine dans les deux voies (ventrale et dorsale) empruntées par les informations visuelles dans le cerveau (Goodale & Milner, 1992; Ungerleider & Mishkin, 1982). Comme décrit dans le Chapitre 1, le flux ventral va du cortex occipital au cortex temporal et transporte des informations sur les caractéristiques des objets ("*Quoi*"), tandis que le flux dorsal va du cortex occipital au cortex pariétal et transporte des informations sur l'emplacement des objets ("*Où*"). Comme l'explique Velichkovsky et al. (2005), ces voies visuelles peuvent être directement observées à travers les mouvements oculaires à l'aide de fixations et de saccades. Une fixation courte suivie d'une saccade de grande amplitude suggère un mode ambient (voie dorsale), tandis qu'une fixation longue suivie d'une saccade de petite amplitude suggère un mode focal (voie ventrale) (Pannasch et al., 2008; Unema et al., 2005; Velichkovsky et al., 2005). Ces deux modes ont été résumés par la littérature à travers deux ratios principaux (Dehais et al., 2015; Krejtz et al., 2016).

Nous avons étudié dans le **Poster 1** comment le ratio proposé par Dehais et al. (2015) pouvait être utilisé pour discriminer les tâches lors de la navigation sur les pages web. Ce rapport a été créé à l'origine par Goldberg and Kotval (1999) pour caractériser l'exploration visuelle dans les interfaces logicielles, et Dehais et al. (2015) l'a modifié pour évaluer les modes de traitement visuel dans le contexte d'un cockpit d'avion. Nous avons appliqué cette version modifiée sur des pages web pour étudier les modes visuels. Nous avons également introduit l'utilisation des seuils de durée de fixation et d'amplitude de saccade de Unema et al. (2005) et Pannasch et al. (2008) pour différencier les fixations/saccades courtes et les fixations/saccades longues. Nous avons montré que le mode ambient (ou mode d'exploration) était plus intense pendant la

tâche de visualisation libre que pendant la tâche de recherche visuelle. De plus, nous avons trouvé des résultats prometteurs montrant qu'un clic pouvait être précédé d'un mode focal (ou mode d'exploitation). Cependant, ce rapport ne différenciait pas le temps passé à explorer (saccades) du temps passé à exploiter rapidement l'information (courtes fixations). Nous avons donc étudié le coefficient K, un ratio proposé par Krejtz et al. (2016) sur les images naturelles. Contrairement au ratio précédent, le coefficient K a été spécifiquement conçu pour décrire les modes ambiant et focal. Dans **Article 1**, nous avons utilisé le coefficient K pour comparer les mouvements des yeux entre une tâche de visualisation libre et une tâche de recherche visuelle. Nous avons constaté que le passage entre les deux modes se faisait à une fréquence élevée, de sorte que la valeur moyenne du coefficient était proche de zéro. Nous avons montré que, même si des différences globales n'apparaissaient pas, la dynamique des modes visuels entre les tâches mettait en évidence des différences dans le temps. De plus, nous avons réussi à différencier globalement les tâches en introduisant de nouvelles variables liées au coefficient K. Enfin, nous avons reproduit ces résultats sur des pages web dans le **Poster 2**. Nous avons montré que le coefficient K ne pouvait pas différencier les tâches globalement, mais en utilisant ces variables liées à K, nous avons pu mieux les différencier globalement et dans le temps.

## Comportement des mouvements oculaires sur les pages web

Le second axe de ce travail avait pour but d'étudier le comportement des mouvements oculaires sur les pages web à travers la relation entre les yeux, le mouvement du pointeur de la souris et le défilement. Les pages web doivent être étudiées différemment des images en raison des interactions possibles avec la page web. Ces interactions peuvent prendre plusieurs formes, notamment les clics, le défilement et le glisser-déposer. Alors que les clics et les glisser-déposer sont un moyen de modifier ou de mettre à jour directement le contenu, le défilement est davantage une découverte du contenu.

La plupart des études sur le comportement de défilement se concentrent sur le comportement des mouvements oculaires lors du défilement et de la lecture de documents

textuels (voir Dyson (2004) pour une revue), mais peu d'études ont examiné ce comportement sur les pages web. La relation entre les yeux et le défilement est encore moins étudiée. Pourtant, le défilement est utilisé par des milliards de personnes lorsqu'elles accèdent à internet ou à un smartphone. C'est pourquoi nous avons étudié la relation entre les yeux et le défilement sur les pages web dans le **Poster 3** et l'**Article 2**. Nous avons montré que l'emplacement des yeux pouvait être utilisé pour déduire les paramètres du défilement en cours ou celui précédent. Par exemple, nous avons observé que lors d'un défilement rapide, nous avions tendance à orienter nos yeux dans la direction opposée du défilement.

Ensuite, nous avons analysé la relation entre le curseur de la souris et les yeux dans l'**Article 2**. Il est intéressant de noter que l'étude de cette relation a suscité beaucoup d'intérêt de la part des moteurs de recherche, tels que Google et Microsoft. En raison de la nature de ces sociétés, la grande majorité des études sur ce sujet ont été réalisées sur des pages de résultats de ces mêmes moteurs de recherche (Search Engine Result Page (SERP)). Le problème est que ces pages web ne sont pas représentatives du web que nous utilisons tous les jours. Dans l'**Article 2**, nous avons étudié cette relation sur des pages web classiques et proposé un modèle pour estimer la position des yeux en fonction de la position du curseur de la souris. Comme dans pour les SERP, nous avons montré que la coordination entre les yeux et le curseur de la souris était meilleure sur l'axe vertical. Cependant, nous avons montré que lorsque les participants étaient sur le point de cliquer, la coordination entre les yeux et la souris sur l'axe horizontal augmentait.

Jusqu'à présent, les relations entre les yeux et le défilement, ou entre les yeux et la souris, ont surtout été étudiées d'un point de vue des zones d'intérêt. Par exemple, les mesures classiques comprennent la durée de la fixation des yeux dans une zone donnée, ou le nombre de clics nécessaires pour atteindre la cible désignée. La littérature sur la description statistique des mouvements des yeux sur les pages web, les mouvements de la souris et le défilement est très rare. Pour ces raisons, nous avons proposé dans l'**Article 3** une description statistique détaillée des mouvements des yeux sur les pages web ainsi que des mouvements de la souris et du défilement. Cette analyse comprenait

des paramètres globaux et leur évolution dans le temps. En outre, afin d'évaluer si les modes ambient et focal peuvent être généralisés à la souris et au défilement, nous avons étendu l'utilisation du coefficient K (Krejtz et al., 2016) à l'analyse de la dynamique de la souris. Nous avons constaté que les paramètres liés à l'œil et à la saccade de la souris diminuaient au fil du temps, tandis que les paramètres de défilement augmentaient. Inversement, les paramètres liés à la fixation des yeux et des souris augmentaient avec le temps, tandis que les paramètres de défilement diminuaient. Dans les deux cas, les paramètres de l'oeil et de la souris ont suivi le même schéma, et les paramètres de défilement ont suivi le schéma opposé. Il est intéressant de noter que ces observations étaient cohérentes d'une tâche à l'autre.

### Modélisation des mouvements oculaires sur les pages web

Nous avons proposé dans cette section l'implémentation d'un mécanisme reproduisant le scroll afin de mieux modéliser les mouvements oculaires lors de l'exploration visuelle de pages webs. Nous avons également proposé la prise en compte des durées des fixations, de l'amplitude de saccade et la direction des saccades à travers le temps. Cela nous à permis d'évaluer et d'analyser la qualité des chemins oculaires globalement et à travers le temps. Nous avons montré que notre modèle obtenait les meilleurs résultats concernant la longueur des saccades, la direction des saccades et la durée des fixations. De plus, nous montrons que bien que notre modélisation soit meilleure sur ces aspects, elle n'est pas constante dans le temps. On remarque particulièrement que la qualité de prédiction des directions des saccades diminue avec le temps alors que celle des longueurs des saccades reste stable. Cependant, malgré un scanpath plus plausible biologiquement, nous n'avons pas amélioré l'état de l'art lorsque comparé avec les métriques de saillance.

## Conclusion

Au carrefour de plusieurs domaines de recherche, le but de cette thèse était de démontrer l'importance de la dynamique dans la compréhension et la prédiction des mouvements

oculaires. Les études jusqu'à présent se sont principalement concentrées sur la prédiction de l'emplacement de la prochaine fixation et sur les facteurs qui pourraient globalement influencer le choix de cette prochaine fixation. Directement héritées du domaine de la modélisation de la saillance, ces études ont négligé la modulation temporelle des paramètres des mouvements oculaires.

Nous avons montré que l'exploration visuelle pouvait être influencée par des éléments dynamiques provenant de sources multiples. La première se référait à la dynamique du stimulus lui-même. Nous avons souligné que lors de l'exploration de pages web, le défilement avait une influence sur les paramètres des mouvements oculaires. Cette influence pouvait se produire soit avant, soit pendant le défilement. La deuxième source se situait dans les paramètres de mouvement des yeux. Nous avons montré que ces paramètres évoluaient au fil du temps lors de la navigation sur une page web. Nous avons ensuite décrit les avantages de l'intégration de ces deux sources de dynamique dans un modèle saccadique, et évalué la plausibilité du parcours oculaire généré grâce à des mesures temporelles dédiées.

Les modèles saccadiques et de saillance sont déjà utilisés pour prédire le comportement des internautes afin d'afficher des informations plus précises ou plus spécifiques là où elles ont le plus de chances d'être vues. Mais, le principal avantage de ces modèles réside dans le fait qu'ils peuvent être adaptés à une grande variété de domaines présentant des problématiques spécifiques, tels que l'ergonomie des logiciels, les jeux, la réalité virtuelle, les outils éducatifs, ou même les utilisations cliniques.

## Articles in journal

- **Milisavljevic, A.**, Abate, F., Le Bras, T., Petermann, C., Gosselin, B., Mancas, M., & Doré-Mazars, K. (Under revision). Similarities and differences between eye and mouse dynamics during web pages exploration. *Frontiers in Psychology*.

## Articles in conference proceedings

- **Milisavljevic, A.**, Le Bras, T., Mancas, M., Petermann, C., Gosselin, B., & Doré-Mazars,K.(2019). Towards a better description of visual exploration through temporal dynamic of ambient and focal modes. *In Proceedings of the 11$^{th}$ ACM Symposium on Eye Tracking Research & Applications - ETRA '19* (pp. 1–4). Denver, Colorado: Association for ComputingMachinery.

- **Milisavljevic, A.**, Hamard, K., Petermann, C., Gosselin, B., Doré-Mazars, K., & Mancas,M.(2018). Eye and Mouse Coordination During Task: From Behaviour to Prediction. *In Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications* (pp. 86–93). Funchal, Madeira, Portugal:SCITEPRESS - Science and Technology Publications.

## Communications in international congresses with review committee

- **Milisavljevic, A.**, Le Bras, T., Abate, F., Gosselin, B., Petermann, C., Mancas, M. and Doré-Mazars, K. (2019). Different visual explorations between free-viewing and target finding tasks in websites: evidence from temporal analyses of ambient and focal modes. 20$^{\text{th}}$ European Conference on Eye Movements, 18-22 August 2018, Alicante, Spain. *Journal of Eye Movement Research* 12(7), p. 390 [Poster]

- **Milisavljevic, A.**, Le Bras, T., Mancas, M., Petermann, C., Gosselin, B., & Doré-Mazars, K. (2019). Towards a better description of visual exploration through temporal dynamic of ambient and focal modes. *In Proceedings of the 11$^{th}$ ACM Symposium on Eye Tracking Research & Applications - ETRA '19* (pp. 1–4). Denver, Colorado: Association for ComputingMachinery [Talk]

- **Milisavljevic, A.**, Le Bras, T., Petermann, C., Mancas, M., Gosselin, B., & Doré-Mazars, K. (2018). A dynamic approach of searching behaviour in webpages. 41$^{\text{th}}$

European Conference on Visual Perception, 26-30 August 2018, Trieste, Italy. *Perception, ECVP 2018*. 48(S1), p.22 [Poster]

- **Milisavljevic, A.**, Hamard, K., Petermann, C., Gosselin, B., Doré-Mazars, K., & Mancas, M. (2018). Eye and Mouse Coordination During Task: From Behaviour to Prediction. *In Proceedings of the 13$^{th}$ International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications* (pp. 86–93). Funchal, Madeira, Portugal:SCITEPRESS - Science and Technology Publications [Talk]

- **Milisavljevic, A.**, Doré-Mazars, K., Gosselin, B., Mancas, M., & Petermann, C. (2017). What scroll can teach us about web users ? 40$^{\text{th}}$ European Conference on Visual Perception , 27-31 August 2017, Berlin, Germany. *Perception, ECVP 2017* [Poster]

## Communications in national congresses

- **Milisavljevic, A.**, Le Bras, T., Mancas, M., Gosselin, G., Petermann, C., & Doré-Mazars, K. (2018). Dynamic behaviour when searching in webpages. *12$^{th}$ GDR Vision Meeting*, 5-6 October 2018, Paris, France [Poster]

# Table of Contents

# List of Tables

# List of Figures

**Figures** **Page**

# Acronyms

**AOI** Area Of Interest. 70, 84, 135

**CC** Correlation Coefficient. 80, 181, 183

**DNN** Deep Neural Network. 78–81, 89, 200, 202, 203

**FD** Fixation Duration. 31, 53–55, 175, 191, 192

**HVS** Human Visual System. 36, 76

**IOR** Inhibition Of Return. 45, 65, 75, 76, 83, 85–90, 92, 175, 178, 179

**KL-Div** Kullback-Leibler Divergence. 80

**LGN** Lateral Geniculate Nucleus. 48

**LSTM** Long Short-Term Memory. 79, 80, 89, 203

**NSS** Normalized Scanpath Saliency. 80, 181, 183

**PCA** Principal Component Analysis. 77, 78

**RNN** Recurrent Neural Network. 89, 203

**ROI** Region Of Interest. 87

**SA** Saccade Amplitude. 31, 53–55, 175, 178, 191, 192

**SERP** Search Engine Result Page. 16, 19, 66, 68–70, 91–93, 95, 135

**WTA** Winner-Take-All. 75, 76, 83, 85, 87–89, 91, 181, 202

# FOREWORD

Visual perception refers to the ability to observe and analyse our environment. Also known as vision for action, our interaction with the world fundamentally depends on acquiring and understanding visual information so that we can act accordingly.

To cope with the abundance of available information we perceive, we developed physiological and cognitive mechanisms. The first mechanism concerns the very way we access visual information: eye movements. Although seeing seems to require little effort, a very large number of underlying cognitive processes and a complex set of muscles are constantly interacting for this sole purpose. For instance, it is not possible for us to see the entire visual environment at the same time. In order to explore a visual scene, our eyes have to make "jumps" called *saccades* to move from one location to another. It is between two saccades that our eyes stabilise to acquire information; we then speak of *fixations*. Thus, the visual exploration of our environment consists of an infinite sequence of fixations and saccades. The second mechanism allowing us to cope with the abundance of information that is continuously available is our attention. Attention acts as a filter that allows us to ignore distracting information to avoid our brain overloading. For example, when we cross the road, we direct our attention to the incoming car rather than to the colour of the sky. During this short period of time, we also give less importance to other visual information.

The complexity of the visual system is of greater importance as our behaviour evolves according to the context or the type of stimulus being visualised. For instance, we are not going to look in the same way at a landscape, a portrait or a web page. But the complexity of the functions related to vision does not stop there. Our behaviour not only changes according to what we see, but also according to our goals. For example, when we browse a website, we are not going to look at the same thing depending on whether we

are looking for a particular photo or a specific paragraph in a text.

To study the complexity of the visual system, we investigate the oculomotor behaviour on web pages. Invented at the beginning of the 1990s, the web has contributed to a massive and rapid diffusion of knowledge accessible to all. This information explosion has totally transformed our society and is still the source of many changes to come. Through websites, we now have access to an unprecedented amount of information, so much that at the end we tend to remember how to access information rather than the information itself. Therefore, understanding how we access information and what draw our attention on this particular medium is of critical importance.

Many researchers have been trying to understand the complexity of visual perception on web pages through two main approaches: experimental psychology and modelling. The aim of the first approach is to set up a scientific experiment to observe the behaviour response of participants. In this manuscript we will present two experimental studies to help us understand oculomotor behaviour when browsing web pages. The aim of the second approach is to approximate human behaviour by using algorithms and mathematics to predict where we look. If a model correctly predicts a behaviour, then we have probably well identified and reproduced the variables that influence this behaviour.

Here, we will evoke models predicting where we orient our attention and where we are gazing. We will particularly focus on models attempting to predict eyes' location. These types of models usually implement factors influencing visual exploration as constant behaviour over the entire visual exploration. What we propose is to take the temporal aspect of these factors into account. More specifically, we focus on eye parameter dynamic modelling to improve the quality of prediction on visual exploration of web pages.

This thesis is organised into three main parts. First of all, the General Introduction (Part I) is dedicated to describe key theoretical concepts in the vision framework used in this work. Chapter 1 introduces the architecture of the Human Visual System (HVS) and its functioning. We also present attentional mechanisms and their links to vision. At the end of this chapter, a particular focus is done on how the study of the dynamic of eyes movements could help to better understand the oculomotor behaviour. Chapter 2

lists the most important factors influencing visual exploration, usually categorised in two different types: bottom-up and top-down. Bottom-up factors designate stimulus-dependent characteristics influencing visual exploration, such as colour, luminance, shape, etc. While top-down factors designate cognitive processes influencing visual exploration, including goal, expertise, emotions, etc. These factors will be addressed on natural images and web pages along with systematic tendencies, or biases in the manner we explore visual scenes. In Chapter 3 a selection of models predicting where we orient our attention are introduced. Then, an extensive number of models predicting eye movements is established with a particular focus on how oculomotor biases are implemented across the different presented approaches.

Then, Part II presents contributions produced during this thesis. More specifically, Chapter 4 outlines the general methodology used in the two experiments of this work. Chapter 5 describes the validation framework developed during this thesis to evaluate our model and compare it to others. Chapter 6 presents results from the first axis of this thesis which consists in investigating how the dynamics of eye movements can be summarised as an indicator. Based on ambient and focal visual processing modes we evaluate the relevance of existing indicators on natural images and web pages. Chapter 7 explores the second axis of this thesis consisting in the study of eye movement behaviour when browsing real web pages. The third axis of this work is described in Chapter 8. We present a model predicting where we look using eye movement parameters evolution over time.

Finally, the General Discussion is introduced in Part III of this document. Our contributions to the improvement of the visual exploration understanding and modelling on web pages are summarised and discussed in Chapter 9. Theoretical and practical perspectives induced by these discussions are presented in Chapter 10.

# Part I

# General introduction

CHAPTER 1

# THE VISUO-MOTOR SYSTEM

## Contents

Vision enables the perception and the understanding of our surrounding environment through physiological and cognitive mechanisms. As anyone can observe, visual information is abundant in our environment, but we are not able to process it all at once. That is why the large quantity of visual information collected by the eyes first needs to be filtered. This is where Attention comes in. While the eyes and more generally the visual system gather visual information, attention and more specifically visual attention acts as a filter to prevent this information from overloading our brain. In this chapter we first explain basic attention mechanisms. Then we address the physiology of the visual system and eye movements. Finally, we tackle eye movements temporal dynamic.

## 1.1 Visual attention

Attention can be described as the focus of mental activity on a specific object. Since William James (James, 1890), numerous theories contributed to a better understanding of attention mechanisms. Nowadays, Attention is typically separated in two systems: exogenous attention and endogenous attention (for a synthesis see Maquestiaux (2013)).

### 1.1.1 Exogenous attention

There are multiple examples of how our attention can be unintentionally captured by an element of our surrounding. For example, when driving a car, we are focusing on the road to ensure our car is securely heading the right way at the right speed and at a correct distance from other cars. But if a pedestrian suddenly crosses the road, there is a high probability that we will notice it almost instantly. In this situation, attention is unintentionally and automatically shifted to an exogenous stimulus, here, the pedestrian. This phenomenon is called attentional capture and is also referred to as bottom-up attention. In visual context, bottom-up attention or attentional capture is generally driven by physical characteristics of a stimulus, such as, luminance, shape,

colour, etc. We tend to be influenced by these characteristics the first couple of seconds of the visual exploration (Buswell, 1935; Karpov et al., 1968).

### 1.1.2 Endogenous attention

Fortunately, attention can also be intentionally directed. For instance, watching driving requires a voluntary endogenous maintenance of attention during a long period of time. Such maintenance is called sustained attention and is involved in any daily activities we have like reading, cooking, working, studying, etc. Tasks however require more or less attentional resources and; while a demanding task such as memorisation requires all attentional resources available, a less demanding task will allow us to divide our attention between two or more simultaneous tasks. This type of attention is called divided attention. It allows us to multitask, but the main disadvantage is that each action will likely be executed with less accuracy. In comparison, selective attention is a voluntary focus of attention on a specific task or stimulus while inhibiting an exogenous distractor. Helmholtz (1896) and Shepherd et al. (1986) showed that it is possible to orient our attention without eye movements, but Shepherd et al. (1986) specified that making a saccade necessarily involved the orientation of our attention to the target location. The link between saccade and attention has been confirmed numerous times since (e.g. Collins & Doré-Mazars, 2006; Deubel & Schneider, 1996; Doré-Mazars & Collins, 2005).

### 1.1.3 Attention orienting

The ability to direct our attention on a part of our visual field without looking at it is called *covert attention*. The opposite behaviour, which consists in moving our eyes and/or our head to put the element of interest on the fovea, is called *overt attention*. The fovea designates the area of the retina where the vision of details is the most accurate. In a set of famous experiments, Posner and collaborators (Posner, 1980; Posner et al., 1978) demonstrated how visuo-spatial attention could be oriented like a spotlight. In the Posner's paradigm described in Figure 1.1, participants are asked to fixate the centre

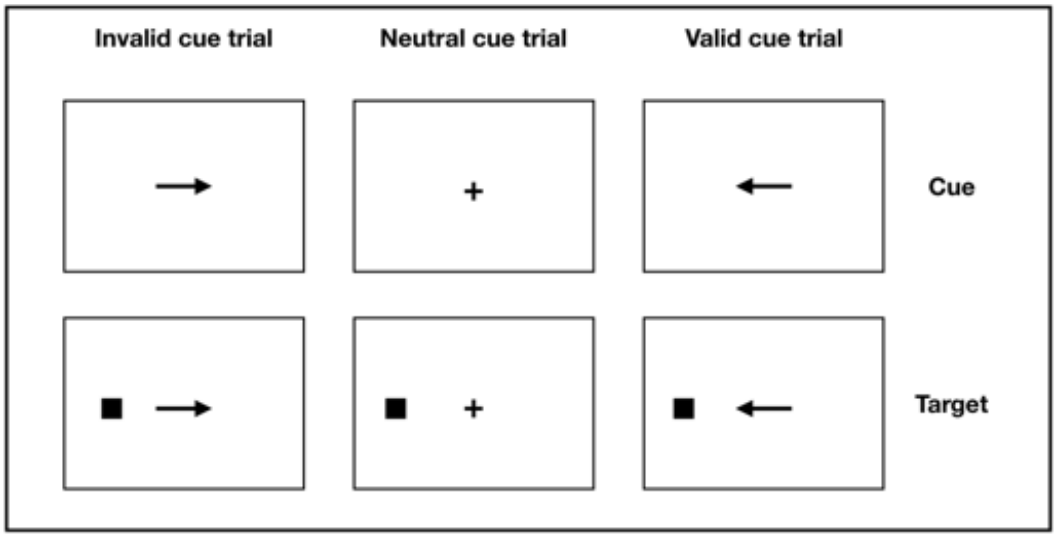



Figure 1.1 – **The Posner central cueing paradigm**. The target is preceded by a cue displayed at the centre of the screen. This cue can either be a plus sign (neutral) or an arrow (valid or invalid) indicating the location where the target may appears. Once the target is displayed, the participant respond as quickly as possible without moving the eyes where the target appeared (covert attention). Adapted from Findlay and Gilchrist (p. 37 2003).

of the screen during the entire experiment (covert attention). A cue is displayed at the centre of the screen, and can either take the form of a plus sign (neutral cue), or of an arrow pointing on the left or right side. Then a target is displayed either on the right or on the left side of the central cue. Once the target appears, the participant is instructed to press, as quickly as possible, on the keyboard the arrow button pointing the same side as the target. During this experiment, two cueing conditions (excepted neutral cue) are presented at the centre of the screen to the participant. In both conditions, an arrow is used as the cue in order to imply an additional cognitive process involving endogenous component of attention. This additional process corresponds to the voluntary and controlled orientation of the participant's attention on the location pointed by the arrow. In 80% of the trials, the target's appearance is correctly cued (valid cue), which means that the cue correctly indicates where the target will appear. In the remaining 20%, the target appears in the opposite direction from the cue (invalid cue). Such large proportion (>50% above chance) of valid cues enables the participant to "trust the cue" and to voluntarily establish a strategy. In this case, endogenous attention takes a greater

part in the orienting of the attention than exogenous. If however, the proportion of valid and invalid cues is set to 50%-50%, the participants will not develop a strategy, and their attention will be captured by the onset of the target. Hence, exogenous attention will take a greater part in the orienting of the attention than endogenous. Results are analysed in terms of costs and benefits. Reaction times in the neutral condition are subtracted to reaction times in the valid and invalid conditions. If the difference is greater than zero, valid or invalid condition is considered beneficial (in terms of reaction times) compared to the neutral condition. If the difference is lower than zero, valid or invalid condition is considered more costly (in terms of reaction times) than the neutral condition. Posner (1980) showed that during the valid cue condition participants indicated the correct side the target would appear faster than during neutral cue condition and invalid cue condition. Later on, Posner and Cohen (1984) reproduced the same experiment but with longer intervals between the disappearance of the cue and the appearance of the target. Contrary to the first series of experiments, they observed longer reaction time in the valid cue condition. They explained this behaviour by a process that would prevent attention from returning to a previously explored location: Inhibition Of Return (IOR). More specifically, this behaviour would be the result of an automatic inhibitor mechanism preventing the oculomotor system from exploring twice a same location, so it can be faster and more efficient during the rest of the visual exploration of the environment. IOR would modify our visual scanning by reducing the number of eye movements directed to the locations previously fixated. In addition, it has been shown that IOR could not occur without eye movements which suggests that IOR could be related to oculomotor system activation (for an extensive review see Klein (2000)).

## 1.2 The visual system

The visual system designates all the physiological structures involved in the capacity of seeing. When focusing our attention on a visible object, the light is reflected on this object to enter the main structure of the visual system, the eyes.

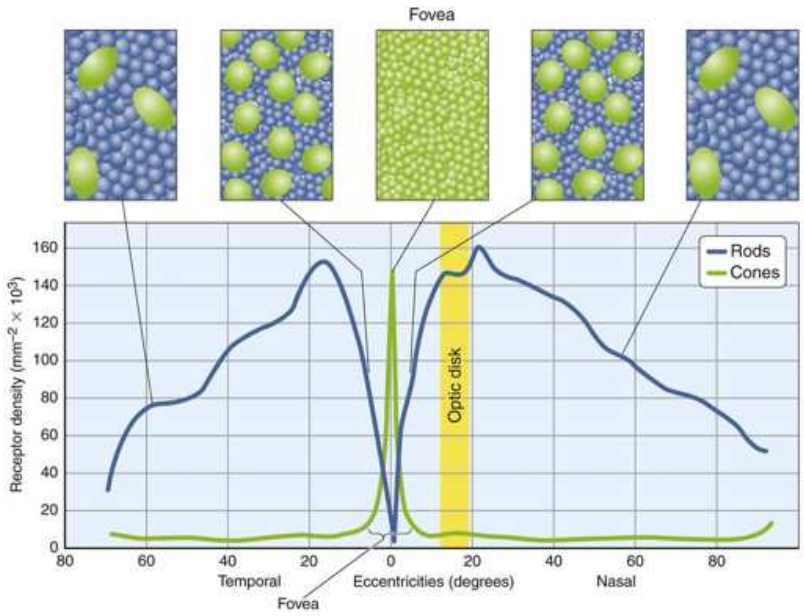



Figure 1.2 – **Photoreceptors distribution on the retina**. The retina is composed of cones and rods. Cones (in green) are mostly found in the fovea where acuity is maximal, and their density decreases with the eccentricity. In contrast, rods (in blue) are mostly found in the periphery and their density increases with the eccentricity. The optic disk corresponds to the area with no photoreceptors, this is where the optic nerve starts. From Mustafi et al. (2009).

### 1.2.1 Neurophysiological architecture

Once the light enter the eyes, it goes through the multiple layers of the eyes to finally running out on the retina. The last layer of the retina is composed of a multitude of cells, called photoreceptors, able to transform a light beam into a nervous signal. They can be categorised into two types: cones and rods (Osterberg, 1935). The cones are responsible for coloured and detailed vision involved, for example, in object recognition, while the rods are responsible for achromatic and global vision involved, for example, in night vision. These two types of photoreceptors are unevenly distributed across the retina (see Figure 1.2). The cones are mostly concentrated in its central region, called the "fovea", corresponding to approximately 2 degrees of visual angle, while the rods are mainly present at the periphery. The visual angle is the common unit to designate the size of an object on the retina.

Once the photoreceptors convert the light beam into a nervous impulse, the signal

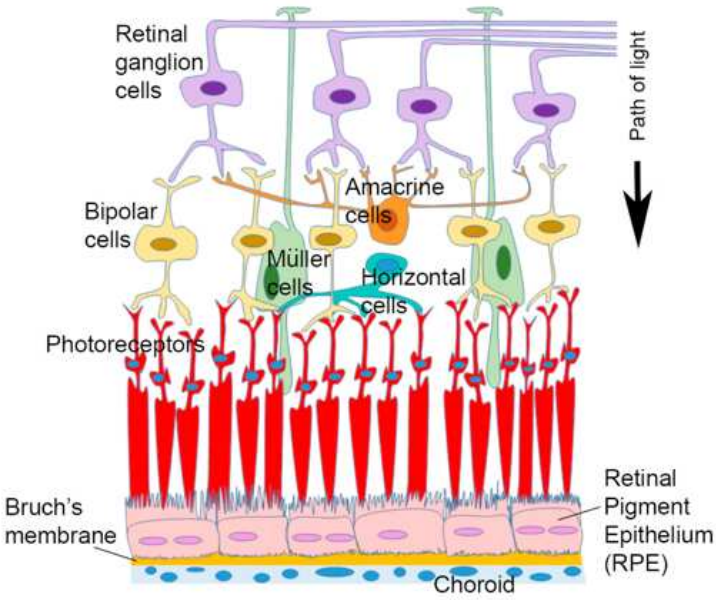



Figure 1.3 – **Retina cells organisation**. Photoreceptors convert the visual information into a nervous impulse successively transmitted to horizontal, bipolar, amacrines and ganglion cells axones finally form to the optic nerve. Image adapted from Keeling et al. (2018).

is sent to several cells within the retina (Figure 1.3). The signal is first transmitted to the bipolar and horizontal cells which relay the visual information to amacrine and ganglion cells (Cowey, 1964). Ganglion cells' axons then gather to constitute the optic nerve which project to the brain. In the fovea, a single cone is connected to a single bipolar cell, which send information to a single ganglion. The further away from the fovea, the more multiple photoreceptors send information to a single bipolar cell and the more bipolar cells relay information to a single ganglion cell. The level of convergence determines our level of visual acuity (Cowey, 1964).

As described in Figure 1.4, the optical nerve does not go straight to the visual cortex area in the occipital brain. First, the nasal hemiretina of the left eye crosses the nasal hemiretina of the right eye without fusing with it. However, hemifields are grouped: what we see on the right side of the retina will go to the right path and what we see on the left side of the retina will go to the left path. It should be noted that the right hemifield of the retina captures the left part of our visual field and the left hemifield of the retina captures the right part of our visual field. This intersection is called the

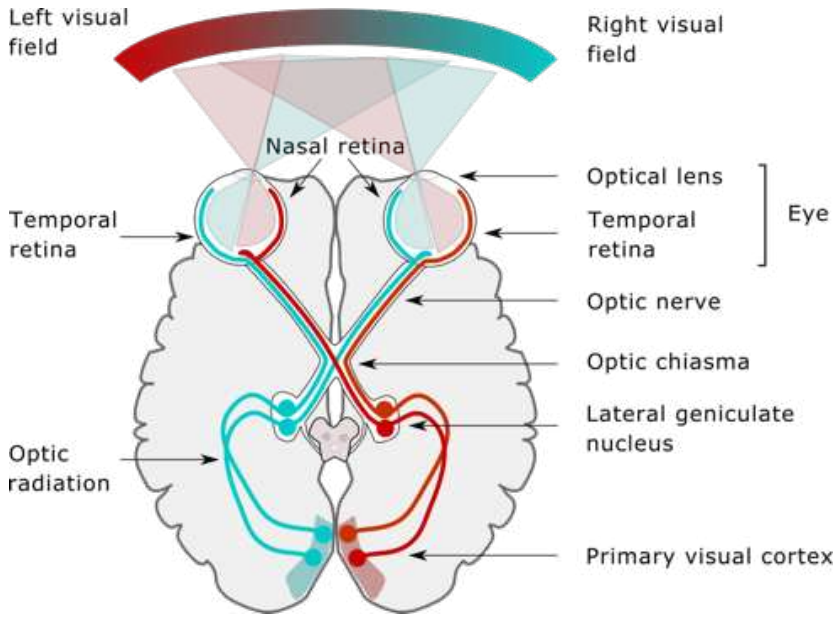



Figure 1.4 – **Visual system neuroanatomic architecture**. The left visual field is projected on the left nasal hemiretina and the right temporal hemiretina (in red). The left nasal hemiretina decussates on the optic chiasma to continue to the right primary visual cortex. Thus, the left visual field is processed by the right primary visual cortex and the right visual field is processed by the left primary visual cortex. Adapted from Gazzaniga et al. (2001).

optic chiasma and the resulting newly grouped optic nerves are called optic tracts. From there, 80 to 90% of the fibre composing the optic tracts project to the Lateral Geniculate Nucleus (LGN) of the thalamus, this is the primary visual pathway. The remaining 10 to 20% break in four different paths, this is the secondary visual pathway. From the LGN, primary visual pathway runs out in the primary visual area through the optic radiation.

The visual cortex consists of different areas distinguished by their functional specialisation. After arriving in the primary visual area (V1), visual information is transmitted to the secondary visual area (V2) which then projects the information on multiple visual areas V3, V4 and V5. These areas are sensitive to different characteristics, such as, colour, movement or shape of stimuli. For instance, V1 area reportedly creates a saliency map of visual inputs based on low-level features (colour, luminance, etc) to guide attentional shifts to salient locations (Zhang et al., 2012), while V2 is more sensitive to orientation and spatial frequency.

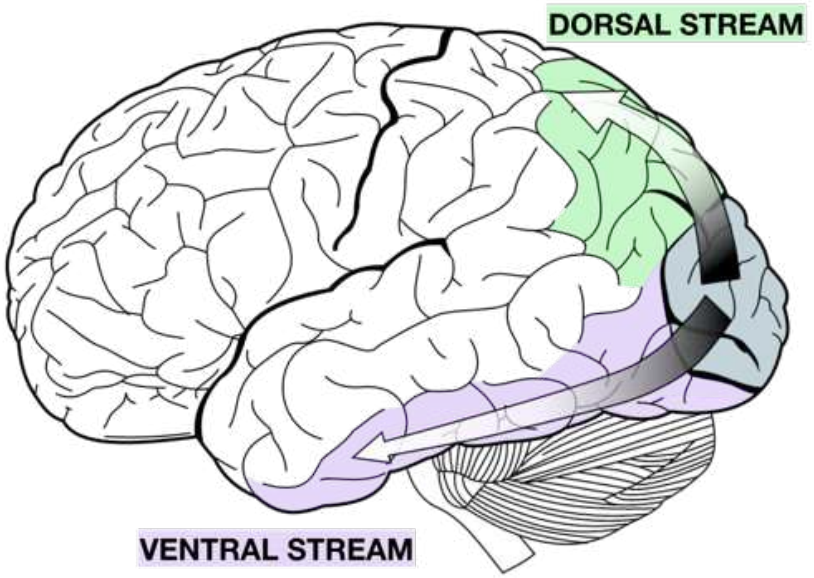



Figure 1.5 – **Illustration of the two visual pathways**. In green, the dorsal stream (vision for action) goes from the visual primary area to the posterior parietal lobe. In purple, the ventral stream (vision for recognition) going from in visual primary area to the inferior temporal lobe. Image adapted from Ungerleider and Mishkin (1982).

To process visual information, these visual areas communicate between each other and with more than 30 different areas of the brain that process visual information. Through the literature, numerous functional dichotomies have been suggested to explain the distribution of these cortical areas. The most notable being ambient-focal (Trevarthen, 1968), noticing-examining (Weiskrantz, 1972), spatial-figural (Breitmeyer & Ganz, 1976) or ambient-foveal (Stone et al., 1979) dichotomy. Ungerleider and Mishkin (1982) demonstrated that these visual pathways are involved in two specific cognitive mechanisms: visual recognition and visuospatial awareness. They found that a bilateral lesion on the occipito-temporal cortex in monkeys disrupted object recognition, while a bilateral lesion of the occipito-parietal cortex disrupted visuospatial awareness. They proposed then that the visual cortex is divided in two main systems: ventral and dorsal pathways (see Figure 1.5). Ventral stream goes from the occipital to the temporal cortex and carry information about which object is being seen, while the dorsal stream goes from occipital to parietal cortex and carry information about the object location. Later, Goodale and Milner (1992) proposed a nuance to the original dichotomy. In their ver-

sion, the ventral stream keeps a similar function as previously described and is called Vision for Recognition, while the dorsal stream is supposedly more related to visually guided movements and is named Vision for Action. Both studies indicated that the two streams are independent and distinct, however recent studies highlighted that they are extensively interconnected (see Milner (2017) and Rossetti et al. (2017) for reviews). A group of studies (Bullier et al., 1996; Morel & Bullier, 1990) suggested an alternative explanation for the dichotomy proposed by Goodale and Milner (1992) based on the time needed to process the visual information. The ventral stream would be the slow pathway involving, for instance, object recognition, and the dorsal stream would be the fast pathway involving, for example, the quick analysis of an object's global shape in the environment.

Independently of the cognitive functions covered by each of the two streams, to recognise or reach an object, it is better to put first this object on the centre of our fovea. To do so, we execute jerky eye movements called saccades.

### 1.2.2 Saccades

Detailed vision is possible when the object of interest is on the fovea. The action of moving the fovea around and enhancing the exploration of the environment is called a saccade. We make around 200 000 saccades a day, which is the most frequent movement a human will make through his life.

As represented in Figure 1.6, each eye is controlled by three pairs of antagonists muscles providing a wide variety of movements to explore our environment (Porter et al., 1995):

- Horizontal movements are the result of the action of the lateral and medial rectus muscles.

- Vertical movements are achieved by the action of the inferior and superior rectus muscles combined with the action of the inferior and superior oblique muscles.

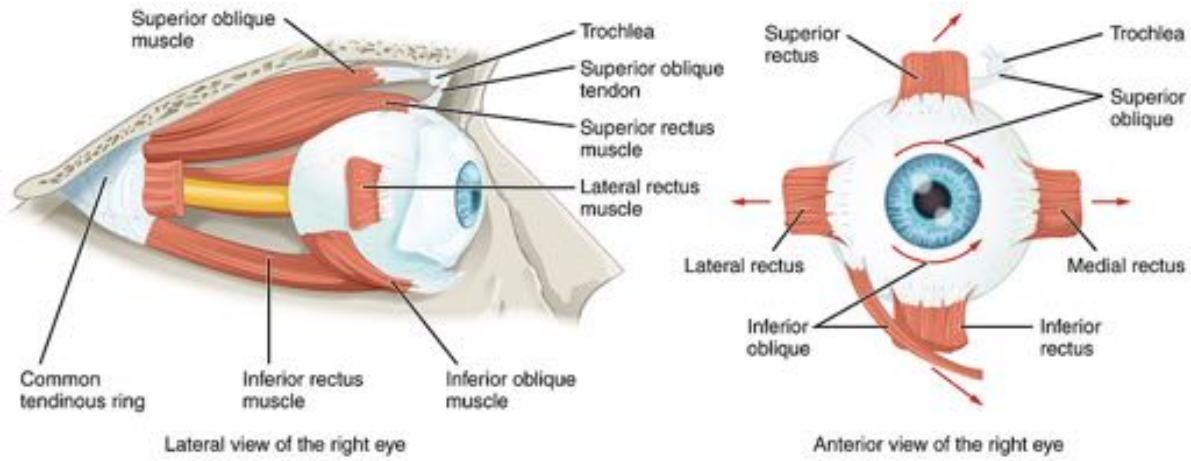



Figure 1.6 – **Extraocular muscles**. Each eye is controlled by three pairs of antagonists muscles. The first pair designates the lateral and medial rectus muscles. The second pair refers to the inferior and superior rectus muscles. The third pair identifies the inferior and superior oblique muscles. Image adapted from Betts et al. (2013).

- Torsion movements are made possible by the action of the inferior and superior oblique muscles.

In order to execute a saccade, extraocular muscles on the same direction as the movement need to be contracted and antagonist muscles need to be released to allow the eyes to move in the desired direction. Saccades can be categorised based on the type of events that can trigger them: exogenous or endogenous (Leigh & Zee, 2006). Exogenous or reactive saccades are triggered by the sudden appearance of a stimulus within the visual field. Hence, they are triggered automatically and quickly. On the contrary, endogenous or voluntary saccades are intentionally triggered to move the fovea toward an object of interest. They are typically controlled and slower than reactive saccades. To understand this difference, we need to go back to the primary visual cortex. In addition to the already cited cortex areas, V1 also transmit information to the Parietal Eye Field (PEF). The PEF then projects to the superior colliculus (SC) which represents an important relay for saccade generation. The V1-PEF-SC pathway is the path taken to generate reactive saccades. The PEF also projects to the Frontal Eye Field (FEF) which projects in turn to the superior colliculus. The V1-PEF-FEF-SC is involved in the generation of voluntary saccades.

Classically, in the saccade domain, the main studied parameters are the latency, the direction and the amplitude of the saccade First, the latency designates the time between the display of a target and the start of the eye movement. The saccade is also described by its direction. Finally, the amplitude corresponds to the difference between the position of the eye at the end of the saccade and its starting position. In ecological situation, saccades amplitude, with no head movements, is generally smaller than 15 degrees of visual angle (Gilchrist, 2011). After the saccade has been executed, ensues most of the time, a stabilisation period, called fixation.

### 1.2.3 Fixations

As we can guess by its name, the role of a fixation is to immobilise the eyes during a long enough period of time to grasp details about a visual stimulus. Although we perceive a steady image when fixating, fixation is more about stabilisation and low speed displacement of the image on the fovea than a proper immobilisation (Fischer et al., 1997). Actually, Ditchburn and Ginsborg (1952) showed that if an image is maintained stable on the retina, vision is fading. It should be noticed that during high-speed movements of the eyes, that is saccades, the acquisition of information is also limited (Burr & Ross, 1982).

To ensure the image is stabilised on the retina during a fixation, three types of movements occur during a fixation. The movements to stabilise an image on the retina during a fixation are called microsaccades. Drifts are slow curvy movements occurring between microsaccades. Finally, tremors are very fast and extremely small oscillations superimposed on drifts (Martinez-Conde et al., 2009).

Another particularity is that during fixations, spontaneous saccades are suppressed to maintain the gaze on the target. Overall, this show us that fixating is not a passive mechanism but rather an active process allowing us to grasp details on our surrounding environment.

To study visual fixation the main parameter is its duration which designates the period of time between two saccades. When observing pictures, fixations' duration is

usually around 300 to 350 milliseconds Mackworth and Morandi, 1967; Yarbus, 1967. However, as for saccade amplitude, fixation duration evolves over time and needs specific analyses. A dynamic approach is then necessary to better understand visual exploration.

## 1.3 Eye movements dynamic

Visual information is acquired through a succession of fixations and saccades distributed in time and space. The role of vision is to aggregate these sequences to build a comprehensive overview of the visual context. Since visual representations are progressively built through saccades and fixations, the study of their parameters is of particular interest. That is why, the evolution of Fixation Duration (FD) and the Saccade Amplitude (SA) over time, has been particularly used to study and better understand visual exploration.

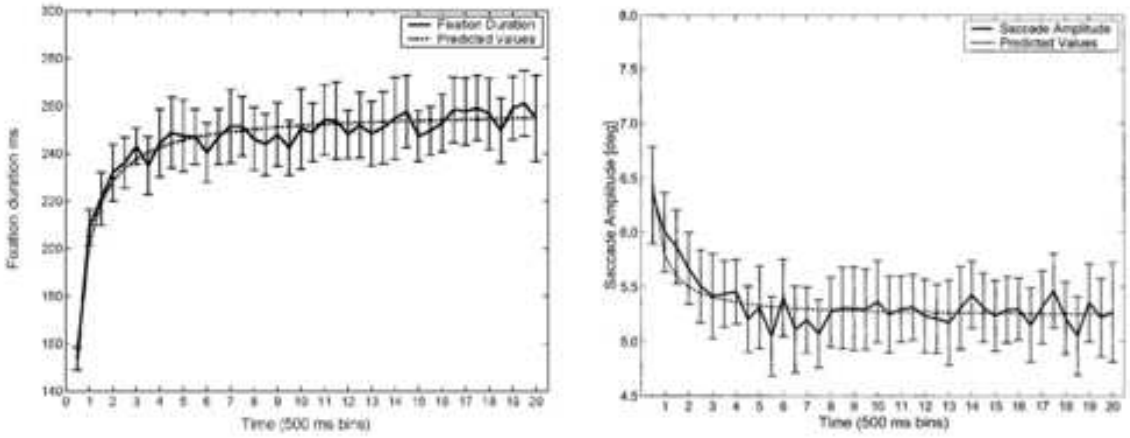



Figure 1.7 – **FD and SA time courses**. Left figure shows the increase of FD as a function of time, while right figure shows the decrease of SA over time. The error bars represent the 95% confidence interval. The dotted line designate predicted values from an exponential function. Adapted from Unema et al. (2005).

### 1.3.1 Fixation duration and saccade amplitude time courses

Buswell (1935) was the first to observe an evolution of FD and SA when exploring paintings. He described two patterns: short fixations during the global scanning of the scene, and longer fixations in more limited areas, usually occurring after the scanning

phase. Comparably, Karpov et al. (1968) noted the presence of an orienting phase during which salient characteristics were explored, followed by a period in which people fixated the most informative elements. Antes (1974), like Buswell (1935), showed a constant increase of mean FD and a constant decrease of mean SA during the exploration of 10 pictures. He also described two phases; participants were progressively shifting from long saccades and short fixations on informative components to longer and more frequent fixations on less informative details. Irwin and Zelinsky (2002) reported a relationship between fixation durations and the duration of the visual exploration: the longer the exploration time was, the longer the fixations were. This relationship was confirmed by Velichkovsky et al. (2005) but only for the first 2 to 3 fixations. Unema et al. (2005) finally replicated previous findings by reporting a steady increase in fixation duration over the first few seconds of the exploration and a following a continuous decrease of saccade amplitude (see Figure 1.7).

### 1.3.2 Two visual processing modes: ambient and focal

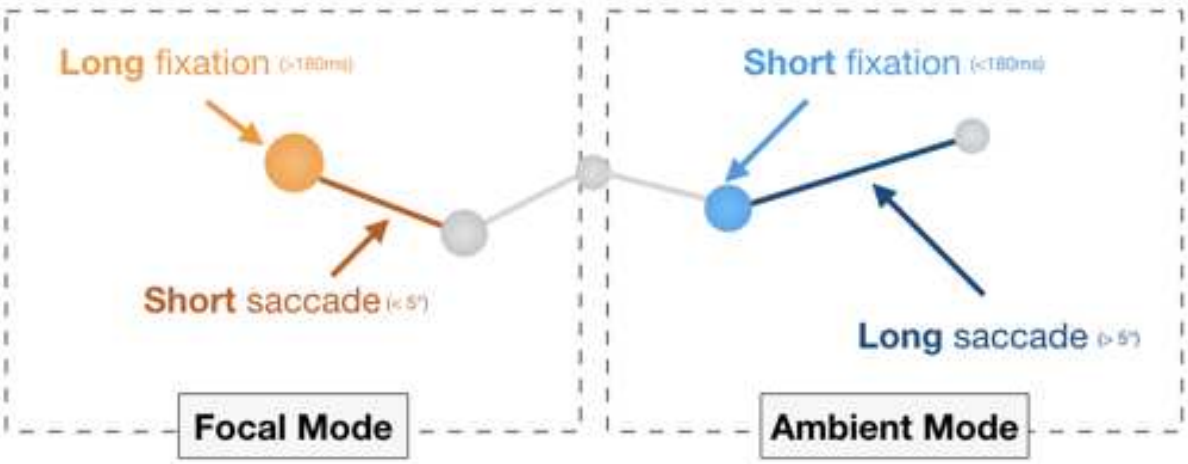



Figure 1.8 – **Visual modes dichotomy as defined by Pannasch et al. (2008)**. The focal visual mode designates a long fixation (> 180ms) followed by a short saccade (<5°) and the ambient visual mode designates a short fixation (< 180ms) followed by a long saccade (>5°).

To better describe the relationship between fixations and saccades, Unema et al. (2005) ran an experiment in which participants were asked to visually explore computer-

generated scenes. Participants were presented with two similar scenes in which 8 or 16 objects were disseminated on a shelf. Participants had to locate and recognise the different objects. They noted that localisation performance was similar in both conditions. They however showed that recognition was faster in the 8 objects condition. This is easily explained by the fact that visual recognition requires more processing. Since more objects were present in the 16 objects condition, less objects could be recognised in the given time. As Velichkovsky et al. (2005) stated, this recognition/location dichotomy is likely directly related to the ventral/dorsal dichotomy of visual pathways. In their work, Velichkovsky et al. were the first to link Trevarthen's focal-ambient dichotomy (Trevarthen, 1968) to eye movements. They identified two distinct segments of eye movements: ambient and focal visual modes (see Figure 1.8). The first one was defined as a short fixation (from 90 to 260ms) related to a large following saccade (>5°), while the second was characterised by a long fixation, 260 to 280ms, followed by a saccade within the parafoveal region. This definition was then clarified by Pannasch et al. (2008). They defined the ambient visual mode as a fixation shorter than 180ms (Unema et al., 2005) followed by a saccade longer than 5° (Velichkovsky et al., 2005) and the focal visual mode as a fixation longer than 180ms followed by a saccade shorter than 5° of visual angle. The ambient mode would allow a wide exploration of the stimulus (dorsal stream), while the focal mode would focus on the information exploitation (ventral stream) (Tatler & Vincent, 2008).

This is the definition of visual modes we chose in this thesis when we describe and interpret the dynamic of eye movements on complex scenes, such as web pages.

### 1.3.3 The evaluation of visual modes

As we have seen in the previous section, the definition of visual modes' is based on a fixation duration and the following saccade amplitude Thus, FD and SA time courses must be assessed in order to determine which visual mode is being used. Nevertheless, the study of both parameters' time courses in multiple participants happens to be extremely complex and hard to interpret. For this reason, Krejtz et al. (2016) designed a synthetic ratio describing visual modes and their intensity: the $K$ coefficient. Even

though, this coefficient is the first one specifically created to describe ambient and focal visual modes, other attempts have been made. The oldest attempt was the ratio proposed by Goldberg and Kotval (1999). In their work, they used a ratio that represented the time spent processing (fixation) by the time spent searching (saccades). This ratio is anterior to the formal definition of visual modes by Velichkovsky et al. (2005), but it is clearly trying to synthesise a phenomenon related to ambient and focal modes. More recently, an enhanced version of this ratio has been developed by Dehais et al. (2015) in the context of surprise in a plane cockpit. They first divided fixations in short and long fixations. Then, short fixations were added up to saccades to represent the time spent scanning a visual scene; and long fixations represented the time spent processing the stimulus on their own.

All ratios presented here share the same approach and provide a global indicator describing which visual mode dominated the visual exploration. The main limit to such an approach, is that the use of a global ratio to describe a dynamic process causes a lost of a large amount of information. Hence, to better estimate the time course of ambient and focal visual modes, we decided to modified Dehais et al. (2015) ratio (**Poster 1**). Unfortunately, this ratio could not take saccade amplitude into account to correctly asses visual processing modes, so we switched to Krejtz et al. (2016) coefficient. Contrary to the first ratio, the K coefficient has been used to report the evolution of the visual modes using time bins. Nevertheless, because web pages are particularly complexes (see Chapter 2) and the duration of exploration important, we proposed to extend this approach by developing complementary K-based indicators (**Article 1** and **Poster 2**). Finally, to investigate if the ambient-focal dichotomy was relevant to other behaviours that occur during the exploration of a web page, we applied the K coefficient and our new indicators on computer mouse dynamic and scrolling (**Article 3**).

## Chapter 1 summary

**Selective attention** allows us to select the most relevant information, inhibiting distracting elements and enhancing the target to be aimed at. The visual information that reaches our retina is extremely rich. After a succession of steps, the visual information is conveyed into the primary visual cortex. It is then transmitted to different cortical areas in order to extract the object's properties or to perform more complex treatments such as *identification* (ventral pathway) or *spatial localisation* (dorsal pathway). These treatments are essentially possible when the object is located in the part of the eye with the highest visual acuity: the fovea. To do this, the eye performs a movement called **saccades** to bring the object of interest to the fovea. When we make a saccade towards an object, our attention is also shifted on it: this is the *overt attention*. These saccades can be triggered reactively or voluntarily and differ in their parameters of latency, direction and amplitude. After the execution of a saccade, a period of stabilisation follows, called **fixation**. Fixation is characterised by its duration and is interspersed with involuntary movements such as microaccades, drifts, and tremors. The various parameters of saccade and fixation mentioned above change over time and make it complex to study. Thus, a **dynamic approach** seems a more precise understanding of visual exploration. To this end, a series of studies have linked the ventral and dorsal pathways with eye movements by associating two visual modes: *ambient* and *focal*. In order to analyse these modes during exploration we used two ratios, the first one is presented in the **Poster 1** and the second one in the **Poster 2**, **Article 1** and **Article 3**.

## Résumé du chapitre 1

**L'attention sélective** permet de sélectionner les informations les plus pertinentes, en inhibant les éléments distracteurs tout en rehaussant la cible à viser. L'information visuelle qui parvient à notre rétine est extrêmement riche. Après une succession d'étapes, l'information visuelle est acheminée dans le cortex visuel primaire. Par la suite elle est transmise à différentes aires corticales afin d'extraire les propriétés de l'objet ou bien d'effectuer des traitements plus complexes tels que *l'identification* (voie ventrale) ou la *localisation spatiale* (voie dorsale). Ces traitements sont essentiellement possibles lorsque l'objet se situe sur la partie de l'oeil ayant la plus haute acuité visuelle : la fovéa. Pour ce faire, l'oeil effectue un mouvement appelé **saccades** permettant d'amener l'objet d'interêt sur la fovéa. Lorsque nous réalisons une saccade vers un objet, notre attention se déplace également sur ce dernier : c'est *l'attention overt*. Ces saccades peuvent être déclenchée de manière réactive ou volontaires et se distinguent dans leurs paramètres de latence, de direction et d'amplitude. Après l'exécution d'une saccade, une période de stabilisation s'ensuit, appelée **fixation**. La fixation se caractérise par sa durée et est entrecoupée de mouvements involontaires tels que les microsaccades, les dérives et les tremblements. Les différents paramètres de la saccade et de la fixation évoqués ci-dessus évoluent à travers le temps ce qui rend complexe leur étude. Ainsi, une **approche dynamique** semble nécessaire afin de mieux appréhender l'exploration visuelle. C'est dans ce but qu'un ensemble d'études a mis en lien les voies ventrales et dorsales avec les mouvements des yeux en y associant deux modes visuels : *ambiant* et *focal*. Afin d'analyser ces modes lors de l'exploration nous avons utilisé deux ratios, le premier est présenté dans le **Poster 1** et le second dans le **Poster 2**, **Article 1** et **Article 3**.

CHAPTER 2

# FACTORS INFLUENCING VISUAL EXPLORATION

## Contents

In Chapter 1, we described how the visual system works and more globally the different mechanisms involved in vision. We also detailed how attention is related to vision and more specifically to eye movements. One of the goals of this thesis is to investigate eye movement behaviour when exploring a web page. A behaviour can be defined as the "*Manner of being and acting of Animals and Humans, objective manifestations of their global activity*" (p. 153 Piéron, 1994). In the context of this thesis, it corresponds to the distinctive characteristics of the visual exploration, defined by factors as fixation duration, saccade amplitude, etc., that change during the exploration of a web page. But, before considering visual behaviour on web pages (Section 2.2), we will first outline factors impacting eye movements on natural visual scenes (Section 2.1).

## 2.1 Eye movements' guidance

Factors that can modulate visual attention and eye movement behaviours can be categorised in three main types. The first are oculomotor biases which are inherent to eye movements and most of the time independent of the stimulus type. The second category, called bottom-up, includes factors directly related to the stimulus itself. The third and last category designates the top-down factors reflecting the cognitive mechanisms involved during visual exploration.

### 2.1.1 Oculomotor biases

As described in Section 1.2.2, the human eye is an organ moving inside the ocular orbit by the means of extraocular muscles. Although we have the sensation to move our eyes almost freely, they remain under physiological constraints arising from these muscles. Oculomotor biases designate the combination of these physical constraints and the behaviours we learned through human evolution. Tatler and Vincent (2009) reviewed in their work some of these oculomotor biases, unrelated to stimulus attributes:

- **Saccade amplitudes**: saccade amplitudes follow a positively-skewed distribution which means that we tend to make short saccades more often than long saccades

(e.g. Gajewski et al., 2005; Tatler et al., 2006). In ecological situations, saccades average amplitude is lower than 15 degrees of visual angle (Gilchrist, 2011). Microsaccades can even be much smaller with amplitudes shorter than 1 degree (Martinez-Conde et al., 2009). Beyond 15 degrees of visual angle, eye movements are usually executed with head movements. It should be noted however that in laboratory, saccades can reach amplitudes up to 40 degrees without head movements. However, saccade amplitudes also depends on their direction. For instance, in addition of being more frequent, horizontal saccades are also longer (Tatler & Vincent, 2008). Moreover, as described in Section 1.3.1, saccade amplitudes are longer at the beginning of the exploration and decrease over time. The amplitude of the current saccade is also related to the previous fixation duration.

- **Saccade directions**: saccades are not uniformly distributed in space. We tend to execute more horizontal saccades than vertical ones and even fewer oblique saccades (e.g. Brandt, 1945; Foulsham et al., 2008; Gilchrist & Harvey, 2006). Although there is no consensus yet in the literature, extraocular muscles are commonly found responsible. As described in Section 1.2.2, whereas only a single pair of muscles is necessary to move our eyes horizontally, two pairs of muscles are required to move the eyes vertically and obliquely. Foulsham et al. (2008) ran two experiments to try to understand the prominence of horizontal saccades. To do so, they showed to participants rotated stimuli. Contrary to the oculomotor-biased theory, they found that participants could easily make vertical or oblique saccades when the stimuli were rotated. They suggested then that previously learned behaviours could be accountable for the predominance of horizontal saccades. These bias could be learned through our environment or culture, such as the importance of the horizon in our evolution or the reading direction.

- **Fixation durations**: several authors noted an evolution of fixations durations over time (e.g. Antes, 1974; Buswell, 1935; Pannasch et al., 2008; Unema et al., 2005). Contrary to saccade amplitude, fixation duration increases over time and

seems to be related to the next saccade amplitude (see Section 1.3 for more details).

Oculomotor biases have a strong influence on visual exploration. Hence, they need to be taken into account when analysing eye movements. That is why, we reported saccade amplitudes and fixation durations globally and dynamically in our experiments presented (**Posters 1, 2, 3** and **Articles 1and 3**). Furthermore, we exploited our findings on these parameters to build our scanpath model described in Chapter 8. This model also takes advantage of saccade orientation time course.

### 2.1.2 Bottom-up

Contrary to oculomotor biases, bottom-up factors are stimulus-dependent. Although some authors have questioned the existence of purely bottom-up influence during visual exploration, it seems that the characteristics of stimuli play an important role in the attention guidance (Theeuwes & Failing, 2020). Numerous models uniquely based on bottom-up features have been created the last 20 years and have been able to get close to human performance. Below, are presented the main features influencing visual exploration and some of the models will be described later in Chapter 3.

- **Basic features**: our understanding of the impact of colour, orientation, luminance and shape on visual attention is closely related to Treisman and Gelade (1980) feature integration theory. They proposed a two-stage architecture of attention in which basic features are first processed in parallel during a pre-attentive phase, to be then processed at a higher level of the vision process, to recognise objects for instance. The feature integration theory led to a wide variety of works including Wolfe and Horowitz (2004), Itti et al. (1998) and more recently Theeuwes and Failing (2020) work.

- **Image size**: von Wartburg et al. (2007) found a positive correlation between mean and median saccade amplitudes with image size. They suggested that stimuli should be at least 20 degrees of visual angle and should not exceed 40 degrees in head-mounted design.

- **Edges**: edges have been reported to be more predictive of fixation spatial behaviour than luminance or contrast (Baddeley & Tatler, 2006). Moreover, Tatler and Vincent (2009) compared a model uniquely based on edges with the Itti et al. (1998) model, and obtained better performances.

- **Faces**: Morrisey et al. (2019) recently showed that faces were fixated first more often than other images (e.g., cars or birds, etc.), and were detected faster than other objects when they were the target. In their experiment, fixations on faces were short indicating that they were processed more efficiently than other stimuli. However, known and unknown faces should be differentiated, the bottom-up features of faces described here mainly applies to unknown faces. The recognition of known faces implies top-down processes which are not covered here.

- **Centre-bias**: first mentioned by Buswell (1935), the centre bias has been observed in numerous studies ever since (Parkhurst et al., 2002; Rothkegel et al., 2017; Tatler, 2007). When exploring visual stimuli, participants tend to look at the centre of the image. The reason for this phenomenon is still discussed but Tatler (2007), and Rothkegel et al. (2017), proposed interesting arguments. Tatler (2007), refuted the idea that low-level salient features are more frequently at the centre of the image and rather suggested that the centre of the screen may be an optimal viewing position. This could be explained by extraocular muscles: the muscle tensions are minimal when the eyes are in the centre of the orbit. Another theory suggests that the centre of an image is actually the optimal position to grasp the entire visual scene, even if not in details. Rothkegel et al. (2017) manipulated the starting position of the eyes and the latency of the initial saccade to investigate the centre bias. They found that the first Tatler (2007) proposition was not possible since the centre bias was decreased with the increase of saccade latency. They proposed a new explanation: the sudden luminance change due to the stimulus appearance may be treated as an object, thus the eyes would be attracted toward the centre.

The "*Feature Integration Theory*" from Treisman and Gelade (1980) led to numerous saliency models which confirmed the importance of basic features. Furthermore, as it will be described in Chapter 4, we used a similar paradigm than Rothkegel et al. (2017) to minimise centre bias (**Poster 1, 2and 3** and **Articles 2and 3**).

### 2.1.3 Top-down

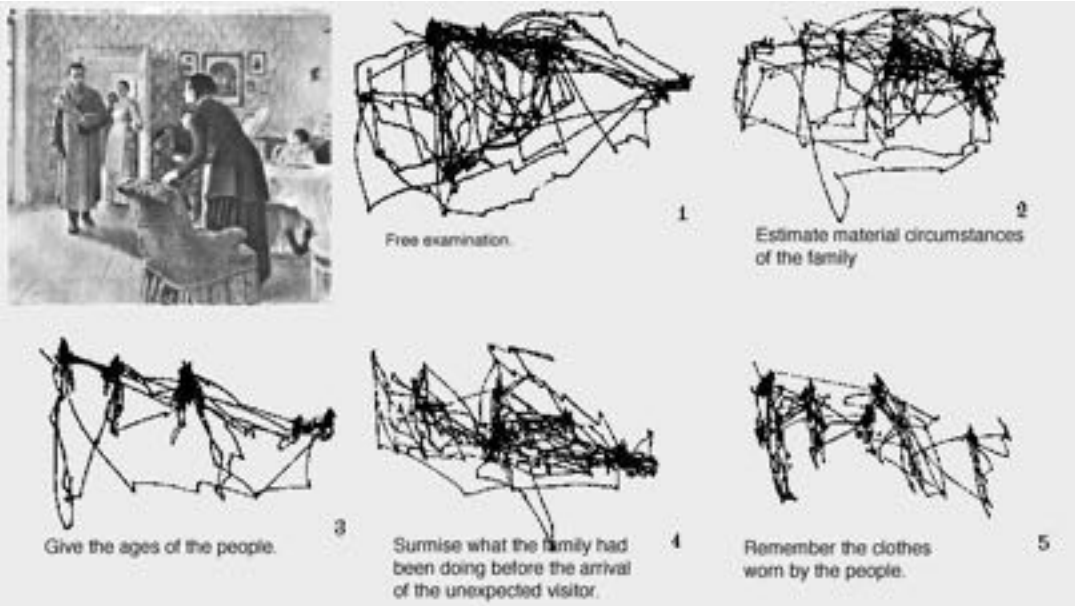



Figure 2.1 – **Task influence on visual exploration**. Visual explorations of participants looking at "The unexpected visitor" painting, according to a given task. From Yarbus (1967).

It is generally assumed that the interaction between bottom-up and top-down factors influence how we orient our attention. In that sense, top-down factors are usually addressed as factors influencing bottom-up ones and are not considered as totally distinct factors. Top-down factors include a wide variety of cognitive processes impacting visual exploration. These processes are mainly related to context, memory, emotions, task, etc. A selection of most common top-down factors is presented below:

- **Age**: young adults make shorter fixations than children. Hence, fixation durations are negatively correlated with age (Helo et al., 2014). For Helo and collaborators, it

could be that fixation duration is related to task complexity (see **Task** below), and because cognitive resources increase with age, children would have less cognitive resources available to process visual information and will take more time. Moreover, Munoz et al. (1998) observed strong age-related effects in performance. They recorded longer saccade durations in elderly subjects than in teenagers ones.

- **Inhibition Of Return (IOR)**: because of IOR, as described in Section 1.1.3, when a specific location has already been, it is less likely to be fixated again for a certain period of time.

- **Task**: Yarbus (1967) showed that it is possible to obtain similar fixation location and saccade amplitude patterns between participants for a same task (see Figure 2.1). For instance, participants with the task to estimate the age of people in the *An Unexpected Visitor* painting, mostly looked at faces. Moreover, according to Mills et al. (2011) fixation durations are longer in free-viewing tasks and information retention tasks than in visual search tasks. In another experiment, Castelhano et al. (2009) have shown that in an information retention task, participants tended to fixate in a more scattered way. Moreover, Soh et al. (2012) and Sharafi et al., 2013 noted that fixation durations could be used to predict task complexity or the intensity of visual processing. Thus, fixation duration may be linked to the difficulty to extract visual information. More globally, Tatler and Vincent (2009) reported empirical evidence of the effect of the task on saccades amplitudes (e.g. Rayner et al., 2007; Tatler et al., 2006).

The broad variety of top-down factors make it difficult to take all of these factors into account when trying to predict where attention will be oriented. Even though top-down models exist, they are far from covering all the endogenous aspects of attention. Most of the time, top-down factors depend on the goal of the study used for acquiring data. Hence, top-down attention cannot be modelled in a single approach as bottom-up attention. In Chapter 3 we describe some top-down models and how IOR is modelled on top of these pre-existing models. Moreover, when we study the effect of the task on

visual exploration we usually compare behaviours on free-viewing and searching tasks. We also investigated eye movements during a visual search task and a free-viewing task (**Article 1 and 3** and **Poster 3**).

## 2.2 Web pages

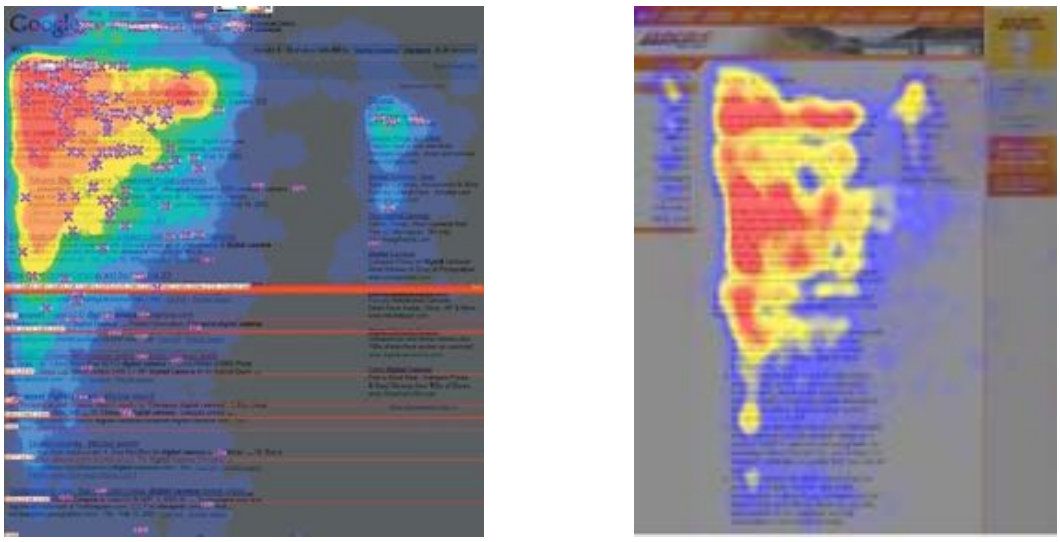
Previously described oculomotor biases are robust across stimuli. Thus, they remain relevant in the understanding of visual exploration behaviours on web pages. In addition to these generic biases, people developed new strategies in order to adapt to this new stimulus type. Some of these biases are described in this chapter. Contrary to static images, a web page is interactive. The use of a computer mouse is required and provides new inputs on exploration behaviour through its movements. This is why, a lot of studies focus on the relationship between mouse movements and eye movements. Interestingly, studies try to predict eye movement locations by using this specific relationship. While this section focuses on this relationship, the use of the mouse as a predictor of eye movements will be tackled in the Chapter 3.

**Websites and web pages**:
a web page is at the core of the web. It is generally what we all see when surfing the internet. Several types of web pages exist, to name a few: Search Engine Result Page (SERP), news, blog, social network, etc. A website is a group of web pages interrelated by hypertext available at the same address, for instance, the www.paris.fr website.

### 2.2.1 Eye movements on web pages

Although bottom-up approach is valid on web pages, Still and Masciocchi (2010) pointed out that most of web-specific biases are top-down. The biases described here are mainly related to learned behaviours. Indeed, because web pages often follow a similar template: a header with main sections of a website, a content with left or right bar and a footer at the end of the web page, users develop strategies to maximise their efficiency in visual

exploration (Buscher et al., 2009). A selection of most common web biases is presented below:



(a) Golden triangle (b) F-shape

Figure 2.2 – **Common web viewing patterns**. (a) Golden triangle represents the area where most of the visual attention is attracted when using Google search engine. (b) F-shape pattern occurs when browsing, for a large variety of web pages. Adapted from Hotchkiss et al. (2005) and Nielsen (2006).

- **Task**: as for natural images, multiple studies found an influence of the task on web pages browsing (Buscher et al., 2009; Cutrell & Guan, 2007).

- **Ad blindness**: as in the real world, advertisers exploit every available technique to capture the users' attention. The famous 'popup' ad or window is a perfect example. This type of solicitation is only based on bottom-up factors, such as sudden appearance, motion, colour, etc. Nowadays, these little "epileptic blinking windows" or banners no longer exist but had given way to more subtle techniques. For instance, in-text ads looking exactly as the content to misguide the user to read it by accident. Nonetheless, whatever the technique is, popup or a more advanced technique, users developed what is called 'ads blindness' (Benway, 1998). This term designates strategies specific to web content we developed to avoid looking at

advertisements. Ad blindness occurs after repeated exposition to advertisements (Cho & Cheon, 2004; Hervet et al., 2011).

- **Left side**: Nielsen (2010) observed that users tend to spend more time on the left part of a web page than on the right part. He also observed this behaviour on right-to-left reading web pages. A more recent study from Fessenden (2017) showed a similar behaviour on SERP but stronger.

- **F-shape/Golden triangle**: they are viewing patterns specific to web specific (see Figure 2.2). The *F-shaped* bias for classic web pages and the *Golden triangle* bias for web pages. Nielsen (2006) ran a usability experiment during which he analysed which part of a web page users were looking at. He observed a recurring viewing pattern in the shape of the F letter. People started their browsing at the top-left corner of the web pages and read horizontally, then they were scrolling down to read a second time horizontally to finally scan the content vertically. Similar behaviour has been reported on Google's results. A golden triangle has been observed by Hotchkiss et al. (2005), describing the areas where search results are the most visible and visited.

### 2.2.2 Scroll and eye movements

Scrolling consists in moving the web page up, down, left or right to see hidden content. Nowadays, its common to visit a web page that needs scrolling to be fully explored. Thus, scrolling is essential when browsing a web page. As described in Braganza et al. (2009), the action of scrolling may be executed in various ways which can be categorised as follows: scrolling with the mouse and scrolling with the keyboard. Scrolling with the mouse includes grabbing and dragging the content or the scrollbar (on the right of the browser) and using mouse's wheel which is the most observed behaviour. In ecological conditions, except on some specific occasions, the keyboard is rarely used. Contrary to the mouse, scrolling can either be used on desktop computers, laptops, smartphones or

tablets. Studying the behaviour related to scrolling is of great interest, because findings on that specific subject might be, in some ways, applicable to multiple devices.

In their work about scroll and reading on web pages, Braganza et al. (2009) showed that participants used a scrolling strategy to minimise vertical eye movements and that fixation were more likely to be made at the bottom part of the web page. In a study on Search Engine Result Page (SERP), Liu et al. (2017) demonstrated that users' strategies were not sensitive to time constraint when browsing. Since scrolling is used in most cases, the study of its behaviour is of primary interest. However, studies mostly focus on scrolling behaviour when reading and does not tackle quantitative analyses (see Dyson (2004) for a review). This is why, we propose the first detailed quantitative analysis of scrolling behaviour (**Article 3**), and introduce a new definition of scrolls and how they can be differentiated. Finally, we analysed eye movement behaviours during scrolling in various conditions (**Poster 3** and **Article 2**).

### 2.2.3 Mouse and eye movements

Chen and Sohn (2001) were the first to specifically investigate the relationship between a computer mouse and the movements of the eyes. In addition to a better understanding of this relationship, their goal was to predict eye movements based on mouse movements. This substitution would allow researchers to collect and process a large amount of data remotely, without interrupting the user. In their work, Chen and Sohn (2001) showed that the presence of the cursor in a particular region of the screen correlated with the probability for the participant to fixate this region. Rodden and Fu (2007) observed that the coordination between the eyes and the computer mouse was higher on the vertical axis of the screen than on the horizontal axis. This behaviour was also observed by Guo and Agichtein (2010). This relationship however remains uncertain, considering that the mouse could be used as a means to mark a potential result previously located with the eyes (Rodden et al., 2008). Furthermore, the amount of time spent by a user on an SERP could affect the gaze and mouse location alignment during the exploration (Huang et al., 2012). Navalpakkam et al. (2013) ran an experiment on non-linear page layouts and

showed that the correlation between the eyes and the mouse was non-linear and user dependent. More specifically, this correlation has been found for time periods during which a user looked at an Area Of Interest (AOI) and when switched between AOIs. However, SERPs are not representative of the web and remain transitional web pages to access a content on a different website. Boi et al. (2016) showed that users spent a significant cumulative amount of time on SERPs, but a cumulative way, in short bursts of time.

Obviously, the study of the relationship between eye movements and mouse movements interested search engines companies, as Google and Microsoft, a lot. It is partly why a vast majority of studies have been done on Search Engine Result Page (SERP). The problem is that those web pages are not representative of the web. We use search engines on a daily basis to search for very different queries from our personal to our professional lives. However, even if we can spend a significant cumulated time on these web pages at the end of the day, we only spent a few consecutive seconds per query. It is why, it is far more interesting and informative to investigate the relationship between eye movements and mouse movements when browsing web pages (**Article 2** and **Article 3**). Furthermore, there are a number of well-established, and ever improving, methods to label raw data from eye recordings. However, mouse and scroll recordings lack such a method, specifically to differentiate two close events. While it is easy to determine if two events separated by two or three seconds are indeed two distinct events, doing the same operation for two events with, for instance, less than one second in between, is much harder. In **Article 3** we propose a new segmentation threshold based on statistics and observed behaviour. This method provided us better data quality to perform dynamic analyses. As eye movements, the study of mouse movements and scrolling focuses on global behaviour or patterns. In **Article 3**, we propose to apply eye movements dynamic analyses to the mouse to better describe how the mouse and the scroll are used.

## Chapter 2 summary

As visual attention, eye movements can be influenced by different types of factors divided in three categories. The first type of factor is **oculomotor bias**. These are a combination of physiological constraints and behaviours learned over the course of evolution. The second factor, known as **bottom-up**, groups together elements specific to the stimulus. The last and third factor, called **top-down** factor, refers to the influence of cognitive processes associated with the context. These top-down factors may also depend on the type of stimulus presented. For instance, web pages follow a similar organisation through the web. This organisation present on the majority of websites has led users to develop strategies to optimise their visual exploration. These are called **web-specific biases**. However, web stimuli have a particularity that distinguishes them from a classic image: they are *interactive* through the mouse. This interactivity has led research to try to better understand the visual behaviour during events such as page scrolling or mouse movement. These aspects are discussed in **Articles 2 and 3** and **Poster 3**. In this sense, the study of mouse movement quickly focused on the coordination between the eye and the mouse in order to try to predict eye movements according to the position of the cursor. The study of this coordination will be discussed in **Article 2**.

## Résumé du chapitre 2

Comme l'attention visuelle, l'exploration oculaire peut-être influencée par différents types de facteurs que l'on peut classer en trois catégories. Les **biais oculomoteurs** constituent le premier type de facteurs influançant cette exploration. Ces derniers sont une combinaison de contraintes physiologiques et de comportements appris au cours de l'évolution. Le second facteur, dit **ascendant**, regroupe les éléments propre au stimulus. Le dernier et troisième facteur, appelé **descendant**, désigne l'influence des processus cognitifs associés au contexte. Ces facteurs descendants peuvent également dépendre du type de stimulus présenté. Par exemple, les pages webs suivent une organisation similaire à travers le web. Cette organisation présente sur la majorité des sites webs a mené les utilisateurs à développer des stratégies afin d'optimiser leur exploration visuelle. On parle alors de **biais spécifiques au web**. Toutefois, les stimuli web possèdent une particularité qui les distingue d'une image classique : ils sont *intéractifs* au travers de la souris. Cette intéractivité a mené les recherches à tenter de mieux comprendre le comportement visuel lors d'évènements tels que le défilement de la page ou le mouvement de la souris. Ces aspects sont abordés dans les **Articles 2 et 3** ainsi que dans le **Poster 3**. En ce sens, l'étude du mouvement de la souris s'est rapidement concentrée sur la coordination entre l'oeil et la souris afin d'essayer de prédire les mouvements des yeux en fonction de la position du curseur. L'étude de cette coordination sera abordée de manière détaillée dans l'**Article 2**.

CHAPTER 3

# SCANPATH AND OCULOMOTOR BIASES MODELLING

## Contents

The eyes are at the core of visual perception: through fixations and saccades, they allow us to acquire information efficiently. The question of where the eyes are located during visual exploration is crucial and gave rise to numerous theories and models. Within the engineering field, these models can be categorised as either saliency models or saccadic models. Saliency models consist in providing a map, similar to a heatmap, that represent pixel per pixel the likelihood that the eyes will be located there. This *Saliency map* holds gaze predictions for the exploration of an entire stimulus independently of the exposure duration. Saccadic models are more about scanning patterns (*Scanpath*). Their goal is to predict in addition to eye locations, the order in which eye fixations will occurr. *Scanpath* holds predictions of the exact coordinates and order of the successive eye fixations for a given stimulus,and is dependent of participants' exposure duration.

Our work focuses on the improvement of saccadic models on web pages. Yet, since modern saccadic models are all based on a saliency map, we will first introduce some bottom-up and top-down computational models. In the previous chapters, we outlined how the visual system works and what biases could influence visual exploration. In this chapter, we will describe how these biases are modelled and how the dynamics of these parameters could be taken into account to improve saccadic models. We will then review the principal saccadic models and how oculomotor biases are modelled. Then, we will assess how the dynamics of these biases could be taken into account to improve saccadic models. Finally, we will review how each approach tackle saliency and scanpath modelling on web pages.

## 3.1 Saliency models

Visual attention and eye movements are influenced by a combination of bottom-up and top-down factors, such as colour, edges, or task (see Chapter 2). Some evidence has shown that the influence of both types of factors fluctuates during visual exploration (Itti & Borji, 2015; Theeuwes & Failing, 2020). The key point of predicting visual attention is

thus to correctly weight each factor depending on the visual context. Because bottom-up factors, based on stimulus' characteristics, are easily accessible, numerous models based on a bottom-up approach of visual exploration have been developed through the last 20 years. Top-down models are less popular and more complex but recently renewed attention has been given to these models. Both types of saliency models are relevant and of interest since most of recent saccadic models can either take a stimulus and/or a saliency map as an input.

### 3.1.1 Bottom-up models

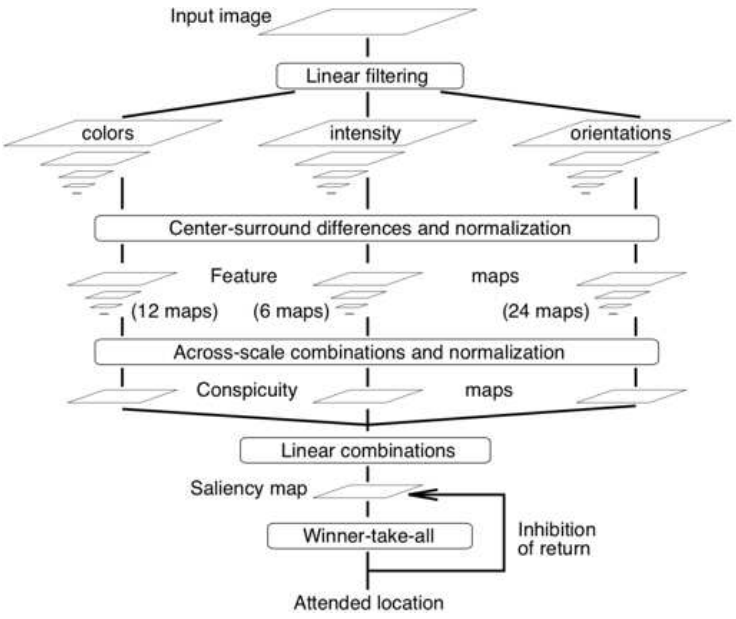



Figure 3.1 – **Overview of the Feature-based Saliency Model**. Features are extracted from the image to be transformed into multi-scale Gaussian pyramids. Then a centre-surround mechanism is applied to each pyramid. Finally, scales of each pyramid are combined and the resulting three maps are again combined. The result is a saliency map. Adapted From Itti et al. (1998).

Early work on computational saliency models was based on Koch and Ullman's framework 1985. In their seminal work, they proposed a biologically plausible framework of Treisman and Gelade (1980) "*Feature Integration Theory*" and introduced the concept of *saliency maps*. They also proposed the Winner-Take-All (WTA) algorithm and implemented the Inhibition Of Return (IOR) phenomenon to select the most salient

locations of a visual scene. The WTA algorithm consists of selecting the most probable location as the solution, while IOR allows the inhibition of the already selected peak to prevent the WTA algorithm from selecting the same location twice. The first complete implementation of this framework was proposed by Itti et al. (1998) on natural scene still images. While, many models are based on the implementation of Itti et al. (1998) implementation (Borji & Itti, 2013), other approaches have since emerged. For instance, models have been proposed based on Information Theory, Bayesian statistics or Pattern Classifications (see Borji and Itti (2013) for an extensive review).

In their "*Feature-based Saliency Model*" (see Figure 3.1), Itti et al. (1998) extract three types of features from the image: colours, intensity and orientations. Each feature is subsampled in Gaussian pyramids. Then centre-surround filter is applied on each pyramid to construct new features maps. These maps are then summed across scales and normalised, which gives a conspicuity map. These three conspicuity maps are finally linearly combined to give a saliency map.

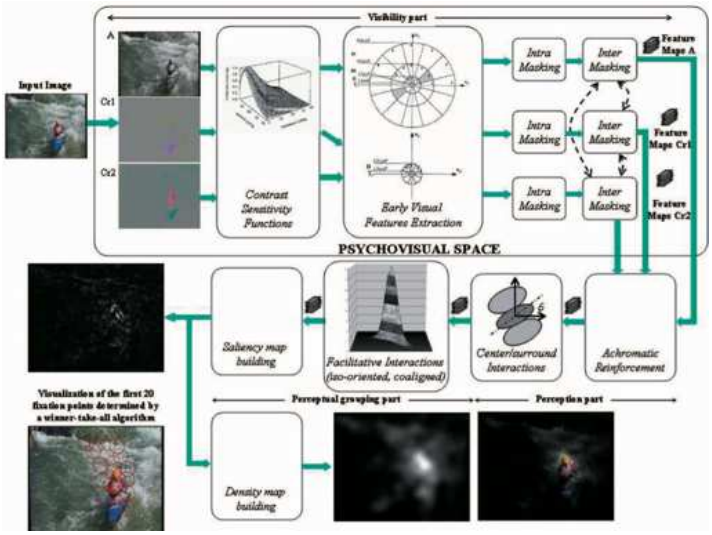



Figure 3.2 – **Overview of the Coherent Computational Saliency Approach**. Image is split into 3 different colour channels. Then visibility, perception and perceptual grouping are performed on each channel to build a coherent saliency map. From Le Meur et al. (2006).

Then, Le Meur et al. (2006) extended this model by tackling attention prediction using mechanisms directly inspired from the Human Visual System (HVS): visibility,

perception, and perceptual grouping. As depicted in Figure 3.2, the image is first separated using Krauskopf's colour space to reflect the three types of cones handling different wavelengths (L-cones, M-cones, and S-cones). Visual features are then extracted from the channels. The role of the visibility part is to simulate the limited sensitivity of our human visual system. The perception phase's goal is to suppress the redundancy of visual information through centre-surround and achromatic mechanisms. Finally, perceptual grouping refers to the ability to group and bind visual features to build a meaningful saliency map.

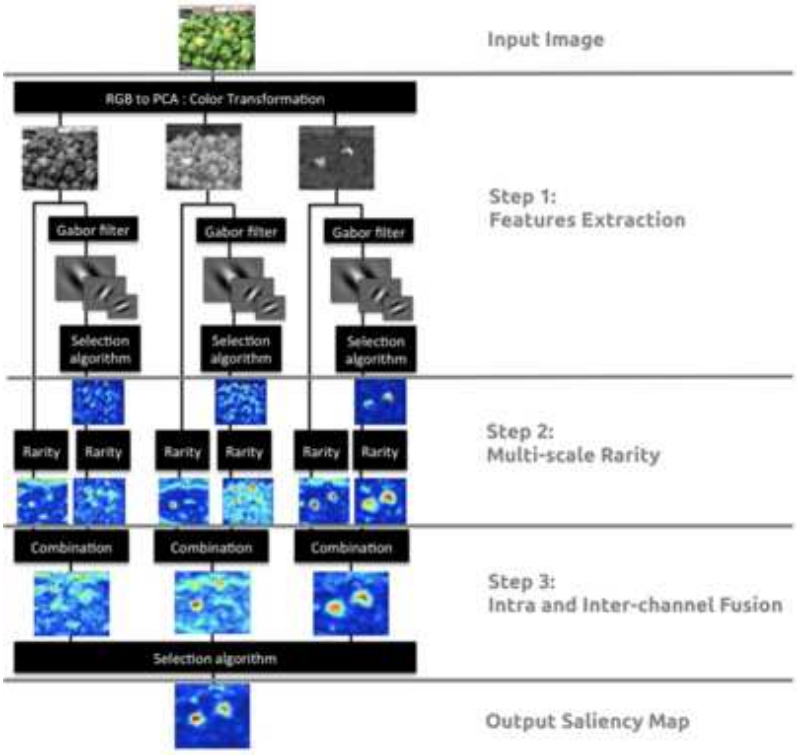



Figure 3.3 – **Overview of the RARE2012 saliency model**. Image is split into 3 different colour channels through PCA. Then a Gabor filter is applied to each channel to detect edges and textures. Then rarity is computed to finally combine all the channels into a single saliency map. From Riche et al. (2013).

Finally, a saliency model based on information theory to find the rare features in a stimulus and predict visual fixations have been proposed by Riche et al. (2013) (see Figure 3.3). Interestingly, they consider that a feature is not necessarily salient alone, but only in a specific context. They thus proposed a multi-scale rarity mechanism to

detect locally contrasted and globally rare features.

During the first stages of the model, low-level features (colours) and then medium-level features (orientations and textures) are extracted. In a first step, authors used a Principal Component Analysis (PCA) to convert the image (RGB) into an alternative colour space (YCbCr) and decompose the image into luminance (Y) and chrominance (Cr and Cb) maps. From there, the model splits in two pathways: the first one use the PCA-based colour transformation to compute its rarity, and the second apply Gabor filters for 3 scales and 8 orientations to extract orientation features' maps. A multi-scale rarity mechanism is then applied. At the end, six rarity maps (3 colour maps and 3 orientation and texture maps) are fused together into a unique saliency map.

### 3.1.2 Top-down models

Attention can either be directed exogenous or endogenous (see Chapter 1). Exogenous attention (or bottom-up) models exclusively focus on predicting salient elements based on stimulus characteristics. This approach can be assimilated to the prediction of attentional capture. However, saliency models trying to predict endogenous attention (or top-down) are based on the view that attention is the result of the modulation of stimuli characteristics by top-down factors (Folk et al., 1992). That is why the majority of top-down models are built on top of bottom-up ones. For instance, Wolfe (1994) presented the Guided Search Model based on Koch and Ullman (1985) framework. In their model, Wolfe (1994) manually defines weights depending on the task to adjust the importance of bottom-up features. According to Borji and Itti (2013), the main problems tackled by top-down models are visual search prediction, the role of scene context and scene layout.

**Deep Neural Network (DNN)**: designates a specific type of artificial neural network with multiple layers of artificial interconnected neurons. The role of an artificial neural network is to simulate how neurons work to solve diverse problems. To do so, the neural network is trained with labelled data and then tested with unknown data. Contrary to classic approaches, important features, their transformation or reduction and their contribution to the result are automatically computed by the neural network.

### 3.1.3 Saliency attentive model

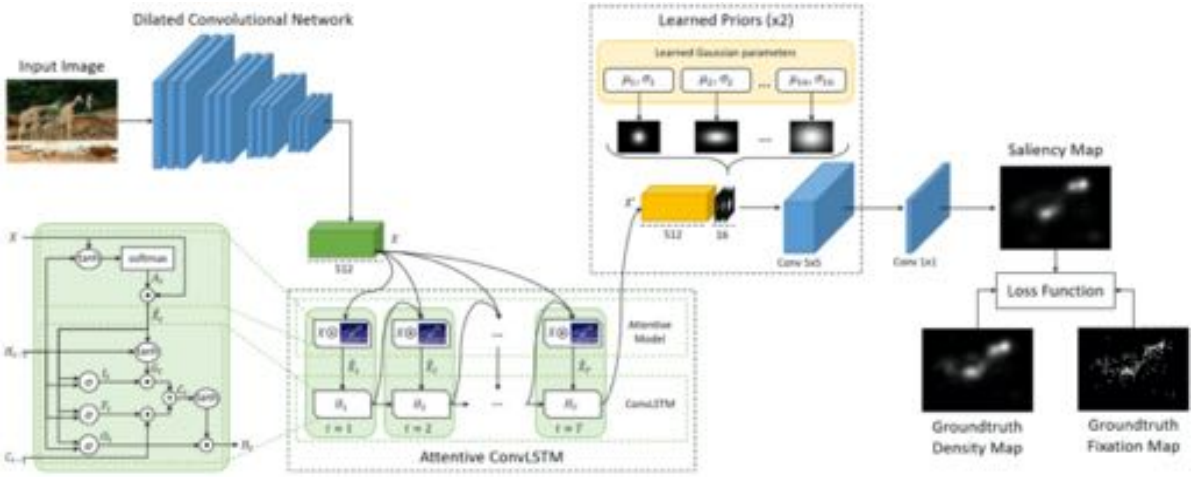



Figure 3.4 – **Overview of Saliency Attentive Model (SAM)**. A Convolutional Network extracts feature maps from the input image, then an Attentive Convolutional LSTM enhances saliency features. Then 16 Gaussian maps simulating centre-bias are combined with previous output. From Cornia et al. (2018).

More recently, advances on Deep Neural Network (DNN) disrupted saliency modelling to the extent that today the top 10 best saliency models from the MIT Benchmark are exclusively DNN models (Kümmerer et al., 2020). The interest of such models is to provide an end-to-end mechanism aware of bottom-up and top-down features. Usually, DNN saliency models are built on top of pre-trained models already able to recognise certain classes of objects (Borji & Itti, 2013). Then the model is augmented with a specific algorithm or information specific to saliency and trained on one or multiple image datasets. Although, DNNs provide new promising results; and have not yet revealed their full potential, some limitations need to be considered.

Cornia et al. (2018) based their model on VGG-16 and ResNet-50 pre-trained models and a modified version of Long Short-Term Memory (LSTM) (see Figure 3.4). VGG-16 and Resnet-50 are both Convolutional Neural Networks able to classify objects on visual scenes. They respectively have 16 and 50 layers in their deep architecture. The interest of using these models is that they have already been trained to recognise objects, so they have already learned the necessary bottom-up features of a wide variety of natural scenes. On top of the saliency part of the Saliency Attentive Model (SAM), there is another kind

of DNN: a Long Short-Term Memory (LSTM). Usually, this type of neural network is used on tasks involving time dependencies and cannot be employed for spatial tasks. Cornia et al. (2018) modified the LSTM network to process features coming from VGG-16 and ResNet-50 iteratively instead of using the model to process the same input changing over time. The iterative processing of salient feature is similar to how attention can be orientated, that is why they called their modified version an "*Attentive ConvLSTM*". Finally, they trained a network to learn the parameters of 16 2D Gaussian functions to simulate centre-bias. These 16 centre bias maps are combined with the output of the Attentive ConvLSTM to produce the final saliency map. Cornia et al. (2018) trained their model using a combination of well known saliency metrics: NSS, CC and KL-Div.

Kummerer et al. (2017) showed that these types of models underestimate bottom-up factors. In their study, they demonstrated that, when applied to datasets with limited top-down elements, state-of-the-art DNN-based models were outperformed by bottom-up ones. Moreover, Kong et al. (2018) quantified what objects were contributing the most to top-down performance. They found that the detection of faces, text and animals were explaining a consistent proportion of DNN-based model performances. They also were able to outperform state-of-the-art models by augmenting a bottom-up model with these features.

### 3.1.4 Web page saliency

A few models are specialised in web content, but these follow the same approach as previous models: combine bottom-up and top-down approaches. Contrary to natural images, web pages are much richer in visual media, text, logos, etc. That is why, specific modelling and biases need to be taken into account. Buscher et al. (2009) performed a linear regression on features extracted from web page contents and used a decision tree to predict visual attention. Shen and Zhao (2014) proposed a revisited version of the Itti-Koch model (Koch & Ullman, 1985) using a new colour-space and web-dependant features. The image is first converted into a Derrington-Krauskopf-Lennie colour-space to extract intensity, colour and orientation. The interest of such a colour space compared

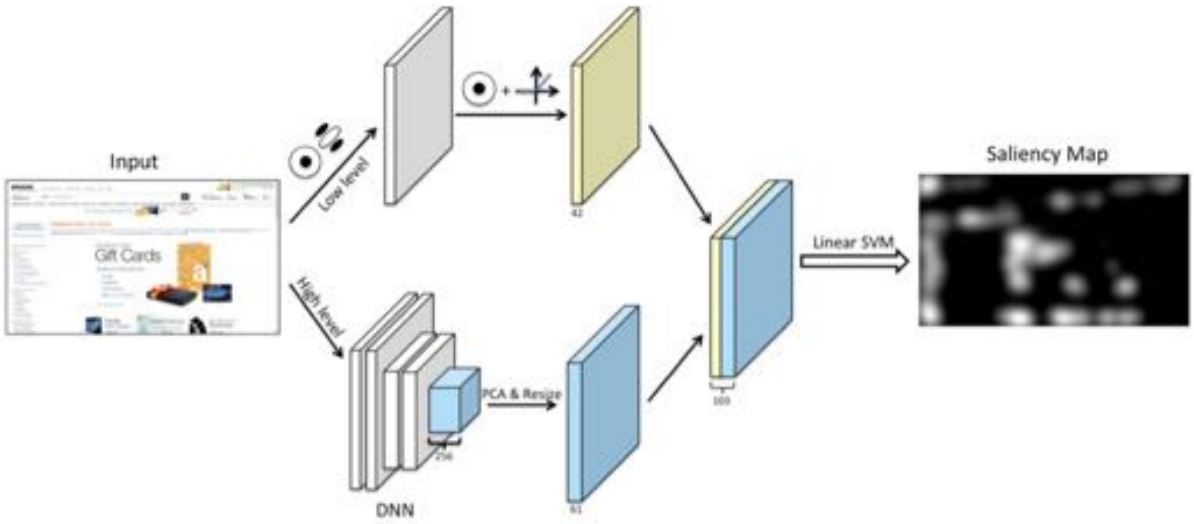



Figure 3.5 – **Overview of the web page saliency model**. From Shen et al. (2015).

to the RGB one, is that each channel reflect the three types of cones handling different wavelengths (L-cones, M-cones, and S-cones). Then, as in the seminal Koch and Ullman (1985) work, Gabor filters, Gaussian pyramids and centre-surround mechanisms are applied to each feature. Afterwards a face and body detector is used, as their data suggest that body parts attract more attention. Next, a positional bias is applied as two maps: one for the top-left bias, which represents the tendency to look at the top-left of a web page, and the centre-bias (see Chapter 2). Finally, they used Multiple Kernel Learning, which is an algorithm combining multiple kernels of support vector machines (SVMs) instead of one, to combine all feature maps. In a later work, Shen et al. (2015) extended their model by adding a Deep Learning approach. As depicted in Figure 3.5, low-level features are processed with a similar approach to the previous work of Shen and Zhao (2014), while high-level features are extracted using a DNN. The neural network used here is AlexNet (Krizhevsky et al., 2017), the predecessor of VGG architecture (Simonyan & Zisserman, 2014). Again, this DNN was trained to differentiate numerous objects on a wide dataset of natural images. The output of low-level and high-level paths are then combined using a linear support vector machine (SVM).

## 3.2 Saccadic models

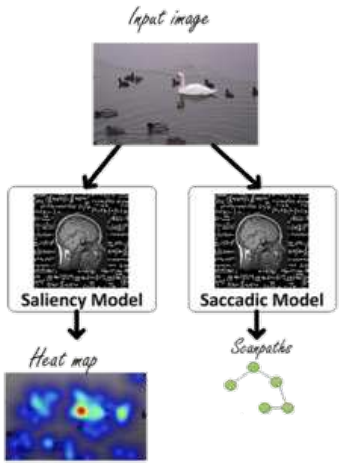



Figure 3.6 – **Differences between saliency and scanpaths models**. A saliency model outputs a saliency map summarising where attention has been oriented. Scanpath model outputs an ordered sequence of fixations of precise locations attended. Adapted From Le Meur et al. (2017).

Saliency models provide average visual attention performances of an image (see Figure 3.6). That is why saliency models are compared to the behaviour of a group of participants (Madelain & Chauvin, 2007, p. 214). Contrary to such models, saccadic modelling is more about the prediction of which locations are fixated on, in which order and sometimes for how long (see Figure 3.6). One of the points of interest of these models is the ability to convert a scanpath back into a saliency map and thus enhance bottom-up or top-down saliency models used for scanpath prediction. Thus, it is common to see scanpath models compared to saliency models through this technique in literature.

Based on the literature, a saccadic model can be synthesised as 3 complementary but essential modules (seeFigure 3.7). The first module's role is to handle the features extraction which can be either bottom-up, top-down or a mix of the two. Most of the time this part is performed by a saliency map which comes from an already existing saliency model (e.g. Boccignone & Ferraro, 2004; Le Meur & Liu, 2015; Wang et al., 2016). The second module models the oculomotor biases mechanisms. As previously described,

biases influence how the next fixation will be selected. Which bias is modelled varies a lot from one model to another, but most of the time IOR and saccade amplitude are used. Finally, the fixation selection module is the one that attracts the most attention in saccadic modelling. Its role is to take the two previous modules to select the next fixation. For instance, in Itti et al. (1998), this part is handled by the WTA algorithm which uses the generated saliency map and IOR to determine the next fixation. The next fixation selection has been widely studied with various approaches including, Markov processes (e.g. Coutrot et al., 2017; Hacisalihzade et al., 1992; Mannaru et al., 2016), Neural Networks (e.g. Chen & Sun, 2018; Ngo & Manjunath, 2017; Simon et al., 2016; Zhang et al., 2011) or Bayesian statistics (e.g. Le Meur & Coutrot, 2016; Le Meur et al., 2017; Le Meur & Liu, 2015; Wang et al., 2016).

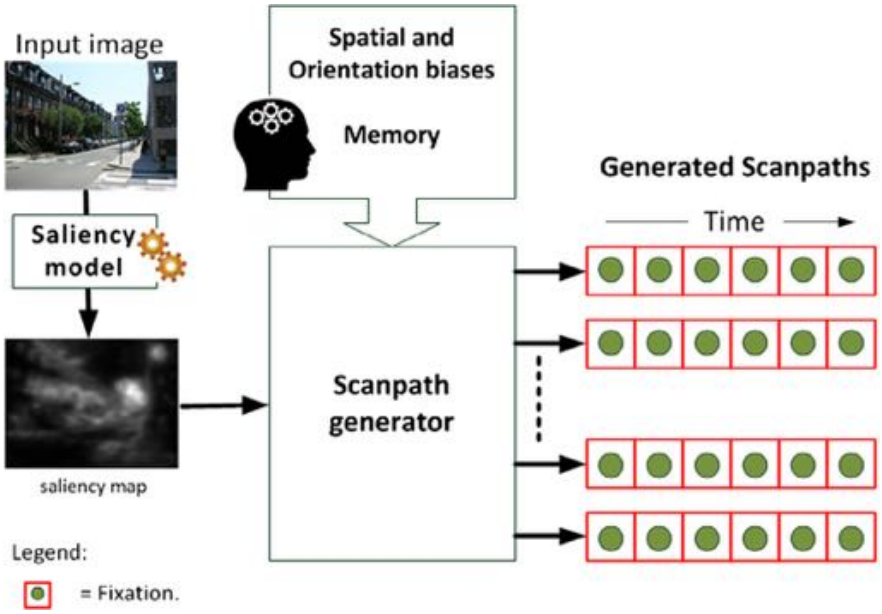



Figure 3.7 – **Modules composing a saccadic model**. A saccadic model can be synthesised as 3 modules: saliency map, oculomotor biases and next fixation selection. The scanpath generator takes these 3 modules and generate scanpaths. Adapted From Le Meur and Liu (2015).

### 3.2.1 Early work on scanpath prediction

Early studies on visual exploration suggested that participants change their visual behaviour according to the task (Yarbus, 1967). Later, Noton and Stark (1971a, 1971b) claimed that a particular visual pattern induced a particular sequence of eye movements called "*Scanpath*". However, further work failed to identify evidence supporting replicable scanning sequences (Findlay & Gilchrist, 2003, p. 134). For instance, Parker (1978) asked participants to memorise a scene containing about six different objects which were always in the same location. Participants looked at each object in their preferred viewing order. Nevertheless, the study of statistical dependencies between eye movement parameters emerged from this approach. For instance, the seminal work of Stark and Ellis (1981), Ellis and Smith (1985) and detailed by Ellis and Stark (1986), suggests that eye movements could be assimilated as a Markov process. They displayed an encounter between their supposed aircraft and an intruder to 8 airlines pilots. They asked to the pilots to use a switch to designate if the intruder was in front or behind their aircraft. The authors then analysed the fixation durations on each element of interest and noted the transitions between them. They noted that instead of being randomly distributed, the pilots looked back and forth between the intruder and intruder's predicted path. They explained this behaviour by a Markov process predicting transitions between two AOIs. A Markov process designates a random variable depending only on the current state. Which means, in this case, the next item expected by the pilots only depends on the one currently viewed. Such a process does not have any memory.

Then, Hacisalihzade et al. (1992) extended this model to predict visual fixations on AOIs of drawings or paintings. They also evaluated their model using the string-edit method (described in Chapter 5). Although, this work led to numerous scanpath prediction models that will be described in detail below, this approach focused more on transitions between manually defined AOIs than the stimulus as a whole. Moreover, the complexity induced by this type of model requires computational power, which was limited at the time.

### 3.2.2 Scanpath and saliency maps

With computational power in mind, Itti et al. (1998) developed the first bottom-up saliency model based on the framework of Koch and Ullman (1985). As described in Figure 3.1, extracted features used for saliency predictions are processed in parallel. While the main contribution of Itti et al.'s work is undoubtedly the saliency model, they also proposed a method reproducing attention orienting. This method is based on two mechanisms: Winner-Take-All (WTA) and Inhibition Of Return (IOR). The Winner-Take-All algorithm consists of selecting the maximum value as the next fixation. But, by just selecting this maximum value, the model could not generate a sequence of expected locations. That is why WTA was combined with an inhibition mechanism. The Inhibition Of Return mechanisms prevent our attention from going back to an already fixated location. This inhibition lasts approximately 500-900ms (Posner & Cohen, 1984). In addition to one of the first bottom-up saliency models, Itti et al. (1998) proposed the first attentional shift model based on a saliency map. Since then, all saccadic models present in the literature are based on at least one saliency map. For instance, Brockmann and Geisel (2000) modelled visual exploration with a "*Levy flight*", which is a type of random walk. A random walk designates a mathematical process describing a 2D path in which each step is selected randomly. For example, a simple random walk starts at 0 and its next step is randomly selected by adding or subtracting 1 with equal probability. The result is a path generated randomly. The "*Levy flight*" random walk refers to a random walk in which the step-length follows a Levy distribution which is positively skewed. Influenced by the work of Itti et al. (1998), Boccignone and Ferraro (2004) adapted Brockmann and Geisel (2000) model by constraining the *Levy flight* through a saliency map.

### 3.2.3 Modelling oculomotor biases

While bottom-up saliency needs to be combined with top-down factors to more accurately describe what will guide attention on a scene, saccadic models cannot be based solely

| Model | Year | IOR | Saccade amplitude | Saccade orientation | Fovea | centre-bias |
|---|---|---|---|---|---|---|
| Zhang et al. | 2009 | *n* last fixs | Gamma distribution | – | Surrounding ROI | – |
| Wang et al. | 2011 | Constant forgetting factor | 20° area | – | Gaussian Pyramid | First fix at image centre |
| Tavakoli et al. | 2013 | Stochastic Mapping Function | Gaussian Mixture Model | – | – | – |
| Engbert et al. | 2015 | Gaussian profile | – | – | Gaussian aperture | – |
| Le Meur and Liu | 2015 | Linearly declines on 8 fixs | Kernel Density Estimation | Kernel Density Estimation | – | No precision |
| Le Meur and Coutrot | 2016 | Linearly declines on 8 fixs | Kernel Density Estimation | Kernel Density Estimation | – | No precision |
| Wang et al. | 2016 | Same as Le Meur and Liu (2015) | Same as Le Meur and Liu (2015) | Same as Le Meur and Liu (2015) | Gaussian pyramid | First fix at image centre |
| Le Meur et al. | 2017 | Linearly declines on 8 fixs | Kernel Density Estimation | Kernel Density Estimation | – | No precision |
| Ngo and Manjunath | 2017 | LSTM | – | – | – | – |
| Shao et al. | 2017 | Only previous fix | Same as Wang et al. (2011) | – | – | – |
| Wloka, Kotseruba, et al. | 2018 | Linearly declines on 100 fixs | – | – | Gaussian pyramid | First fix at image centre |
| Chen and Sun | 2018 | LSTM | – | – | Gaussian pyramid | – |
| Han and Xiao | 2018 | Linear model | Statistics | Statistics | Gaussian pyramid | – |
| Xia et al. | 2019 | No precision | Weighted map encouraging short saccades | – | – | Centered weighted map |
| Xia and Quan | 2020 | Based on distance with previous fix | 2D Gaussian function | Gaussian function | – | – |

Table 3.1 – **Summary of how common oculomotor biases are modelled** in the literature. This includes Inhibition Of Return (IOR), saccade amplitude, saccade orientation, the fovea and the centre-bias.

on saliency maps. The main reason is that biases influence our visual exploration. For instance, we tend to do more short saccades than long ones, and more horizontal saccades than vertical ones (see Chapter 2). Since saliency is a global view of where attention is directed, saccade amplitude or orientation does not influence saliency models. However, the biases directly influence how the next fixation is selected, thus saccadic models are dependant on such biases. The first illustration of important biases in saccadic modelling, but marginal in saliency modelling, is the Inhibition Of Return (IOR), as presented in Itti et al. (1998). In their model, the couple WTA-IOR comes at the end of the saliency model and does not affect salient area prediction but is at the core of scanpath pattern prediction. The WTA algorithm selects the saliency maximum as the next fixation and IOR prevents the WTA from selecting the same maximum twice.

Zhang et al. (2009) were some of the first to introduce additional oculomotor biases in their model, see Table 3.1 for an overview of oculomotor biases modelling in the literature. Zhang et al. (2009) based their model on Itti's saliency map, but added saccade amplitude and fovea modelling as biases, and used a genetic algorithm to select the next fixation. To mimic our tendency to do shorter saccades than long ones, they used a Gamma distribution which was combined with saliency map. Then, they modelled the fovea/retina as a Region Of Interest (ROI) around the previous fixation. The fixation selection algorithm was forced to stay in a certain ROI, and when too many fixations were generated in this ROI, the algorithm was again forced to jump to another one, and so on. Wang et al. (2011) implemented a model based on information maximisation with fovea, centre-bias, visual memory and saccade amplitude biases. Contrary to Zhang et al. (2009), they modelled the fovea using a Gaussian pyramid as proposed by Geisler and Perry (2002). The first level of the Gaussian pyramid designated the original image. The second level was obtained by blurring and down-sampling the original image. The next levels were then built the same way as the previous level. The number of levels or scales was determined beforehand. Finally, the region which needed to be foveated was determined (where the fixation is), and each level of the pyramid was then blended together. To simulate centre bias, which describes the tendency to look at the centre of

an image, Wang et al. (2011) placed the first fixation at the centre of the image. Visual memory (or IOR) was modelled as a constant decaying factor without further precision. Finally, saccade amplitude was modelled as a 20-degree area around the fixation in which the next fixation would be selected using information maximisation.

Tavakoli et al. (2013) took the Ellis and Stark (1986) idea that visual exploration could be approximated as a markov model and combined it with Itti's approach, basing their model on saliency maps and oculomotor biases. They implemented saccade amplitude as a Gaussian Mixture Model fitted to eye-tracking data, and IOR as a stochastic mapping function to avoid immediate return to fixation. Moreover, this Markovian approach was extended by Liu et al. (2013) and Coutrot et al. (2017), but without biases modelling. Le Meur and Liu (2015) proposed a modular Bayesian approach. They built a model on the junction of probabilities between a bottom-up saliency map, a saccade amplitude and orientation probability map and an inhibition map. The next fixation was selected as the location with the highest saliency gain. Saccade orientation and saccade amplitude maps were inferred using Kernel Density Estimation from fixations location distribution of four publicly available datasets. Inhibition Of Return was modelled as a simple linear model with the IOR disappearing after eight fixations. Later on, Le Meur and Coutrot (2016) applied this approach by adapting each module to the content type. For instance, they computed saccade amplitude and orientation probability maps depending on the stimulus type, such as web pages, natural scenes and conversation videos. Le Meur et al. (2017), demonstrated again the advantages of such a modular approach by estimating saccade orientation and amplitude distributions by the participant's age. Wang et al. (2016) and Han and Xiao (2018) extended Le Meur and Liu (2015) model by adding fovea computation and then applying saliency. To better mimic scanpath's dynamic Wloka, Kotseruba, et al. (2018) proposed a new approach involving periphery-fovea separation and a dynamic computing of saliency maps. The image was foveated using a Gaussian pyramid, then the fovea and the periphery were separated and a bottom-up saliency model was run on the periphery part and a top-down saliency model on the fovea part. Then both parts were combined and the next fixation was selected with WTA-IOR

algorithms. This approach is a step towards better scanpath dynamics. So far, studies have focused on the dynamics of the fixation selection process. In most of the models presented here, the fovea system is used to update the existing saliency map and adapt it for each fixation. However, Wloka, Kotseruba, et al. (2018) were the first to entirely distinguish fovea and periphery and apply different types of models to each iteration. An interesting evolution of this approach would be to update the saliency models used, add oculomotor biases and combine this with another fixation selection strategy than WTA. Thus, scanpath prediction could become even more dynamic and more plausible.

Recently, as in the saliency modelling fields, Deep Neural Network have been used to tackle scanpath prediction. Ngo and Manjunath (2017) used an already trained VGG-16 Convolutional Network (Simonyan & Zisserman, 2014) connected to a Recurrent Neural Network (RNN) to simulate IOR. VGG-16 is a 16-layer Convolutional network able to correctly classify and identify 1000 different objects over 14 million images with an accuracy of 92.7%. The idea behind an RNN is that its output is connected with its input, so that short-term memory can influence next iteration of the network. The RNN used in this model and the following ones is the Long Short-Term Memory (LSTM). Its particularity lies in the fact that, in addition to its output connected to its input for short-term memory, a hidden state reproducing long-term memory follows the same schema. The interest of this two-DNN approach is to use a Convolutional Network to extract features from the image and compute saliency, while using eye-tracking data to train an LSTM to model IOR, saccade amplitude and orientation biases. Chen and Sun (2018) used a similar approach, but used different kinds of LSTM and added foveated input images through a Gaussian pyramid.

As we have described in previous sections, existing models try to reproduce the eye dynamics based on static saliency maps and distributions of oculomotor biases. Fixation selection dynamics has been mainly modelled as Bayesian statistics, Levy flight or Winner-Take-All. Static stimuli, generally natural scenes, and static saliency maps have been made dynamic using different techniques, such as centre-surround (Xia et al., 2019), the fovea mechanism (Chen & Sun, 2018; Engbert et al., 2015; Han & Xiao, 2018;

Wang et al., 2011; Wang et al., 2016; Zhang et al., 2009) or the dynamic fovea-periphery mechanism (Wloka, Kotseruba, et al., 2018). Even though the use of biases is not systematic (Boccignone & Ferraro, 2004; Brockmann & Geisel, 2000; Mannaru et al., 2016; Simon et al., 2016; Zhang et al., 2011), numerous models implement them. The most used one would be Inhibition Of Return, followed by the fovea mechanism and then saccade amplitude and orientation. Even though different implementations do not agree on how long the IOR should last, a consensus seems to have been reached on its need and the fact that it is limited in time. Regarding the fovea mechanism, all models seem to have converged on the use of a Gaussian pyramid. However, there are still divergences on how and which oculomotor biases to implement. For instance, Zhang et al. (2009) used a Gamma distribution to model both bias. Wang et al. (2011) chose to simulate saccade amplitude as an area of 20 degrees around the previous fixation in which the next would be picked. Tavakoli et al. (2013) applied a Gaussian Mixture Model on saccade amplitude, while Le Meur and Coutrot (2016), Le Meur et al. (2017), Le Meur and Liu (2015), and Wang et al. (2016) used Kernel Density Estimation to model saccade amplitude and orientation. Nevertheless, the use of such biases has not been done dynamically. Saccade parameters are estimated or computed according to different techniques but remain on the globality of the exploration. Such an approach does not tackle the dynamic of eye movement parameters described by Pannasch et al. (2008), Unema et al. (2005), and Velichkovsky et al. (2005). That is why we propose a dynamic approach of oculomotor biases in Chapter 8. In this approach, how these parameters evolve over time is considered to enhance saccadic models. Moreover, we introduce in a new methodology to analyse models based on the time-dynamic of the generated scanpath and compare it to the human scanpath.

### 3.2.4 Scanpath on web stimuli

Scanpath modelling on web pages has been tackled using two distinct approaches. The first is the classic saccadic model, but adapted to web content complexity and diversity. The second approach takes advantage of the use of the computer mouse, which

is not usually available on natural images. These models focuses on the eye-mouse coordination to model the location of the eyes based on the mouse location. Although these two methods are of great interest, they have been addressed by the literature quite differently. For instance, the eye-mouse modelling resulted in models specialised in Search Engine Result Page (SERP), while scanpath modelling resulted in a very limited number of models.

Guo and Agichtein (2010) were the first to propose a computer-mouse-based model of the scanpath. They first used their own study to compute the euclidean distance between the eyes and the mouse. Then, they labelled each mouse-eye sample with two classes: "*InFocus*" and "*Away*". The category was defined if the distance was below or above a predefined threshold. Finally, they trained a LogitBoost algorithm on these labeled data and were able to predict the position of the eyes within a 100-pixel area around the mouse with a mean accuracy of 77%. Huang et al. (2012) modelled eye-mouse coordination as a Multiple Linear Regression using 4 different features. The first was the cursor position coordinates, defined as a tuple $(x, y)$. The second feature was the behaviour feature reflecting the cursor's activity. Their analyse showed that an active cursor was better aligned with the eyes, than an inactive cursor. Therefore, they measured idle time following the last movement as a behaviour feature. The third feature corresponded to the time elapsed since the loading of the page, this is the dwell feature. The fourth and final feature was the future feature, which represents the position of the mouse within the next 10 seconds. Using Euclidean Distance to evaluate their model, they reached an average 181-pixel accuracy. Which means, the predicted eye position was, in average, 181 pixels from the mouse's position.

Regarding the saccadic modelling classic approach, we found very few models tackling web pages. In addition of Le Meur and Coutrot (2016), we only found a recent work of Xia and Quan (2020) tackling this question. Xia and Quan (2020) proposed a model based on their own saliency algorithm, with biases modelling, and WTA as the fixation selection mechanism. They introduced a top-left-bias as a weighted map to highlight our tendency to do more fixations on the top-left regions of a web page at the beginning of

exploration. To model saccade amplitude and orientation, they trained a 2D Gaussian function on dataset's statistics and set asymmetric standard deviations to encourage horizontal saccades. Moreover, they implemented an IOR mechanism, but instead of using time to recover already seen area, they computed it as a function of the distance between the current fixation and the previous one.

Each approach suffers from distinct major disadvantages. For instance, models predicting the eyes based on the computer mouse are mainly trained with SERP data. An SERP layout is very specific and can be recognised among many others. We use it every day for both our professional queries and our personal ones, so we could state that we are highly expert in these types of web page layout. Moreover, this page type does not reflect other web page types, such as news sites, blogs, videos, forums, mixed content, etc. In the end, we cannot generalise these models. That is why we proposed a simple model trained on classic dynamic web pages in **Article 2**. We used a Gaussian model based on learned statistics with Euclidean Distance between the eyes and the mouse. Concerning classic saccadic models, the first main disadvantage also lies in the generalisation problem. As we described in the previous section, many methods have been explored to predict scanpath. However, these methods were created based on natural images and it is still unsure how they could be generalised since web pages are much more complex. The second main disadvantage concerns about the stimulus itself. Contrary to eye-mouse models, which are on dynamic SERP, scanpath prediction has been only predicted on static web pages (Le Meur & Coutrot, 2016; Xia & Quan, 2020). Yet, this dynamic fro web pages generates much interest. With all these elements in mind, we developed a saccadic model predicting scanpath on dynamic web pages, described in **Chapter 8**. We implemented specific mechanisms to handle web page dynamics and related statistics. Moreover, we used the findings of **Article 3**, on the evolution of eye movements and mouse parameters over time, to propose time-dependant bias modelling.

## Chapter 3 summary

The eyes are at the centre of visual perception, and through the sequence of fixations and saccades, they enable visual information to be acquired efficiently. This is why many models have tried to reproduce this mechanism. Two types of models have emerged: saliency models and saccadic models. The former aim to provide a summary of the places visited by a participant when exploring a visual scene. These models are often categorised according to the types of factors they take into account. On the one hand, we have the so-called bottom-up models, which have given rise to a flourishing literature, predict the areas visited based on the characteristics of the stimulus. On the other hand, we have the so-called top-down models. Most of the time, these models start from a bottom-up model which they "increase" with parameters related to the task, the presence of faces, objects, etc. Although less extensive than the literature on bottom-up models, top-down models are nowadays at the heart of researchers' interest. As for the second type of models, their aim is no longer to provide a summary of the fixed regions, but the detailed and ordered sequence of the areas seen. In this approach the oculomotor parameters are of much greater importance as they can influence the choice of the next fixation. This is why this type of model, in addition to reproduce mechanisms for selecting the next fixation, also models these biases. The biases most often modelled are the saccade amplitude, saccade orientation, fovea, and inhibition of return. However, these biases are modelled most of the time in the same way. Both saliency and saccadic models have been developed most of the time to predict eye movements when exploring natural scenes. When we look at web pages, models become much rarer. However, web page modelling has given rise to another type of approach: the modelling of the eye position as a function of the mouse position. Several models have been proposed following this approach but they have focused on Search Engine Result Page (SERP) rather than classical web pages. In the light of all these elements, we propose in the **Chapter 8** a saccadic model on dynamic web pages. In addition, we apply the temporal analyses of ocular parameters carried out in the **Article 3** to propose a model of oculomotor biases based on their evolution over time.

## Résumé du chapitre 3

Les yeux sont au centre de la percetion visuelle, et à travers l'enchainement des fixations et des saccades, ils permettent d'acquérir des informations visuelle de manière efficiente. C'est pourquoi de nombreux modèles ont tenté de reproduire cette mécanique. Deux types de modèles ont alors émergés: les modèles de saillances et les modèles saccadiques. Les premiers ont pour objectif de fournir un résumé des endroits visité par un participant lors de l'exploration d'une scène visuelle. On catégorise souvent ces modèles selon les types de facteurs qu'ils prennent en compte. D'un côté, nous avons les modèles dits ascendants (ou bottom-up), qui ont donné lieux à une littérature fleurissante, prédisent les régions visitées en se basant sur les caratéristiques du stimulus. De l'autre côté, nous avons les modèles dits descendants (ou top-down). La majorité du temps, ces modèles partent d'un modèle ascendant qu'ils "augmentent" avec des paramètres liés à la tâche, les visages présents, les objets, etc. Bien que moins étendu que la littérature des modèles ascendants, les modèles descendants sont aujourd'hui au coeur de l'intérêt des chercheurs. Quant aux seconds types de modèles, leur but est, non plus de fournir un résumé des régions fixées, mais la séquence détaillée et ordonnée des zones vues. Ainsi, les paramètres oculomoteurs ont une bien plus grande importance car ils peuvent influencer le choix de la fixation suivante. C'est pourquoi ce type de modèle, en plus de modéliser le mécanisme de sélection de la prochaine fixation, modélise également ces biais. Les biais les plus souvent modélisés étant l'amplitude de saccade, l'orientation des saccades, la fovéa, et l'inhibition de retour. Toutefois, ces biais sont la plus part du temps modélisés de la même manière. A la fois les modèles de saillance et les modèles saccadiques ont été développés pour prédire les mouvements des yeux lors de l'exploration de scènes naturelles. Lorsque ne nous intéressons aux pages webs, les modèles se font beaucoup plus rares. Cependant, la modélisation sur page web a donné lieu à un autre type d'approche : la modélisation de la position de l'oeil en fonction de la position de la souris. Plusieurs modèles ont été proposés suivant cette approche mais ils se sont focalisés sur les pages de recherches de moteur de recherche plutôt que sur les pages webs classiques. A la lumière de tous ces éléments, nous proposons dans le **Chapitre 8** un modèle saccadique sur des pages webs dynamiques. De plus, nous appliquons les analyses temporelles des paramètres oculaires réalisées dans l'**Article 3** afin de proposer une modélisation des biais oculomoteurs basée sur leur évolution dans le temps.

# THESIS GOALS

According to the literature presented in this first part, saccadic modelling aims to reproduce eye movement dynamic and which factors influence it. The principal motivation behind this thesis is to demonstrate how eye movement parameters time courses and the dynamic induced by the scroll are important in saccadic modelling on web pages.

The **first axis** of this thesis focused on the investigation of how eye movement dynamic could be summarised as an indicator. Based on ambient and focal visual processing modes definition (Pannasch et al., 2008; Unema et al., 2005; Velichkovsky et al., 2005) we evaluated the relevance of existing ratios to describe these visual processing modes (Dehais et al., 2015; Goldberg & Kotval, 1999; Krejtz et al., 2016) on natural images and web pages.

Several research focused on the coordination between the eyes and the mouse cursor, but these studies involved specific web pages (SERP) and neglected the scroll (Guo & Agichtein, 2010; Huang et al., 2012; Rodden et al., 2008). For these reasons, the **second axis** of this thesis aims to analyse eye movement behaviour when browsing more ecological web pages of our daily life. To this end, we set up two experimental studies, and we examined the relationship between the eyes, the mouse cursor and the scroll.

The two first axes provided a better understanding of eye movements dynamic and the relationship between the eyes, the mouse cursor and the scroll. We used some of our findings in the **third axis** in order to improve saccadic modelling. Thus, we proposed a saccadic model including a scroll mechanism and modelling the temporal evolution of eye movement parameters. This model was evaluated on high-quality data from the experimental studies presented in the next chapter.

# Part II

# Experimental contributions

CHAPTER 4

# GENERAL METHOD

## Contents

Two experimental studies have been run during this thesis. The principal motivation behind these two studies is the lack of available high-quality datasets containing real web pages. In the usability field, most studies are done on web pages with a limited sample size, while datasets available in the modelling field only contain screenshots. That is why we set up two studies to provide high-quality data on French real web pages with a large sample size.

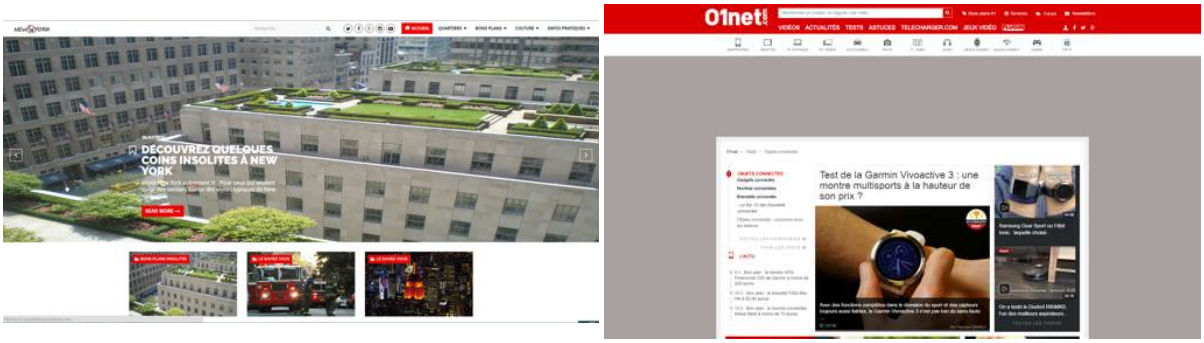
The results of the first study are shown in **Article 2** (eye-mouse coordination and scrolling behaviour) and **Poster 3** (scrolling behaviour) which tackle the eyes and the mouse coordination and scrolling behaviour. The results of the second study are described in **Poster 1** (first ratio for ambient and focal modes), **Poster 2** (second ratio for ambient and focal modes), **Article 3** (eye-mouse-scroll coordination) and **Chapter 8** (scanpath modelling). The purpose of this chapter is to introduce similarities and differences between these two studies.

## 4.1 Participants

In **Article 2** and **Poster 3**, participants were PhD students in the engineering laboratory of the University of Mons.

In **Articles 3**, **Posters 1 and 2** and **Chapter 8**, participants were bachelor students from the Institute Psychology of Paris University or members of the Vision Action Cognition laboratory. In exchange to their participation, bachelor students received academic credits. In addition to students, external participants were recruited for the second study, through the "Réseau d'Information en Sciences Cognitives", and were compensated with a 15€ voucher for their participation. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and national research committee (local Ethics Committee of Paris University, No. CER-PD:2018-77) and with the 1964 Helsinki declaration and its later amendments. All participants gave written informed consent before the experiments.

## 4.2 Stimuli




(a) First experiment

(b) Second experiment

Figure 4.1 – **Websites examples**. Examples of websites displayed during experiment 1 (a) and experiment 2 (b). Only headers of the web pages are shown here, the full page is in average 5000 pixels long (about 4-5 times screen's height). In (b) the gray area allows to normalise website content to ensure websites layouts were the same.

In the study of **Article 2** and **Poster 3**, stimuli were displayed on a 17-inch computer screen set to a resolution of 1920x1080 pixels. Instructions and websites were displayed in a Google Chrome browser with a resolution of 1920x955 pixels. Ten web pages were displayed presented in the same order to each participant. In the study of **Articles 3** and **Posters 1 and 2**, stimuli were displayed on a 24.5 inch LCD computer screen with a 1920x1080 pixels resolution and a 144Hz refresh rate. Participants placed their head in a chin-rest at 57cm of the screen. Eighteen web pages (e.g. Figure 4.1) from eighteen different websites were randomly presented to the participants with a mean height of 6405 pixels. However, in **Chapter 8**, only 9 web pages of the presented stimuli were used to avoid web pages with sticky header and a good distribution of participants. A sticky header being a header of a web page staying on top of the screen while scrolling.

In both studies, participants were allowed to freely move the mouse, scroll or click without restriction, but hyperlinks were deactivated, thus participants could not leave the displayed web page.

## 4.3 Eye movement recordings

In **Article 2** and **Poster 3**, eye movements were recorded using a FaceLAB 5 eye-tracker sampled at 60Hz without head constraint. Mouse movements were recorded through the browser and uploaded on the fly on a local NodeJS server using browser's extension system. In **Articles 3**, **Posters 1 and 2** and **Chapter 8**, eye movements were recorded using an Eye-Link 1000 Plus (SR Research Ltd., Canada) at a 1000Hz sampling rate with 0.05°precision. The right eye of the participants was recorded with a 35mm monocular lens. Mouse movements were recorded with a standard USB optical mouse with a 125Hz polling rate.

In both studies, the experiment started with a calibration phase during which participants had to fixate successive points in 5 different locations in**Article 2** and **Poster 3** and 9 different locations in **Articles 3**, **Posters 1 and 2** and **Chapter 8**. In the second study, this calibration phase was repeated between each trial with a 5 points calibration and a 9 points calibration at the half of the experiment. Then, instruction was displayed in **Article 2** and **Poster 3**. However, participants had to first go through a training phase, before displaying instructions. During the second experiment, if a participant's calibration reached an error greater than 1°, calibration was started over.

## 4.4 Tasks

In **Article 2** and **Poster 3**, a set of three types of tasks was presented to participants: free-viewing, visual search and reading. As the websites, the tasks were displayed in the same order for all participants. During the free-viewing task, participants were asked to freely browse the web page as long as they wanted. During the reading task, participants were asked to read the a specific paragraph. In the visual search participants had to either find an item, or an hyperlink.

In **Article 3** and **Posters 1 and 2**, participants had to perform two types of tasks: free-viewing and visual search. Tasks were randomly displayed nine times each. During the free viewing task, the participants were instructed to explore the web page freely for

exactly 60s. During the visual search task, participants were asked to find a target in maximum two minutes. The targets could be anywhere on the web page, but they were distributed between the top, middle and bottom of the web page across web pages. The participants were ignorant about the number of targets presented

## 4.5 Data analyses

In **Poster 3**, we investigated if the saccade preceding a scroll was on the same direction as the scroll. We analysed average scroll speed with and without anticipation during both tasks. Moreover, we analysed whether there was a correlation between eye position before scrolling and the amplitude of the scroll. Finally, we evaluated where participants fixated during scroll and how the area of the screen fixated changed depending on the scroll speed.

In **Article 2**, we compared eye and mouse spatial distribution using Pearson's Correlation Coefficient. Then, we dynamically assessed the distance and the correlation between the position of the eyes and the mouse. Moreover, we ran similar analyses as in **Poster 3** on scroll speed influence and scroll amplitude but using different screen division. Finally, we proposed a gaussian model to predict eye position depending on the location of the mouse.

In **Poster 1**, ambient and focal visual processing modes were described using Dehais et al. (2015) ratio. This analysis has been performed on data of the second study. In **Article 1**, again ambient and focal visual processing modes were described, but using Krejtz et al. (2016) coefficient. Contrary to **Poster 2**, in this article the goal was to discriminate two tasks of an external dataset of scene viewing. To do so, we used variables derived from Krejtz et al. (2016) coefficient. In **Poster 2**, the same coefficient has been applied on the second study's data.

In **Article 3**, we ran an extensive quantitative analysis on mouse movements, scrolling and eye movements on web pages. In addition, we investigated the relationship between parameters, such as, saccade amplitude, scroll amplitude, fixation duration,

scroll idle, etc, over time.

Finally, in **Chapter 8**, we investigated the influence of eye movements dynamic parameters, such as, saccades orientation, fixation duration and saccades amplitude on scanpath modelling.

CHAPTER 5

# VALIDATION FRAMEWORK

## Contents

As seen in Chapter 3, lots of saliency and saccadic models have been released. In order to assess how good a model is, there are two prerequisites: a dataset of stimuli including ground-truth experimental data to compare the models with, and a metric to objectively measure how close models are close to ground-truth. However, these models were developed using various technologies. Some were made available to researchers while the others are harder or impossible to access. Hence, it became difficult to reproduce results and compare new models to past ones. To address this problem, the MIT Benchmark was created (Kümmerer et al., 2020). The role of the MIT benchmark is to provide a platform where researchers can compare their saliency models to others with common metrics and predefined datasets. One of these dataset is the MIT dataset (Judd et al., 2009). It relies on two essential parts: a public part and an hidden part. In the public part of the dataset, stimuli with ground-truth data are publicly available and can be freely downloaded on the website (Judd et al., 2009). The hidden part only consists of stimuli without ground-truth data. Thus, researchers can train or test their models on the public part, while the hidden part of the dataset is used to benchmark how good is a model on unknown data. The MIT benchmark (Kümmerer et al., 2020) is a fantastic initiative and provides a great platform to compare saliency models. Yet, this platform is focused on saliency models and leave out saccadic models. Of course, it is possible to convert generated scanpaths to a saliency map and then compare this map to saliency models. But, the very essence of saccadic models lies in their dynamics which is suppressed with such approach. That is why a counter part of the MIT benchmark specific to scanpath models would be a great step forward towards the openness of the modelling community.

A validation framework is the combination of one or more datasets and some metrics, as the MIT benchmark is. But, when researchers want to compare their model, they need to develop their own validation framework in order to load data, run the evaluation metrics, and compile the results. This work is very tedious and is, to some extent, a waste of time. That is why, open source validation frameworks recently emerged. The first initiative was proposed by Zanca et al. (2018) with FixaTons. Focused on scanpath

modelling, its role is to gather in one place a tool to download public datasets, run metrics and compute statistics. In addition to these, FixaTons propose a structure to encourage others to add their own datasets and metrics. Around the same time, another framework was created: the Saliency Model Implementation Library for Experimental Research (SMILER) (Wloka, Kunić, et al., 2018). Contrary to FixaTons, SMILER provides interfaces to run the maximum available saliency models within a single tool. In addition, Wloka, Kunić, et al. (2018) suggested a common data format for models parameters to facilitate models compatibility. More recently, the ownership of the MIT Benchmark changed from the Massachusetts Institute of Technology, United-States to Tuebingen University, Germany. With it, a brand new validation framework has been developed (Kummerer, 2019). Called PySaliency, this framework provides a more complete approach for saliency models including dataset, metrics and models. It should be noted that PySaliency provides a bridge to transform scanpaths into saliency maps. However, the SMILER and PySaliency frameworks are saliency-oriented, while FixaTons mainly focuses on scanpath metrics. Moreover, none of the above extensively focus on scanpath metrics, benchmarking or dynamic datasets. In our case, we need to evaluate models on dynamic web pages taking into account the scroll, which is not currently possible in available frameworks. For all these reasons, we developed the first end-to-end validation framework handling the steps when creating a new saliency and saccadic model from training to benchmarking. Named *SalScan*, this framework aims to be modular, open-sourced and developer-oriented to encourage contributions from the community.

## 5.1 Datasets

When evaluating a model, the data source is of primary interest. That is why, instead of standardising data files so that every dataset followed the same guidelines (Kummerer, 2019; Zanca et al., 2018), *SalScan* comes as a software layer on top of the original data without modifying it. Thus, original data can instantly be used as a data source to train or evaluate an already existing model. Many datasets are publicly available (Borji & Itti,

2015; Le Meur et al., 2006; Xu et al., 2014), but we describe here the main ones: MIT (Judd et al., 2009), Toronto (Bruce & Tsotsos, 2007) and Fiwi (Shen & Zhao, 2014).

### 5.1.1 MIT datasets

Figure 5.1 – **Illustration of the MIT dataset**. Row 1: the original stimulus, row 2: fixation points, and row 3: density map. From Judd et al. (2012).

The MIT dataset depicted in Figure 5.1, is the dataset on which models are evaluated when submitted to the MIT Benchmark. Actually, as previously evoked, there is two MIT datasets. The MIT300 (Judd et al., 2012) used for the benchmark evaluation and the MIT1003 used by researchers to train their models. Both have the same characteristics, but ground-truth from MIT300 is not publicly available. They respectively contain 1003 and 300 random images from Flickr creative commons and LabelMe (Russell et al., 2008). Images were displayed on a computer screen during 3 seconds to 15 participants for MIT1003 and 39 for MIT300. Presented images included text, faces, indoor, outdoor, landscape and portrait images.

### 5.1.2 Toronto dataset

Figure 5.2 – **Illustration of the Toronto dataset**. Rows 1 and 2: indoor scenes and row 3: outdoor scene with no particular regions of interest. Column 1: the stimulus, column 2: the fixation points, column 3: the density map, and column 4: the heat map. Adapted From Riche (2015).

Bruce and Tsotsos (2007) proposed a dataset called Toronto containing 120 natural scene images as displayed in Figure 5.2. They were randomly presented on a computer screen for 4 seconds. Images included indoor and outdoor scenes collected from 20 participants.

### 5.1.3 Fiwi dataset

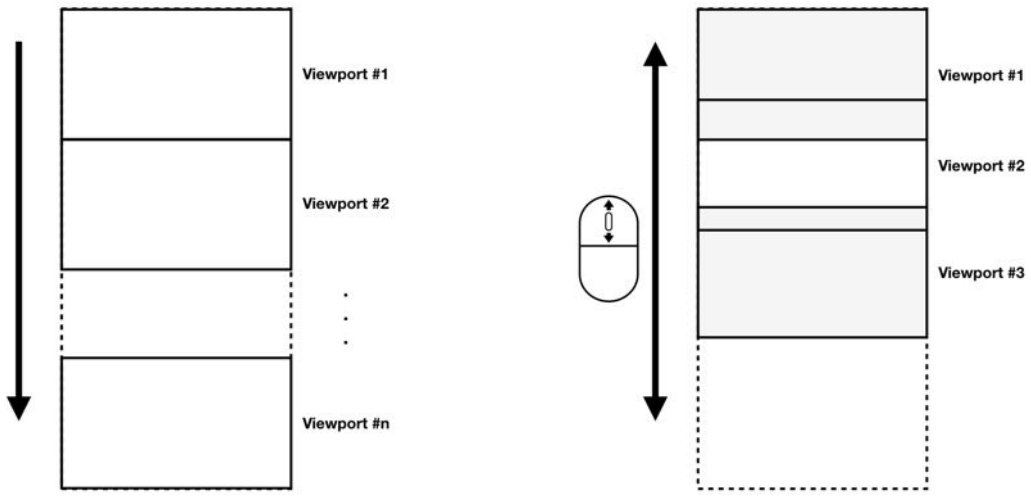

Figure 5.3 – **Illustration of the Fiwi dataset**. Pictorial: web pages occupied by one dominant picture or several large thumbnail, Text: web pages containing informative text with high density, and Mixed: web pages with a balanced mix of the two previous categories. From Shen and Zhao (2014).

The Fiwi (Fixations in Webpage Images) dataset (Shen & Zhao, 2014) is the only large web pages dataset available. It consists of 149 screenshots of various sources on the web categorised as pictorial, text and mixed content (see Figure 5.3). Stimuli were displayed to 11 participants during 5 seconds.

### 5.1.4 EMDW dataset

The Eye Movements Dataset on Web pages (EMDW) is a subset from the experimental data collected during this thesis (see Chapter 4). We extracted 204 trials from the original data including 122 participants and 9 websites. Websites were freely browsed during 60s. Contrary to Fiwi dataset, web pages were entirely scrollable.

## 5.2 Viewport engine




(a) Vertical viewport (b) Scroll viewport

Figure 5.4 – **Viewport engines**. (a) each viewport directly follows the previous one without overlapping. (b) reproduce scroll events from recorded data.

All models mentioned in Chapter 3 assume the input to be a static image. This is not a problem since all datasets in the saliency and saccadic modelling fields are static images. However, while these models can perfectly work on web pages screenshots, they cannot correctly process dynamic web pages. Hence, we implemented in the *SalScan*

framework a viewport engine. Its role is to provide the models with a mechanism to process unconventional stimuli such as a 360 degrees images or a web page. For instance, when we browse a web page, only a limited part of the web page is displayed on the screen, this is the viewport. That is why we need to use the scroll to fully explore it. Thus, a viewport designates a static image extracted from a larger stimulus given an exploration strategy. For now, two viewport strategies are implemented: vertical and scroll. As depicted in Figure 5.4 (a), the vertical viewport strategy consists of browsing the stimulus vertically. The viewport fits the width of the image, while its height is manually predefined. Each viewport directly follows the previous one without overlapping. Once at the bottom of the stimulus, the exploration ends. The scroll viewport strategy is based on scrolling information. The viewport is determined according to the scroll events (Figure 5.4 (b)). Depending on how the participant scrolls, viewports can overlap each other and can be either below or above the previous one.

## 5.3 Metrics

When evaluating a model's performance, it needs to be tested on a dataset with metrics. The role of a metric is to give a numerical value assessing the quality of model's predictions. In the saliency field, a wide variety of metrics exist, but there is no gold metric standard. The saliency metrics that are presented bellow can be divided in two categories: metrics focusing on saliency map values at fixation positions and metrics comparing saliency map and ground-truth fixation statistical distributions.

As explained in Chapter 3, a lot of saccadic models were based on saliency maps. Since many saliency metrics were initially available and due to the complexity of evaluating scanpath patterns, the scanpath modelling literature used saliency metrics for evaluation. To do so, they convert generated scanpaths back to saliency map and use related metrics. Recently, specific metrics emerged (Jarodzka et al., 2010). Contrary to saliency maps, which designate 2D maps of where visual attention is directed, a scanpath is much more complex and involve many additional dimensions. For instance, the duration of

each fixation, the order on fixated locations, etc. As in the saliency fields, scanpaths comparison metrics tried to evaluate models by summarising their performances with a single value (Brandt & Stark, 1997; Shepherd et al., 2010). However, metrics trying to tackle the complexity of scanpath emerged by proposing multiple values evaluating multiple aspects of a scanpath (Jarodzka et al., 2010).

### 5.3.1 Saliency metrics

Are implemented in *SalScan* one saliency metric of each category: The Normalized Scanpath Saliency (NSS) evaluates saliency map values at fixation positions, while Correlation Coefficient (CC) compares saliency map with ground-truth statistical distributions. Both presented metrics are computed globally as in the literature and over time. To do so, each metric is computed for each second to provide the metric temporal evolution.

#### 5.3.1.1 Normalized Scanpath Saliency (NSS)

Introduced by Peters et al. (2005), the idea of the NSS metric is to sum all values from saliency maps where ground-truth fixations are located. The result is then normalised following:

$$NSS = \frac{1}{N} * \sum_{p=1}^{N} \frac{SM(p) - \mu_{SM}}{\sigma_{SM}} \tag{5.1}$$

where $p$ designates the location of a ground-truth fixation, $SM$ is the saliency map and $N$ the total number of fixations. The higher the score is, the better the saliency map prediction is.

#### 5.3.1.2 Correlation Coefficient (CC)

The CC metric has been first applied to saliency map evaluation by Ouerhani et al. (2004). The linear CC is obtained following equation:

$$CC = \frac{cov(SM,GT)}{\sigma_{SM} * \sigma_{GT}} \tag{5.2}$$

where $SM$ designates the saliency map and $GT$ the ground-truth saliency map. The output goes from -1 to 1. When the result is close to -1 or 1, there is an almost perfect linear relationship between the two saliency maps.

### 5.3.2 Scanpath metrics

Scanpath-specific metrics intend to to take into account the dynamic aspect of visual exploration. Multiple approaches has been proposed over the years. The two most common are Dynamic Time Warping (DTW) and string-edit algorithms. They are both described below. More recently, the MultiMatch metric has been proposed to assess various aspects of the scanpath comparison.

#### 5.3.2.1 Dynamic Time Warping (DTW)

DTW is a common algorithm to evaluate the similarity between two paths with potential temporal shift. To compare two scanpaths, the distance between each fixation of the same order is computed. The result is a distance matrix. Then the path with the minimum cost is selected. The final score is computed as the sum of all distance constituting this path.

#### 5.3.2.2 String-edit

String-edit or the Levenshtein distance is an algorithm introduced by Levenshtein (1966). Originally, this algorithm quantifies the difference between two word, sentences or any characters strings. It has been adapted to scanpath differences by Brandt and Stark (1997). In this version, the image is divided in $N$ equal regions labeled with the letters of the alphabet beginning with "$A$". Then, each fixation of the scanpath is assigned the

letter corresponding to the region in which the fixation is located. The difference between scanpaths is then computed using the string-edit algorithm.

#### 5.3.2.3 MultiMatch

MultiMatch is an algorithm proposed by Jarodzka et al. (2010) to compare two scanpaths on five dimensions: scanpath shape, saccades length, fixations location, fixations duration and saccades orientation. Each dimension can be studied separately or in an global MultiMatch score averaging all dimensions. Before computing each measure, the scanpath is pre-processed through two steps: simplification and temporal alignment. The simplification phase consists of deleting small saccades and merging consecutive long saccades with the same direction. Thus, the noise induced by the individuality of each participants is reduced. Then, the temporal alignment step aims to transform simplified scanpath in a graph from which the shortest path is computed using Dijkstra algorithm (Dijkstra, 1959).

As for NSS and CC metrics, we implemented a dynamic version of MultiMatch. Each dimension is computed for each second of the exploration. However, due to the temporal alignment, results can be biased. That is why, the dynamic MultiMatch metric is computed without alignment.

## 5.4 Evaluation and comparison

The major advantage of the *SalScan* framework is to provide an easy way to evaluate any model with any metric on any dataset. Usually, different frameworks provide complementary tools to manipulate each component separately, the researcher must then combine each element. In this framework the validation process can first be executed through *Sessions*. To do so, a model, a dataset and metric(s) are povided to the *Session* module. First, the *Session* takes care of running the model on every image of the dataset. A viewport engine can also be provided. In that case, the session will use the viewport engine and directly send the viewport to the model. Thus, models not developed

for unusual stimuli, such as dynamic web pages, can be used transparently. Finally, the *Session* evaluates every generated scanpath or saliency map with provided metrics. If saliency metrics are given to evaluate a scanpath model, saliency maps are built from generated scanpaths. The result is a table summarising all metrics scores.

The second tool included in this framework is the *Benchmark*. The *Benchmark* can have two roles. The first one is to test different configurations of a model on one or multiple datasets with one or multiple metrics. This is convenient when developing a new model to find the best parameters to reach the best results. The second role of the *Benchmark* tool, is to compare a wide quantity of models on multiple datasets. Both uses can be run at the same time.

CHAPTER 6

# AMBIENT AND FOCAL AS AN INDICATOR OF EYE MOVEMENT DYNAMIC

## Contents

**References:**

- **Poster 1** (page 120): *Milisavljevic, A., Le Bras, T., Petermann, C., Mancas, M., Gosselin, B., & Doré-Mazars, K.(2018). A dynamic approach of searching behaviour in webpages. 41$^{th}$ European Conference on Visual Perception, 26-30 August 2018, Trieste, Italy. Perception, ECVP 2018. 48(S1), p.22*
- **Article 1** (page 123): *Milisavljevic, A., Bras, T. L., Mancas, M., Petermann, C., Gosselin, B., & Doré-Mazars,K. (2019). Towards a better description of visual exploration through temporaldynamic of ambient and focal modes. InProceedings of the 11$^{th}$ ACM Symposium*
- **Poster 2** (page 129): *Milisavljevic, A., Le Bras, T., Abate, F., Gosselin, B., Petermann, C., Mancas, M. and Doré-Mazars, K.(2019). Different visual explorations between free-viewing and target finding tasks in websites: evidence from temporal analyses of ambient and focal modes. 20$^{th}$ European Conference on Eye Movements, 18-22 August 2018, Alicante, Spain. Journal of Eye Movement Research 12(7), p. 390*

## 6.1 Contributions presentation

The first axis of this thesis is to determine how ambient and focal visual processing can describe eye movement dynamic. These visual modes find their origin in the two pathways (ventral and dorsal) taken by visual information in the brain (Goodale & Milner, 1992; Ungerleider & Mishkin, 1982). As described in Chapter 1, the ventral stream goes from occipital to temporal cortex and carries information about object features ("*what*"), while the dorsal stream goes from occipital to parietal cortex and carries information about object locations ("*where*"). As Velichkovsky et al. (2005) explained, these visual pathways can be directly observed through eye movements using fixations and saccades. A short fixation followed by a high-amplitude saccade suggests an ambient mode (dorsal pathway), while a long fixation followed by a small-amplitude saccade suggests a focal mode (ventral pathway) (Pannasch et al., 2008; Unema et al., 2005; Velichkovsky et al., 2005). These two modes have been summarised by the literature through two main ratios (Dehais et al., 2015; Krejtz et al., 2016).

We studied in **Poster 1** how the ratio proposed by Dehais et al. (2015) could be used to discriminate tasks during web pages browsing. This ratio was originally created by

Goldberg and Kotval (1999) to characterise visual exploration in software interfaces, and Dehais et al. (2015) modified it to assess visual processing modes in the context of surprise in a plane cockpit. We applied this modified version on web pages to investigate visual modes. We also introduced the use of Unema et al. (2005) and Pannasch et al. (2008) fixation duration and saccade amplitude thresholds to differentiate short fixations/saccades and long fixations/saccades. We showed that ambient mode (or explore mode) was more intense during free viewing task than during visual search task. Moreover, we found promising results showing that a click could be preceded by a focal mode (or exploit mode). However, this ratio did not differentiate the time spent exploring (saccades) from the time spent exploiting information quickly (short fixations). Hence, we investigated the K coefficient, a ratio proposed by Krejtz et al. (2016) on natural images. Contrary to previous ratio, the K coefficient was specifically designed to describe ambient and focal modes. In **Article 1**, we used the K coefficient to compare eye movements between a free viewing task and a visual search task. We found that the switch between the two modes occurred at a high frequency so that the average value of the ratio was close to zero. We showed that, even though global differences did not emerge, the dynamic of the visual modes between tasks highlighted differences over time. Moreover, we succeeded to globally differentiate tasks by introducing new K-related variables. Finally, we replicated these results on web pages in **Poster 2**. We showed that the K coefficient could not discriminate tasks globally, but by using these K-related variables, we were able to better discriminate tasks globally and over time.

## 6.2 Poster 1: "A dynamic approach of searching behaviour in webpages"

*Milisavljevic, A., Le Bras, T., Petermann, C., Mancas, M., Gosselin, B., & Doré-Mazars, K.(2018). A dynamic approach of searching behaviour in webpages. 41$^{th}$ European Conference on Visual Perception, 26-30 August 2018, Trieste, Italy. Perception, ECVP 2018. 48(S1), p.22*

### Poster 1 summary

In this study we analysed the effect of the type of task on global parameters of ocular exploration as well as on its dynamics. We asked 16 participants to browse 18 websites in ecologically valid conditions. They had to perform two types of task: free exploration and visual search. Preliminary results showed an influence of the task on visual exploration parameters, such as the length of the scanpath, the horizontal dispersion of fixations and the saccade amplitude. However, to better understand the behaviour of the participants, we studied the influence of the task on the dynamics of visual exploration. To do so, we used the ratio proposed by Dehais et al. (2015). We found that the "*Explore*" (ambient) mode was more intense during the free-viewing compared to the visual search task. Furthermore, we found that a mouse clicks were often preceded by a "*Exploit*" (focal) mode.

## Résumé du poster 1

Dans cette étude nous avons analysé l'effet de la tâche sur les paramètres globaux de l'exploration oculaire ainsi que sa dynamique. Nous avons demandé à 16 participants de parcourir 18 sites web en conditions écologiques. Ils devaient réaliser deux types de tâches : exploration libre et recherche visuelle. Les résultats préliminaires montrent une influence de la tâche sur les paramètres de l'exploration visuelle tels que la longueur du chemin oculaire, la dispersion horizontale des fixations et l'amplitude des saccades. Cependant, afin de mieux comprendre le comportement des participants, nous avons étudié l'influence de la tâche sur la dynamique de l'exploration. Pour ce faire, nous avons utilisé le ratio proposé par Dehais et al. (2015). Nous avons observé que le mode "*Explore*" (ambient) était plus intense en parcours libre qu'en recherche visuelle. De plus, nous avons noté qu'un click était précédé la plupart du temps par un mode "*Exploit*" (focal).

# A dynamic approach of searching behaviour in webpages

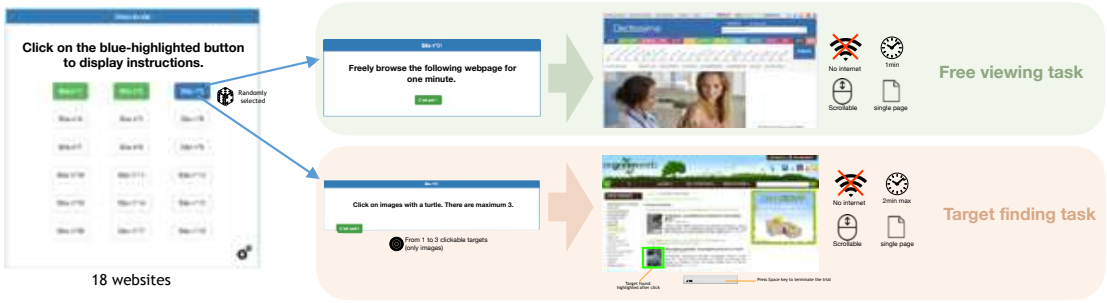
Alexandre Milisavljevic[1,2,3], Thomas Le Bras[1], Matei Mancas[2], Coralie Petermann[3], Bernard Gosselin[2], Karine Doré-Mazars[1]

[1] Laboratoire Vision Action Cognition EA 7326, Institut de Psychologie, INC, Université Paris Descartes, Sorbonne Paris Cité, Boulogne-Billancourt, France
[2] Numediart institute, University of Mons, Mons, Belgium
[3] Research and Development department, Sublime Skinz, Paris, France

## INTRODUCTION

The visit of a webpage is driven by multiple bottom-up and top-down factors, such as the inner characteristics of the webpage, the goal or the user profile. In the present experiment, we studied static and dynamic goal's effects on participants' visual behaviour while browsing webpages. In order to achieve this, we asked them to carry out two kinds of tasks: Free Viewing task and Target Finding task.

## METHODS




### ▶ Participants
- 16 participants.
- Normal or corrected-to-normal vision.
- 6 ♂ ; 10 ♀ ; ~23 y.o.

### ▶ Stimuli
- Mean height of 6505px.
- Width of 1920px.
- Scrollable.
- Fully offline.
- Single page.

### ▶ Apparatus
- EyeLink 1000 (SR Research®).
- Temporal resolution of 1KHz.
- Spatial resolution of 0.05°.
- 24.5-inch screen size.
- 144Hz screen's refresh rate.
- 1920x1080 of screen resolution

## RESULTS

### ▶ Does the task influence eye movements ?

As shown in **table 1**, the scanpath length is longer during Free Viewing condition. This can be explained by the fact that this condition was standardized to 1 minute exactly while Target Finding condition duration was up to the participant finding the target(s). At the opposite, other variables are higher during Target Finding but no conclusion can be given.

Global analyses show the influence of the task on eye movements but the scanpath is dynamic and change over time. With this type of analysis it is difficult to understand the behaviour of a participant while browsing.

| | Free Viewing | Target Finding | F | PR(>F) |
|---|---|---|---|---|
| Scanpath spatial length | 998° | 975° | 5.32e-07 | PR << F |
| Fixation dispersion | 370px | 390px | 0.000019 | PR << F |
| Saccade amplitude | 4.33° | 6.03° | 0.000822 | PR << F |
| Fixation duration | 228ms | 242ms | 6.86e-08 | PR << F |

*Table 1: Influence of task on eye movements variables*

### ▶ Explore or Exploit ?

**Visual exploration modes**

**Exploit/Focal:** long fixations (>180ms) followed by short saccades (<5°).

**Explore/Ambient:** short fixations (<180ms) and large saccades (>5°).

*(Unema et al., 2005)*

In order to better understand this dynamic, we looked for visual exploration modes during the browsing of the webpage. As formalized by (Unema et al., 2005), there is two visual exploration modes: Exploit or Focal and Explore or Ambient. These modes could help to have a more precise understanding of the scanpath's dynamic. To do so, we used (Dehais et al., 2015) version of (Goldberg and Kotval, 1999) ratio as described in *Figure 1*.

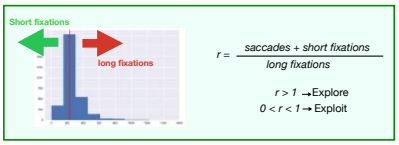



*Figure 1: Classification methods and explore/exploit ratio computing.*

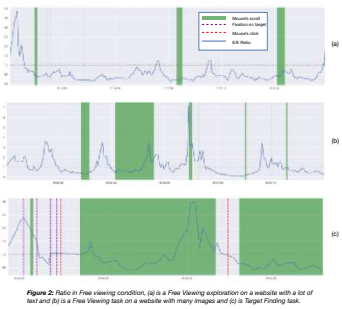


*Figure 2: Ratio in Free viewing condition, (a) is a Free Viewing exploration on a website with a lot of text and (b) is a Free Viewing task on a website with many images and (c) is Target Finding task.*

As you can see in *Figure 2*, the switch between the two modes changes in frequency and intensity according to the task. This switch occurs for two kind of reasons in both case: **bottom-up** or **top-down** stimulation. What change is the importance of one compared to the other depending on the task and stimulus.

We observed the following behaviours:

**Target Finding**
- Clicks are 80% of the time performed during exploit mode ($r < 1.5$).
- Neither the scroll or the click on a target could explain the switch between the two modes.

**Free Viewing**
- Explore mode during the first 2 seconds is not systematic, specially when the user scroll within these 2 seconds.
- Explore mode is more intense during Free Viewing (higher peaks).
- In addition of the task and content, eye position in the page influence the dynamic of the exploration (F=0.004, PR<F).

### ▶ What implies target detection ?

Finally, as shown in *Figure 3*, we found that most of the time, clicks were surrounded by abnormal long fixations thus describing a change in the exploration dynamic. This could be explained by the fact that the participant wanted to be sure to click on the right target.

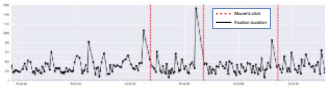


*Figure 3: Clicks and fixation duration.*

## CONCLUSION

Preliminary results showed the influence of the task on the scanpath length, the spatial dispersion of the fixations and the amplitude of the saccades. However, scanpath's characteristics evolve during the navigation which highlighted explore/exploit modes. Further analyses suggest that the scanpath's dynamic is also influenced by the target detection. In our future work, we will investigate the bottom-up influence and mix it with the top-down to better explain participants behaviour during target-finding condition.

**References**

Goldberg, J., and Kotval, X.. 1999. Computer Interface Evaluation Using Eye Movements: Methods and Constructs. International Journal of Industrial Ergonomics 24 (6): 631–45.

Dehais, F., Peysakhovich, V., Scannella, S. and Gateau, T.. 2015. Automation Surprise' in Aviation: Real-Time Solutions. In 33rd Annual ACM Conference on Human Factors in Computing Systems, 2525–34. Seoul.

Unema, P. J. A., Pannasch, S., Joos, M., & Velichkovsky, B. M. (2005). Time course of information processing during scene perception. Visual Cognition 12(3). 473-494.

alexandre.milisavljevic@etu.parisdescartes.fr

*41th European Conference on Visual Perception, Trieste, Italy, 2018, 26th to 30th August*

## 6.3 Article 1: "Towards a better description of visual exploration through temporal dynamic of ambient and focal modes"

*Milisavljevic, A., Bras, T. L., Mancas, M., Petermann, C., Gosselin, B., & Doré-Mazars,K. (2019). Towards a better description of visual exploration through temporaldynamic of ambient and focal modes. InProceedings of the 11$^{th}$ ACM Symposium*

### Article 1 summary

In **Poster 1**, we showed that the type of task influenced the overall parameters of visual exploration as well as the intensity of the "Explore-Exploit" (ambient-focal) modes. The examination of the dynamics also allowed us to observe that in addition to the difference in maximum intensity from one task to another, the intensity distribution in time was also different. It is in the continuity of this work that the **Article 1** intervenes. In this article, we first changed the ratio in order to use the coefficient K (Krejtz et al., 2016). The interest of this change lies in the fact that the coefficient K is dedicated to describe the ambient and focal visual modes. Then, after having failed to observe significant differences in dominant visual mode between tasks, we proposed new K-based complementary measures: the number of switches between visual modes and the average time spent in one mode. Finally, we were able to distinguish between both tasks using these new variables. We showed that visual mode changes were very frequent during exploration. So much that the intensity of the ratio was close to zero which made difficult to differentiate the two tasks. Finally, we analysed the dynamic of the K coefficient and the number of mode changes to narrow the description of each mode over time. Contrary to the global analysis, the dynamic of the K coefficient turned out to be a better tool to differentiate the two tasks.

## Résumé de l'article 1

Dans le **Poster 1**, nous avons montré que la tâche avait une influence sur les paramètres généraux de l'exploration visuelle ainsi que sur l'intensité des modes "Explore-Exploit" (ambient-focal). L'étude de cette dynamique nous a également permis d'observer qu'en plus de la différence d'intensité maximale d'une tâche à l'autre, leur distribution dans le temps était également différente. C'est dans la continuité de ce travail qu'intervient l'**Article 1**. Dans cet article, nous avons d'abord modifié le ratio afin d'utiliser le coefficient K (Krejtz et al., 2016). L'intérêt de ce changement réside dans le fait que ce dernier est dédié à la description des modes visuels ambiant et focal. Après ne pas avoir observé de différences significatives du mode visuel dominant entre les tâches, nous avons proposé de nouvelles mesures complémentaires basées sur K : le nombre de changements de mode et le temps moyen passé dans un mode. Nous avons finalement pu distinguer les deux tâches à l'aide de ces nouvelles variables. Nous avons montré par la suite que les changements de mode visuel étaient très fréquents au cours de l'exploration. A tel point que l'intensité du ratio était proche de zéro, ce qui rendait difficile la différenciation entre les deux tâches. Enfin, nous avons analysé la dynamique du coefficient K et le nombre de changements de mode pour affiner la description de chaque mode dans le temps. Contrairement à l'analyse globale, la dynamique du coefficient K s'est révélée être un meilleur outil pour différencier les deux tâches.

# Towards a better description of visual exploration through temporal dynamic of ambient and focal modes

Alexandre Milisavljevic
Paris Descartes University
Paris, France

Thomas Le Bras
Paris Descartes University
Paris, France

Matei Mancas
Mons University
Mons, Belgium

Coralie Petermann
Sublime Skinz
Paris, France

Bernard Gosselin
Mons University
Mons, Belgium

Karine Doré-Mazars
Paris Descartes University
Paris, France

## ABSTRACT

Human eye movements are far from being well described with current indicators. From the dataset provided by the ETRA 2019 challenge, we analyzed saccades and fixations during a free exploration of blank or natural scenes and during visual search. Based on the two modes of exploration, ambient and focal, we used the K coefficient [Krejtz et al. 2016]. We failed to find any differences between tasks but this indicator gives only the dominant mode over the entire recording. The stability of both modes, assesses with the switch frequency and the mode duration allowed to differentiate gaze behavior according to situations. Time course analyses of K coefficient and switch frequency corroborate that the latter is a useful indicator, describing a greater portion of the eye movement recording.

## CCS CONCEPTS

• **Mathematics of computing** → *Statistical paradigms*; • **Applied computing** → *Psychology*.

## KEYWORDS

visual processing, temporal analysis, ambient, focal

**ACM Reference Format:**
Alexandre Milisavljevic, Thomas Le Bras, Matei Mancas, Coralie Petermann, Bernard Gosselin, and Karine Doré-Mazars. 2019. Towards a better description of visual exploration through temporal dynamic of ambient and focal modes. In *2019 Symposium on Eye Tracking Research and Applications (ETRA '19), June 25–28, 2019, Denver , CO, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3314111.3323075

## 1 INTRODUCTION

Nowadays, we know that human gaze behavior is influenced by multiple aspects of a stimulus such as faces, shapes, colors and so on [Coutrot and Guyader 2014; Tatler et al. 2003, 2008]. In that sense, two categories of factors emerged: bottom-up and top-down [Helo et al. 2014; Yarbus 1967]. Bottom-up factors are low-level features describing stimuli's physical characteristics like luminance, contrast or edges [Tatler et al. 2008]. The influence of these factors appeared to be higher at the beginning of visual exploration [Tatler et al. 2008]. In contrast, top-down factors are high-level features which represent the wide scope of cognitive processes [Henderson and Hollingworth 1999] including the task, semantic, memory, emotions, etc [Le Meur and Coutrot 2016; Yarbus 1967]. Contrary to bottom-up factors, the influence of top-down ones is more complex to understand and to predict because of its nature inherent to each person [Borji and Itti 2013; Le Meur and Coutrot 2016]. Thus, bottom-up and top-down factors alternatively influence the visual exploration during its course [Henderson 2003; Torralba et al. 2006]. The reasons of the switch between the two remain uncertain but [Unema et al. 2005] showed the existence of two visual processing modes related to the two visual pathways. These visual processing modes were define according to fixation duration and saccade amplitude parameters. The first mode is the ambient one which is defined by a short fixation (<180ms) followed by a large saccade (>5°) [Pannasch and Velichkovsky 2009; Velichkovsky et al. 2002]. Its role is to contextualize elements present within the visual scene [Pannasch et al. 2008; Velichkovsky et al. 2002]. The second one is the focal mode which is characterized by a long fixation (>180ms) followed by a short saccade (<5°) [Helo et al. 2014; Velichkovsky et al. 2002]. It allows to identify specific elements of the visual scene. As reported by [Velichkovsky et al. 2002], ambient mode is dominant during the two first seconds of the exploration while the focal mode gradually becomes dominant overt time. Therefore ambient mode was associated with bottom-up factors while focal mode seems to be related to top-down ones [Helo et al. 2014].

The interest in the dynamic of the ocular exploration and more recently in these two visual processing modes led researchers to implement several ratios to describe and exploit these aspects. To our knowledge, Goldberg and Kotval [Goldberg and Kotval 1999] were the first to try to represent this dynamic by describing its diversity. In their study, these authors proposed a ratio separating fixation duration and saccade duration, visual information processing taking place only during the fixation. However, they had limited results with this method. More recently, [Dehais et al. 2015] introduced an improved version of this ratio based on the distinction between short and long fixations.

Nevertheless, such ratios are not directly related to the two visual processing modes described here but report an interest of researchers to explain the complexity of visual exploration. To our


*ETRA '19, June 25–28, 2019, Denver , CO, USA*

ACM ISBN 978-1-4503-6709-7/19/06…$15.00
https://doi.org/10.1145/3314111.3323075

knowledge, the first and only ratio created in order to represent ambient and focal modes is the one proposed by [Krejtz et al. 2016]. In their work, the authors exploited the oculomotor parameters defined by [Unema et al. 2005] to identify the two visual processing modes and their respective intensity. Thereby, this ratio called K coefficient, seems to be a good global estimator while respecting the definition of the two modes given by [Unema et al. 2005].

The main goal of the present work is to assess whether the K coefficient described in [Krejtz et al. 2016] could be a good tool to understand visual behavior. That is why we introduce new analyses in which we use this coefficient and extend it with two new variables in order to give more insights into the understanding of visual exploration. These analyses can be independently used for global and temporal analyses. To demonstrate their utility, we use them to discriminate free-viewing and visual search tasks.

## 2 DATASET

As part of the 2019 ACM Eye-Tracking Research and Application (ETRA) conference, a new dataset composed of images and raw eye movements recordings has been released. We only describe here the subset we used for our analyses, see [McCamy et al. 2014; Otero-Millan et al. 2008] for further details.

### 2.1 Participants

The head of participants was placed on a chin-rest 57 cm from the video monitor (75 Hz refresh rate). Eye-tracking data were recorded from eight participants (2 women and 6 men) in three experimental sessions of 60 minutes each.

### 2.2 Stimuli

Stimuli were split in four categories: Blank scene, Natural scene, where is Waldo scene and picture puzzle but we only are interested in the first three. The first one was a plain 50% grey displayed on the whole screen. The second category contained 15 images of multiple scenes from flower bed to school bus. Some scenes included people and animals but never at the same time. The third category was composed of 15 where is Waldo scenes.

### 2.3 Tasks

Participants were asked to perform three tasks: free-viewing, visual search and fixation but we only are interested by the first two. For Natural stimuli, participants had to complete a free-viewing task while they had to search a visual target (Waldo, or another character or item) into the Waldo scene.

### 2.4 Experimental Design

All Participants performed all the conditions, 4 in free-viewing task and 4 in fixation task. For all the conditions, stimuli were presented during 45 seconds. In the fixation task, participants received an auditory feedback when their gaze left an area of 2 deg around the fixation cross for more than 500ms. In the visual search task for puzzle scenes and "*Where is Waldo ?*" scenes, they had to click where they thought the differences or the targets were, after the 45 seconds.

### 2.5 Data cleaning

Our interest in this research is to study the dynamics of ocular exploration. For this reason we kept data from Blank, and Natural scenes in the free-viewing condition and *where is Waldo ?* in the visual search condition. Provided data are samples of events recorded by the eye-tracker every 2 milliseconds. In order to aggregate and identify fixations and saccades, we used the identification velocity-based algorithm (I-VT) from [Salvucci and Goldberg 2000]. Then, we set a velocity threshold of 100°/s to separate fixations and saccades. Next, based on fixation duration distributions, we removed outliers by deleting every fixation under 100 milliseconds and greater than 1000 ms. Finally, we removed fixations outside the screen and re-computed saccades.

## 3 ANALYSES AND RESULTS

We first globally computed the K coefficient's intensity as defined in [Krejtz et al. 2016] to understand the general tendencies across stimuli and tasks. We then completed it with other variables such as mean duration in each mode to illustrate the wide variety of possibilities brought by this same coefficient. Finally, as the visual exploration is dynamic, we selected the most interesting variables and observed them through time. We first checked that our basic statistics of eye movements were in accordance with McCamy (2014) and Otero (2008) [McCamy et al. 2014; Otero-Millan et al. 2008]. We observed similar fixation durations for Waldo stimuli (M=282.7, SD=131.8), Natural stimuli (M=287.7, SD=150.5) and Blank stimuli (M=360.7, SD=201) as well as saccades amplitudes for Waldo stimuli (M=4.41, SD=4.25), Natural stimuli (M=5.58, SD=4.81) and Blank stimuli (M=7.89, SD=6.81).

### 3.1 Intensity

We compute the K coefficient's intensity as described in [Krejtz et al. 2016]. To this end and as shown in equation 1, the K coefficient is the z-scored difference between fixation duration and next saccade amplitude where $\mu$ and $\sigma$ are respectively the mean and the standard deviation of fixations duration or next saccades amplitude within a trial.

$$K = \frac{1}{n}\sum_{n}\frac{d_i - \mu_d}{\sigma_d} - \frac{a_{i+1} - \mu_a}{\sigma_a} \tag{1}$$

Thus, a negative value of the K coefficient means that the fixation duration $d_i$ deviates from the mean duration and the next saccade amplitude $a_{i+1}$ is a long saccade (>5°) which deviates from the mean amplitude of the trial. On the contrary, a positive value indicates that the fixation $d_i$ and the next saccade $a_{i+1}$ correspond to a focal mode.

We did not found significant differences between Blank, Natural and Waldo stimuli, $F(2,14) = 1.38, p > .05$. As seen in Table 1, means of K coefficient are very similar and close to 0, hence the fact that there is no dominant mode. For this reason, the variation of gaze behavior during the exploration added to the characteristics of tasks and stimuli do not allow to differentiate visual explorations through K coefficient.

**Table 1: Means and standard deviations (std) of K coefficient, number of switches, ambient and focal durations as a function of the tasks and visual stimuli: free-viewing in Blank and Natural scenes and visual search in Waldo scenes**

| | Blank | | Natural | | Waldo | |
|---|---|---|---|---|---|---|
| | mean | std | mean | std | mean | std |
| Coeff. K | 4.43e-18 | 3.34e-16 | 3.15e-17 | 4.03e-16 | -4.55e-17 | 3.83e-16 |
| Number of switches | 34.21 | 12.86 | 62.81 | 12.56 | 69.42 | 11.70 |
| Ambient duration | 330.12 | 51.73 | 253.94 | 26.64 | 250.36 | 23.56 |
| Focal duration | 529.56 | 104.81 | 381.35 | 57.38 | 360.89 | 45.05 |

### 3.2 Stability

Here we extend the coefficient by computing two new variables which are the average duration per mode (ms) and the number of switches between modes. The first allows us to know the mean duration during which participants stayed in the same mode. The higher the value is, the more the time spent in each mode increases and the more stable the participant's exploration is. To do so, we determined when a mode session started and when it finished. Then we calculated the mean duration by adding the fixation durations and saccade durations for each session. Next, we calculated the mean duration of each session in ambient mode and in focal mode as shown in Table 1. Unlike with the K coefficient, we found a significant main effect of the stimulus type on mean ambient duration. Moreover, differences on the mean ambient duration were significant between the three stimuli $F(2,14) = 39.61$, $p<.001$. A Tukey test showed that differences between Blank and Waldo stimuli was significant, $t(7) = 56.44$, $p<.001$; as between Blank and Natural stimuli, $t(7) = 46.35$,p.<.001. However, difference between Natural and Waldo stimuli, $t(7) = 1.89$,$p>.05$, was not significant. We observed the same significant effect on the mean duration in focal $F(2,14) = 49.75$,$p<.001$. Tukey test analyses showed again that Blank and Waldo stimuli were significantly different, $t(7) = 47.24$,$p<.001$; as well as the difference between Blank and Natural stimuli, $t(7) = 36.95$, $p.<.001$. Difference between Natural and Waldo stimuli $t(7) = 4.83$, $p>.05$, was not significant. Such analyses based on the mode stability reveal differences between visual explorations, in particular for the Blank stimulus but not between Waldo and Natural scenes.

The second variable allows us to investigate another aspect of the mode stability: the number of mode switches during recording. This variable corresponds to the number of times K coefficient switches from positive to negative values or the reverse. As for mean duration per mode, we found a significant main effect of the stimulus type on the number of switches $F(2,14) = 100.21$,$p<.001$, see Table 1). Interestingly, Tukey analyses revealed differences between the three stimuli. Blank stimulus differed from Waldo stimulus, $t(7)= 231.46$,$p<.001$) as well as from Natural stimulus $t(7) = 112.85$,$p<.001$. However, the difference was significant here between Natural and Waldo stimuli, $t(7) = 5.23$,$p<.05$. Differences emerged when indicators of stability were taken into account, suggesting to turn to other analyses to better explain the dynamics of visual explorations and differentiate them as a function of each stimulus presented to the participants.

### 3.3 Dynamics

The mean duration of each mode and the number of switches between modes change over time across the three stimuli. This provides more information than global analyses which does not take into account the temporal dynamic. Thus we need to consider the dynamic of the exploration by dividing our data in time sequences and observe the time course of our variables. To minimize our data loss, we removed every records after 34s which corresponds to the shortest trial after cleaning. We then divided each trial in eight sequences of 4.25s.

As shown in Figure 1, there is a significant effect of time on the number of switches and K coefficient $F(7,49) = 7.80$, $p<.001$ which increases over time. Moreover, the exploration of all the three stimuli begins with an ambient mode which then tend to focal mode over time. We noticed that in the first sequence, K coefficient was significantly different between Waldo stimulus and Natural stimulus $t(7) = 6.82$,$p<.05$, Waldo stimulus and Blank stimulus $t(7) = 8.20$,$p<.05$ but not between Blank stimulus and Natural stimulus $t(7) =0.02$,$p>.05$. These differences are not significant from the next sequence until the end of exploration, respectively $t(7) = 1.66$,$p>.05$ ;$t(7) = 1.06$,$p>.05$ ;$t(7) = 3.32$,$p>.05$.

When considering the number of mode switches, differences between Natural and Waldo stimuli were not significant for the first sequence $t(7)=0.56$,$p>.05$, the seventh sequence $t(7) = 2.49$, $p>.05$ and the eighth time sequence $t(7) = 3.36$,$p>.05$. For each other sequences, differences between Blank stimulus, Natural and Waldo stimuli were significant $t(7) = 28.29$, p<.001 ; $t(7) = 106.79$,$p<.001$.

If we put these observations in perspective, it becomes clear why the K coefficient did not discriminate visual exploration between stimuli. There were two sequences of approximately 9s where K coefficient differentiated stimuli against the last six sequences covering 25.5s which did not differ. Therefore, the analysis of K coefficient through time gave new insights on how to discriminate stimuli and tasks. However, the coefficient could only discriminate 25% of the exploration. As shown in Figure 1 (right), the difference between *Where's Waldo* and Natural conditions remained significant during the first half of the exploration (i.e. for each of the first four sequences). It is interesting to note that the difference is visible longer than for the K coefficient.

## 4 DISCUSSION AND CONCLUSION

The analyses of the K coefficient showed that a global approach is too coarse to emerge significant relationships between stimuli and modes. When analyzed statistically, the values of coefficient K for the three stimuli are close to the origin. This could be explained

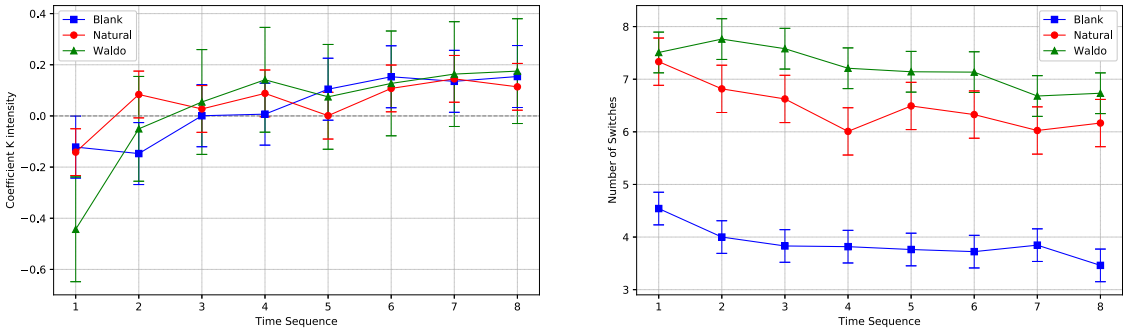



**Figure 1: K Coefficient (left panel) and Number of switches (right panel) as a function of stimuli and time sequence ranges.**

by the fact that the gaze behavior is dynamic in essence and is constantly changing between ambient and focal modes. These fluctuations could cancel each other and result in a coefficient near 0. This assumption is supported by the fact that significant differences appeared when the coefficient was analyzed over time. In addition, we observed a dominant ambient mode during the first sequence which turned into a focal mode during the second sequence except for Blank stimulus probably due to the fact that no visual stimulus is available and thus does not require cognitive resources. Thus, participants could stay much longer in each mode. This hypothesis is supported by the results shown in Figure 1 (right).

The analyses of mean duration in ambient and focal modes are to put in perspective with the number of switches. When the mean duration in each mode increases, the number of switches decreases and reverse. A higher mean duration in ambient implies more contextualization from the participant and less processing but does not necessarily mean the dominant mode is ambient.

The investigation on the number of mode switches allowed us to discriminate the stimuli up to 50% of the exploration. This improvement suggests the number of switches could help to better explain bottom-up and top-down influences during the visual exploration.

In this study, it is important to note that we were limited by the missing information about the given tasks and the target identity in *Where's Waldo* condition. Indeed, we do not know when the participant found the target, impeding to take into account only the period where the participant was really performing a visual search rather than the entire recording.

We think that future works should take into account these variables based on the K coefficient and their dynamic analyses, as they provide very interesting tools to better understand ocular behavior in situations differing as for visual inputs or goals.

## ACKNOWLEDGMENTS

This work is supported by the French Research and Technology Association (ANRT) under Grant No.2016/0957 and the company Sublime Skinz.

## REFERENCES

Ali Borji and Laurent Itti. 2013. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence* 35, 1 (2013), 185–207.

Antoine Coutrot and Nathalie Guyader. 2014. How saliency, faces, and sound influence gaze in dynamic social scenes. *Journal of vision* 14, 8 (2014), 5–5.

Frédéric Dehais, Vsevolod Peysakhovich, Sebastien Scannella, Jennifer Fongue, and Thibault Gateau. 2015. "Automation Surprise" in Aviation: Real-Time Solutions. In *CHI*.

Joseph H Goldberg and Xerxes P Kotval. 1999. Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics* 24, 6 (1999), 631 – 645. https://doi.org/10.1016/S0169-8141(98)00068-7

Andrea Helo, Sebastian Pannasch, Louah Sirri, and Pia Rämä. 2014. The maturation of eye movement behavior: Scene viewing characteristics in children and adults. *Vision research* 103 (2014), 83–91.

John M Henderson. 2003. Human gaze control during real-world scene perception. *Trends in cognitive sciences* 7, 11 (2003), 498–504.

John M Henderson and Andrew Hollingworth. 1999. High-level scene perception. *Annual review of psychology* 50, 1 (1999), 243–271.

Krzysztof Krejtz, Andrew Duchowski, Izabela Krejtz, Agnieszka Szarkowska, and Agata Kopacz. 2016. Discerning Ambient/Focal Attention with Coefficient K. *ACM Transactions on Applied Perception* 13, 3 (2016), 1–20. https://doi.org/10.1145/2896452

Olivier Le Meur and Antoine Coutrot. 2016. Introducing context-dependent and spatially-variant viewing biases in saccadic models. *Vision research* 121 (2016), 72–84.

Michael B McCamy, Jorge Otero-Millan, Leandro Luigi Di Stasi, Stephen L Macknik, and Susana Martinez-Conde. 2014. Highly informative natural scene regions increase microsaccade production during visual scanning. *Journal of neuroscience* 34, 8 (2014), 2956–2966.

Jorge Otero-Millan, Xoana G Troncoso, Stephen L Macknik, Ignacio Serrano-Pedraza, and Susana Martinez-Conde. 2008. Saccades and microsaccades during visual fixation, exploration, and search: foundations for a common saccadic generator. *Journal of vision* 8, 14 (2008), 21–21.

Sebastian Pannasch, Jens R Helmert, Katharina Roth, Ann-Katrin Herbold, Henrik Walter, et al. 2008. Visual fixation durations and saccade amplitudes: Shifting relationship in a variety of conditions. *Journal of Eye Movement Research* 2, 2 (2008), 1–19.

Sebastian Pannasch and Boris M Velichkovsky. 2009. Distractor effect and saccade amplitudes: Further evidence on different modes of processing in free exploration of visual images. *Visual Cognition* 17, 6-7 (2009), 1109–1131.

Dario D Salvucci and Joseph H Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*. ACM, 71–78.

Benjamin W Tatler, Iain D Gilchrist, and Jenny Rusted. 2003. The time course of abstract visual representation. *Perception* 32, 5 (2003), 579–592.

Benjamin W Tatler, Benjamin T Vincent, et al. 2008. Systematic tendencies in scene viewing. *Journal of Eye Movement Research* 2, 2 (2008), 1–18.

Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson. 2006. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review* 113, 4 (2006), 766.

Pieter J.A. Unema, Sebastian Pannasch, Markus Joos, and Boris M. Velichkovsky. 2005. Time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration. *Visual Cognition* 12, 3 (2005), 473–494. https://doi.org/10.1080/13506280444000409

Boris M Velichkovsky, Alexandra Rothert, Mathias Kopf, Sascha M Dornhöfer, and Markus Joos. 2002. Towards an express-diagnostics for level of processing and hazard perception. *Transportation Research Part F: Traffic Psychology and Behaviour* 5, 2 (2002), 145–156.

Alfred L Yarbus. 1967. Eye movements during perception of complex objects. (1967), 171–211.

## 6.4 Poster 2: "Different visual explorations between free-viewing and target finding tasks in websites: evidence from temporal analyses of ambient and focal modes"

*Milisavljevic, A., Le Bras, T., Abate, F., Gosselin, B., Petermann, C., Mancas, M. and Doré-Mazars, K.(2019). Different visual explorations between free-viewing and target finding tasks in websites: evidence from temporal analyses of ambient and focal modes. $20^{th}$ European Conference on Eye Movements, 18-22 August 2018, Alicante, Spain. Journal of Eye Movement Research 12(7), p. 390*

### Poster 2 summary

In **Article 1**, we showed that the K coefficient and our k-based variables (Krejtz et al., 2016) provided more precise tools to differentiate tasks and understand oculomotor behaviour. **Poster 2** aimed to show that the investigation of ambient and focal mode carried out in **Article 1** could be generalised to web pages. Similarly to **Poster 1**, we asked 116 participants to browse 18 websites in ecologically valid conditions. They had to perform two types of task: free viewing and visual search. We then reproduced the methodology used in **Article 1**: we first used the global intensity of the K coefficient (Krejtz et al., 2016) to differentiate tasks, and we then analysed its dynamic to refine the results. We replicated our previous results: the global intensity described by the K coefficient did not allow us to differentiate the tasks, whereas the k-based variables provided more precise results. Furthermore, although the number of changes between visual modes was similar for both tasks at the beginning of the exploration, our results showed that it decreased within the two first seconds of visual search.

## Résumé du Poster 2

Dans l'**Article 1**, nous avons montré que la dynamique du coefficient K et de nos variables basées sur K (Krejtz et al., 2016) fournissaient des outils plus précis pour différencier les tâches et comprendre le comportement oculomoteur. Le **Poster 2** visait à montrer que l'étude de la dynamique des modes visuels ambient et focal réalisée dans l'**Article 1** pouvait être généralisée aux pages web. Comme pour le **Poster 1**, nous avons demandé à 116 participants de parcourir 18 sites web dans des conditions écologiques. Ils devaient effectuer deux types de tâches : une exploration libre et une recherche visuelle. Nous avons ensuite reproduit la méthodologie utilisée dans l'**Article 1** : différencier d'abord les tâches en utilisant l'intensité globale du coefficient K (Krejtz et al., 2016), puis analyser sa dynamique dans le but d'affiner les résultats. Nous avons montré que, comme précédemment, l'intensité globale décrite par le coefficient K ne permettait pas de différencier les tâches, alors que les variables basées sur K donnaient des résultats plus précis. De plus, nous avons observé que bien que le nombre de changements de mode était similaire pour les deux tâches au début de l'exploration, il diminuait dans les deux premières secondes lors d'une recherche visuelle.

# Different visual explorations between free-viewing and target finding tasks in websites: evidence from temporal analyses of ambient and focal modes

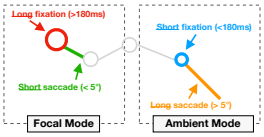
Alexandre Milisavljevic[1,2,3], Thomas Le Bras[1], Fabrice Abate[1], Matei Mancas[2], Coralie Petermann[3], Bernard Gosselin[2], Karine Doré-Mazars[1]

[1] Laboratoire Vision Action Cognition EA 7326, Institut de Psychologie, Université Paris Descartes, Boulogne-Billancourt, France
[2] Numediart institute, University of Mons, Mons, Belgium
[3] Research and Development department, Sublime Skinz, Paris, France

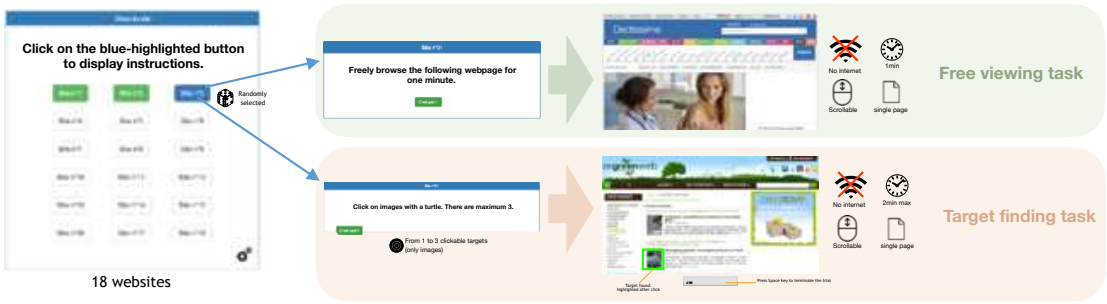
alexandre.milisavljevic@etu.parisdescartes.fr

## INTRODUCTION

Two visual exploration modes were highlighted by Unema et al. (2005) from Trevarthen's work (1968) on the two visual pathways. An ambient mode which is influenced by bottom-up factors and a focal mode which is influenced by top-down factors. Kreijtz et al. (2016) proposed the K coefficient to measure these two modes.
In the present study we use K-derived new variables as described by Milisavljevic et al. (2019) in order to study task's effects on gaze during the exploration of webpages.




## METHODS




**Apparatus**
- EyeLink 1000 (SR Research®).
- Temporal resolution of 1KHz.
- Spatial resolution of 0.05°.
- 24.5-inch screen size.
- 144Hz screen's refresh rate.
- 1920x1080 of screen resolution

**Participants**
- 116 participants.
- Normal or corrected-to-normal vision.
- 19 ♂ ; 97 ♀ ; ~25.5 y.o.

**Stimuli**
- Mean height of 6505px.
- Width of 1920px.
- Scrollable.

**Data Cleaning**
- Records >= 35s are kept.
- 15 bins of 2.33 seconds each.
- Blinks and outliers removed.

## RESULTS

### Global analyses

First, we ran global analyses on classic eye movements variables which are *fixation duration* and *saccade amplitude*.
We found significant differences between Target Finding and Free Viewing tasks.

Then, we computed the K coefficient (Krejtz, et al., 2016; **see formula on the right**) from which we computed 3 new variables from (Milisavljevic et al., 2019):
- number of mode switches
- mean duration in focal mode
- mean duration in ambient mode

$$K = \frac{1}{n}\sum_{n}\frac{d_i - \mu_d}{\sigma_d} - \frac{a_{i+1} - \mu_a}{\sigma_a}$$

$K > 0 \rightarrow$ *Focal*
$K < 0 \rightarrow$ *Ambient*

| | Free Viewing | Target Finding | F | p-value |
|---|---|---|---|---|
| Fixation duration (ms) | 233.11 | 249.13 | 58.27 | **p < .001** |
| Amplitude (°) | 4.79 | 7.01 | 415.80 | **p < .001** |
| K coefficient | 9.96e-19 | -1.63e-17 | 0.26 | p > .05 |
| Number of mode switches | 61.1 | 58.8 | 6.07 | **p < .05** |
| Mean duration in focal mode (ms) | 308.07 | 333.75 | 55.94 | **p < .001** |
| Mean duration in ambient mode (ms) | 219.56 | 238.73 | 185.40 | **p < .001** |

All the new variables related to the K coefficient (**see table on the left**), in contrast to the K coefficient itself, are able to significantly better differentiate the gaze behavior dynamic between the tasks.

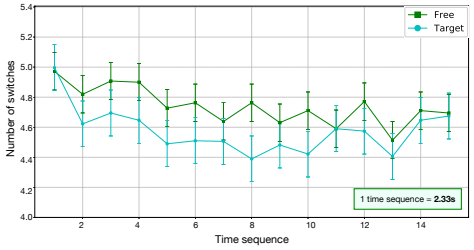
Global analyses show the task's influence on all our variables.
New variables show that participants stay longer in both focal and ambient modes during the target finding task than during the free viewing task.
The significant task's effect on the number of switches highlights that gaze behavior dynamic variates as a function of the task given to the participant.

### Temporal dynamics analyses

The number of mode switches (**see figure on the right**) is different between tasks only for the first part of the visual exploration.

However, Target Finding task seems to have a smaller number of switches, for every time sequences, compared to Free Viewing task.

Taken together, these results can be explained by the fact that in Target Finding task, participants have to stay more often in focal mode to discriminate the target.




## CONCLUSION

Global analyses show a significant effect of the task on all variables except the K coefficient. This result highlights that the new variables we propose are more useful than the K coefficient to discriminate tasks based on gaze behavior. Further dynamical analyses show that our new variables are more robust and thus reveal significant task's effects on the K coefficient.
Future work should focus on visual processing modes when the target is detected by the participant. These analyses should also be put in perspective of the mouse's scroll during the exploration of the webpage.

### References

Trevarthen, C.B. (1968). Two mechanisms of vision in primates. Psychol. Forsch, 31: 299.
Unema, P. J. A., Pannasch, S., Joos, M., & Velichkovsky, B. M. (2005). Time course of information processing during scene perception. Visual Cognition 12(3). 473-494.
Kreijtz, K., Duchowski, A., Krejtz, I., Szarkowska, A., & Kopacz, A. (2016). Discerning ambient/focal attention with coefficient K. ACM Transactions on Applied Perception (TAP), 13(3), 11.
Milisavljevic, A., Le Bras, T., Mancas, M., Petermann, C., Gosselin, B. & Doré-Mazars, K. (2019). Towards a better description of visual exploration through temporal dynamic of ambient and focal modes. ACM Eye-Tracking Research and Application (ETRA), Denver, United-States.

*20th European Conference on Eye Movements, Alicante, Spain, 2019, 18th to 22th August*

CHAPTER 7

# EYE MOVEMENT BEHAVIOUR ON WEB PAGES

## Contents

**References:**

- **Poster 3** (page 136): *Milisavljevic, A., Doré-Mazars, K., Gosselin, B., Mancas, M., & Petermann, C. (2017). What scroll can teach us about web users ? $40^{th}$ European Conference on Visual Perception , 27-31 August 2017, Berlin, Germany. Perception, ECVP 2017*

- **Article 2** (page 139): *Milisavljevic, A., Hamard, K., Petermann, C., Gosselin, B., Doré-Mazars, K., & Mancas, M. (2018). Eye and Mouse Coordination During Task: From Behaviour to Prediction. In Proceedings of the $13^{th}$ International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications(pp. 86–93). Funchal, Madeira, Portugal: SCITEPRESS - Science and Technology Publications*

- **Article 3** (page 149): *Milisavljevic, A., Abate, F., Le Bras, T., Petermann, C., Gosselin, B., Mancas, M., & Doré-Mazars, K. (Under revision). Similarities and differences between eye and mouse dynamics during web pages exploration. Frontiers in Psychology*

## 7.1 Contributions presentation

The purpose of the second axis of this work was to investigate eye movement behaviour on web pages through the relationship between the eyes, the movement of the mouse pointer and the scroll. Ecologically valid web pages need to be studied differently than images because of possible interactions with the web page itself. These interactions can take multiple forms, including clicks, scrolling, and drags and drops. While clicks and drags and drops are a mean to directly change or update the content, scrolling is more about the discovery of the content.

Most studies about scrolling behaviour focus on eye movement behaviour while scrolling and reading text documents (see Dyson (2004) for a review), but few studies have examined this behaviour on web pages. The relationship between the eyes and the scroll is even less studied. Yet, the scroll is used by billions of people when accessing internet or a smartphone. Hence, we investigated the relationship between the eyes and the scroll on web pages in **Poster 3** and **Article 2**. We showed that the eyes location could be used to infer next or current scroll parameters. For instance, we observed that

when quickly scrolling, we tended to orient our eyes towards the opposite direction of the scroll.

Next, we analysed the relationship between the mouse cursor and the eyes in **Article 2**. Interestingly, the study of this relationship attracted a lot of interest from search engine companies, such as Google and Microsoft. Due to the nature of these companies, the vast majority of studies on this subject were done on Search Engine Result Page (SERP). The problem is that those web pages are not representative of the web we use every day. In **Article 2**, we investigated this relationship on classic web pages and proposed a model to estimate eye position based on the mouse cursor position. As in SERP, we showed that the coordination between the eyes and mouse cursor was better on the vertical axis. However, we showed that when participants were about to click, the coordination between eyes and mouse on the horizontal axis increased.

So far, the relationships between eyes and scroll, or between eyes and mouse, have mostly been studied from an Area Of Interest (AOI) point of view. For instance, classical measures include the duration of eye fixations in a given area, or the number of clicks needed to reach the designated target. The literature on the statistical description of eye movements on web pages, mouse movements and scrolling is very sparse. For these reasons, we proposed in **Article 3** a detailed statistical description of eye movements on web pages along with the mouse movements and the scroll. This analysis included global parameters and their time courses. Furthermore, in order to evaluate if the ambient and focal modes can be generalised to the mouse and scroll, We extended the use of K coefficient (Krejtz et al., 2016) to the analysis of mouse dynamics. We found that eye and mouse saccade-related parameters decreased over time, while scrolling parameters increased. Conversely eye and mouse fixation-related parameters increased over time, while scroll parameters decreased. In both cases, eye and mouse parameters followed the same pattern, and the scroll parameters followed the opposite one. Interestingly, these observations were consistent across tasks.

## 7.2 Poster 3: "What scroll can teach us about web users ?"

*Milisavljevic, A., Doré-Mazars, K., Gosselin, B., Mancas, M., & Petermann, C. (2017). What scroll can teach us about web users ? 40$^{th}$ European Conference on Visual Perception , 27-31 August 2017, Berlin, Germany. Perception, ECVP 2017*

### Poster 3 summary

In this preliminary study, we asked 5 participants to browse 10 websites in ecologically valid conditions. They were asked to perform three types of task: free exploration, visual search and text reading. The aim was to analyse the influence of web page scrolling on visual exploration. We show that participants do not systematically anticipate scrolling by orienting their gaze in the same direction as the upcoming scroll. However, when this anticipation happens, participants do scroll faster. We also show that the amplitude of the scroll is related to the last known position of the eyes before the scroll. Furthermore, once the scroll is triggered, gaze position varies as a function of the speed of the scroll. This poster is also an opportunity to define what a scroll is: a continuous set of scrolls ending with a mouse movement.

## Résumé du poster 3

Dans cette étude préliminaire, nous avons demandé à 5 participants de parcourir 10 sites webs en conditions écologiques. Trois types de tâches leur ont été demandé : exploration libre, recherche visuelle et lecture de texte. Le but était d'analyser l'influence du défilement de la page internet sur l'exploration oculaire. Nous montrons que les participants n'anticipent pas systématiquement le défilement en dirigeant leur regard dans la même direction vers laquelle ils s'apprêtent à défiler. Toutefois, quand cela arrive le défilement est plus rapide. Nous avons également montré que l'amplitude du scroll est liée à la dernière position connue des yeux avant que celui-ci ne soit déclenché. De plus, une fois le défilement déclenché, les participants positionnement leurs yeux en fonction de la vitesse de celui-ci. Ce poster est également l'occasion de définir ce qu'est un scroll : un ensemble continu de défilement terminés par un mouvement de souris.

# What scroll can teach us about web users ?

**Alexandre Milisavljevic[1,2,3], Karine Doré-Mazars[1], Bernard Gosselin[2], Matei Mancas[2], Coralie Petermann[3]**

[1] Psychology institute, Vision Action Cognition EA 7326 lab - Paris Descartes University, Boulogne-Billancourt, France
[2] Numediart institute, University of Mons, Mons, Belgium
[3] Research and Development department, Sublime Skinz, Paris, France

## Introduction

Understanding why a user is on a webpage is a good way to deduce his or her interest in the content. To measure this interest, Eye-tracking is a precise tool that allows to estimate **goal impact on user's eye-gaze** (Yarbus, 1967). However, this method is hard to scale up.

### ▶ State of the art

Thus, mouse-tracking models emerged as an efficient proxy to determine user's attention (Rodden, 2007; Navalpakkam, 2013; Guo, 2010; Huang & White, 2012; Huang et al. 2012). These same models mainly use mouse movements, mouse clicks and hovered page elements while considering scrolling as a simple model feature. In addition to these analyses, other studies focused on how the eye behave during onscreen reading using scroll (Sharmin, 2013).

### ▶ Definitions

**Mouse movement:** Physical mouse shift resulting in a change of cursor position on the screen.
**Scroll:** use of the mouse's wheel or equivalent to scroll up or down
**Scroll session:** set of continuous scrolling events on the same webpage ended with a mouse movement

## Methods

### ▶ Set-up

- Two types of webpage to display : **instruction** or **website** bound to a task. (cf Figure 1)
- Instructions and websites links were stored on a local server.
- All instructions were stored locally and websites were online.

### ▶ Participants and apparatus

- 5 participants.
- Normal or corrected-to-normal vision.
- 4 ♂ ; 1 ♀ ; 24±2 y.o.
- Monocular recording with a FaceLab 5 eye-tracker.
- Google Chrome maximized at 1920x955.

### ▶ Tasks

- At most 2 webpage visits.
- No time limit.
- Full scroll possibilities.

| Tasks | Description |
|---|---|
| Free viewing | Browse the website by visiting at least two other pages |
| | Browse two articles of your choice |
| Target finding | Browse the following pages: calendar, team and news |
| | Buy the specific given item |
| Text reading | Read the two first paragraphs |

*Table 1: Tasks list by category*

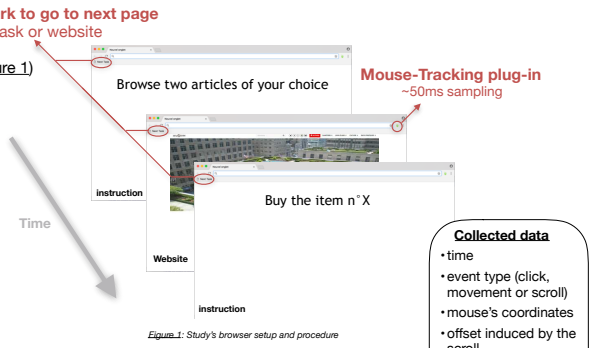



*Figure 1: Study's browser setup and procedure*

### ▶ Procedure

- Click on the bookmark situated in the browser's top bar request to the server the next link to load (cf Figure 1).
- Then website or instruction is displayed.
- When the user finishes reading or doing the task, he/she clicks on the bookmark again (cf Figure 1).
- Etc.

## Results

- For our analyses we divided the screen in 2 or 6 areas numbered as follows to have two-levels accuracy:

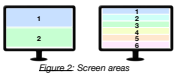


*Figure 2: Screen areas*

- We consider Free Viewing and Target Finding tasks because Reading task was too localized and did not require scroll.

### ▶ Anticipation

Analyses showed no significant effect of scroll anticipation by the eyes for both types of task in six-areas and two-areas configurations.

However we found that:

- **scroll is faster** when eyes anticipates the scroll.
- when the user begin to scroll, there is a higher probability that his/her eye position was on the same half of the screen than the direction of his/her scroll.

| | Mean speed with anticipation | Mean speed without anticipation |
|---|---|---|
| Target Finding | 1154px/s | 1038px/s |
| Free Viewing | 996px/s | 783px/s |

*Table 2: Speed in pixels per seconds with and without anticipation and its effects on scroll speed*

### ▶ Amplitude

Users adapted their eyes position before scrolling.

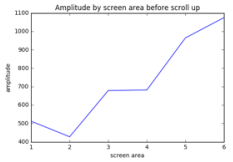



*Figure 3: Scroll up amplitude according to pre-scroll eyes position*

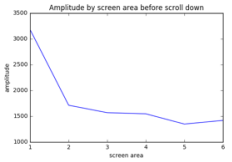



*Figure 4: Scroll down amplitude according to pre-scroll eyes position*

**correlation** of 0.94

- When the eyes were located on the bottom screen area before scrolling up the scroll amplitude increased.

**anti-correlation** of -0.77

- When the eyes were located on the top screen area before scrolling down the scroll amplitude was much higher.

### ▶ Speed

We observed that participants adapted their eyes position according to their scrolling speed (cf Figures 5, 6 and 7).

While looking for **specific information**, users scroll slower to be able to differentiate elements (text, blocs, titles, etc).
While looking for a more **generic information** like an image, a colored box surrounding a website category, etc, users positioned their eyes on the opposite side of the scroll direction. They used their peripheral vision to detect bottom-up elements.


*Figure 5: Scroll down speed according to eyes position while scrolling*

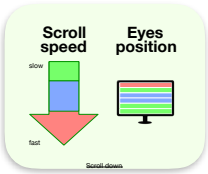



*Figure 6: Illustration of eyes position according to scroll speed while scrolling down*

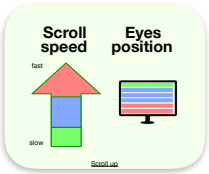



*Figure 7: Illustration of eyes position according to scroll speed while scrolling up*

## Conclusion

Users adapt their eyes position according to their intent:

- **Before scroll**, eyes position are a clue to guess how fast and how far the user will scroll
- **While scrolling**, users position their eyes according to where they think the information could be. While scrolling fast, they position their eyes at the **opposite direction** to be able to detect bottom-up characteristics through peripheral vision. Furthermore, when users are looking for a specific information or one that need more attention, they scroll more slowly and position their eyes in the center of the screen or in the same direction as the scroll.

## References

Guo, Q., & Agichtein, E. (2010). International Conference Extended Abstracts on Human Factors in Computing Systems
Huang, J., & White, R. (2012). Conference on Human Factors in Computing Systems
Huang, J., White, R. W., Buscher, G., & Wang, K. (2012). Conference on Research and Development in Information Retrieval
Navalpakkam, V. et al. (2013). Conference on World Wide Web
Rodden, K., & Fu, X. (2007). Conference on Research and Development in Information Retrieval
Sharmin, S. et al. (2013). Conference on Eye Tracking South Africa
Yarbus AL. (1967). Plenum Press

## Acknowledgment

We thank French Research and Technology Association and Sublime Skinz for supporting this work and Kevin Hamard for his help on this work.

*40th European Conference on Visual Perception, Berlin, Germany, 2017, August 27th - August 31st*

## 7.3 Article 2: 'Eye and Mouse coordination during task: from behaviour to prediction"

*Milisavljevic, A., Hamard, K., Petermann, C., Gosselin, B., Doré-Mazars, K., & Mancas, M. (2018). Eye and Mouse Coordination During Task: From Behaviour to Prediction. In Proceedings of the 13$^{th}$ International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications(pp. 86–93). Funchal, Madeira, Portugal: SCITEPRESS - Science and Technology Publications*

### Article 2 summary

In **Article 2**, we investigate the coordination between the mouse and the eye position, and the scroll and the eye position on web pages. As in **Poster 3**, we asked 5 participants to browse 10 websites in ecologically valid conditions. They were asked to perform three types of task: free viewing, visual search and text reading. We first analyse the eye-mouse coordination by comparing their density maps on the entire web page. Then, we analyse how the euclidean distance between the two effectors fluctuates with the task. We show that the coordination between the eyes and mouse cursor is better on the vertical axis. Based on these fluctuations, we propose a model predicting eye position based on mouse position. Finally, we show that the scroll amplitude varies with eye position before the start of the scroll, which confirms findings of **Poster 3**.

## Résumé de l'article 2

Dans l'**Article 2**, nous étudions la coordination entre la souris et l'oeil ainsi que le défilement et l'oeil sur les pages web. Comme dans le **Poster 3**, nous avons demandé à 5 participants de parcourir 10 sites web dans des conditions écologiques. Nous leur avons demandé d'effectuer trois types de tâches : exploration libre, recherche visuelle et lecture de texte. Nous analysons d'abord la coordination oeil-souris en comparant leurs cartes de densité sur l'ensemble de la page web. Ensuite, nous analysons comment la distance euclidienne entre les deux fluctue en fonction de la tâche. Nous montrons que la coordination entre les yeux et le curseur de la souris est meilleure sur l'axe vertical. Sur la base de ces fluctuations, nous proposons un modèle prédisant l'emplacement des yeux en fonction de la position de la souris. Enfin, nous montrons que l'amplitude du défilement varie en fonction de la position des yeux avant que celui-ci ne commence, ce qui confirme les résultats exposés dans le **Poster 3**.

# Eye and Mouse coordination during task: from behaviour to prediction

Alexandre Milisavljevic[1,2,3], Kevin Hamard[3], Coralie Petermann[3], Bernard Gosselin[1] Karine Dor-Mazars[2] and Matei Mancas[1]

[1]*Numediart institute, University of Mons, Mons, Belgium*

[2]*Psychology institute, VAC EA7326 team, Paris Descartes University, Paris, France*

[3]*Research and Development department, Sublime Skinz, Paris, France*

Keywords: Behaviour, Visual Attention, Webpages, Mouse-tracking

Abstract: The study of web users' behaviour is of crucial importance for understanding people reaction when browsing websites. Eye-tracking is a precise tool for this purpose, but it is hard to scale up when trying to apply it to a wide range of situations and websites. On the other hand, mouse-tracking fulfills these requirements. Unfortunately, mouse data provides a limited approximation of the eye position as it was shown in the literature. In this paper, we investigated the relationship between mouse and eye behaviour on several kind of websites with three different tasks to create models based on these behaviours. Our findings were that 1) saliency Pearson's correlation is not suitable to analyse eye and mouse coordination, 2) this coordination is altered according to the task, 3) scroll speed directly influence where the eyes are during the scroll, 4) amplitude vary according to eyes position before the scroll and 5) by using the $X$ axis variations it is possible to find the moments where it is easier to model eyes location from mouse location.

## 1 Introduction

Understanding why a user visits a webpage has been a central question since the beginning of the twenty first century. To answer this question, Eye-tracking has been used as a precise tool to estimate intention impact on users gaze. However, these kinds of studies are hard to scale up and apply it to a wide user panel is difficult. That is why mouse-tracking emerged as an efficient proxy to determine users attention. Since then, correlation between mouse movements and eye movements has been found (Mueller and Lockerd, 2001; Chen, 2001; Rodden and Fu, 2007; Rodden et al., 2008; Cooke, 2006; Guo and Agichtein, 2010; Huang and White, 2012; Navalpakkam et al., 2013) and modelling attempts followed (Guo and Agichtein, 2010; Huang and White, 2012; Navalpakkam et al., 2013; Boi et al., 2016).

Nevertheless, a majority of these studies focused on SERP (SEarch Result Pages) from search engines putting aside the rest of the web and tasks. In addition, eye-mouse and eye-task relationships have been studied separately (Yarbus, 1967; Castelhano et al., 2009; Mills et al., 2011) but rarely together. That is why, the goal of this study was to explore the eye-mouse-task relationship in a more diversified environment.

Chen (2001) was the first to show that areas visited by the mouse were also visited by the eye in free-viewing condition. Rodden and Fu (2007) also showed that regions visited by the mouse were also visited by the eye but they were the first to highlight the better correspondence on $Y$ axis between mouse and eye. Unlike previous work, they set-up an experiment with pre-defined search queries on a search engine. Guo and Agichtein (2010) confirmed Rodden and Fu (2007) results about more accurate correlation on $Y$ axis. Their main contribution was the first attempt to automatically infer the user's eye position using mouse movements. They also suggested the presence of images did not have a significant effect on eye-mouse coordination. Huang and White (2012) presented that amount of time spent on a search web page by a participant can affect where they were pointing and looking and then used this finding to enhance their algorithm. They showed that gaze-cursor alignment was distinct for each participant but did not highlight significant difference among women and men. Navalpakkam et al. (2013) updated previous work by investigating more recent SERP which now includes images and more complex content. They showed that this content induced different behaviour. Then they proposed a non-linear model

outperforming state-of-the-art models because of its non-linearity.

Our exploratory study aimed to investigate effect(s) of user's goal on eye and mouse coordination in ecological conditions (different categories of web sites, no scroll limitations). Our hypothesis was the following: there is a direct link between the eye, the mouse movements and the task which fluctuates according to the task. Thus, we could enhance the precision of the current models.

This paper was structured as follows: experiment set-up was described in section 2 followed by the results of the static and dynamic analyses in section 3. Finally we discussed and concluded the paper in section 4.

## 2 Method

We recruited five participants with normal or corrected-to-normal vision (4 males and 1 female) aged between 24 and 25 years from the local signal processing department. All participants were right-handed and fluent with computer operations. They were tested on 10 various websites including blogs, e-commerce platforms, etc. From the calibration phase at the beginning of the study to the end, the whole process took about 20 minutes per person.

### 2.1 Tasks

We used a set of five tasks distributed in three classes as presented in Table 1: free viewing, target finding and text reading. The number of pages that could be visited during Free viewing tasks was limited to two but was not limited in time and participants had full scroll possibilities. The reading task was specific enough to prevent any free interpretation in order to simulate participants' willingness to read a specific paragraph. Finally, we chose two types of target finding tasks: one in which participants were instructed to find and buy an item, the second in which they had to find a given page.

### 2.2 Set-up

To record eye movements, we used a FaceLAB 5 eye-tracker at 60Hz without head constraint on a 17-inch screen set to a resolution of 1920X1080. Instructions and websites were displayed in Google Chrome maximized with a resolution of 1920X955.

To record Mouse movements we developed a plug-in using WebExtensions[1] standard. It took the form of an ON/OFF button on the browser top bar and has only been used by the operator. The extension monitored the following metrics: time-stamp, event type (click, movement or scroll), mouse's coordinates, offset induced by the scroll, URL and screen size. The plug-in developed in Javascript was uploading all mentioned metrics on the fly or every 50-60ms to a NodeJS[2] server via a socket connection. The same server inserted in real time the data in a MySQL database without further processing. The server also kept track of the page to deliver to the participant.

| Category | Description |
|---|---|
| Free viewing | Browse the website by visiting at least two other pages |
| | Browse two articles of your choice |
| Target finding | Browse the following pages: calendar, team and news |
| | Buy the specific given item |
| Text reading | Read the two first paragraphs |

Table 1: Generic tasks used in the study.

### 2.3 Procedure

Participants started on a homepage describing the context of the study. To visit the next planned page by the study they had to click on a Javascript bookmark situated in the browser's bookmark top bar. The first click on it led them to the first task instruction. All tasks were stored in HTML format locally. After reading the instruction, participants could once again click on the bookmark and begin the task. When the task was completed participants had to click again on the bookmark to read the next instruction and so on. At the end of the study, participants were asked to answer to a survey about there knowledge of the websites.

## 3 Results

We ran three sets of analyses in order to highlight coordination between eye and mouse movements. First, we used 2D saliency metric *PCC* (Pearson's Correlation Coefficient) to check consistency between overall eye and mouse movements. Then we repeated the same analysis between participants' eyes movements. Second, we applied literature's temporal and distance estimation to our task-related context to bring out tasks' influence on eye and mouse co-

[1] https://developer.mozilla.org/Add-ons/WebExtensions

[2] https://nodejs.org/en/

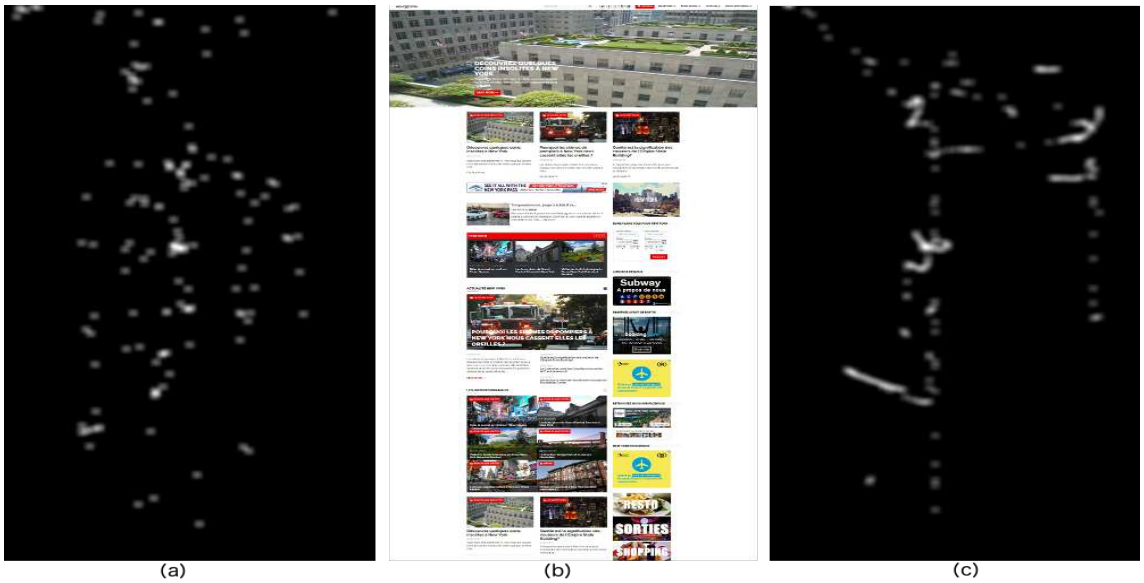



Figure 1: (a) eye fixation density map, (b) original website and (c) mouse fixation density.

ordination. Third, we analysed the participants' behaviour while scrolling because - at our knowledge - it has not been treated by the literature whereas it could be an essential information to the understanding of eye and mouse coordination. Finally, we used the results of the two first sets of observations to create two Gaussian-based models to approximate eye position from mouse position.

While the first approach focused on a static and spatial analysis, the second and third aimed for a dynamic analysis taking into account the temporal evolution of both eye and mouse tracks.

## 3.1 Static fixation densities comparison

Pearson's Correlation Coefficient (PCC) also known as the Pearson Product-Moment Correlation (1), is a metric used in saliency maps comparison by authors like Ouerhani et al. (2004) and Le Meur et al. (2007) and used to compare fixations and mouse movements by Tavakoli et al. (2017). *PCC* has a value between -1 and 1. When the coefficient is almost equal to 1, there is a strong relationship between the two variables. The goal was to apply this metric to highlight eye and mouse coordination changes between tasks. The originality of this metric lies in the fact that it uses probability densities instead of raw variables values. To do so, we computed PCC between eyes density map and mouse density map. To obtain these maps, fixations from eye-tracking and mouse-tracking were convolved with a Gaussian filter. Thus *PCC* was computed between images (a) and (c) as shown in Figure 1.

$$P_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{1}$$

We obtained for our three tasks classes (free viewing, target finding and text reading) defined in section 2.1 correlation scores as in Table 2, "inter" column. Both classes correlation and their relative difference remained small which showed that mouse-tracking could not be directly used to model eye movements. For this reason, we decided to refine the investigation based on motion dynamics in the next sections.

Furthermore, when comparing eye-tracking results between different participants on the same stimulus, we obtained results in Table 2, "intra-eye" column which showed a higher correlation for "Text reading" task than for the two others. This result confirmed that if the task and its location were precise, then most of the participants would produce similar eye-gaze patterns. We observed the same behaviour for mouse tracks in Table 2, "intra-mouse" column, but with a lower overall correlation which showed that mouse behaviour remained less consistent than eye behaviour.

## 3.2 Dynamic analyses

Considering the dominant use of scroll in our experiment, modern vertically-based designs and the tend of Human eye to be more efficient horizontally, we separated $X$ and $Y$ coordinates to enhance granularity

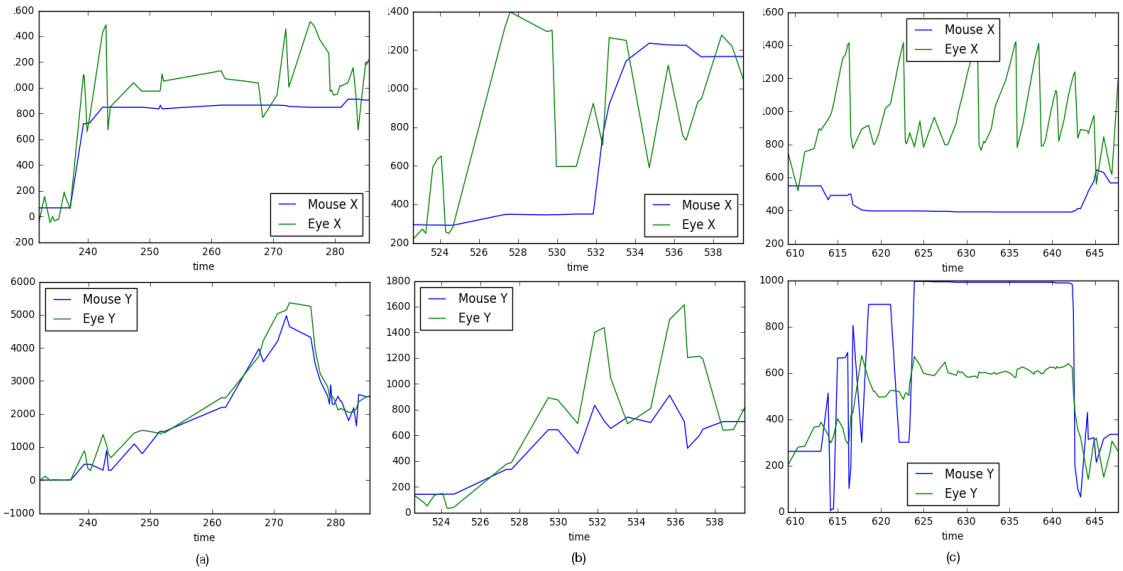



Figure 2: First column (a) is a free viewing task, second column (b) is a target finding task and third column (c) is a text reading task

| Task | inter | intra-eye | intra-mouse |
|---|---|---|---|
| Free viewing | 0.132 | 0.036 | 0.082 |
| Target finding | 0.171 | 0.028 | 0.107 |
| Text reading | 0.176 | 0.440 | 0.162 |

Table 2: Pearson's correlation coefficients for intra and inter analyses.

in our dynamic analyses. For each $X$ and $Y$ coordinate we got temporal vectors which were synchronized between mouse and eye. To do so, we matched eye fixations with mouse events and then down-sampled mouse data to fit eye data. We chose to not interpolate as in Deng et al. (2016) because it could have generated non-existing fixations and wrong results.

### 3.2.1 Temporal and distance estimation

We observed for some participants a time shift on $Y$ axis between mouse and eye with the mouse being delayed as in Figure 2 (a) and (c) right columns. This finding joined Huang and White (2012) previous work in which they detected a lag between the mouse and the eye. This could be explained by the fact that, in visual exploration context, the eye is the only mean of perception and leads the hand movements.

We computed euclidean distance (3) and obtained an eye-mouse distance of 554 pixels. This result was not in accordance with the average 229 pixels from state-of-the-art (Rodden and Fu, 2007; Guo and Agichtein, 2010; Huang and White, 2012). We then refined our analysis by separating the two axes. Using formula (2) we got a mean distance of 409 pixels for $X$ axis and 291 pixels for $Y$ axis. With this results we began to have a better consistency on $Y$ axis as expected. However, Bejan (2009) demonstrated that our eyes scan horizontally faster than in the vertical dimension. Based on our results, we could assumed that participants kept their mouse vertically stationary to scroll down or up and used it as a vertical pointer, allowing them to horizontally browse without difficulties. Thus the participant could easily move his eye on $X$ axis more often. That is why the participant tended to move it's eye on $X$ axis more often.

We then continued with separate axes to compute correlation. Compared to distances, correlation coefficients between mouse and eye were drastically different. Chen (2001) obtained a correlation of 0.58 with more than 50% of the pages associated with correlations larger than 0.8. In our study, we measured a mean correlation of 0.64 on $Y$ axis and 0.18 on $X$ axis.

Difference between axes got even more significant when we examined these correlations coefficients according to their corresponding task. As exposed in Table 3, free viewing task had the best correlation on $Y$ with 0.9. This result reflected a greater trend to use the mouse as a vertical pointer as in other tasks. Coefficients for target finding were more balanced with an increased correlation on $X$ but a decrease on $Y$. Finally, text reading correlations expressed the fact that participants did not used much the mouse during this task. We could assume that more the cognitive load of the task is important more the correlation drop on

| Task type | $r_x$ | $r_y$ |
|---|---|---|
| Free viewing | 0.176 | 0.921 |
| Target finding | 0.383 | 0.699 |
| Text reading | 0.006 | 0.32 |

Table 3: Pearson's correlation on X and Y axes.

both axes.

$$d(i) = |x_m(i) - x_e(i)| \quad (2)$$

$$d(X,Y) = \sqrt{(x_m - x_e)^2 + (y_m - y_e)^2} \quad (3)$$

### 3.2.2 Scroll's speed and direction influence

As we previously exposed, mouse and eyes were more correlated on $Y$ axis. In addition, scroll events are a barely studied subject while it is a common behaviour in all webpages browsing. We based all our calculus on scroll sessions which corresponds to a set of continuous scroll events ended with a mouse movement. Scroll is an important feature providing good information about the degree of participants' interest on a website. Another advantage is that the scroll is measurable on desktop and mobile. Through the following analyses, we highlighted influence of behaviour on scroll's speed and amplitude.
We collected for each scroll session the direction (up or down) and the absolute speed. After empirical tries and errors and after taking into account the amount of data, we also separated the browser screen into 3 equals categories as in Figure 5 to detect patterns.

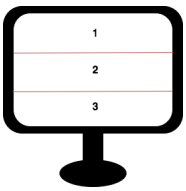



Figure 5: Screen's three areas

For the current analyses we removed text reading tasks because it did not included enough scroll events. For both amplitude and speed influence test, we performed a one-way independent ANOVA (analysis of variance) (4) test. The ANOVA examines if the mean of numeric variables differs across levels of categorical variables. After checking all assumptions (normality of errors, equal error variance across category, independence of errors), we hypothesized:

$$H_0 : \mu_0 = \mu_1 = \mu_2 \quad (4)$$

$H_1$ : At least one mean is not equal to the others.

As shown in Table 4, we considered that all means were equal to each other. The statistic test we ran was the ratio of the between-category variance and the within-category variance. If this ratio was greater than the critical probability distribution F, we could reject the null hypothesis. After obtaining a p-value below the 0.05 threshold, we could affirm the rejection of the null hypothesis with a confidence rate of 95 %. Thus, we can conclude that there is an effect of scroll speed on eyes category position.

| Indicator | Task | Down | Up |
|---|---|---|---|
| F-test | Free viewing | 4.26 | 7.07 |
| | Target finding | 3.76 | - |
| P-value | Free viewing | 0.017 | 0.001 |
| | Target finding | 0.031 | - |

Table 4: Result test ANOVA with significance level (p-value) and F-score.

To go further, we had to determine and define this influence. We focused on means for each tasks using a Tuckey's test. We observed that while scrolling quickly, participants positioned their eyes at the opposite side of the scroll's direction to be able to detect bottom-up characteristics through peripheral vision as shown in Figure 3. Furthermore, when participants were looking for a specific information (target finding task), they tended to quickly look towards the center of the screen when the scroll speed decreased.

Then we focused on scroll's amplitude, which is the distance between the start and the end of a scroll session. We wanted to know if participants adapted their eyes position before scrolling. Here again we differentiated the target finding and the free viewing tasks and calculated the means distance for each area before the participant scroll.

| Indicator | Task | Down | Up |
|---|---|---|---|
| F-test | Free viewing | 3.08 | 10.44 |
| | Target finding | 0.09 | - |
| P-value | Free viewing | $< 0.001$ | $< 0.001$ |
| | Target finding | 0.9 | - |

Table 5: Result test ANOVA with significance level (p-value) and F-score.

We could conclude using ANOVA test that for the free viewing task, when the eyes were located at the bottom of the screen and before scrolling up, the scroll amplitude increased with p-value $< 0.05$ as shown in Table 5. As expected, when the eyes were

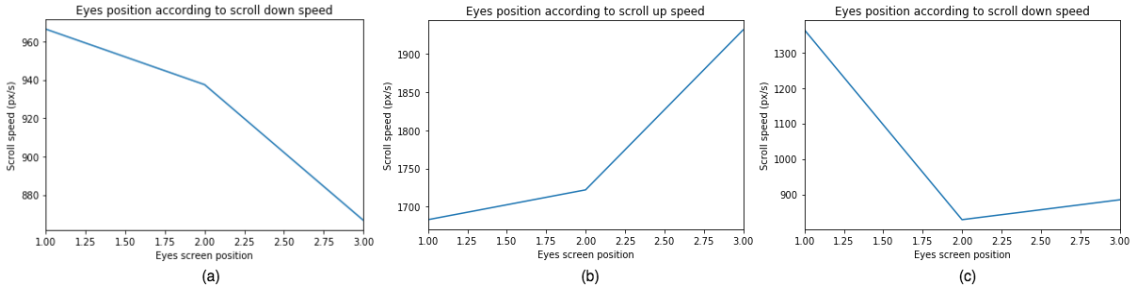



Figure 3: Eyes position according to scroll speed, (a) and (b) corresponds to free viewing task, (c) corresponds to target finding task

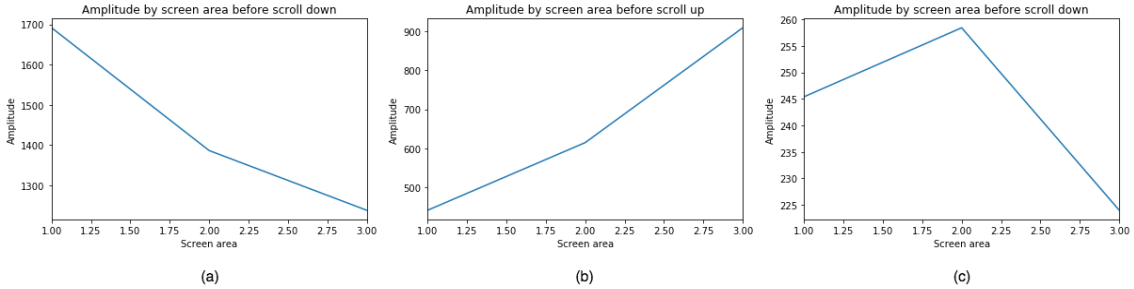



Figure 4: Scroll amplitude according to screen area before scroll, (a) and (b) corresponds to free viewing task, (c) corresponds to target finding task

located on the top of the screen before scrolling down the scroll amplitude were much higher, see Figure 4 (a) and (b). About the target finding task, there was no significant impact (c) of amplitude on the screen area before scrolling (p-value $> 0.05$). However, we noticed that when searching specific information, participants did not have a long scroll amplitude in order to not miss an element (text, blocs, titles, etc) and to differentiate them.

The scroll event could improve the prediction of the localization of the eyes on Y axis using the combination of direction, amplitude and speed variables.

## 3.3 Model

Previous analyses provided several insights about users behaviours on webpages given more or less specific tasks. We built our models from these, more particularly from the eyes movements standard deviations. As in section 3.2.1, we separated $X$ and $Y$ axes to infer the parameters of a Gaussian model which predicted the eyes position based on the mouse position and cognitive load of the task. From these standard deviations we were able to define a confidence area around the mouse in which the eyes had a 70% probability to be in it. We chose to base our calculus on the 70$^{th}$ percentile because it was the minimum confidence rate we observed in the state of the art. As shown in Table 6, columns "x std." and "y std.", the 70$^{th}$ percentile (5) gave a first coarse pixel area around the mouse cursor.

$$percentile = \mu \pm Z\sigma \tag{5}$$

But we were interested in a better model, so we focused on specific behaviours during tasks. As shown in Figure 2 (a) target finding class, we identified sudden changes on $X$ axis. After analysing participants' videos and comparing with several target finding tasks among them, we found that these sudden changes matched participant's interest. When the participant had a target finding goal and found his target, he quickly moved his mouse to the point of interest.

Thus, we manually defined a threshold at the beginning of each sudden changes and we computed the standard deviation before and after every final sudden change on $X$. We obtained better results as shown in Table 6, column "x std. thrs." and "x std. thrs.". Area covered by the 70$^{th}$ was reduced by around 150 pixels for both axes. With this second model, we were able to increase the accuracy but only by focusing on a specific event.

| Task | x std. | y std. | x std. thrs. | y std. thrs. |
|---|---|---|---|---|
| Free viewing | 558.0 | 416.4 | - | - |
| Target finding | 486.4 | 403.8 | 361.3 | 251.0 |
| Text reading | 627.8 | 257.9 | - | - |

Table 6: Standard deviation (percentile 70%) normal and using only sudden *X* changes.

## 4 Conclusion and Discussion

We first compared eye and mouse data with the saliency metric *PCC*. We did not find significant consistency between participants' eyes and mouse positions (inter) and between participants' eyes (intra). However, results showed that participants behaved in a more similar way when they had the same task with the same location (reading task).

Then, we got deeper with dynamic analyses. We showed that using distance and correlation, we were able to highlight more interesting coordinations between eyes and mouse. We had better results on *Y* axis than *X* axis and succeed to demonstrate behaviour differences between tasks. In addition, scroll analyses, clearly showed a relation between eyes position and scroll speed while browsing and amplitude before the scroll.

Finally, we made a model for each task able to predict the area around the mouse's cursor in which the eyes had 70% chances to be located in. However, eyes location uncertainty compared to mouse position remained high, even if we succeed to enhance the model during target finding task by observing brutal changes on *X* axis.

In this paper, we presented results of a preliminary study, used as a validation to conduct a bigger experiment, including more participants. This will allow us to analyse the impact of participants' age on their mouse movements. Moreover, we did not use scroll events analyses to enhance our models. In future work, we think that doing so, could boost the precision of the model by reducing the area around the mouse's cursor. We could also investigate new relations between the scroll and the eyes by analysing scroll in 2D. Then, we could use machine learning models to integrate new features and more user behaviours such as mouse patterns. Finally, our main objective is to propose the most accurate model in order to use it in real time to predict web user behaviours.

## 5 Acknowledgement

We thank French Research and Technology Association (ANRT) and Sublime Skinz for supporting this work.

## REFERENCES


Bejan, A. (2009). The goldenratio predicted: Vision, cognition and locomotion as a single design in nature. *International Journal of Design & Nature and Ecodynamics*, 4(2):97–104.

Boi, P., Fenu, G., Davide Spano, L., and Vargiu, V. (2016). Reconstructing User's Attention on the Web through Mouse Movements and Perception-Based Content Identification. *Transactions on Applied Perception*, 13(3):1–21.

Castelhano, M. S., Mack, M. L., and Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision*, 9(3):1–15.

Chen, M.-c. (2001). What can a mouse cursor tell us more? Correlation of eye / mouse movements on web browsing. In *Conference on Human Factors in Computing Systems*, pages 281–282.

Cooke, L. (2006). Is the Mouse a" Poor Man's Eye Tracker"? *Annual Conference-Society for Technical Communication*, 53:252 – 255.

Deng, S., Chang, J., Kirkby, J. A., and Zhang, J. J. (2016). Gazemouse coordinated movements and dependency with coordination demands in tracing. *Behaviour & Information Technology*, 35(8):665–679.

Guo, Q. and Agichtein, E. (2010). Towards predicting web searcher gaze position from mouse movements. In *Extended Abstracts of Conference on Human Factors in Computing Systems*, pages 3601–3606.

Huang, J. and White, R. (2012). User See, User Point: Gaze and Cursor Alignment in Web Search. In *Special Interest Group of Conference on Human Factors in Computing Systems*, pages 1341–1350.

Le Meur, O., Le Callet, P., and Barba, D. (2007). Predicting visual fixations on video based on low-level visual features. *Vision Research*, 47(19):2483–2498.

Mills, M., Hollingworth, A., and Dodd, M. D. (2011). Examining the influence of task set on eye movements and fi xations. *Journal of Vision*, 11(8):1–15.

Mueller, F. and Lockerd, A. (2001). Cheese: tracking mouse movement activity on websites, a tool for user modeling. In *Extended Abstracts of Conference on Human Factors in Computing Systems*, pages 279–280.

Navalpakkam, V., Jentzsch, L. L., Sayres, R., Ravi, S., Ahmed, A., and Smola, A. (2013). Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *International Conference on World Wide Web*, pages 953–964.

Ouerhani, N., Wartburg, R. V., and Heinz, H. (2004). Empirical Validation of the Saliency-based Model of Visual Attention. *Electronic Letters on Computer Vision and Image Analysis*, 3(1):13–24.

Rodden, K. and Fu, X. (2007). Exploring how mouse movements relate to eye movements on web search results pages. In *Special Interest Group on Information Retrieval Workshop on Web Information Seeking and Interaction*, pages 29–32.

Rodden, K., Fu, X., Aula, A., and Spiro, I. (2008). Eye-mouse coordination patterns on web search results pages. In *Extended Abstracts of Conference on Human Factors in Computing Systems*, pages 2997–3002.

Tavakoli, H. R., Ahmed, F., Borji, A., and Laaksonen, J. (2017). Saliency Revisited: Analysis of Mouse Movements versus Fixations. In *Conference on Computer Vision and Pattern Recognition*, pages 4321–4329.

Yarbus, A. L. (1967). Eye Movements. *New York: Plenum Press*.

## 7.4 Article 3: "Similarities and differences between eye and mouse dynamics during web pages exploration"

*Milisavljevic, A., Abate, F., Le Bras, T., Petermann, C., Gosselin, B., Mancas, M., & Doré-Mazars, K. (Under revision). Similarities and differences between eye and mouse dynamics during web pages exploration. Frontiers in Psychology*

### Article 3 summary

In the this article, we extend to the scroll, the analyses on eyes and mouse coordination started in **Article 2**. The purpose of **Article 3** is to statistically define a mouse movement and a scroll when exploring web pages. In addition, we study the oculomotor behaviour and examine the relationship between the eyes, the mouse and the scroll. To this end, we recorded the eye, mouse and scroll movements of 151 participants exploring 18 dynamic web pages while performing free viewing and visual search tasks for 20 seconds. The data revealed significant differences of eye, mouse and scroll parameters over time which stabilise at the end of exploration. This suggests the existence of a task-independent relationship between the eye, the mouse and the scroll parameters which is characterised by two distinct patterns: one common pattern for movement parameters and a second for dwelling/fixation parameters. Within these patterns, mouse and eye movements remained consistent with each other, while the scrolling behaved oppositely.

## Résumé de l'article 3

Dans cet article, nous appliquons au défilement, les analyses sur la coordination des yeux et de la souris qui ont été présentés dans l'**Article 2**. Le but de l'**Article 3** est de définir statistiquement un mouvement de souris ainsi qu'un défilement lors de l'exploration de pages web. En outre, nous étudions le comportement oculomoteur et examinons la relation entre les yeux, la souris et le défilement. À cette fin, nous avons enregistré les mouvements des yeux, de la souris et du défilement de 151 participants qui ont exploré 18 pages web dynamiques tout en effectuant des tâches de visualisation et de recherche visuelle pendant 20 secondes. Les données ont révélé des différences significatives des paramètres de l'oeil, de la souris et du défilement au fil du temps, qui se stabilisent à la fin de l'exploration. Cela suggère l'existence d'une relation indépendante de la tâche entre l'œil, la souris et les paramètres de défilement qui sont caractérisés par deux schémas distincts : un schéma commun pour les paramètres de mouvement et un second pour les paramètres de pause/fixation. Dans ces schémas, les mouvements de la souris et de l'oeil sont restés cohérents l'un par rapport à l'autre tout le long de l'exploration, tandis que le défilement s'est comporté de manière opposée.

# Similarities and differences between eye and mouse dynamics during web pages exploration

**Alexandre Milisavljevic** [1,2,3,*], **Fabrice Abate,**[1], **Thomas Le Bras,**[1], **Bernard Gosselin,**[2], **Matei Mancas,**[2] **and Karine Doré-Mazars** [1]

[1] *Vision Action Cognition Laboratory, Psychology Institute, Université de Paris, Boulogne-Billancourt , France*
[2] *Information, Signal and Artificial Intelligence Laboratory, Numediart Institute, University of Mons, Mons, Belgium*
[3] *Research and Development Department, Sublime Skinz, Paris, France*

Correspondence*:
Alexandre Milisavljevic
alexandre.milisavljevic@etu.parisdescartes.fr

## ABSTRACT

The study of eye movements is a common way to non-invasively understand and analyse human behaviour. However, eye-tracking techniques are very hard to scale, and require expensive equipment and extensive expertise. In the context of web browsing, these issues could be overcome by studying the link between the eye and the computer mouse. Here, we propose new analysis methods, and a more advanced characterisation of this link. To this end, we recorded the eye, mouse and scroll movements of 151 participants exploring 18 dynamic web pages while performing free viewing and visual search tasks for 20 seconds. The data revealed significant differences of eye, mouse and scroll parameters over time which stabilise at the end of exploration. This suggests the existence of a task-independent relationship between eye, mouse and scroll parameters which are characterised by two distinct patterns: one common pattern for movement parameters and a second for dwelling/fixation parameters. Within these patterns, mouse and eye movements remained consistent with each other, while the scrolling behaved the opposite way.

**Keywords: eye movement, behaviour, computer mouse, scroll, web page**

7941 words, 4 Figures and 4 Tables. British English.

## 1 INTRODUCTION

Websites, and more particularly web pages, designate a type of stimulus we potentially see every day. Such stimuli are rarely entirely visible, hence the fact that we cannot fully explore them using only our eyes. That is one of the reasons web browsing on a desktop computer requires the use and coordination of the eyes and the computer mouse. On the one hand, the eyes are used to explore and extract information of interest, such as the location of items. On the other hand, the mouse is used to interact with the content. This interaction can take multiple forms, including clicks, scrolling, and drags and drops. While clicks and

drags and drops allow the user to perform actions on the visible content, scrolling drives which part of the web page is displayed. These specific characteristics proper to web pages induce more complex behaviours, as well as more challenging issues to address. One particularly interesting aspect is how the eyes and the mouse are related.

Conveniently, eye movements have been extensively studied. We know that visual exploration is modulated by bottom-up and top-down factors regardless of the stimulus type (Yarbus, 1967; DeAngelus and Pelz, 2009; Helo et al., 2014; Itti and Borji, 2015). Bottom-up factors are characterised by low-level features of the stimulus, such as luminance, contrast or edges (Tatler and Vincent, 2008). In comparison, top-down factors are characterised by high-level properties representing cognitive processes (Henderson and Hollingworth, 1999). Both factors have been widely investigated during website exploration in order to better understand user behaviour and thus improve the usability of web pages. For instance, Pan et al. (2004) showed differences in visual exploration depending on the type of website, their presentation order and the gender of the user. They did not find any difference between a memorisation and a free viewing task, highlighting the importance of adapting a website to its targeted audience. In his work, Tullis (2007) found that older users spent more time looking at a page content, especially navigational areas, compared to younger users. Additionally, Roth et al. (2013) showed that user expectations had an influence on visual exploration, and, more particularly, less fixations were needed to find items in expected locations compared to unexpected ones.

These studies clearly show an influence of bottom-up and top-down factors. However, Tatler and Vincent (2008) and Anderson et al. (2015) show that bottom-up influence was higher at the beginning of visual exploration. Thus, both factors alternatively influence visual exploration (Henderson, 2003; Torralba et al., 2006). As such, Cronin et al. (2020) encouraged the need to focus more on the dynamic of eye movements. They showed that comparing experimental conditions on the basis of global eye movement parameters did not necessarily allow them to be distinguished them. To do so, they compared fixation durations and saccade amplitudes between a memorisation task and an aesthetic judgment task. While they did not find differences in the mean level analyses, the use of temporal and distributional analyses allowed them to discriminate the two tasks.

Previous research already highlighted the dynamic of eye movements (Unema et al., 2005; Pannasch et al., 2008; Pannasch and Velichkovsky, 2009). They found that the amplitude of saccades decreased while the duration of fixations increased over time. Pannasch and Velichkovsky (2009) and Velichkovsky et al. (2002) defined two visual exploration modes based on the relationship between saccade amplitudes and fixation durations. The ambient mode corresponds to short fixations ($<$180ms) followed by saccades with an amplitude greater than 5°, while the focal mode corresponds to long fixations ($>$180ms) followed by saccades with an amplitude of less than 5°. Generally, visual exploration begins in ambient mode before gradually switching to focal mode (Pannasch and Velichkovsky, 2009; Velichkovsky et al., 2002). Our knowledge on these visual modes is growing but still incomplete. A closer understanding of these two modes could help to better grasp the dynamic of eye movements when looking at complex stimuli, such as web pages. More specifically, in addition to eye movements, it would also be of interest to use these two visual modes to investigate the dynamic of mouse movements.

To our knowledge, despite the fact that the use of the computer mouse is well studied, its dynamic is rarely considered. Generally, research on the computer mouse focuses on how mouse movements could reveal users' intentions. Its availability and its potential for scalability enable innovative applications, such as authentication (Zheng et al., 2011), the prediction of the users' cognitive load (Rheem et al., 2018), the prediction of users' intentions (Guo and Agichtein, 2010a; Fu et al., 2017) or pattern behaviour analysis

(Tzafilkou and Protogeros, 2018). One of the most studied topics is the computer mouse movement patterns commonly used by participants when browsing. Tzafilkou and Protogeros (2018) reviewed six patterns: the straight pattern (Griffiths and Chen, 2007), the hesitation pattern (Mueller and Lockerd, 2001), the horizontal reading pattern (Rodden et al., 2008), the vertical reading pattern (Rodden et al., 2008), the random pattern (Ferreira et al., 2010) and the fixed pattern (Griffiths and Chen, 2007).

Whether it is necessary to describe mouse movement patterns or their dynamic, mouse movements are not limited to moving the mouse, and include scrolling as well. However, contrary to mouse movements, scrolling behaviour has, to our knowledge, not been closely examined. For instance, Liu et al. (2017) investigated users' strategies when navigating Search Engine Results Pages (SERP) through their scrolling behaviour. An SERP consists of a list of links corresponding to a query entered by a user in a search engine. Liu et al. (2017) analysed the number of scrolls and their direction. In their work, Braganza et al. (2009) evaluated user preferences depending on the web page layout and the scrolling mechanism using the number of scrolls and their total duration. More generally, these studies show that the mouse is a convenient and cheap way to infer users' cognitive processes, such as intentions or reading strategies, but neglect mouse and scroll parameters.

These limitations can also be found when it comes to the relationship between the eye and the computer mouse. To this day, one of the most studied web stimuli for investigating this relationship is the Search Engine Results Page (SERP). On this type of web page, the coordination between the eyes and the computer mouse is higher for the vertical axis of the screen than for the horizontal axis (Rodden and Fu, 2007; Guo and Agichtein, 2010b). However this relationship remains uncertain, considering that the mouse could be used as a means to mark a potential result previously located with the eyes (Rodden et al., 2008). Furthermore, the amount of time spent by a user on an SERP can affect the location of the gaze and the mouse during the exploration (Huang et al., 2012). Navalpakkam et al. (2013) designed a model to predict the location of the eyes based on the mouse location and showed that the correlation between the eyes and the mouse is nonlinear and user dependant. More specifically, this correlation has been found for time periods during which a user looked at an Area Of Interest (AOI) and when switched between AOIs. However, SERPs are not representative of the web and remain transitional web pages to access a content on a different website. As a matter of fact, users spend a significant cumulative amount of time on SERPs, but in short bursts of time. When focusing on common web pages, the eyes and the mouse are also coordinated on the vertical axis, and the scroll speeds influence the position of the eyes during scrolling (Milisavljevic et al., 2018). The participant is looking at the opposite part of the screen when scrolling at a high speed. Moreover, the presence of the cursor in a region of the screen correlates with the probability that the participant is fixating on this region (Chen et al., 2001). To better estimate if the eyes and the mouse are coordinated, Boi et al. (2016) generalised the work of Navalpakkam et al. (2013) by defining that the eyes and mouse must be positioned over the same content. This new definition allowed them to improve the predictive power of the models of Guo and Agichtein (2010b) and Huang et al. (2012) when applied to classic web pages. Finally, when it comes to the coordination of the eyes and scrolling, web pages are not of primary interest. That is why, to our knowledge, no studies tackle the coordination between the two outside the reading field (Kumar et al., 2007; Sharmin et al., 2013).

The goal of our study was to contribute to this growing area of research by exploring the similarities and differences between movement of the eyes and computer mouse on web pages. First, we introduced a new segmentation threshold in order to differentiate two mouse movements or scrolls as precisely as possible. Then, with this new segmentation, analyses from eye movement methodology were applied to mouse movement and scrolling parameters. This methodology allowed us to investigate the influence of the tasks

(free viewing and visual search) on eye, mouse and scroll parameters. Beyond these global analyses, we also considered the influence of time on the dynamic of each type of movement through visual exploration modes.

## 2 MATERIAL AND METHODS

### 2.1 Participants

We recruited one hundred and fifty-one participants (127 females and 24 males) aged between 18 and 56 ($M = 22.77$, $SD = \pm 6.33$). Participants reported normal or corrected to normal vision and were naive about the purpose of the study. They were right-handed or accustomed to using a computer mouse with the right hand. A majority were undergraduate students from the psychology institute at the Université de Paris. Participants were compensated either by course credit or a 15 euro gift card. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee (local Ethics Committee of Paris Descartes University, No. CER-PD: 2018-77) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. All subjects gave written informed consent.

### 2.2 Apparatus

Eye movements were recorded using an Eye-Link 1000 Plus (SR Research Ltd., Canada) at a 1000Hz sampling rate with 0.05°precision. We recorded the right eye of the participants with a 35mm monocular lens. Mouse movements were recorded with a standard USB optical mouse with a 125Hz polling rate. Stimuli were displayed on a 24.5 inch LCD computer screen with a 1920x1080 pixel resolution and a 144Hz refresh rate. The experiment was run using Python 2.7 with Pylink from the manufacturer and Chromium 64.

### 2.3 Stimuli

In this experiment eighteen web pages (see Figure 1) from eighteen different websites were randomly presented to the participants. The web pages has a width of 1920 pixel and their total height was between 5000 pixels and 19230 pixels ($M = 6405px$, $SD = \pm 2673px$). Participants were allowed to freely move the mouse, scroll or click without restriction. However, hyperlinks and content animations were deactivated, thus participants could not leave the displayed web page. The web pages were chosen according to several criteria to minimise biases. The two first criteria were the popularity and language of the website. We ensured stimuli were from French websites with differing popularity. The third criterion was about the websites' news content. Since this study was run over several months, a web page could not have any content referring to current events or content related to a season, date, holiday, celebration, etc. As the fourth criterion, we checked that the web pages did not have any external advertising. In contrast to the first four criteria, which were respected on all web pages, the following criteria were counterbalanced between web pages. As Bruyer et al. (1987) explained, faces are handled differently by our brain during visual exploration. To this end we made sure that we keep a balance of faces between the web pages. We also made sure that a balance was maintained for images, texts, general layout and total length of the web page to have stimuli with different content types and organisation. Finally, as described in the following paragraph, we gave targets already present within the original web page. Thus, we checked the number of targets available on the web page and their distribution across the page.

### 2.4 Tasks

Participants had to perform two types of tasks during this study. Both tasks were randomly displayed nine times each. During the free viewing task, the participants were instructed to explore the web page freely for exactly sixty seconds. In the visual search task, participants were asked to find a target in maximum

two minutes. The participants did not know how many targets there were but we informed them that there were up to three targets, with at least one, per web page. As previously defined, the targets were icons or images already present on the original web page. Moreover, the targets were distributed between the top, middle and bottom of the web page, and could be found on the sides, or in the content, header or footer.

### 2.5 Procedure

In a quiet room, with constant luminosity, the participants were instructed to position their head on a chin rest in front of a computer screen at a viewing distance of 57cm. The experiment then began with practice trials, one for each task. After this phase, the participants' right eye was calibrated at nine points and this was repeated until the error value was below 1°. Once the calibration was successfully complete, the participants had to click on the next trial with the mouse on a 3 by 6 table, as shown in Figure 2. Then the instructions were displayed on a new screen with a button to launch the trial. The position of this button was randomly chosen in order to avoid bias related to the first fixation commonly being at the same position as the button launching the trial. Furthermore, to ensure the web page would have completely loaded before the trial started a 3-second countdown was added to the button launching the trial. The countdown only began after the page entirely loaded, thus visual elements displayed after few seconds could be avoided. During this phase, the participants were informed of the presence of maximum three targets when carrying out the visual search task. After clicking on the button, the web page was displayed for 60 or 120s, depending on the task. During the visual search task, the participants had to click on the targets when they founded them. If the image clicked was one of the targets, a green rectangle surrounded the target to indicate that one of the targets had been found. The participants were instructed to press the space bar on the keyboard when they thought they had found all the targets. After 1 minute of the free viewing task, and 2 minutes or after the space bar was pressed in the visual search, the recording was stopped, and the 3 by 6 table displayed at the beginning was displayed again. Between each trial a 5-point calibration was performed. A 9-point calibration was initiated after the ninth trial, or if any problems occurred during the experiment.

### 2.6 Data Analysis

#### 2.6.1 Data cleaning

Data from 12 participants who did not finished the experimental protocol due to calibration problems were discarded. Among the remaining 139 participants (2502 trials), due to problems encountered during the experiment, such as calibration problems, participants talking during a trial, external noise, etc, we removed 4.88% of all trials (122 trials). The remaining data (2380 trials) was then pre-processed and cleaned in three steps. The first step was only applied to the visual search task. The last 2 last seconds of recording were removed in order to deal with the moment the participant looked at the keyboard when pressing the space bar. In addition, and for the same reason, residual fixations below the screen at the end of the exploration were removed. Throughout the second step, blinks and fixations under 100ms around a blink were cleaned (Holmqvist et al., 2011). During the third and final step, fixations with a visual angle of more than 3°from the screen's border were deleted. Fixations outside the screen, but below the 3°threshold, were reset to the corresponding border of the screen. These three steps led to deletions within all the trials. All 139 participants, and 95% of the initial trials (2378 trials), were kept. In total, 91.74% of all records were retained for analyses. Finally, only the first 20 seconds were selected for this work, and 18 more trials were deleted due to insufficient mouse moves or scrolling events (2360 trials remaining). It should be noted that eye movement analyses were run on aggregated data, and scrolling and mouse events on raw data. All analyses were carried out using Python 3.6.

### 2.6.2 Events Segmentation

There are a number of well-established, and ever improving, methods to label raw data from eye recordings. However, mouse and scroll recordings lack such a method, specifically to differentiate two close events. While it is easy to determine if two events separated by two or three seconds are indeed two distinct events, doing the same operation for two events with, for instance, less than one second in between, is much harder. In the literature, we can find multiple attempts to define a threshold allowing the differentiation of idle time and movement of the mouse. Since the mouse is a pointing device, a simple threshold seems to be appropriate, contrary to eye movements which are more complex. In their attempt to define a new behavioural biometric technique based on mouse movements, Gamboa and Fred (2004) differentiated two mouse movements as a pause in the user's interaction when the two consecutive events were separated by more than 100ms. In their work, Reeder and Maxion (2006) arbitrarily considered a threshold of 3s with to the user being silent and inactive (with both the mouse and the keyboard) in order to propose a method to detect user difficulties when using an interface. On the other hand, Feher et al. (2012) empirically set this threshold to 500ms to categorise mouse movements and thus uniquely identify users. More recently, Seelye et al. (2015) studied cognitive impairment using computer mouse movement patterns. They mentioned a median idle time, which is the time spent idling or pausing between mouse movements, of 310ms. In the continuity of the work of Gamboa and Fred (2004), Antal and Egyed-Zsigmond (2019) used a threshold of 10 seconds to segment mouse movements and used them to detect intruders on a computer.

Moreover, some studies focused specifically on scroll segmentation. In their study into the scrolling behaviour, Braganza et al. (2009) determined that two scrolls recorded within one second of each other were considered as a single scroll. To set this threshold, they tried values ranging from 200ms to 4s, with increments of 100ms. They did not find any major differences between these timings, and consequently chose 1s as a threshold. In their study, Milisavljevic et al. (2018) defined a scroll session as a set of continuous scroll events ended with a mouse movement. On the topic of scrolling when reading, Brady et al. (2018) sampled a frame every 100ms to check if the displayed text had moved. If it had moved more than half a line between one sentence and the next, it was counted as a scroll.

Even though the presented techniques try to segment scrolling or mouse events, they are mostly arbitrary thresholds. If we take a closer look at our previous attempt to segment events, we defined a threshold based on the events number rather than the time (Milisavljevic et al., 2018). This definition does not take into account all parameters that come into play when interacting using mouse or scroll. The main parameter is the fact that, on a desktop, it is possible to move the mouse during a scroll. In such a case, a single scroll would be labelled as two different scrolls. The bias will remain if the participant uses the browser scroll bar, which allows the user to grab a bar on the right of the browser and scroll by moving it up or down. Furthermore, Brady et al. (2018) used a spatial threshold of 40 pixels to identify when a user was scrolling, but this is applicable to mouse movements. In addition to highlighting the need to use a time-based threshold, all previously mentioned studies did not correctly handle stops and micro-stops. A stop is a period of time during which the user does not move the mouse or scroll. During this idle time, the user explores the web page and processes it. But based on this definition, a new question arises: what is the minimal length of this period of time to give the user enough time to process the stimulus and make the decision to keep moving, scrolling, or stop entirely? In other terms, how can we differentiate micro-stops from the movement itself? A micro-stop is an interruption during the action which is long enough to allow the user to make a decision, but this is not visible to the eye. To differentiate micro-stops from movements we looked at the study from Moher and Song (2019) in which they compared behaviours between a 3D

reach tracker, a computer mouse and a stylus. Among multiple conditions, they measured the average response latency of 220ms when displacing a target. This could be considered as the minimum time to visualise a target's new position and make the decision to reorient the movement. Thus, a micro-stop could not be less than 220ms, and a stop below this threshold should be considered as the continuity of the previous action. We used a unified threshold to segment mouse movements and scrolls. We chose a threshold of 300ms to differentiate two distinct movements or scrolls. This corresponds to the average visual fixation duration in a scene viewing (Henderson and Hollingworth, 1998). Despite the fact that visual fixations can be shorter than 300ms, this does not apply to ecological conditions and semantic-rich stimuli, such as web pages.

### 2.6.3 Variables

After all cleaning processes, we ran our analyses on a wide range of new parameters. In the state-of-the-art, the same types of parameters are frequently used. For the use of the mouse, these include curvature, trajectory, clicks, dwells or the number of movements (Zheng et al., 2011; Tzafilkou and Protogeros, 2018; Rheem et al., 2018; Fu et al., 2017), and for scrolls, amplitude, speed and number (Liu et al., 2017; Braganza et al., 2009; Milisavljevic et al., 2018). In comparison, eye-mouse studied parameters are more related to their respective positions, but are not limited to this factor. For instance, eye-mouse distance, content hovered, lag, percentage of regions visited by both the eyes and mouse, etc., have been studied (Chen et al., 2001; Rodden and Fu, 2007; Rodden et al., 2008; Guo and Agichtein, 2010b; Huang et al., 2012; Navalpakkam et al., 2013; Boi et al., 2016).

In this paper, we propose a more complete set of parameters directly inspired from eye movement analyses. These parameters include dwell duration, movement duration, movement amplitude and number of events. It should be noted that duration variables are expressed in seconds or milliseconds, while amplitude variables are expressed in degree of visual angle. Furthermore, in order to better characterise the dynamic of the exploration through ambient and focal visual modes, we apply, for the first time, the K coefficient defined by Krejtz et al. (2016) to mouse and scroll events. This coefficient is calculated by averaging the differences in z-scores between the duration of each fixation and the next saccade, as shown in equation 1. A negative value indicates that the fixation $d_i$ is short and the next saccade $a_{i+1}$ is long ($>5°$). In contrast, a positive value suggests that the fixation $d_i$ is long and the next saccade $a_{i+1}$ is short ($<5°$) which corresponds to a focal mode.

$$K = \frac{1}{n}\sum_{n}\frac{d_i - \mu_d}{\sigma_d} - \frac{a_{i+1} - \mu_a}{\sigma_a} \tag{1}$$

Milisavljevic et al. (2019) introduced two new variables to better capture the dynamic of focal and ambient modes. While the K coefficient did not discriminate between the different stimuli used in their study, the number of switches between modes did. It is for this reason that we are using these parameters to more precisely describe the dynamic of the exploration for both the eyes and mouse.

### 2.6.4 Mouse and Scroll Overlap

Participants were able to independently move the mouse and scroll. Consequently, this led to overlaps between mouse movements and scrolls. We found that this overlap occurred only 10% (*SD* = *±4.83%*) of the total mouse movement time and 15% (*SD* = *±10.59%*) of the total scrolling time. During these overlaps we observed mouse movements with an amplitude of 0.02°(*SD* = *±0.02°*) and a duration of 240ms (*SD* = *±195.53ms*) for a total duration of 570ms (*SD* = *±430ms*). As described, during overlaps, movements represented a negligible part of the exploration. Moreover, these overlaps followed three main patterns: move-scroll, scroll-move and move-scroll-move. The move-scroll pattern refers to a scroll

that began while already moving the mouse. This pattern occurred 43% of the time and was the most frequent. The second pattern we observed was the scroll-move pattern. This pattern is the exact opposite: the participant began to move the mouse while already scrolling. This pattern happened 25% of time. The move-scroll-move pattern is when the participant scrolled within a single mouse move. This was less common and occurred 21% of the time. Finally, the 11% remaining was exotic patterns, such as move-scroll-move-scroll or move-scroll-move-scroll-move which represent 2% each, etc. Due to the low frequency of overlaps between scrolls and mouse movements, we can safely conclude that these specific movements are residual movements or involuntary micro-movements generated by the use of the mouse wheel. For this reason, we did not take overlaps into account in the following analyses.

## 3 RESULTS

To study the similarities and differences between eye movements, mouse movements and scrolling, we ran two types of analyses. We first described eye, mouse and scroll parameters globally, to clearly define what a mouse or scroll movement was, and summarised them in Table 1. Then, we examined the role of tasks and time, by performing a 2 (free viewing and visual search) X 4 (0-5s time-bin, 5-10s time-bin, 10-15s time-bin and 15-20s time-bin) repeated measures analyses of variance (ANOVAs). Post-hoc analyses were run using pair-wised student's t-test with a Bonferroni correction.

### 3.1 Eye Movements Analysis

We measured a rather stable distribution between fixations and saccades across the different conditions. During the exploration of a website, participants spent approximately 14% ($SD = \pm 1.72\%$) of the time doing a saccade (see Table 1). Although this proportion was maintained across the tasks, we found a task effect on the distribution of fixations/saccades ($F(1,138)=231.98$, $p < 0.001$). Participants spent 13.6% ($SD = \pm 1.79\%$) of the time doing a saccade in the free viewing task and 15% ($SD = \pm 1.84\%$) during the visual search task. Furthermore, we found a time effect ($F(3,414)=685.59$, $p < 0.001$) present between the first and second time-bins ($t=-29.50$, $p < 0.001$), and between the second and third time-bins ($t=8.98$, $p < 0.001$), but not between the third and fourth time-bins ($t=-2.33$, $p > 0.05$). We also found a significant interaction effect between task and time ($F(1,138)=3.48$, $p < 0.05$), but post-hoc analyses confirmed that main effects were preserved (see Table 2).

#### 3.1.1 Number of Fixations and Saccades

Globally, participants performed an average of 72 ($SD = \pm 6.5$) fixations and saccades during the exploration of a website for 20s. The task had an effect on the number of fixations and saccades ($F(1,138)=424.29$, $p < 0.001$) with less fixations and saccades during the visual search ($M = 68.4$, $SD = \pm 6.31$) compared to the free viewing task ($M = 75.16$, $SD = \pm 7.08$). We found a time effect ($F(3,414)=27.86$, $p < 0.001$), but there were no significant differences between the first and second time-bins ($t=0.32$, $p > 0.05$). However, there was a significant decrease in the number of fixations and saccades between the second and third time-bins ($t=-4.84$, $p < 0.001$), as well as between the third and fourth time-bins ($t=-2.85$, $p < 0.05$). The interaction between the time and task was also significant ($F(3,414)=3.29$, $p < 0.05$). The main task effect was maintained for each time-bin (all $p < 0.001$). In free viewing task, there were no significant differences between the successive time-bins (all $p > 0.05$). However, in visual search, the only difference with the main time effect was the absence of a reduction between the third and fourth time-bins ($p > 0.05$) (see Table 2).

#### 3.1.2 Fixation Duration

As expected, we observed an average fixation duration of 236ms ($SD = \pm 24.45ms$). The average fixation duration varied according to the task ($F(1,138)=195.75$, $p < 0.001$), being shorter in the free viewing task ($M = 229ms$, $SD = \pm 24.59ms$) than the visual search task ($M = 247.17ms$, $SD = \pm 26.41ms$). The average

fixation duration significantly increased over time ($F(3,414)=297.65$, $p<0.001$) up to the third time-bin. More precisely, the first time-bin was significantly different from the second time-bin ($t=20.91$, $p<0.001$), and this second time-bin was significantly different from the third time-bin ($t=6.80$, $p<0.001$). However, the third time-bin was not significantly different from the fourth ($p>0.05$. There was also an interaction effect between task and time ($F(3,414)=3.29$, $p<0.05$), but post-hoc analyses confirmed that main effects were preserved (see Table 2).

### 3.1.3 Saccade Amplitude

We measured an average saccade amplitude of 6.1°($SD = \pm 0.67°$). We found a significant difference between the tasks ($F(1,138)=1314.42$, $p<0.001$), saccade amplitudes were shorter during the free viewing task ($M = 5.08°$, $SD = \pm 0.77°$) than during the visual search task ($M = 7.36°$, $SD = \pm 0.77°$). We also observed a time effect ($F(3,414)=378.60$, $p<0.001$) up to the third time-bin. The average saccade amplitude decreased from the first to the second time-bin ($t=-21.27$, $p<0.001$), and from the second to the third time-bin ($t=-8.45$, $p<0.001$), but not between the third and fourth time-bins ($t=-1.55$, $p>0.05$). However, there was no significant interaction between the time and task ($F(3,414)=2.11$, $p>0.05$) (see Table 2).

### 3.1.4 Dominant Mode

Finally, to understand the dynamic of visual exploration, we computed the K coefficient and its associated variables, as defined by Krejtz et al. (2016) and Milisavljevic et al. (2019), and described in the Methodology section. Globally, we found a dominance of the ambient mode with a K coefficient below zero ($M = -0.13$, $SD = \pm 0.2$). There was a significant difference between tasks ($F(1,138)=313.8$, $p<0.001$) which indicated a higher dominance of the ambient mode in the visual search task ($M = -0.28$, $SD = \pm 0.23$) than in the free viewing task ($M = -0.01.$, $SD = \pm 0.21.$) We also found a significant time effect ($F(3,414)=579.66$, $p<0.001$). The K coefficient, beginning with negative values, got significantly closer to 0 between the first and second time-bins ($t=-27.10$, $p<0.001$), became positive between the second and third time-bins ($t=-10.23$, $p<0.001$), but did not significantly change between the third and fourth time-bins ($t=1.94$, $p>0.05$). Post-hoc analyses did not show a significant interaction between the task and time ($F(3,414)=1.97$, $p>0.05$) (see Table 2).

### 3.1.5 Visual Modes Switches

As described in the Methodology section, the number of visual modes switches corresponds to how many times a participant switched from ambient to focal and focal to ambient during a trial. Participants switched between visual modes 33.15 ($SD = \pm 3.25$) times and this amount varied according to the task ($F(1,138)=63.06$, $p<0.001$). There were more switches in the free viewing task ($M = 34.26$, $SD = \pm 4.30$) than in the visual search task ($M = 31.67$, $SD = \pm 3.22$). There was also a time effect ($F(3,414)=22.69$, $p<0.001$). The number of visual mode switches significantly increased between the first and second time-bins ($t=8.05$, $p<0.001$), but significantly decreased between the second and third time-bins ($t=-4.05$, $p<0.001$). It was not, however, significantly different between the third and fourth time-bins ($t=-1.24$, $p>0.05$). Furthermore, we found a significant interaction between the task and time ($F(3,414)=6.33$, $p<0.001$). The main task effect was maintained except for the third time-bin ($t=4.33$, $p>0.05$). Similarly, the main time effect was preserved for the free viewing task, but not in the visual search task, during which there were no significant differences between the second and third, and the third and fourth time-bins (all $p>0.05$) (see Table 2).

### 3.1.6 Visual Modes Proportions

The participants spent, in total, 43% ($SD = \pm 6.81\%$) of the time in ambient mode. This proportion significantly varied according to the task ($F(1,138)=358.75$, $p<0.001$). It was higher in the visual search task ($M = 48.35\%$, $SD = \pm 7.33\%$) than in the free viewing task ($M = 38.21\%$, $SD = \pm 7.65\%$). There was a significant time effect ($F(3,414)=638.94$, $p<0.001$). The proportion of time spent in ambient

mode significantly decreased between all successive time-bins: between the first and second time-bins ($t$=-31.30, $p$ <0.001), between the second and third time-bins ($t$=-9.32, $p$ <0.001) and between the third and fourth time-bins ($t$=-1.44, $p$ >0.05). We also found a significant interaction between the time and task ($F(3,414)$=8.75, $p$ <0.001), but post-hoc analyses confirmed that main effects were preserved (see Table 2).

To summarise, we found a task and time effect on all the variables of eye movements parameters. Most of the parameters increased over time to then stabilise starting at the third time-bin (after 10-15s). More specifically, fixation-related variables increased and movement-related variables decreased over time. Moreover, ambient mode was predominant during the exploration but progressively switched to focal mode as time went by.

### 3.2 Mouse Analysis

The participants freely used the mouse during their exploration, and spent 20.85% ($SD$ = ±8.33%) of the time moving it. We found a significant task effect ($F(1,138)$=37.66, $p$ <0.001), the proportion of time spent moving the mouse was significantly higher in the visual search task ($M$ = 23.33%, $SD$ = ±8.48%) than in the free viewing task ($M$ = 18.94%, $SD$ = ±10.11%). We also observed a time effect ($F(3,414)$=420.24, $p$ <0.001) with a significant decrease between the first and second time-bins ($t$=-24.14, $p$ <0.001), and between the second and third time-bins ($t$=-3.25, $p$ <0.01). However, there was no significant difference between the third and fourth time-bins ($t$=-1.68, $p$ >0.05). There was a significant interaction between time and task ($F(3,414)$=7.75, $p$ <0.001). The main task effect was maintained excepted for the second time-bin ($t$=1.2, $p$ >0.05). The main time effect was preserved in the free viewing, but not entirely during the visual search task, there was no significant difference between the second and third time-bins ($p$ >0.05) (see Table 3).

#### 3.2.1 Number of Mouse Movements

The participants did 6.04 ($SD$ = ±1.78) movements on average. We found a task effect ($F(1,138)$=73.45, $p$ <0.001) with more mouse movements during the visual search task ($M$ = 6.77, $SD$ = ±2.01) than during the free viewing task ($M$ = 5.43, $SD$ = ±1.97). We found an influence of time ($F(3,414)$=183.46, $p$ <0.001) with a significant decrease between the first and second time-bins ($t$=-14.34, $p$ <0.001), and between the second and third time-bins ($t$=-4.70, $p$ <0.001). However, there was no significant difference between the third and fourth time-bins ($t$=-1.79, $p$ >0.05). We also found a significant interaction between time and task ($F(3,414)$=14.15, $p$ <0.001). The main task effect was preserved excepted for the second time-bin ($p$ >0.05). In the free viewing task, the main time effect was preserved, but, in the visual search task this main effect was maintained only between the first and second time-bins ($p$ <0.001) (see Table 3).

#### 3.2.2 Duration of Mouse Movements

The participants moved the mouse for 768ms ($SD$ = ±342.55ms) on average. We found a task effect ($F(1,138)$=15.63, $p$ <0.001) with significantly longer mouse movements in the free viewing task ($M$ = 772.68ms, $SD$ = ±362.58ms) than in the visual search task ($M$ = 767.43ms, $SD$ = ±386.39ms). Moreover, we found a time effect ($F(3,414)$=269.83, $p$ <0.001) with a significant decrease between the first and second time-bins ($t$=-19.53, $p$ <0.001), but no significant difference between the second and third time-bins ($t$=-2.56, $p$ >0.05) nor between the third and fourth time-bins ($t$=0.74, $p$ >0.05). We also found a significant interaction between time and task ($F(3,414)$=3.69, $p$ <0.05). However, the main task effect was preserved only for the two last time-bins (all $p$ <0.005), while the main time effect was only preserved for the visual search task. During the free viewing task, we observed significant differences between the first and second time-bins, and between the second and third time-bins (all $p$ >0.05) (see Table 3).

#### 3.2.3 Amplitude of Mouse Movements

The participants performed mouse movements of 0.27°(*SD* = ±*0.23*°) on average. We found a significant differences between the two tasks (*F(1,138)=24.16*, *p <0.001*). The average amplitude slightly decreased from the free viewing task (*M = 0.26*°, *SD* = ±*0.2*°) to the visual search task (*M = 0.3*°, *SD* = ±*0.3*°). We also found a time effect (*F(3,414)=235.57*, *p <0.001*). There was a significant decrease between the first and second time-bins (*t=-17.57*, *p <0.001*)amp, but no significant differences between the second and third time-bins (*t=-2.42*, *p >0.05*) nor between the third and fourth time-bins (*t=0.22*, *p >0.05*). We did not find any interaction effect (*F(3,414)=1.61*, *p >0.05*) (see Table 3).

#### 3.2.4 Dynamic of Mouse Movements

Here, K coefficient is used to better understand the mouse movement dynamic. The K coefficient showed a dominance of the ambient mode (*M = -0.35*, *SD* = ±*0.63*). We found significant differences between tasks (*F(1,138)=15.27*, *p <0.001*) which was slightly higher in the free viewing task (*M = -0.31*, *SD* = ±*0.58*) than in the visual search task (*M = -0.39*, *SD* = ±*0.77*). There also was a significant time effect (*F(3,414)=410.86*, *p <0.001*). We found a significant increase between all successive time-bins (all *p <0.001*). However, there was no significant interaction effect (*F(3,414)=2.48*, *p >0.05*) (see Table 3).

#### 3.2.5 Mode switches

On average, 3.78 (*SD* = ±*0.89*) switches occurred between modes given by the K coefficient. There was a significant task effect (*F(1,138)=70.08*, *p <0.001*) which was characterised by a lower number of mode switches during the free viewing task (*M = 3.44*, *SD* = ±*1.04*) than during the visual search task (*M = 4.19*, *SD* = ±*1.07*). There was also a significant time effect (*F(3,414)=109.86*, *p <0.001*). The number of switches significantly increased between the first and second time-bins (*t=11.68*, *p <0.001*), and between the second and third time-bins (*t=3.72*, *p <0.005*), but there was no significant difference between the third and fourth time-bins (*t=1.42*, *p >0.05*). We also found a significant interaction between the time and task (*F(3,414)=11.93*, *p <0.001*). The main task effect was preserved excepted for the first time-bin (*p <0.05*). Furthermore, the main time effect was maintained for the free viewing task, but, for the visual search task, the first and second time-bins were significantly different (*p <0.001*), while remaining time-bins did not have significant differences (all *p >0.05*) (see Table 3).

To summarise, we found a task and time effect for all the mouse parameters. As found for eye movements, most of the mouse parameters stabilised at the end of the exploration. Interestingly, the mouse parameters behaved similarly to eye movements parameters. Finally, ambient mode was the prevailing mode for mouse movements, but, as for the eyes, progressively switched to the focal mode over time.

### 3.3 Scroll Analysis

The participants, globally, spent 16.58% (*SD* = ±*5.32%*) of a trial scrolling. There was a task effect (*F(1,138)=469.10*, *p <0.001*). The proportion of time spent scrolling was higher in the visual search task (*M = 23.80%*, *SD* = ±*8.28%*) compared to the free viewing task (*M = 10.86%*, *SD* = ±*4.87%*). We also found a time effect (*F(3,414)=239.92*, *p <0.001*). There was a significant increase between the first and second time-bins (*t=20.74*, *p <0.001*), as well as between the third and fourth time-bins (*t=3.70*, *p <0.005*), while there was no significant differences between the second and third time-bins (*t=0.06*, *p >0.05*). We found a significant interaction between the time and task (*F(3,414)=11.94*, *p <0.001*). The main task effect was maintained for all time-bins (all *p <0.001*). However, the time effect was not preserved. In both tasks, the first and the second time-bins were significantly different (*t=-20.5*, *p <0.001*), but we did not find significant differences between other time-bins (*p >0.05*) (see Table 4).

### 3.3.1 Number of Scrolls

During the trial, the participants scrolled on average 8.77 ($SD = \pm 2.04$) times. We found a task effect ($F(1,138)=512.15$, $p < 0.001$). We measured lower numbers in the free viewing task ($M = 6.62$, $SD = \pm 2.25$) compared to the visual search task ($M = 11.44$, $SD = \pm 2.63$). We also found a time effect ($F(3,414)=282.94$, $p < 0.001$). There was a significant increase between the first and second time-bins ($t=24.37$, $p < 0.001$). However there was no significant differences between the second and third time-bins ($t=0.19$, $p > 0.05$) nor between the third and fourth time-bins ($t=-0.62$, $p > 0.05$). There was a significant interaction between the time and task ($F(3,414)=6.03$, $p < 0.001$). However, post-hocs analyses showed that the main effects were maintained (see Table 4).

### 3.3.2 Scroll Duration

Scrolls lasted on average 367.57ms ($SD = \pm 121.65ms$). We found a task effect ($F(1,138)=205.20$, $p < 0.001$). Scroll was shorter in the free viewing task ($M = 328.64ms$, $SD = \pm 99.57ms$) compared to the visual search task ($M = 417.24ms$, $SD = \pm 186.17ms$). Additionally, we found a time effect ($F(3,414)=55.49$, $p < 0.001$). There was a significant increase between the first and second time-bins ($t=9.34$, $p < 0.001$), as well as between the third and fourth time-bins ($t=3.39$, $p < 0.01$). However there was no significant difference between the second and third time-bins ($t=1$, $p > 0.05$). We did not find any interaction ($F(3,414)=1.94$, $p > 0.05$) (see Table 4).

### 3.3.3 Scroll Amplitude

A scroll was on average 8.52°($SD = \pm 2.35$) long. The task had an influence on scroll amplitude ($F(1,138)=389.81$, $p < 0.001$). Scrolls were longer in the visual search task ($M = 10.58°$, $SD = \pm 3.12°$) than in the free viewing task ($M = 6.91.$, $SD = \pm 2.6.$) The time also had an influence ($F(3,414)=34.04$, $p < 0.001$). There was a significant increase between the first and second time-bins ($t=9.44$, $p < 0.001$), but not between the second and third time-bins ($t=0.77$, $p > 0.05$), nor between the third and fourth time-bins ($t=1.20$, $p > 0.05$). There was a significant interaction between the time and task ($F(3,414)=6.51$, $p < 0.001$), but post-hoc analyses confirmed that main effects were preserved (see Table 4).

### 3.3.4 Scrolling Dynamic

In contrast to eye and mouse dynamics, scrolling dynamic was dominated by the focal mode ($M = 0.43$, $SD = \pm 0.45$). There was a task effect on the K coefficient ($F(1,138)=454.64$, $p < 0.001$), which was significantly more indicative of the focal mode in the free viewing task ($M = 0.92$, $SD = \pm 0.67$) than in the visual search task ($M = -0.17$, $SD = \pm 0.47$). There was also a time effect ($F(3,414)=5.58$, $p < 0.001$), the K coefficient significantly decreased between the first and second time-bins ($t=-4.29$, $p < 0.001$), but did not between the following successive time-bins (all $p > 0.05$). We found an interaction between the time and task ($F(3,414)=39.55$, $p < 0.001$). The main task effect was maintained (all $p < 0.001$). However, maintained during the free viewing task, the main time effect, was not maintained in the visual search task. We measured a significant reduction between the first and second time-bins, and the second and third time-bins (all $p < 0.05$, but not between the third and fourth time-bins ($p > 0.05$) (see Table 4).

### 3.3.5 Modes Switches

The participants switched between modes an average of 3.63 ($SD = \pm 0.74$) times. There was a significant task effect ($F(1,138)=257.59$, $p < 0.001$). The number of switches between modes was significantly lower in the free viewing task ($M = 2.99$, $SD = \pm 0.94$) than in the visual search task ($M = 4.37$, $SD = \pm 1$). We also found a significant time effect ($F(3,414)=109.40$, $p < 0.001$). There was a significant decrease in the number of switches between the first and the second time-bins ($t=-15.27$, $p < 0.001$), but no significant differences between the following successive time-bins (all $p > 0.05$). The interaction of the time and task was also significant ($F(3,414)=4.60$, $p < 0.001$), but post-hoc analyses confirmed that main effects were preserved (see Table 4).

To summarise, we found a task and time effect for all scrolling parameters. As with the eyes and mouse parameters, most of the scrolling parameters stabilised at the end of the exploration. However, this evolution was in the opposite sense of that for the eye and mouse movements. While the eye and mouse fixation or dwelling parameters increased over time, scrolling dwells decreased. Inversely, while the eye and mouse movement parameters decreased over time, scrolling increased. As such, the focal mode was predominant in the global exploration, but tended to ambient mode over time.

## 4 DISCUSSION

In the present study, a large set of participants explored 18 web pages for 20 seconds each and were asked to perform two tasks (a free viewing task and a visual search task). Our study is the first to ally the description of eye, mouse and scroll movement parameters on web pages, and how they evolved between tasks and over time.

### 4.1 Eye Movement Parameters

We first found a task effect for all eye variables which replicated several studies in the literature (Yarbus, 1967; DeAngelus and Pelz, 2009; Itti and Borji, 2015). Fixation-related variables were higher in the free viewing task compared to the visual search task, while movement-related variables were higher in the visual search task. We also found a time effect on all variables. Fixation-related variables increased over time for both tasks while movement-related variables decreased. Participants did fewer fixations and saccades, but longer fixations and shorter saccades over time (Unema et al., 2005). As a result, we observed a global domination of ambient mode (i.e. short fixations with long saccades), but over time the dominant mode progressively switched to focal mode (i.e. long fixations with short saccades). This behaviour could indicate that participants try to contextualise the stimulus at the beginning of the exploration to then focus more and more on content as time goes by.

### 4.2 Mouse Parameters

Then we ran the same analyses on mouse movements and scrolls. We found a task effect for all parameters of the mouse exploration, except for the average amplitude and duration of the mouse movements. As for the eye movements, dwell-related variables were higher in the free viewing task compared to the visual search task, while movement-related variables were higher in the visual search task. Again, we found a time effect on all variables. Comparably to eye movement parameters, dwell-related variables increased over time and movement-related variables decreased over time for both tasks. This behaviour is similar to that of eye movements and suggests strong similarities between the two. Hence, we applied visual mode concepts to mouse movements. However, it is worth noting that the number of mouse movements was broadly inferior to the number of eye movements, so these results should be discussed with caution. Despite the difference in the number of events, we observed similar behaviour in the mouse dynamic, which began in ambient mode to progressively switch to focal model over the course of the exploration.

Regarding scrolling, all parameters varied according to the task. Comparably to eye and mouse movement parameters, we found a task effect for all parameters. We also found an time effect on all the variables, but dwell-related variables decreased over time while scroll-related variables increased. However, the stabilisation of scroll parameters began earlier than for mouse parameters (see Figures 3 and 4). Although there were fewer scroll movements than eye movements their frequency remained slightly higher than that of mouse movements. Therefore, we conducted analyses of dominant modes and found that, globally, scrolling was in focal mode. However, when looking over time, we observed that the focal mode was more dominant at the beginning of the exploration and ambient mode at the end. Since participants scrolled

increasingly over time but did longer eye fixations, they seemed to balance the natural emergence of the focal mode of the eyes, by scrolling to keep changing and contextualising the newly displayed content.

### 4.3 Similarities and Differences

Next, we separated computed variables into two distinct categories: variables related to movements and variables related to fixations or dwells. In Figure 3 A we can observe a clear relationship between the fixation-related variables of the eyes, mouse and scroll. On the one hand, eye and mouse parameters behaved similarly. Fixation or dwell durations, and percentages of fixations or dwells, were at their lowest at the beginning of the exploration and increased up to the end of exploration. On the other hand, scrolling behaved exactly the opposite way. Scroll dwell was at its highest at the beginning of the exploration and decreased overtime. These observations are consistent in both the free viewing and visual search tasks (Figure 3 B, C). Yet we observed a stabilisation of mouse and scroll dwell durations starting from the second time-bin. In Figure 4 A we can observe the opposite pattern for movement-related variables. Eye and mouse movement variables decreased over time and scroll variables increased. Eye and mouse parameters behaved in the opposite way to scroll parameters, just as with fixation-related variables. Furthermore, this relationship was maintained across both tasks (Figure 4 B, C).

Our results show a clear relationship between eye, mouse and scroll parameters. Previous studies have already shown the spatial coordination of the eyes and mouse (Guo and Agichtein, 2010b; Boi et al., 2016; Huang et al., 2012) and some coordination between the eyes and scroll speed (Milisavljevic et al., 2018). However, here we show that this relationship is even deeper than expected, and can be identified through analysing eye, mouse and scroll parameters. Indeed, coordination is not only between the eyes and the mouse, or, between the eyes and the scroll, but clearly between all three. Our findings show, for the first time, that eye and mouse parameters behave similarly, which confirms the interest of using mouse behaviour to predict eye behaviour. Yet the interaction described here does not take spatial coordinates into account that could be combined with relationship parameters to better predict eye movements from mouse events.

## REFERENCES

Anderson, N. C., Ort, E., Kruijne, W., Meeter, M., and Donk, M. (2015). It depends on when you look at it: Salience influences eye movements in natural scene viewing and search early in time. *Journal of Vision* 15, 9–9

Antal, M. and Egyed-Zsigmond, E. (2019). Intrusion Detection Using Mouse Dynamics. *IET Biometrics* 8, 285–294

Boi, P., Fenu, G., Spano, L. D., and Vargiu, V. (2016). Reconstructing User's Attention on the Web through Mouse Movements and Perception-Based Content Identification. *ACM Transactions on Applied Perception* 13, 1–21

Brady, K., Cho, S. J., Narasimham, G., Fisher, D., and Goodwin, A. (2018). Is Scrolling Disrupting While Reading? , 8

Braganza, C., Marriott, K., Moulder, P., Wybrow, M., and Dwyer, T. (2009). Scrolling behaviour with single- and multi-column layout. In *Proceedings of the 18th international conference on World wide web - WWW '09* (Madrid, Spain: ACM Press), 831–840

Bruyer, R., Abdi, H., and Benoit, J. (1987). Stimulus versus face recognition in laterally displayed stimuli. *The American journal of psychology* 100, 117–121

Chen, M. C., Anderson, J. R., and Sohn, M. H. (2001). What can a mouse cursor tell us more? correlation of eye/mouse movements on web browsing. In *CHI '01 Extended Abstracts on Human Factors in*

*Computing Systems* (New York, NY, USA: Association for Computing Machinery), CHI EA '01, 281—-282

Cronin, D. A., Hall, E. H., Goold, J. E., Hayes, T. R., and Henderson, J. M. (2020). Eye Movements in Real-World Scene Photographs: General Characteristics and Effects of Viewing Task. *Frontiers in Psychology* 10

DeAngelus, M. and Pelz, J. B. (2009). Top-down control of eye movements: Yarbus revisited. *Visual Cognition* 17, 790–811

Feher, C., Elovici, Y., Moskovitch, R., Rokach, L., and Schclar, A. (2012). User identity verification via mouse dynamics. *Information Sciences* 201, 19–36

Ferreira, S., Arroyo, E., Tarrago, R., and Blat, J. (2010). Applying mouse tracking to investigate patterns of mouse movements in web forms. *Universitat Pompeu Fabra*

Fu, E. Y., Kwok, T. C., Wu, E. Y., Leong, H. V., Ngai, G., and Chan, S. C. (2017). Your Mouse Reveals Your Next Activity: Towards Predicting User Intention from Mouse Interaction. In *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)* (Turin: IEEE), 869–874

Gamboa, H. and Fred, A. (2004). A behavioral biometric system based on human-computer interaction. In *Biometric Technology for Human Identification* (International Society for Optics and Photonics), vol. 5404, 381–392

Griffiths, L. and Chen, Z. (2007). Investigating the Differences in Web Browsing Behaviour of Chinese and European Users Using Mouse Tracking. In *Usability and Internationalization. HCI and Culture* (Berlin, Heidelberg: Springer Berlin Heidelberg), vol. 4559. 502–512

Guo, Q. and Agichtein, E. (2010a). Towards predicting web searcher gaze position from mouse movements. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA: Association for Computing Machinery), CHI EA '10, 3601–3606

Guo, Q. and Agichtein, E. (2010b). Towards predicting web searcher gaze position from mouse movements. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems - CHI EA '10* (Atlanta, Georgia, USA: ACM Press), 3601

Helo, A., Pannasch, S., Sirri, L., and Rämä, P. (2014). The maturation of eye movement behavior: Scene viewing characteristics in children and adults. *Vision Research* 103, 83–91

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences* 7, 498–504

Henderson, J. M. and Hollingworth, A. (1998). Eye movements during scene viewing: An overview. In *Eye guidance in reading and scene perception* (Elsevier). 269–293

Henderson, J. M. and Hollingworth, A. (1999). High-level scene perception. *Annual review of psychology* 50, 243–271

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Halszka, J., and van de Weijer, J. (2011). *Eye Tracking : A Comprehensive Guide to Methods and Measures* (Oxford University Press)

Huang, J., White, R., and Buscher, G. (2012). User see, user point: gaze and cursor alignment in web search. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12* (Austin, Texas, USA: ACM Press), 1341–1350

Itti, L. and Borji, A. (2015). Computational models: Bottom-up and top-down aspects. *arXiv:1510.07748 [cs]*

Krejtz, K., Duchowski, A., Krejtz, I., Szarkowska, A., and Kopacz, A. (2016). Discerning Ambient/Focal Attention with Coefficient *K*. *ACM Transactions on Applied Perception* 13, 1–20

Kumar, M., Winograd, T., and Paepcke, A. (2007). Gaze-enhanced scrolling techniques. In *CHI'07 Extended Abstracts on Human Factors in Computing Systems*. 2531–2536

Liu, C., Liu, J., and Wei, Y. (2017). Scroll up or down?: Using Wheel Activity as an Indicator of Browsing Strategy across Different Contextual Factors. In *Proceedings of the 2017 Conference on Human Information Interaction and Retrieval - CHIIR '17* (Oslo, Norway: ACM Press), 333–336

Milisavljevic, A., Bras, T. L., Mancas, M., Petermann, C., Gosselin, B., and Doré-Mazars, K. (2019). Towards a better description of visual exploration through temporal dynamic of ambient and focal modes. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications - ETRA '19* (Denver, Colorado: ACM Press), 1–4

Milisavljevic, A., Hamard, K., Petermann, C., Gosselin, B., Doré-Mazars, K., and Mancas, M. (2018). Eye and Mouse Coordination During Task: From Behaviour to Prediction:. In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications* (Funchal, Madeira, Portugal: SCITEPRESS - Science and Technology Publications), 86–93

Moher, J. and Song, J.-H. (2019). A comparison of simple movement behaviors across three different devices. *Attention, Perception & Psychophysics* 81, 2558–2569

Mueller, F. and Lockerd, A. (2001). Cheese: tracking mouse movement activity on websites, a tool for user modeling. In *CHI'01 extended abstracts on Human factors in computing systems*. 279–280

Navalpakkam, V., Jentzsch, L., Sayres, R., Ravi, S., Ahmed, A., and Smola, A. (2013). Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *Proceedings of the 22nd international conference on World Wide Web - WWW '13* (Rio de Janeiro, Brazil: ACM Press), 953–964

Pan, B., Hembrooke, H. A., Gay, G. K., Granka, L. A., Feusner, M. K., and Newman, J. K. (2004). The determinants of web page viewing behavior: an eye-tracking study. In *Proceedings of the 2004 symposium on Eye tracking research & applications*. 147–154

Pannasch, S., Helmert, J. R., Roth, K., Herbold, A.-K., and Walter, H. (2008). Visual fixation durations and saccade amplitudes: Shifting relationship in a variety of conditions. *Journal of Eye Movement Research* 2

Pannasch, S. and Velichkovsky, B. M. (2009). Distractor effect and saccade amplitudes: Further evidence on different modes of processing in free exploration of visual images. *Visual Cognition* 17, 1109–1131

Reeder, R. and Maxion, R. (2006). User Interface Defect Detection by Hesitation Analysis. In *International Conference on Dependable Systems and Networks (DSN'06)* (Philadelphia, PA, USA: IEEE), 61–72

Rheem, H., Verma, V., and Becker, D. V. (2018). Use of Mouse-tracking Method to Measure Cognitive Load. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 62, 1982–1986

Rodden, K. and Fu, X. (2007). Exploring How Mouse Movements Relate to Eye Movements on Web Search Results Pages. In *30th Annual International ACM SIGIR Conference*

Rodden, K., Fu, X., Aula, A., and Spiro, I. (2008). Eye-mouse coordination patterns on web search results pages. In *Proceeding of the twenty-sixth annual CHI conference extended abstracts on Human factors in computing systems - CHI '08* (Florence, Italy: ACM Press), 2997–3002

Roth, S. P., Tuch, A. N., Mekler, E. D., Bargas-Avila, J. A., and Opwis, K. (2013). Location matters, especially for non-salient features–An eye-tracking study on the effects of web object placement on different types of websites. *International Journal of Human-Computer Studies* 71, 228–235

Seelye, A., Hagler, S., Mattek, N., Howieson, D. B., Wild, K., Dodge, H. H., et al. (2015). Computer mouse movement patterns: A potential marker of mild cognitive impairment. *Alzheimer's & Dementia (Amsterdam, Netherlands)* 1, 472–480

Sharmin, S., Špakov, O., and Räihä, K.-J. (2013). Reading on-screen text with gaze-based auto-scrolling. In *Proceedings of the 2013 Conference on Eye Tracking South Africa* (New York, NY, USA: Association for Computing Machinery), ETSA '13, 24–31

Tatler, B. W. and Vincent, B. T. (2008). Systematic tendencies in scene viewing. *Journal of Eye Movement Research* 2, 1–18

Torralba, A., Oliva, A., Castelhano, M. S., and Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review* 113, 766

Tullis, T. S. (2007). Older Adults and the Web: Lessons Learned from Eye-Tracking. In *Universal Acess in Human Computer Interaction. Coping with Diversity* (Berlin, Heidelberg: Springer Berlin Heidelberg), vol. 4554. 1030–1039

Tzafilkou, K. and Protogeros, N. (2018). Mouse behavioral patterns and keystroke dynamics in End-User Development: What can they tell us about users' behavioral attributes? *Computers in Human Behavior* 83, 288–305

Unema, P. J. A., Pannasch, S., Joos, M., and Velichkovsky, B. M. (2005). Time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration. *Visual Cognition* 12, 473–494

Velichkovsky, B. M., Rothert, A., Kopf, M., Dornhöfer, S. M., and Joos, M. (2002). Towards an express-diagnostics for level of processing and hazard perception. *Transportation Research Part F: Traffic Psychology and Behaviour* 5, 145–156

Yarbus, A. L. (1967). Eye Movements and Vision. *New York: Plenum*

Zheng, N., Paloski, A., and Wang, H. (2011). An efficient user verification system via mouse movements. In *Proceedings of the 18th ACM conference on Computer and communications security - CCS '11* (Chicago, Illinois, USA: ACM Press), 139–150

**Figure 1.** Example of website displayed during the experiment.


**Figure 2.** Example of screen on which participants had to click the next item to get the instruction. The white button indicates a website not yet visited, the green button a website already visited, and the blue button the next website to visit. Only the blue button was clickable.

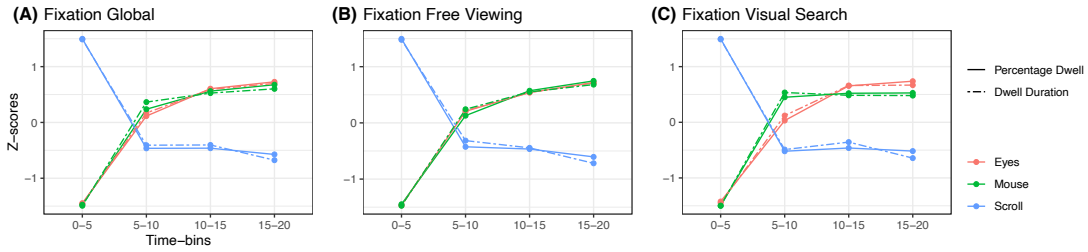



**Figure 3.** Relationship between fixation-related variables of the eyes, mouse and scroll. (A) global z-scored averages. (B) z-scored averages over time in the free viewing condition. (C) z-scored averages over time in the visual search condition.

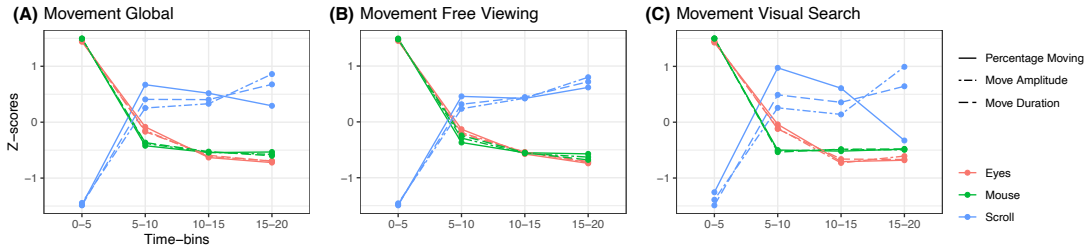



**Figure 4.** Relationship between saccade-related variables of the eyes, mouse and scroll. (A) global z-scored averages. (B) z-scored averages over time in the free viewing condition. (C) z-scored averages over time in the visual search condition.

**Table 1.** Global means and standard deviation of all studied variables (139 participants on 18 web pages for 20s each).

| | **Eye** | | **Mouse** | | **Scroll** | |
|---|---|---|---|---|---|---|
| | mean | std. | mean | std. | mean | std. |
| **Fixations/Dwells** | | | | | | |
| % of time | 85.76% | ±1.72% | 79.15% | ±8.33% | 83.42% | ±5.32% |
| Avg. count | 72.18 | ±6.5 | 8.71 | ±2.57 | 10.07 | ±1 |
| Avg. duration | 236.97ms | ±24.45ms | 2.49s | ±1s | 2.3s | ±0.71s |
| Tot. duration | 16.83s | ±0.33s | 15.68s | ±1.65s | 16.52s | ±2.33s |
| **Movements** | | | | | | |
| % of time | 14.24% | ±1.72% | 20.85% | ±8.33% | 16.58% | ±5.32% |
| Avg. count | 72.18 | ±6.5 | 6.04 | ±1.78 | 8.77 | ±2.04 |
| Avg. duration | 39.04ms | ±4.29ms | 768.24ms | ±342.55ms | 367.57ms | ±121.65ms |
| Tot. duration | 2.8s | ±0.34s | 4.13s | ±1.65s | 3.28s | ±1.39s |
| Avg. amplitude | 6.10° | ±0.67° | 0.27° | ±0.23° | 8.52° | ±2.35° |
| Tot. amplitude | 435.22° | ±52.74° | 1.6° | ±0.71° | 70.79° | ±19.70° |
| **Dynamic** | | | | | | |
| Avg. K coeff. | -0.13 | ±0.2 | -0.35 | ±0.63 | 0.43 | ±0.45 |
| Avg. nb. switches | 33.15 | ±3.25 | 3.78 | ±0.89 | 3.63 | ±0.74 |
| % time in ambient | 42.82% | ±6.81% | - | - | - | - |
| % time in focal | 57.37% | ±6.8% | - | - | - | - |

**Table 2.** Means and standard deviations of all studied variables as a function of tasks and time-bins for the eye (139 participants on 18 web pages for 20s each). The Non-significant column regroups all post-hocs with a p-value $>0.05$. Not mentioned post-hocs have a p-value

| | **Free Viewing** | | | | **Visual Search** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $T1_F$ | $T2_F$ | $T3_F$ | $T4_F$ | $T1_V$ | $T2_V$ | $T3_V$ | $T4_V$ | *Non-significant* |
| **Fixations/Dwells** | | | | | | | | | |
| Amount of time (%) | 84.06±2.10 | 86.69±1.92 | 87.20±1.93 | 87.47±1.83 | 82.90±2.11 | 85.13±2.07 | 85.90±1.92 | 85.91±1.95 | $[T3_F\text{-}T4_F, T3_V\text{-}T4_V]$ |
| Avg. count | 18.93±1.83 | 18.93±1.87 | 18.71±1.94 | 18.60±1.87 | 17.39±1.62 | 17.32±1.74 | 16.98±1.79 | 16.70±1.74 | $[T1_F\text{-}T2_F, T2_F\text{-}T3_F, T3_F\text{-}T4_F, T1_V\text{-}T2_V, T3_V\text{-}T4_V]$ |
| Avg. duration (s) | 211.27±23.27 | 234.24±28.06 | 238.83±30.34 | 241.28±30.33 | 227.61±23.75 | 251.11±30.66 | 261.18±33.71 | 262.53±33.60 | $[T3_F\text{-}T4_F, T3_V\text{-}T4_V]$ |
| **Movements** | | | | | | | | | |
| Amount of time (%) | 15.94±2.10 | 13.31±1.92 | 12.80±1.93 | 12.53±1.83 | 17.10±2.11 | 14.87±2.07 | 14.10±1.92 | 14.09±1.95 | $[T3_F\text{-}T4_F, T3_V\text{-}T4_V]$ |
| Avg. count | 18.93±1.83 | 18.93±1.87 | 18.71±1.94 | 18.60±1.87 | 17.39±1.62 | 17.32±1.74 | 16.98±1.79 | 16.70±1.74 | $[T1_F\text{-}T2_F, T2_F\text{-}T3_F, T3_F\text{-}T4_F, T1_V\text{-}T2_V, T3_V\text{-}T4_V]$ |
| Avg. duration (ms) | 39.51±5.14 | 35.28±4.81 | 34.39±4.85 | 33.91±4.74 | 46.07±4.99 | 42.86±4.97 | 41.58±4.71 | 41.83±5.21 | $[T3_F\text{-}T4_F, T3_V\text{-}T4_V]$ |
| Avg. amplitude (°) | 6.26±1.03 | 4.97±0.88 | 4.61±0.89 | 4.47±0.88 | 8.41±0.97 | 7.32±0.96 | 6.82±0.96 | 6.84±1.01 | All |
| **Dynamic** | | | | | | | | | |
| Avg. K coeff. | -0.36±0.25 | 0.05±0.25 | 0.15±0.26 | 0.20±0.26 | -0.62±0.25 | -0.24±0.28 | -0.08±0.31 | -0.07±0.29 | All |
| Avg. switch nb. | 8.85±1.11 | 9.31±1.36 | 8.90±1.48 | 8.89±1.43 | 7.79±1 | 8.70±1.18 | 8.53±1.18 | 8.28±1.24 | $[T3_F\text{-}T3_V, T3_F\text{-}T4_F, T2_V\text{-}T3_V, T3_V\text{-}T4_V]$ |
| % time in ambient (%) | 52.27±8.96 | 37.01±8.38 | 33.36±9.14 | 32.02±8.75 | 59.82±8.30 | 47.93±9.12 | 43.75±9.46 | 44.03±8.92 | $[T3_F\text{-}T4_F, T3_V\text{-}T4_V]$ |
| % time in focal (%) | 48.51±8.91 | 63.67±8.34 | 67.32±9.09 | 68.64±8.72 | 41.14±8.26 | 52.92±9.10 | 57.10±9.43 | 56.83±8.88 | $[T3_F\text{-}T4_F, T3_V\text{-}T4_V]$ |

**Table 3.** Means and standard deviations of all studied variables as a function of tasks and time-bins for the mouse (139 participants on 18 web pages for 20s each). The Non-significant column regroups all post-hocs with a p-value $>0.05$. Not mentioned post-hocs have a p-value $<0.05$ or less.

| | **Free Viewing** | | | | **Visual Search** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $T1_F$ | $T2_F$ | $T3_F$ | $T4_F$ | $T1_V$ | $T2_V$ | $T3_V$ | $T4_V$ | *Non-significant* |
| **Fixations/Dwells** | | | | | | | | | |
| % of time (%) | 63.68±12.82 | 79.74±13.30 | 82.44±12.71 | 83.85±12.24 | 60.14±12.15 | 79.39±11.25 | 77.84±10.76 | 77.70±10.53 | $[T2_F\text{-}T2_V, T3_F\text{-}T4_F, T2_V\text{-}T3_V, T3_V\text{-}T4_V]$ |
| Avg. count | 3.52±0.51 | 3.38±0.73 | 3.22±0.66 | 3.13±0.66 | 3.71±0.59 | 3.51±0.79 | 3.54±0.70 | 3.57±0.78 | $[T3_F\text{-}T4_F, T2_V\text{-}T3_V, T3_V\text{-}T4_V]$ |
| Avg. duration (s) | 1.00±0.29 | 1.35±0.43 | 1.45±0.43 | 1.52±0.43 | 0.89±0.27 | 1.30±0.39 | 1.25±0.35 | 1.24±0.35 | $[T3_F\text{-}T4_F, T2_V\text{-}T3_V, T3_V\text{-}T4_V]$ |
| **Movements** | | | | | | | | | |
| % of time (%) | 36.43±12.79 | 22.32±13.38 | 20.22±12.96 | 18.31±12.88 | 39.90±12.19 | 22.97±11.24 | 23.88±11.01 | 23.96±10.22 | $[T2_F\text{-}T2_V, T3_F\text{-}T4_F, T2_V\text{-}T3_V, T3_V\text{-}T4_V]$ |
| Avg. count | 2.19±0.44 | 2.04±0.62 | 1.95±0.62 | 1.79±0.55 | 2.34±0.48 | 2.12±0.71 | 2.11±0.59 | 2.15±0.60 | $[T2_F\text{-}T2_V, T3_F\text{-}T4_F, T2_V\text{-}T3_V, T3_V\text{-}T4_V]$ |
| Avg. duration (ms) | 965±399 | 570±431 | 511±358 | 517±423 | 1000±452 | 572±334 | 589±328 | 592±303 | $[T1_F\text{-}T1_V, T2_F\text{-}T2_V, T3_F\text{-}T4_F, T2_V\text{-}T3_V, T3_V\text{-}T4_V]$ |
| Avg. amplitude (°) | 0.43±0.27 | 0.15±0.16 | 0.11±0.12 | 0.13±0.17 | 0.46±0.31 | 0.18±0.24 | 0.19±0.16 | 0.20±0.25 | All |
| **Dynamic** | | | | | | | | | |
| Avg. K coeff. | -1.02±0.74 | -0.05±0.71 | 0.34±0.64 | 0.63±0.79 | -1.13±0.81 | -0.14±0.99 | 0.08±0.55 | 0.12±0.64 | All |
| Avg. switch nb. | 1.61±0.32 | 1.51±0.37 | 1.45±0.34 | 1.43±0.37 | 1.65±0.32 | 1.60±0.40 | 1.60±0.39 | 1.64±0.38 | $[T1_F\text{-}T1_V, T3_F\text{-}T4_F, T2_V\text{-}T3_V, T3_V\text{-}T4_V]$ |

**Table 4.** Means and standard deviations of all studied variables as a function of tasks and time-bins for scrolling (139 participants on 18 web pages for 20s each). The Non-significant column regroups all post-hocs with a p-value $>0.05$. Not mentioned post-hocs have a p-value $<0.05$ or less.

| | Free Viewing | | | | Visual Search | | | | *Non-significant* |
|---|---|---|---|---|---|---|---|---|---|
| | $T1_F$ | $T2_F$ | $T3_F$ | $T4_F$ | $T1_V$ | $T2_V$ | $T3_V$ | $T4_V$ | |
| **Fixations/Dwells** | | | | | | | | | |
| Amount of time (%) | 91.40±3.51 | 86.19±6.16 | 85.58±5.96 | 84.58±6.46 | 82.14±8.23 | 73.54±9.35 | 74.09±8.66 | 72.56±9.20 | $[T2_F\text{-}T3_F, T3_F\text{-}T4_F, T2_V\text{-}T3_V, T3_V\text{-}T4_V]$ |
| Avg. count | 2.50±0.45 | 3.24±0.62 | 3.28±0.66 | 3.35±0.67 | 3.35±0.73 | 4.36±0.80 | 4.35±0.72 | 4.29±0.71 | $[T2_F\text{-}T3_F, T3_F\text{-}T4_F, T2_V\text{-}T3_V, T3_V\text{-}T4_V]$ |
| Avg. duration (s) | 1.91±0.29 | 1.50±0.33 | 1.48±0.31 | 1.44±0.32 | 1.39±0.39 | 0.95±0.30 | 0.95±0.26 | 0.93±0.27 | $[T2_F\text{-}T3_F, T3_F\text{-}T4_F, T2_V\text{-}T3_V, T3_V\text{-}T4_V]$ |
| **Movements** | | | | | | | | | |
| Amount of time (%) | 8.68±3.53 | 13.84±6.18 | 14.46±5.94 | 15.45±6.43 | 17.90±8.19 | 26.46±9.35 | 25.91±8.66 | 27.58±9.57 | $[T2_F\text{-}T3_F, T3_F\text{-}T4_F, T2_V\text{-}T3_V, T3_V\text{-}T4_V]$ |
| Avg. count | 1.50±0.44 | 2.22±0.62 | 2.26±0.64 | 2.33±0.67 | 2.35±0.73 | 3.35±0.81 | 3.32±0.72 | 3.26±0.68 | $[T2_F\text{-}T3_F, T3_F\text{-}T4_F, T2_V\text{-}T3_V, T3_V\text{-}T4_V]$ |
| Avg. duration (ms) | 292±125 | 305±99 | 316±104 | 333±148 | 385±172 | 397±129 | 394±141 | 429±158 | All |
| Avg. amplitude (°) | 5.88±3.08 | 6.48±3.08 | 6.46±2.42 | 6.78±2.56 | 9.29±3.56 | 9.78±2.87 | 10.12±3.58 | 10.47±3.62 | $[T2_F\text{-}T3_F, T3_F\text{-}T4_F, T2_V\text{-}T3_V, T3_V\text{-}T4_V]$ |
| **Dynamic** | | | | | | | | | |
| Avg. K coeff. | 0.75±0.39 | 0.96±0.61 | 0.92±0.77 | 0.88±0.89 | 0.16±0.51 | -0.06±0.54 | -0.19±0.56 | -0.31±0.68 | $[T2_F\text{-}T3_F, T3_F\text{-}T4_F, T3_V\text{-}T4_V]$ |
| Avg. switch nb. | 1.24±0.26 | 1.46±0.32 | 1.46±0.31 | 1.47±0.31 | 1.47±0.29 | 1.65±0.36 | 1.65±0.34 | 1.63±0.34 | $[T2_F\text{-}T3_F, T3_F\text{-}T4_F, T2_V\text{-}T3_V, T3_V\text{-}T4_V]$ |

CHAPTER 8

# MODELLING EYE MOVEMENTS ON WEB PAGES

## Contents

The third axis of this work is to integrate our findings from previous axes in web pages scanpath modelling in order to improve prediction accuracy. Existing scanpath models rarely address web pages, but when they do, they consider web pages as static screenshots without the need to scroll (Le Meur & Coutrot, 2016; Shen et al., 2015; Shen & Zhao, 2014; Xia & Quan, 2020). To tackle this problematic, we propose the first saccadic model including scrolling. Furthermore, we propose a dynamic approach of oculomotor biases. Our approach models oculomotor biases as a function of time to enhance saccadic models. We also introduce a new methodology to analyse models based on the time-dynamic of generated scanpath and compare it to human scanpath. Thus, our work highlights the importance of dynamics in the prediction of eye movements when exploring web pages.

## 8.1 Data

The model presented in this chapter has been trained and evaluated on a subset of the study used in **Articles 3** and **Posters 1 and 2** (see Chapter 4). In this study, 18 web pages were displayed on a computer screen with a 1920x1080 pixels resolution. In this chapter, 9 web pages were selected to train and evaluate the model during free viewing task. We only selected 9 web pages over the 18 originally presented to avoid web pages with sticky header and a good distribution of participants. A sticky header being a header of a web page staying on top of the screen while scrolling.

## 8.2 Model architecture

Based on the literature, a saccadic model can be synthesised as 3 complementary but essential modules. The first module's role is to handle the features extraction which can be either bottom-up, top-down or a mix of the two. Most of the time this part is performed by a saliency map which comes from an already existing saliency model (e.g. Boccignone & Ferraro, 2004; Le Meur & Liu, 2015; Wang et al., 2016). The second

module models the oculomotor biases mechanisms. Which bias is modelled vary a lot from one model to another, but most of the time IOR and saccade amplitude are included. Finally, the fixation selection module is the one that attracted the most attention in saccadic modelling. Its role is to take the two previous modules to select the next fixation. Furthermore, since we model ocular behaviour on web pages, we had a new module called a viewport engine. Described in Chapter 5, its role is to reproduce the scroll behaviour and to extract viewport as an input to the model for prediction.

### 8.2.1 Top-down saliency map

It is generally assumed that attention orienting is the result of an interaction between bottom-up and top-down factors. Some authors have even questioned the existence of purely bottom-up influence during visual exploration (Theeuwes & Failing, 2020). Moreover, top performing saliency models of the MIT Benchmark (Kümmerer et al., 2020) are learning-based which are mostly top-down. Therefore, the saliency map is predicted using the state-of-the-art Saliency Attentive Model (SAM) top-down model (Cornia et al., 2018) (see Chapter 3 for more details).

### 8.2.2 Oculomotor biases

#### 8.2.2.1 Dynamic fixation duration

Fixating is an active process allowing us to grasp details on our surrounding environment. The Fixation Duration (FD), which designates the period of time between two saccades, is usually around 300 to 350 milliseconds Mackworth and Morandi, 1967; Yarbus, 1967. However, FD is not stable during the exploration and evolves over time. Buswell (1935) was the first to observe an evolution of FD and SA when exploring paintings. He described two patterns: short fixations during the global scanning of the scene, and longer fixations in more limited areas, usually occurring after the scanning phase. Antes (1974), like Buswell (1935), showed a constant increase of mean FD and a constant decrease of mean SA during the exploration of 10 pictures. Unema et al. (2005) reported a steady increase

in fixations duration over the first few seconds. They also modelled these fluctuations by an exponential function described in Equation 8.1.

$$FD(t) = b * e^{a/t} \tag{8.1}$$

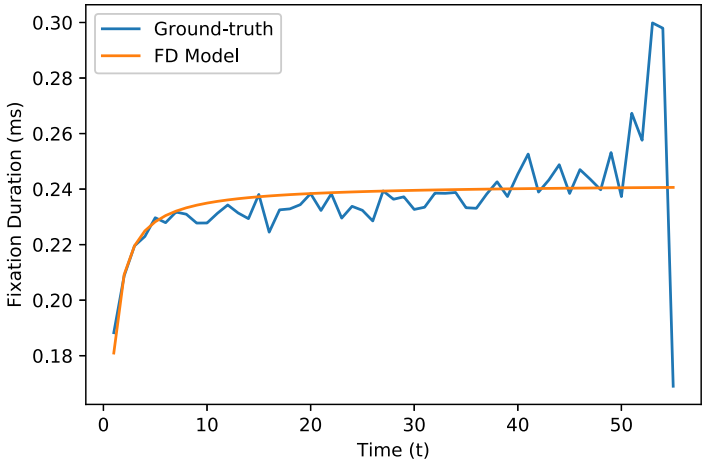



Figure 8.1 – **Fixation duration modelling**. Predicted and observed average fixation durations as a function of time.

Where $b$ is the asymptotic value, $a$ the acceleration rate of the function, and $t$ the time elapsed since the beginning of the exploration. In our model, we first grouped fixations in 1s time bin across all participants. We then averaged fixations duration of each time bins to have a single value for each second. Finally, we fitted the resulting average fixations duration using Equation 8.1. We then used the fitted exponential function to predict fixation duration after for each predicted fixation. The result of our model is showed in Figure 8.1. Here, the fixation duration prediction is based on global behaviour observed across literature. Thus, it is not directly related to the content.

#### 8.2.2.2 Dynamic direction map

Saccades are not uniformly distributed in space. We tend to execute more horizontal saccades than vertical ones and even fewer oblique saccades (e.g. Brandt, 1945; Gilchrist

& Harvey, 2006). Although there is no consensus yet in the literature, extraocular muscles are commonly found responsible. As described in Section 1.2.2, whereas only a single pair of muscles is necessary to move our eyes horizontally, two pairs of muscles are required to move the eyes vertically and obliquely. Foulsham et al. (2008) found that participants could easily make vertical or oblique saccades when the stimuli were rotated.

To model saccade orientation over time, we follow a similar approach as Le Meur and Coutrot (2016). The screen is first divided in 9 areas of equal size (640x360 pixels). Then, every saccade origin initiated in a given area is shifted to the area's centre, while keeping the correct saccade direction. Finally, a line is drawn for each saccade and a Gaussian filter of 1 degree of visual angle is applied. This operation is repeated for each area every second. The result is a set of direction maps for every area.

During prediction, the direction map corresponding to the area of the previous predicted fixation is selected. The direction map is then centered to the previous fixation and merged with the top-down saliency map.

#### 8.2.2.3 Dynamic saccade amplitude

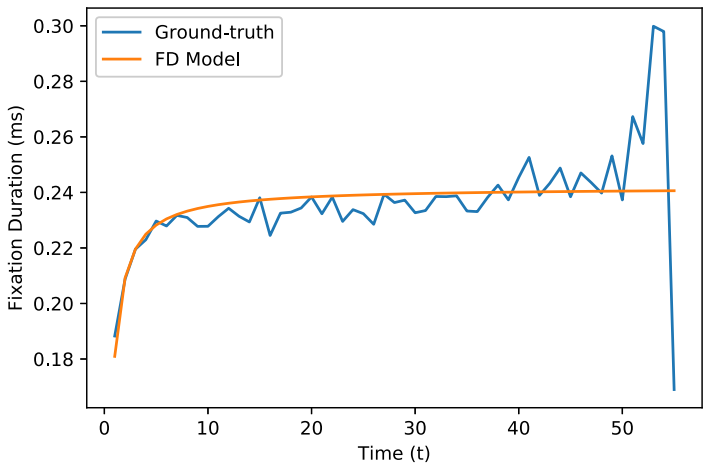



Figure 8.2 – **Saccade amplitude modelling**. Predicted and observed average saccade amplitudes as a function of time.

Saccade amplitudes (SA) follow a positively-skewed distribution which means that we tend to make short saccades more often than long saccades (e.g. Gajewski et al., 2005; Tatler et al., 2006). In ecological situations, saccades average amplitude is lower than 15 degrees of visual angle (Gilchrist, 2011). Moreover, as described in Section 1.3.1, SA are longer at the beginning of the exploration and decrease over time. Unema et al. (2005) reported a steady decrease in saccade amplitude over the first few seconds. They modelled these fluctuations with the same exponential function as fixation duration described in Equation 8.2.

$$SA(t) = b * e^{a/t} \tag{8.2}$$

Where $b$ is the asymptotic value, $a$ the acceleration rate of the function, and $t$ the time elapsed since the beginning of the exploration. As for fixation duration, saccades are first grouped in 1s time bins across all participants. We then averaged saccades amplitude of each time bin to have a single value for each second. Finally, we fitted the resulting average saccades amplitudes using Equation 8.2. The saccade amplitude is determined by computing the time elapsed since the first predicted fixation. The result of our model is showed in Figure 8.2.

#### 8.2.2.4 Inhibition Of Return

Posner and Cohen (1984) observed an automatic inhibitor mechanism preventing the organism from exploring twice a same location, so it can be faster and more efficient during the rest of the visual exploration of the environment: Inhibition Of Return (IOR). IOR is directly involved in the modulation of our visual scanning by reducing the number of eye movements directed to the locations previously fixated. Furthermore, it has been shown that IOR could not occur without eye movements in the orienting of endogenous attention which suggests that IOR could be related to oculomotor system activation (for an extensive review see Klein (2000)).

$$P_{ior}(t) = \frac{T - N + t + 1}{T} \tag{8.3}$$

We define $P_{ior}(t)$ as the inhibition state of a fixation at time $t$ from a scanpath with $N$ fixations (see Equation 8.3). Based on Mannan et al. (1997) and Samuel and Kat (2003) works, Le Meur and Liu (2015) suggested the disappearance of IOR after 2.4s which corresponds to $T = 8$ fixations with an average duration of 300ms. Thus, location fixated at time $t$ and an area of 2 degrees of visual angle around it is inhibited and linearly recovers after 8 fixations. All fixations inhibited at time $t$ are represented in a probability map $P_{ior}$.

### 8.2.3 Scroll mechanism

A web page must be scrolled for its content to be fully explored. For this reason, we implemented a scroll mechanism to always provide to the model a realistic representation of what a user would see. Using this mechanism, an image with the size of the screen (1920x1080 pixels) is extracted from the web page given scrolling information. The scrolling strategy is directly based on real scrolling sessions recorded during the experimental study. The part of the web page is updated when at least 10% of the screen has changed due to the scroll, below this threshold the viewport does not change. Next, the time before the next 10% change is computed. The model then predicts as many fixations as needed to reach the computed time. The saliency map is computed for each pause between scrolls, while the inhibition map $P_{ior}(t)$ is computed for the entire web page.

### 8.2.4 Strategy to choose the next fixation

To determine the next fixation, we consider the three computed probability maps at time $t$ described in Equation 8.4.

$$P(x,t) = P_{sal}(x) \times P_{dm}(x_{t-1}, t-1) \times P_{ior}(t) \tag{8.4}$$

Where $P(x,t)$ is the probability for each pixel to be selected as the new fixation, $P_{sal}(x)$ the top-down saliency map, $P_{dm}(x_{t-1},t)$ the direction map at previous location and time and $P_{ior}(t)$ represents the memory state of the location $x$ at time $t$. Then the next fixation is determined using Equation 8.5.

$$\begin{aligned} x &= \text{argmin}(d(d_x, s(t))) \\ d_x &= d(x_{t-1}, P(x,t) > \alpha) \end{aligned} \tag{8.5}$$

All values from $P(x,t)$ greater than the threshold $\alpha$ are kept. Then, every possible amplitude between the previous fixation $x_{t-1}$ and the selected probabilities $P(x,t) > \alpha$ are computed. Next, another euclidean distance is computed between possible amplitudes and the predicted saccade amplitude $s(t)$ at time $t$. The next fixation corresponds to the point from $P(x,t)$ with the minimum distance between its amplitude with the previous fixation $d_x$ and the predicted saccade amplitude $s(t)$.

## 8.3 Results

We evaluate the relevance of the oculomotor biases modelling as a function of time through the MultiMatch metric (Jarodzka et al., 2010). The contribution of each oculomotor bias to the overall performance is examined. MultiMatch is an algorithm to compare two scanpaths on five dimensions: scanpath shape, saccade length, fixations location, fixations duration and saccades orientation. Each dimension can be studied separately or in an global MultiMatch score averaging all dimensions. Before computing each measure, the scanpath is pre-processed through two steps: simplification and temporal alignment. The simplification phase consists of deleting small saccades and merging consecutive

| Method | MM | $MM_{dir}$ | $MM_{shape}$ | $MM_{len}$ | $MM_{pos}$ | $MM_{dur}$ | DurDiff | LenDiff | DirDiff | CC | NSS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| WTA+IOR | 0.847 | 0.752 | 0.992 | 0.991 | 0.846 | 0.655 | 0.664 | 0.908 | 0.255 | 0.175 | 0.520 |
| Random | 0.823 | 0.691 | 0.976 | 0.968 | 0.827 | 0.657 | 0.664 | 0.787 | 0.279 | 0.171 | 0.432 |
| Le Meur | 0.849 | 0.780 | 0.983 | 0.979 | 0.846 | 0.656 | 0.664 | 0.859 | 0.250 | **0.374** | **0.999** |
| **Fixation duration** | | | | | | | | | | | |
| const. fix. dur. | 0.848 | 0.756 | 0.992 | 0.991 | 0.844 | 0.655 | 0.664 | 0.582 | 0.257 | 0.171 | 0.512 |
| exp. fix. dur. | 0.848 | 0.756 | 0.992 | 0.991 | 0.844 | **0.754** | **0.745** | 0.582 | 0.257 | 0.171 | 0.512 |
| **Saccade amplitude** | | | | | | | | | | | |
| sac. amp. top 100% | 0.872 | 0.783 | 0.990 | 0.988 | 0.845 | 0.754 | 0.745 | 0.928 | 0.255 | 0.192 | 0.554 |
| sac. amp. top 50% | 0.872 | **0.784** | 0.990 | 0.988 | 0.845 | 0.755 | 0.745 | 0.928 | 0.255 | 0.191 | 0.552 |
| sac. amp. top 10% | 0.872 | 0.775 | 0.992 | 0.991 | 0.846 | 0.755 | 0.745 | 0.922 | 0.256 | 0.178 | 0.521 |
| **Saccade orientation** | | | | | | | | | | | |
| dir. map. | 0.860 | 0.684 | 0.989 | 0.986 | **0.885** | 0.754 | 0.745 | 0.821 | 0.321 | 0.237 | 0.696 |
| dir. map + sac. amp. top 100% | **0.877** | 0.764 | **0.993** | **0.992** | 0.882 | 0.754 | 0.745 | **0.937** | **0.339** | 0.269 | 0.768 |

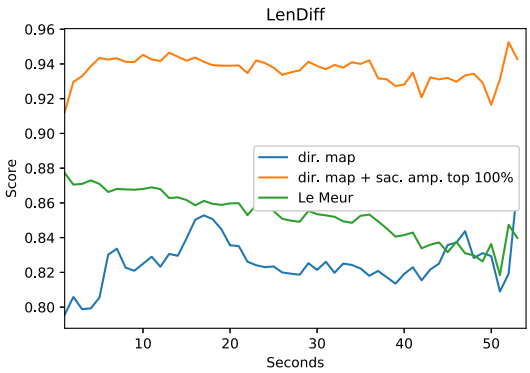
Table 8.1 – **Models performance** using MultiMatch (MM) which average values from direction ($MM_{dir}$), shape ($MM_{shape}$), length ($MM_{len}$), position ($MM_{pos}$) and duration ($MM_{dur}$). Duration difference (DurDiff), Length difference (LenDiff) and Direction difference (DirDiff) are equivalent of corresponding MultiMatch metrics without temporal alignment. Moreover, saliency metrics Correlation Coefficient (CC) and Normalized Scanpath Saliency (NSS) are computed. The model "*Le Meur*" corresponds to the model presented in Le Meur and Coutrot (2016).

long saccades with the same direction. Thus, the noise induced by the individuality of each participants is reduced. Then, the temporal alignment step aims to transform simplified scanpath in a graph from which the shortest path is computed using Dijkstra algorithm (Dijkstra, 1959).

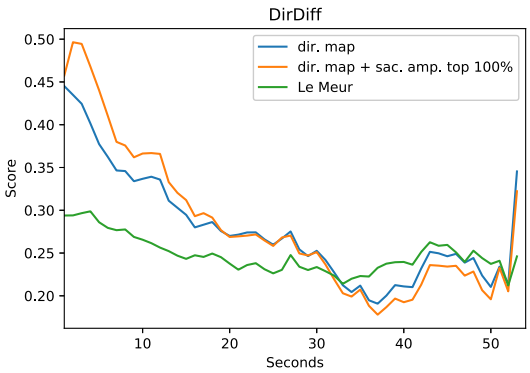
Since we assess the evolution of biases over time, we implemented a dynamic version of MultiMatch without the alignment phase. In the original version, temporal alignment is using fixation location using the shortest path. Thus, a fixation contribution to the shortest path may be removed or moved in the fixation sequence, which would interfere with the scanpath time course evaluation. Each dimension is computed for each second of the exploration. In addition, we use the Normalized Scanpath Saliency (NSS) and the Correlation Coefficient (CC) to compare scanpath-based saliency map with ground-truth statistical distributions. Again, both metrics are computed globally and over time.

As shown in Table 8.1 and Figure 8.3, the combination of direction map modelling direction and the model implementing saccade amplitude performs the best. However, *Le Meur* model still performs better when comparing with CC and NSS. As shown in Figure 8.3 (c) and (d), our approach outperform other models, but only for the few first

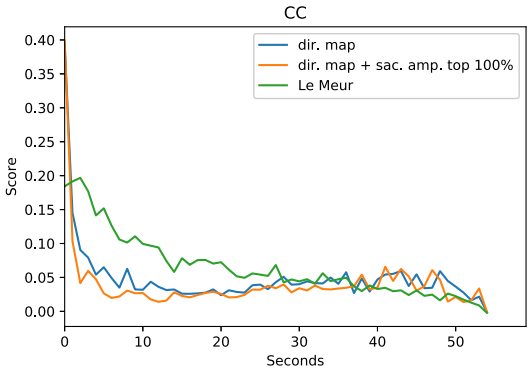
seconds. Then the scores converge around the half of exploration.



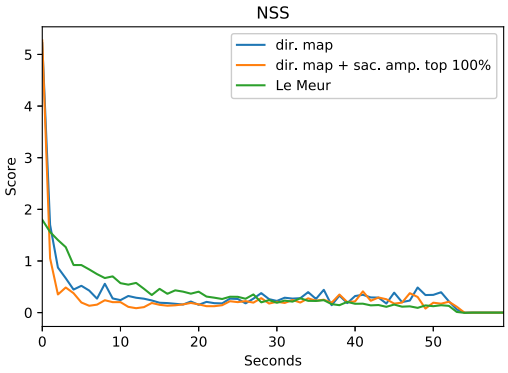

(a) LenDiff




(b) DirDiff




(c) CC




(d) NSS

Figure 8.3 – **Metrics time courses**. (a) depicts the time course of top models using LenDiff metric, (b) describe the time course of top models using DirDiff metric, (c) shows the time course of top models using CC saliency metric, (d) shows the time course of top models using NSS saliency metric. The higher the score is, the better the model performed.

## 8.4 Conclusions and perspectives

Our goal was to investigate how oculomotor biases could improve scanpath modelling. The fixation duration approximation over time was at the core of our approach. The vast majority of existing models try to reproduce eye movement dynamics but neglect fixation duration. Without fixation duration, the analysis of model performances over time is not

possible. We demonstrated that using a fixed average fixation duration of 300ms was an easy and reasonable approximation, but a more refined approximation as we introduced in this chapter, provided significant improvement of the proposed model. Furthermore, we showed that the modelling of saccade amplitude dynamics contributed to improve the model. As shown in Figure 8.3 (a) this improvement is rather stable across time when combined to direction maps. Interestingly, in addition to increase overall scores, the implementation of direction maps obtained the best result in fixation position. However, when investigated over time (see Figure 8.3 (b)), we can observe that saccades orientation is better modelled during the first 30 seconds and then slowly decreases over time. The complete model provided more realistic scanpaths but failed to outperform classic models on spatial saliency metrics CC and NSS. These limited global and temporal performances may be explained by the interest or the inner goal of the participants. For the presented model, only the free viewing task has been kept. Initially, participants had to browse the web page during 60 seconds. Thus, each participant may have had a different inner goal that may have been influenced by different factors. Furthermore, the modified version of MultiMatch we used turned out to be more informative than the original ones. We suggest that eye movements cannot be summarised and compared as a simple vector as in the MultiMatch metric. Also, the normalisation of each sub-metrics is most of the time done using screen's dimensions, which result in scores often close to the maximum. Such results, interfere with the understanding of model comparison. Finally, further work is needed to confirm the interest of our dynamic approach on other datasets.

## Chapter 8 summary

In this chapter, we proposed the implementation of a scroll-like mechanism to better model eye movements on web pages. We also proposed to take into account the fixation duration, the saccade amplitude and the saccade direction through time. Our approach allowed us to evaluate and analyse the quality of the scanpath globally and over time. We have shown that our model gave the best results in terms of saccade length, saccade direction and fixation duration. In addition, we showed that although our model was better in these aspects, it was not constant over time. In particular, we showed that the prediction accuracy of the saccade direction decreased over time, while prediction accuracy of the saccade length remained stable. However, despite a more biologically plausible scanpath, we have not improved the state of the art when compared with the salience metrics.

## Résumé du chapitre 8

Nous avons proposé dans ce chapitre l'implémentation d'un mécanisme reproduisant le scroll afin de mieux modéliser les mouvements oculaires lors de l'exploration visuelle de pages webs. Nous avons également proposé la prise en compte des durées des fixations, de l'amplitude de saccade et la direction des saccades à travers le temps. Cela nous à permis d'évaluer et d'analyser la qualité des chemins oculaires globalement et à travers le temps. Nous avons montré que notre modèle obtenait les meilleurs résultats concernant la longueur des saccades, la direction des saccades et la durée des fixations. De plus, nous montrons que bien que notre modélisation soit meilleure sur ces aspects, elle n'est pas constante dans le temps. On remarque particulièrement que la qualité de prédiction des directions des saccades diminue avec le temps alors que celle des longueurs des saccades reste stable. Cependant, malgré un scanpath plus plausible biologiquement, nous n'avons pas amélioré l'état de l'art lorsque comparé avec les métriques de saillance.

# Part III

# General discussion

CHAPTER 9

# THESIS RESULTS AND INTERPRETATIONS

## Contents

This thesis presented several advances in the domain of vision and scanpath modelling. In addition to contribute to a better understanding of ocular behaviour on web pages, axes developed in this document provided improvements in computational scanpath modelling. The original contributions and their interpretation are presented in this chapter.

## 9.1 Eye movement dynamic as a single indicator

To process visual information, visual areas of the brain communicate between each others and with more than thirty different visual cortical areas. Starting from the primary visual area, these cortical regions forms two visual pathways. The ventral stream carries information about what object is seen (Vision for Recognition), while dorsal stream is related to action guided by vision (Vision for Action) (Goodale & Milner, 1992; Trevarthen, 1968).

Velichkovsky et al. (2005) were the first to link Trevarthen's focal-ambient visual pathways dichotomy to eye movements. They identified two distinct segments of eye movements: ambient and focal modes of processing. The first one was defined as short fixation (from 90 to 260ms) related to larger following saccade (>5°), while the second was characterised by fixation longer than 260-280ms followed by a saccade within the parafoveal region (<5°).

The first goal of this thesis was to find an indicator summarising the manifestation of these two visual modes in eye movements. In **Poster 1** we used the ratio introduced by Dehais et al. (2015). This ratio differentiated the time spent searching (saccades and short fixations) and time spent processing (long fixations), but it lacked of formal and physiological definitions. For this reason, we switched to another ratio, directly inspired from Velichkovsky et al. (2005) definition of ambient and focal visual modes (Krejtz et al., 2016). In **Article 1** we used this ratio to differentiate free viewing and visual search tasks. We first evaluated the dominant mode over the visual exploration, but failed to find differences between tasks. That is why we introduced two new variables derived

from the ratio: number of mode switches and average mode durations. These two new variables provided better results to distinguish participants' tasks. However, we showed in **Article 1** and **Poster 2** that the temporal study of ambient and focal modes was more informative than the global score provided by the ratio. This dynamic analysis provided new evidence explaining why a global approach of visual modes might not produce significant results. Ambient mode is predominant at the beginning of exploration, while focal mode is predominant at the end. However, this predominance is not constant within these time periods. We are constantly switching back and forth between the two. This explain why the global score could not indicate which visual mode was dominant during visual exploration. However, the use of complementary measures help to better summarise the dynamic. This is mainly explained by the fact that these measures denote a frequency and a duration, while the original indicator denotes an intensity. When investigated over time, even the ratio performs better to explain visual exploration. Moreover, we can observe that even though the switch between visual modes is clearly visible, there remain a period of time without clear visual mode. In these cases, the ratio is really close to zero. Back to Velichkovsky et al. (2005) and Unema et al. (2005) definition of visual modes, they described ambient mode as an association of a short fixation and a high-amplitude saccade, and focal mode as an association of a long fixation and a low-amplitude saccade. They did not address the two remaining associations: short fixation and low-amplitude saccade association, as well as long fixation and high-amplitude saccade association. These two associations are neither covered by visual modes nor visual pathway theories from Goodale and Milner (1992).

Recent work from Follet et al. (2011) found that saccade amplitude was more representative of the visual modes than fixation duration. This would explain the FD-SA missing associations. Then, only the saccade amplitude would account for the visual mode. Further research should examined these observations with stronger tasks in order to confirm this hypothesis. However, they also suggested that focal fixations relied more on low-level features than ambient fixations. This is in contradiction with our results presented in **Article 1** in which we found a dominant ambient mode at the beginning

of the exploration which is also the period of time during which bottom-up factors have more influence on visual exploration. However, the idea that visual modes might be the result of an interplay between low-level and high-level features seems interesting. As we showed in **Article 1**, the intensity of the visual modes varies with the participant and the stimulus during the exploration. Thus, we should maybe consider ambient and focal visual modes as a continuum and not a dichotomy. In this continuum, a short fixation followed by a high-amplitude saccade may be labelled as ambient, and a long fixation followed by a low-amplitude saccade may be labelled as focal. Similarly to the interaction between bottom-up and top-down factors, the variation between the two visual processing modes would reflect the interplay between the task influence and stimulus' features. Thus, when the ratio is near zero, it may suggests a state during which we switch from one mode to another. One of the marker (FD or SA) from the current mode would then be combined to a marker (FD or SA) from the next mode. This might be explained why in these cases we observed a short fixation normally associated to ambient mode, with a low-amplitude saccade normally related to focal mode. We hypothesise that the decision to switch from one mode to another may be initiated in parallel of the end of the current visual mode. This transition might take some time to be effective, and could be observed within the ratio's value. Further work is needed to assess the proportion of these "transition modes" and how long they last.

## 9.2 The relationship between the eyes and the computer mouse

The second axis of this thesis was to investigate eye movement behaviour on web pages through the relationship between the eyes, the movement of the mouse pointer and the scroll. Contrary to images, ecological web pages need to be studied differently due to possible interactions. These interactions can take multiple forms, including clicks, scrolling, and drags&drops. While clicks and drags&drops are a mean to directly interact with the web page's content, scrolling is more about content discovery. In this work, we specifically focused on the relationship between eye movements, mouse movements and

scrolling parameters and the influence of scroll on eye movements.

Thus, we investigated in **Article 2** the coordination between the mouse cursor and the eye position. We found a better correlation on the vertical axis than the horizontal one. Yet, the coordination on the horizontal axis increased when the participant intended to click. Furthermore, we created a Gaussian model based on experimental statistics, predicting where the eyes were located given the position of the mouse.

Next, we studied how eye movements were influenced by the scroll. In **Article 2** and **Poster 3**, we found that eye positions could help to predict the next scroll amplitude. For instance, when the eyes were positioned at the top of the screen before the start of the scroll, its amplitude was higher than when the eyes were positioned at the bottom. Moreover, we showed a similar relation between eye fixation location and scroll speed during the scroll. When scrolling fast, participants tended to position their eyes at the opposite of the scroll direction.

We proposed in **Article 3** a detailed statistical description of eye movements on web pages along with the mouse movements and the scroll. We found that eyes and mouse parameters related to the movement, such as amplitude of the movement and percentage of time moving, decreased over time, while scroll parameters increased. Conversely eye and mouse parameters related to the fixation/pause, such as fixation/pause duration and percentage of time spent fixating/idling, increased over time, while scroll parameters decreased. In both cases, eyes and mouse parameters followed the same pattern, while the scroll parameters followed the opposite one. Interestingly, these observations were consistent across tasks.

These findings demonstrate that the relationship between the eyes and the mouse is deeper than expected. In **Article 2** we showed a relative link between the eyes and the mouse position, but this spatial coordination varies across studies and does not remain constant. Even though further studies are needed to confirm our results, the relationship between eyes and mouse parameters seems consistent over time. This may be related to similar processing in the ventral and dorsal streams. For instance, Stone and Gonzalez (2015) reported several studies in which ventral and dorsal streams of congenitally

blind individuals were preserved during pointing and grasping tasks. Thus, we can assume that the important role of both streams involved in hand movements and eye movements may explain why the eyes and the mouse parameters behave similarly during the exploration. However, this hypothesis does not address why the scroll parameters behave oppositely.

The opposite behaviour we observed for the scroll may be explained by the "*the sensory weighting hypothesis*" (Ernst & Banks, 2002). This theory states that during a task involving sensory competition, here the presence of both vision and haptic, we tend to rely on the optimal one to complete the task. For instance, before reaching an object whose position is unknown, we first need to look at it, but there are occasions when we reach objects without looking at them because we already know their exact position. In our case, the task is to browse the page with or without a target. At the beginning of the exploration, the optimal sensory input to fulfill this task would be the eyes. As time goes by, we discover the web page more and more until we browsed it entirely. The scroll would gradually become the optimal way to browse the web page, since fixation duration is increasing and saccade amplitude decreasing, and the scroll would then replace large saccades.

Further research is necessary to better understand what mechanisms are involved in the eyes and mouse coordination during web pages exploration. For instance, we did not differentiate scroll up from scroll down in our analyses. When we scroll down, we usually discover the content for the first time. But a scroll up is necessary to re-examine an already seen area of the web page. Differentiating the two directions might provide finer results on what cognitive processes are involved.

## 9.3 Including visual exploration and stimulus-dependant dynamics improve existing models performance

The third axis goal was to integrate findings from previous axes in web pages scanpath modelling in order to improve prediction accuracy. Existing scanpath models rarely

address web pages, but when they do, they consider web pages as static screenshots without the need to scroll (Le Meur & Coutrot, 2016; Shen et al., 2015; Shen & Zhao, 2014; Xia & Quan, 2020). To tackle this problematic, we proposed the first saccadic model including scrolling. Furthermore, scanpath modelling usually includes some oculomotor biases, which are mostly considered as stable through visual exploration. In our approach we addressed these biases through their evolution over time. Thus, this work highlighted the importance of dynamics in the prediction of eye movements when exploring web pages.

In **Chapter 8** we introduced the use of the scroll mechanism in a scanpath model, and showed significant improvement of the prediction accuracy. Our modular approach allowed us to apply this mechanism either on saliency models or scanpath models. Regarding saliency models, the use of such mechanism provided more precise saliency maps. Without it, these models were resizing the stimulus to match their expected input size, which resulted in blurry saliency maps indicating vague and confused areas. Regarding scanpath modelling, these type of models originally predicted scanpaths around the centre of the web page, without being able to fully explore the stimulus. In both cases, the scroll provided promising results, and confirmed the need to integrate such mechanism in future models on web pages.

Based on scanpath metrics, our spatio-temporal approach of oculomotor biases, including fixation durations, saccade amplitudes, and saccade directions, resulted in more realistic scanpaths. In addition to be modelled as a function of time, saccade directions have also been implemented depending on the screen area similarly to Le Meur and Coutrot (2016). However, based on saliency metrics, our scanpath model did not significantly perform better than existing scanpath models (using scroll mechanism). Saliency metrics usually describe spatial distribution of fixations during the entire exploration. Thus, they provide a summary of how good the model predicted eye positions. This may be explained by the underlying dynamic involved in vision. Attention and more specifically eye movements cannot be reduced to where we look at. The time-dependency and the order of fixations and saccades are essential to create better scanpath models.

This is an additional evidence to the importance of assessing such models separately from saliency ones.

CHAPTER 10

# PERSPECTIVES

## Contents

The results obtained in this interdisciplinary work contribute to the understanding and modelling of ocular behaviour on web pages. We propose new methods to investigate web pages and we illustrate through concrete examples how the dynamic of the scroll plays an undeniable role in ocular behaviour. Yet, we also open novel perspectives. In this chapter, we address some future works that could follow this thesis.

## 10.1 Ocular behaviour on web pages

As described in Chapter 2, the studies of eye movement behaviour on web pages are very sparse. Although our results on the description of the relationship between eye, mouse and scroll parameters are promising, we did not investigate stimulus-related factors that could influence this relationship. Further research is necessary to better characterise what influence each element separately and the impact on their relationship. We present here two preliminaries analyses on factors possibly influencing visual exploration on web pages. The first preliminary one focuses on the differences between scrolling up and down. The second preliminary analysis investigates the scroll behaviour at the beginning of the exploration. Finally, we discuss the interest of more extensive scrolling behaviour investigations on other devices than desktop computer.

### 10.1.1 Differences between scrolling up and down

The first preliminary analysis assessed the influence of scroll direction on its parameters, such as its amplitude, frequency or pause. We observed less scrolls up than scrolls down in both free viewing task and visual search task. However, we found that participants did more scrolls up in visual search task compared to free viewing task. We also observed a greater scroll amplitude when scrolling up again in both tasks. As expected we found scroll parameters differences according to its direction. When we scroll down, we usually discover the content for the first time, while a scroll up is necessary to re-examine an already displayed area of the web page, even if it has been briefly seen. However, we

observed more scrolls up in visual search task than in free viewing task. It could be explained by the requirements of the task itself. In visual search, the target can be anywhere in the web page, thus, when participants did not find target(s), they tend to double check by scrolling up and down. Accordingly, we found that scrolls up were executed in average later during the exploration compared to scrolls down.

These results show interesting differences between scrolls up and down. They support the idea that both scrolls could respond to different goals. We can also assume that a scroll down at the beginning of the exploration does not have the same purpose of a scroll down at the end. Thus, different purposes could result in different ocular behaviours. Further studies on this subject could provide a more precise understanding of the underlying dynamic of scrolling behaviour and its influence on eye movements.

### 10.1.2 Insights of the first scroll on eye movement dynamic

To consider the influence of scrolling behaviour differences during browsing, the second preliminary analysis focused on the relationship between early scrolls and eye movement dynamic. Since scrolls up occur later during web page exploration, we investigated the first five scrolls down. We found that participants scrolled for the first time after 4 seconds in average, while scrolling after 1.7 seconds afterward. In both cases, the participants scrolled around 26% of the screen height. Thus, the scroll amplitude is not influenced by the duration of the preceding pause, but the first scroll is done after a longer pause compared to following scrolls. This preliminary result may suggest that during the first few seconds of visual exploration, participants explored in more detail the first screen of the web page. This may also support previous studies on natural images and paintings that the first seconds could be an orienting phase (Karpov et al., 1968), the extraction of most informative regions (Antes, 1974) or an adaptation strategy to the task (Scinto et al., 1986). More studies are needed to investigate the influence of the beginning of the exploration on scrolling behaviour, but these preliminary results suggest that eye movement dynamic directly influence the scrolling dynamic. Thus, both dynamics should not be addressed separately but as a whole. We suggest that this

behaviour should be investigated with dedicated study implementing a free viewing task of defined duration.

### 10.1.3 Multi-device

All the studies presented in this work aimed to understand the ocular behaviour when browsing web pages on desktop computers. Yet, since the commercialisation of the iPhone in 2007 and the iPad in 2010, people started consuming more and more content on these new devices. Marginal at the beginning, more than 50% of the entire web traffic is done on mobile since 2017 (StatCounter, 2020). This new distribution of devices highlights the need to investigate web behaviour on mobiles. However, as we described in this thesis, the literature lacks of studies on ocular behaviour parameters when browsing web pages. That is why we suggest to orient new studies on the scroll. Contrary to the mouse, scrolling can either be used on desktop computers, laptops, smartphones, tablets or TVs. Although each device has different size, results on a specific device might be, in some ways, applicable to others. For instance, we hypothesise that the behaviour we observed about the eyes positioned at the top of the screen while quickly scrolling down and vice versa (see Chapter 7), may be reproducible on smartphones and tablets. We think that parameters might change between devices, including scroll amplitude, fixation duration, etc, but the behaviour may remain the same.

## 10.2 Scanpath modelling

Scanpath models attract more and more attention, but they are only a few to tackle web pages. The main disadvantage of these models is that they use web pages screenshots. As we showed in this thesis, contrary to screenshots, real web pages need to be addressed with specific tools and analyses. That is why we discuss here the urgent need to create a high-quality open source database of eye movements on web pages. Furthermore, we discuss the interest of using techniques, such as Deep Neural Network (DNN) in eye movements modelling.

### 10.2.1 Open source dataset

The first step of our work has been to collect a high-quality dataset, in order to use them to analyse behaviours and train our model. Models need data to be trained and/or evaluated. The quality of collected data can influence how good a model will be. We already evoked the MIT benchmark initiative in saliency modelling field which in addition to providing a platform to compare models between each others, they gather many available datasets in one place (Kümmerer et al., 2020). From this endeavour emerged a wide variety of models proposing numerous approaches to saliency. However, we identified only one dataset publicly available containing web pages stimuli (Shen & Zhao, 2014). Described in Chapter 5, it contains web pages screenshots displayed during 5 seconds. To help future research on web pages behaviour and eye movements prediction, we suggest the creation of a large dataset containing recordings of eye movements, mouse movements and scroll on web pages. From our experience we suggest important details that need particular attention:

- **Layout**: the importance of the web page layout needs to be properly handled. For instance, it is common to browse web pages with a layout filling the entire screen width. But, it is also common to visit web pages with a layout centered on the screen with wings containing advertisements or nothing. Since both are possible, only one should be selected to avoid layout-biases. Moreover, homepage layout needs to be differentiated from layout with content. Generally, the homepage of a website summarises available content and main categories in an appealing setting. Thus, its purpose is not the same as web pages with content.

- **Computer mouse**: we described in Chapter 2 and Chapter 7 the rising interest in estimating the eye positions from the mouse. Thus, the use of the mouse represent a unique opportunity to improve eye movement modelling.

- **Length**: the length of displayed web pages should also be taken into account. Similar web page length may cause an adaptation of the participants, which may result in stereotyped behaviour.

- **Content**: some biases could emerge if the content of a web page is not selected carefully. For instance, all content related to news or incoming or past events should be avoided. Moreover, in free viewing condition, participants may give poor interest in certain contents. To reduce this bias we recommend to mix content's topics.

### 10.2.2 DNN as a model

The model described in Chapter 8 is a computational model reproducing behaviour observed from experimental studies. These types of models are usually a combination of algorithms in order to obtain the best scanpath prediction accuracy. Their objective is to find which features are essential and how they can be approximated by an algorithm, statistics or both. The last ten years have seen the rise of Deep Neural Network (DNN) in numerous research fields to solve various problems, such as hand writing recognition, speech recognition, translation, object recognition, etc. The use of DNN paired with an increasing processing power, led to new breakthroughs which encouraged researchers from many other fields to try this solution. The interest of such algorithm lies in its unsupervised learning capabilities. Given a large amount of data and an evaluation metric (or loss function), a DNN can learn by itself which feature is important for prediction. A model that is able to learn from an input and predict an output without human action is called an end-to-end model. It should be noted that the creation of the model itself is not trivial and requires an extensive expertise in DNN.

A good example of how these models can perform compared to computational models is the accuracy achieved by DNN models in saliency map prediction. Every top-performing models of the MIT Benchmark are now all DNN models. As explained in Chapter 3 these models neglect bottom-up features, but they remain very accurate. Recently, the use of Deep Neural Network emerged in scanpath modelling (Chen & Sun, 2018; Ngo & Manjunath, 2017; Simonyan & Zisserman, 2014; Xia & Quan, 2020). However, none of these models are end-to-end yet. They are still based on computational parts, such as the Winner-Take-All (WTA) algorithm to select the optimal fixation. Further work

should focus on creating an end-to-end deep learning model.

We tried to develop an end-to-end model using Recurrent Neural Network (RNN), but we faced different problems. To lower the complexity of our approach we ran our tests on natural images. We implemented an RNN with two consecutive gated recurrent units (Cho et al., 2014). The idea behind a RNN is that its output is connected with its input, so that short-term memory can influence next iteration of the network. The RNN used in this model is similar to the Long Short-Term Memory (LSTM). The particularity of an LSTM lies in the fact that, in addition to its output connected to its input for short-term memory, a hidden state reproducing long-term memory follows the same schema. It is called a hidden state because it is not directly influenced by the input. A gated recurrent units is similar to a LSTM but with a forgetting mechanism. Unfortunately, using Mean Square Error (MSE) as a loss function, our model failed to converge to a solution.

Some models from the literature succeeded in using LSTM for scanpath prediction (Chen & Sun, 2018), which support the need to use these DNN. However, an end-to-end solution requires further research. One of our biggest problem was similar to all deep learning models: the amount of data. DNN requires a large amount of data in order to be able to learn which feature is important. We did not find a large enough dataset to train our model. This problematic got even worse for web pages. Thus, the need to build large public datasets for scanpath prediction has never been more important.

# General conclusion

At the crossroad of several research fields, the goal of this thesis was to demonstrate the importance of the dynamic in the understanding and the prediction of eye movements. Previous studies mainly focused on the prediction of where the next fixation could be located and what factors could globally influence the selection of this next fixation. Directly inherited from saliency modelling field, these studies neglected the temporal modulation of eye movements parameters.

We showed that visual exploration could be influenced by dynamic elements from multiple sources. The first one referred to the dynamic of the stimulus itself. We highlighted that during web pages exploration, scrolling had an influence on eye movement parameters. This influence could occur either before or during the scroll. The second source lied in the eye movement parameters. We showed that these parameters evolved over time when browsing a web page. We then described the benefits of incorporating these two sources of dynamic in a scanpath model, and evaluated the plausibility of generated scanpath through dedicated temporal metrics.

Saccadic and saliency models are already used to predict web users behaviour to display more accurate or specific information where it has the best chances to be seen. But, the main advantage of these models lies in the fact that they can be adapted to a wide variety of domains with specific problematic, such as software ergonomic, gaming, virtual reality, educational tools, or even clinical uses.

# BIBLIOGRAPHY


Antes, J. R. (1974). The time course of picture viewing. *Journal of Experimental Psychology*, *103*(1), 62–70.

Baddeley, R. J., & Tatler, B. W. (2006). High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Vision Research*, *46*(18), 2824–2833.

Benway, J. P. (1998). Banner Blindness: The Irony of Attention Grabbing on the World Wide Web: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *42*(5), 463–467.

Betts, J. G., DeSaix, P., Johnson, E., Johnson, J. E., Korol, O., Kruse, D. H., Poe, B., Wise, J. A., Womble, M., & Young, K. A. (2013). *Anatomy and physiology*. OpenStax.

Boccignone, G., & Ferraro, M. (2004). Modelling gaze shift as a constrained random walk. *Physica A: Statistical Mechanics and its Applications*, *331*(1-2), 207–218.

Boi, P., Fenu, G., Spano, L. D., & Vargiu, V. (2016). Reconstructing User's Attention on the Web through Mouse Movements and Perception-Based Content Identification. *ACM Transactions on Applied Perception*, *13*(3), 1–21.

Borji, A., & Itti, L. (2013). State-of-the-Art in Visual Attention Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 185–207.

Borji, A., & Itti, L. (2015). CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research. *arXiv:1505.03581 [cs]*.

Braganza, C., Marriott, K., Moulder, P., Wybrow, M., & Dwyer, T. (2009). Scrolling behaviour with single-and multi-column layout. *Proceedings of the 18th international conference on World Wide web - WWW '09*, 831–840.

Brandt, H. F. (1945). *The psychology of seeing*. The Philosophical Library.

Brandt, S. A., & Stark, L. W. (1997). Spontaneous Eye Movements During Visual Imagery Reflect the Content of the Visual Scene. *Journal of Cognitive Neuroscience*, *9*, 27–38.

Breitmeyer, B. G., & Ganz, L. (1976). Implications of sustained and transient channels for theories of visual pattern masking, saccadic suppression, and information processing. *Psychological Review*, *83*(1), 1–36.

Brockmann, D., & Geisel, T. (2000). The ecology of gaze shifts. *Neurocomputing*, *32-33*, 643–650.

Bruce, N., & Tsotsos, J. (2007). Attention based on information maximization. *Journal of Vision*, *7*(9), 950–950.

Bullier, J., Schall, J. D., & Morel, A. (1996). Functional streams in occipito-frontal connections in the monkey. *Behavioural Brain Research*, *76*(1-2), 89–97.

Burr, D. C., & Ross, J. (1982). Contrast sensitivity at high velocities. *Vision Research*, *22*(4), 479–484.

Buscher, G., Cutrell, E., & Morris, M. R. (2009). What Do You See When You're Surfing? Using Eye Tracking to Predict Salient Regions of Web Pages, 21–30.

Buswell, G. T. (1935). *How people look at pictures: a study of the psychology and perception in art*. University of Chicago Press.

Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision*, *9*(3), 6–6.

Chen, M.-C., & Sohn, M.-H. (2001). What can a mouse cursor tell us more? Correlation of eye/mouse movements on web browsing. *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, 281–282.

Chen, Z., & Sun, W. (2018). Scanpath prediction for visual attention using IOR-ROI LSTM. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 642–648.

Cho, C.-H., & Cheon, H. J. (2004). Why Do People Avoid Advertising on the Internet? *Journal of Advertising*, *33*(4), 89–97.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv:1406.1078 [cs, stat]*.

Collins, T., & Doré-Mazars, K. (2006). Eye movement signals influence perception: Evidence from the adaptation of reactive and volitional saccades. *Vision Research*, *46*(21), 3659–3673.

Cornia, M., Baraldi, L., Serra, G., & Cucchiara, R. (2018). Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. *IEEE Transactions on Image Processing*, *27*(10), 5142–5154.

Coutrot, A., Hsiao, J., & Chan, A. (2017). Scanpath modeling and classification with Hidden Markov Models. *Behavior Research Methods*, *50*, 1–26.

Cowey, A. (1964). Projection of the retina on to striate and prstriate cortex in the squirrel monkey, saimiri sciureus. *Journal of Neurophysiology*, *27*, 366–393.

Cutrell, E., & Guan, Z. (2007). What are you looking for?: an eye-tracking study of information usage in web search. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '07*, 407–416.

Dehais, F., Peysakhovich, V., Scannella, S., Fongue, J., & Gateau, T. (2015). "Automation Surprise" in Aviation: Real-Time Solutions. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2525–2534.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, *36*(12), 1827–1837.

Dijkstra, E. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*.

Ditchburn, R. W., & Ginsborg, B. L. (1952). Vision with a Stabilized Retinal Image. *Nature*, *170*(4314), 36–37.

Doré-Mazars, K., & Collins, T. (2005). Saccadic adaptation shifts the pre-saccadic attention focus. *Experimental Brain Research*, *162*(4), 537–542.

Dyson, M. C. (2004). How physical text layout affects reading from screen. *Behaviour & Information Technology*, *23*(6), 377–393.

Ellis, S. R., & Smith, D. J. (1985). Patterns of statistical dependency in visual scanning. *Eye movements and human information processing chapter eye movements and human information processing* (Elsevier Science Publishers BV, pp. 221–238). North Holland Press.

Ellis, S. R., & Stark, L. (1986). Statistical Dependency in Visual Scanning. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *28*(4), 421–438.

Engbert, R., Trukenbrod, H. A., Barthelmé, S., & Wichmann, F. A. (2015). Spatial statistics and attentional dynamics in scene viewing. *Journal of Vision*, *15*(1), 14–14.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433.

Fessenden, T. (2017). Horizontal Attention Leans Left. Retrieved August 20, 2020, from https://www.nngroup.com/articles/horizontal-attention-leans-left/

Findlay, J., & Gilchrist, I. (2003). *Active Vision: The Psychology of Looking and Seeing*. Oxford University Press.

Fischer, B., Gezeck, S., & Hartnegg, K. (1997). The analysis of saccadic eye movements from gap and overlap paradigms. *Brain Research Protocols*, *2*(1), 47–52.

Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(4), 1030–1044.

Follet, B., Le Meur, O., & Baccino, T. (2011). New insights into ambient and focal visual fixations using an automatic classification algorithm. *i-Perception*, *2*(6), 592–610.

Foulsham, T., Kingstone, A., & Underwood, G. (2008). Turning the world around: Patterns in saccade direction vary with picture orientation. *Vision Research*, *48*(17), 1777–1790.

Gajewski, D. A., Pearson, A. M., Mack, M. L., Bartlett, F. N., & Henderson, J. M. (2005). Human Gaze Control in Real World Search. In L. Paletta, J. K. Tsotsos, E. Rome, & G. Humphreys (Eds.), *Attention and Performance in Computational Vision* (pp. 83–99). Springer.

Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (2001). *Neurosciences Cognitives : La Biologie de l'Esprit*. De Boeck Université.

Geisler, W. S., & Perry, J. S. (2002). Real-Time Simulation of Arbitrary Visual Fields. *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications*, 83–87.

Gilchrist, I. D. (2011). Saccades. *The Oxford Handbook of Eye Movements* (S. P. Liversedge, I. D. Gilchrist, & S. Everling, pp. 85–94). Oxford University Press.

Gilchrist, I. D., & Harvey, M. (2006). Evidence for a systematic component within scan paths in visual search. *Visual Cognition*, *14*(4-8), 704–715.

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics*, *24*(6), 631–645.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in neurosciences*, *15*(1), 20–25.

Guo, Q., & Agichtein, E. (2010). Towards predicting web searcher gaze position from mouse movements. *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems - CHI EA '10*, 3601.

Hacisalihzade, S., Stark, L., & Allen, J. (1992). Visual perception and sequences of eye movement fixations: a stochastic modeling approach. *IEEE Transactions on Systems, Man, and Cybernetics*, *22*(3), 474–481.

Han, R., & Xiao, S. (2018). Human Visual Scanpath Prediction Based on RGB-D Saliency. *Proceedings of the 2018 International Conference on Image and Graphics Processing*, 180–184.

Helmholtz, H. (1896). *Physiological Optics*.

Helo, A., Pannasch, S., Sirri, L., & Rämä, P. (2014). The maturation of eye movement behavior: Scene viewing characteristics in children and adults. *Vision Research*, *103*, 83–91.

Hervet, G., Guérard, K., Tremblay, S., & Chtourou, M. S. (2011). Is banner blindness genuine? Eye tracking internet text advertising. *Applied Cognitive Psychology*, *25*(5), 708–716.

Hotchkiss, G., Alston, S., & Edwards, G. (2005). Eye Tracking Study. Retrieved August 20, 2020, from https://searchengineland.com/figz/wp-content/seloads/2007/09/hotchkiss-eye-tracking-2005.pdf

Huang, J., White, R., & Buscher, G. (2012). User see, user point: gaze and cursor alignment in web search. *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, 1341.

Irwin, D. E., & Zelinsky, G. J. (2002). Eye movements and scene perception: Memory for things observed. *Perception & Psychophysics*, *64*(6), 882–895.

Itti, L., & Borji, A. (2015). Computational models: Bottom-up and top-down aspects. *arXiv:1510.07748 [cs]*.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259.

James, W. (1890). *The Principles of Psychology*. Henry Holt; Company.

Jarodzka, H., Holmqvist, K., & Nyström, M. (2010). A vector-based, multidimensional scanpath similarity measure. *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications - ETRA '10*, 211.

Judd, T., Durand, F., & Torralba, A. (2012). A Benchmark of Computational Models of Saliency to Predict Human Fixations. *CSAIL Technical Reports*.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. *2009 IEEE 12th International Conference on Computer Vision*, 2106–2113.

Karpov, B. A., Luria, A. R., & Yarbuss, A. L. (1968). Disturbances of the structure of active perception in lesions of the posterior and anterior regions of the brain. *Neuropsychologia*, *6*(2), 157–166.

Keeling, E., Lotery, A. J., Tumbarello, D. A., & Ratnayaka, J. A. (2018). Impaired Cargo Clearance in the Retinal Pigment Epithelium (RPE) Underlies Irreversible Blinding Diseases. *Cells*, *7*(2), 16.

Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Sciences*, *4*(4), 138–147.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, *4*, 219–227.

Kong, P., Mancas, M., Thuon, N., Kheang, S., & Gosselin, B. (2018). Do Deep-Learning Saliency Models Really Model Saliency? *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2331–2335.

Krejtz, K., Duchowski, A., Krejtz, I., Szarkowska, A., & Kopacz, A. (2016). Discerning Ambient/Focal Attention with Coefficient K. *ACM Transactions on Applied Perception*, *13*(3), 1–20.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90.

Kummerer, M. (2019). PySaliency [original-date: 2015-11-25T23:08:26Z]. Retrieved September 21, 2020, from https://github.com/matthias-k/pysaliency

Kummerer, M., Wallis, T. S., Gatys, L. A., & Bethge, M. (2017). Understanding Low- and High-Level Contributions to Fixation Prediction. *2017 IEEE International Conference on Computer Vision (ICCV)*, 4799–4808.

Kümmerer, M., Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., & Torralba, A. (2020). MIT/Tübingen Saliency Benchmark. https://saliency.tuebingen.ai/

Le Meur, O., Le Callet, P., Barba, D., & Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(5), 802–817.

Le Meur, O., & Coutrot, A. (2016). Introducing context-dependent and spatially-variant viewing biases in saccadic models. *Vision Research*, *121*, 72–84.

Le Meur, O., Coutrot, A., Liu, Z., Roch, A. L., Helo, A., & Rama, P. (2017). Computational Model for Predicting Visual Fixations from Childhood to Adulthood. *arXiv:1702.04657 [cs]*.

Le Meur, O., & Liu, Z. (2015). Saccadic model of eye movements for free-viewing condition. *Vision Research*, *116*, 152–164.

Leigh, J. R., & Zee, D. S. (2006). *The Neurology of Eye Movements*. Oxford University Press.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, *10*, 707.

Liu, C., Liu, J., & Wei, Y. (2017). Scroll up or down?: Using Wheel Activity as an Indicator of Browsing Strategy across Different Contextual Factors. *Proceedings of the 2017 Conference on Human Information Interaction and Retrieval - CHIIR '17*, 333–336.

Liu, H., Xu, D., Huang, Q., Li, W., Xu, M., & Lin, S. (2013). Semantically-Based Human Scanpath Estimation with HMMs. *2013 IEEE International Conference on Computer Vision*, 3232–3239.

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception & Psychophysics*, *2*(11), 547–552.

Madelain, L., & Chauvin, A. (2007). Saccades et attention spatiale. *Neuroscience cognitive de l'attention visuelle* (G. A. Micheal, pp. 203–227). SOLAL Editeur.

Mannaru, P., Balasingam, B., Pattipati, K., Sibley, C., & Coyne, J. (2016). On the use of hidden Markov models for gaze pattern modeling. In B. D. Broome, T. P. Hanratty, D. L. Hall, & J. Llinas (Eds.), *Next-Generation Analyst IV* (pp. 252–258). SPIE.

Maquestiaux, F. (2013). *Psychologie de l'attention*. De Boeck.

Martinez-Conde, S., Macknik, S. L., Troncoso, X. G., & Hubel, D. H. (2009). Microsaccades: a neurophysiological analysis. *Trends in Neurosciences*, *32*(9), 463–475.

Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L., & Dodd, M. D. (2011). Examining the influence of task set on eye movements and fixations. *Journal of Vision*, *11*(8), 17–17.

Milner, A. D. (2017). How do the two visual streams interact with each other? *Experimental Brain Research*, *235*(5), 1297–1308.

Morel, A., & Bullier, J. (1990). Anatomical segregation of two cortical visual pathways in the macaque monkey. *Visual Neuroscience*, *4*(6), 555–578.

Morrisey, M. N., Hofrichter, R., & Rutherford, M. D. (2019). Human faces capture attention and attract first saccades without longer fixation. *Visual Cognition*, *27*(2), 158–170.

Munoz, D. P., Broughton, J. R., Goldring, J. E., & Armstrong, I. T. (1998). Age-related performance of human subjects on saccadic eye movement tasks. *Experimental Brain Research*, *121*(4), 391–400.

Mustafi, D., Engel, A. H., & Palczewski, K. (2009). Structure of cone photoreceptors. *Progress in Retinal and Eye Research*, *28*, 289–302.

Navalpakkam, V., Jentzsch, L., Sayres, R., Ravi, S., Ahmed, A., & Smola, A. (2013). Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. *Proceedings of the 22nd international conference on World Wide Web - WWW '13*, 953–964.

Ngo, T., & Manjunath, B. S. (2017). Saccade gaze prediction using a recurrent neural network. *2017 IEEE International Conference on Image Processing (ICIP)*.

Nielsen, J. (2006). F-Shaped Pattern For Reading Web Content. Retrieved August 20, 2020, from https://www.nngroup.com/articles/f-shaped-pattern-reading-web-content-discovered/

Nielsen, J. (2010). Horizontal Attention Leans Left (Early Research). Retrieved August 20, 2020, from https://www.nngroup.com/articles/horizontal-attention-original-research/

Noton, D., & Stark, L. (1971a). Scanpaths in Eye Movements during Pattern Perception. *Science*, *171*(3968), 308–311.

Noton, D., & Stark, L. (1971b). Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision Research*, *11*(9), 929–942.

Osterberg, G. (1935). Topography of the layer of the rods and cones in the human retina. *Acta Ophthalmol*, *13*(6), 1–102.

Ouerhani, N., Wartburg, R. V., & Heinz, H. (2004). Empirical Validation of the Saliency-based Model of Visual Attention. *Electronic Letters on Computer Vision and Image Analysis*, *3*(1), 13–24.

Pannasch, S., Helmert, J. R., Roth, K., Herbold, A.-K., Walter, H., et al. (2008). Visual fixation durations and saccade amplitudes: Shifting relationship in a variety of conditions. *Journal of Eye Movement Research*, *2*(2), 1–19.

Parker, R. E. (1978). Picture processing during recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(2), 284–293.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*(1), 107–123.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, *45*(18), 2397–2416.

Piéron, H. (1994). Grand dictionnaire de la psychologie. Larousse.

Porter, J. D., Baker, R. S., Ragusa, R. J., & Brueckner, J. K. (1995). Extraocular muscles: Basic and clinical aspects of structure and function. *Survey of Ophthalmology*, *39*(6), 451–484.

Posner, M. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25.

Posner, M., & Cohen, Y. (1984). Components of visual orienting. *Attention and performance* (D. Bouwhuis & H. Bouma, pp. 531–556). Lawrence Erlbaum.

Posner, M., Nissen, M., & Ogden, W. (1978). Attended and unattended processing modes: The role of set for spatial location. *Modes of Perceiving and Processing Information*, *137*.

Rayner, K., Li, X., Williams, C. C., Cave, K. R., & Well, A. D. (2007). Eye movements during information processing tasks: Individual differences and cultural effects. *Vision Research*, *47*(21), 2714–2726.

Riche, N. (2015). *VISION: Video and Image Saliency Detection* (Doctoral dissertation). Université de Mons. Mons.

Riche, N., Mancas, M., Duvinage, M., Mibulumukini, M., Gosselin, B., & Dutoit, T. (2013). RARE2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication*, *28*(6), 642–658.

Rodden, K., & Fu, X. (2007). Exploring How Mouse Movements Relate to Eye Movements on Web Search Results Pages. *Proceedings of the Special Interest Group on Information Retrieval Workshop on Web Information Seeking and Interaction*, 29–32.

Rodden, K., Fu, X., Aula, A., & Spiro, I. (2008). Eye-mouse coordination patterns on web search results pages. *Proceeding of the 26th CHI conference extended abstracts on Human factors in computing systems - CHI '08*, 2997.

Rossetti, Y., Pisella, L., & McIntosh, R. D. (2017). Rise and fall of the two visual systems theory. *Annals of Physical and Rehabilitation Medicine*, *60*(3), 130–140.

Rothkegel, L. O. M., Trukenbrod, H. A., Schütt, H. H., Wichmann, F. A., & Engbert, R. (2017). Temporal evolution of the central fixation bias in scene viewing. *Journal of Vision*, *17*(13), 18.

Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision*, *77*(1), 157–173.

Scinto, L. F. M., Pillalamarri, R., & Karsh, R. (1986). Cognitive strategies for visual search. *Acta Psychologica*, *62*(3), 263–292.

Shao, X., B, Y. L., Zhu, D., Li, S., Itti, L., & Lu, J. (2017). Scanpath Prediction Based on High-Level Features and Memory Bias. *Proceeding of the 2017 International Conference on Neural Information Processing*, *1*, 3–13.

Sharafi, Z., Marchetto, A., Susi, A., Antoniol, G., & Guéhéneuc, Y.-G. (2013). An empirical study on the efficiency of graphical vs. textual representations in requirements comprehension. *2013 21st International Conference on Program Comprehension (ICPC)*, 33–42.

Shen, C., Huang, X., & Zhao, Q. (2015). Predicting Eye Fixations in Webpages with Multi-scale Features and High-level Representations from Deep Networks. *IEEE Transaction on Multimedia*, *17*(11), 2084–2093.

Shen, C., & Zhao, Q. (2014). Webpage Saliency. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014* (pp. 33–46). Springer International Publishing.

Shepherd, M., Findlay, J. M., & Hockey, R. J. (1986). The Relationship between Eye Movements and Spatial Attention: *The Quarterly Journal of Experimental Psychology*, *38A*, 475–491.

Shepherd, S. V., Steckenfinger, S. A., Hasson, U., & Ghazanfar, A. A. (2010). Human-monkey gaze correlations reveal convergent and divergent patterns of movie viewing. *Current biology: CB*, *20*(7), 649–656.

Simon, D., Sridharan, S., Sah, S., Ptucha, R., Kanan, C., & Bailey, R. (2016). Automatic scanpath generation with deep recurrent neural networks. *Proceedings of the ACM Symposium on Applied Perception - SAP '16*, 130–130.

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*.

Soh, Z., Sharafi, Z., Van den Plas, B., Porras, G. C., Guéhéneuc, Y.-G., & Antoniol, G. (2012). Professional status and expertise for UML class diagram comprehension: An empirical study. *2012 20th IEEE International Conference on Program Comprehension (ICPC)*, 163–172.

Stark, L., & Ellis, S. R. (1981). Scanpaths revisited Cognitive models, direct active looking. *Eye Movements: Cognition and Visual Perception* (D. F. Fisher, R. A. Monty, and I. W. Senders). Lawrence Erlbaum.

StatCounter. (2020). Desktop vs Mobile vs Tablet Market Share Worldwide. Retrieved October 9, 2020, from https://gs.statcounter.com/platform-market-share/desktop-mobile-tablet/worldwide/

Still, J. D., & Masciocchi, C. M. (2010). A saliency model predicts fixations in web interfaces. *Proceedings of the 5th international workshop on model-driven development of advanced user interactions*, *617*, 25–28.

Stone, J., Dreher, B., & Leventhal, A. (1979). Hierarchical and parallel mechanisms in the organization of visual cortex. *Brain Research Reviews*, *1*(3), 345–394.

Stone, K. D., & Gonzalez, C. L. R. (2015). The contributions of vision and haptics to reaching and grasping. *Frontiers in Psychology*, *6*.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14), 16.

Tatler, B. W., Baddeley, R. J., & Vincent, B. T. (2006). The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision Research*, *46*(12), 1857–1862.

Tatler, B. W., & Vincent, B. T. (2008). Systematic tendencies in scene viewing. *Journal of Eye Movement Research*, *2*(2), 1–18.

Tatler, B. W., & Vincent, B. T. (2009). The prominence of behavioural biases in eye guidance. *Visual Cognition*, *17*(6-7), 1029–1054.

Tavakoli, H. R., Rahtu, E., & Heikkilä, J. (2013). Stochastic bottom–up fixation prediction and saccade generation. *Image and Vision Computing*, *31*(9), 686–693.

Theeuwes, J., & Failing, M. (2020). *Attentional Selection: Top-Down, Bottom-Up and History-Based Biases*. Cambridge University Press.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Trevarthen, C. B. (1968). Two mechanisms of vision in primates. *Psychologische Forschung*, *31*(4), 299–337.

Unema, P. J. A., Pannasch, S., Joos, M., & Velichkovsky, B. M. (2005). Time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration. *Visual Cognition*, *12*(3), 473–494.

Ungerleider, L., & Mishkin, M. (1982). Two cortical visual systems. *Analysis of visual behavior* (Ingle D.J., Goodale M.A., Mansfield R.J.W., pp. 549–586). MIT Press.

Velichkovsky, B. M., Joos, M., Helmert, J. R., & Pannasch, S. (2005). Two Visual Systems and their Eye Movements: Evidence from Static and Dynamic Scene Perception. *Proceedings of the XXVVII conference of the cognitive science society*, 2283–2288.

von Wartburg, R., Wurtz, P., Pflugshaupt, T., Nyffeler, T., Lüthi, M., & Müri, R. M. (2007). Size Matters: Saccades during Scene Perception. *Perception*, *36*(3), 355–365.

Wang, W., Chen, C., Wang, Y., Jiang, T., Fang, F., & Yao, Y. (2011). Simulating human saccadic scanpaths on natural images. *CVPR 2011*, 441–448.

Wang, Y., Wang, B., Wu, X., & Zhang, L. (2016). Scanpath estimation based on foveated image saliency. *Cognitive Processing*, *18*(1), 87–95.

Weiskrantz, L. (1972). Behavioural analysis of the monkey's visual nervous system. *Proceedings of the Royal Society of London. Series B, Biological sciences*, *182*(1069), 427–455.

Wloka, C., Kotseruba, I., & Tsotsos, J. K. (2018). Active Fixation Control to Predict Saccade Sequences. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wloka, C., Kunić, T., Kotseruba, I., Fahimi, R., Frosst, N., Bruce, N. D. B., & Tsotsos, J. K. (2018). SMILER: Saliency Model Implementation Library for Experimental Research. *arXiv:1812.08848 [cs]*.

Wolfe, J. M. (1994). Guided Search 2.0 A revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202–238.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, *5*(6), 495–501.

Xia, C., Han, J., Qi, F., & Shi, G. (2019). Predicting Human Saccadic Scanpaths Based on Iterative Representation Learning. *IEEE Transactions on Image Processing*, *28*(7), 3502–3515.

Xia, C., & Quan, R. (2020). Predicting Saccadic Eye Movements in Free Viewing of Webpages. *IEEE Access*, *8*, 15598–15610.

Xu, J., Jiang, M., Wang, S., Kankanhalli, M. S., & Zhao, Q. (2014). Predicting human gaze beyond pixels. *Journal of Vision*, *14*(1), 28–28.

Yarbus, A. L. (1967). *Eye movements and vision*. Plenum Press.

Zanca, D., Serchi, V., Piu, P., Rosini, F., & Rufa, A. (2018). FixaTons: A collection of Human Fixations Datasets and Metrics for Scanpath Similarity. *arXiv:1802.02534 [cs]*.

Zhang, X., Zhaoping, L., Zhou, T., & Fang, F. (2012). Neural Activities in V1 Create a Bottom-Up Saliency Map. *Neuron*, *73*(1), 183–192.

Zhang, Y., Fu, H., Liang, Z., Chi, Z., Zhao, X., Feng, D., & Zhao, X. (2009). Eye movement data modeling using a genetic algorithm. *2009 IEEE Congress on Evolutionary Computation, CEC 2009*, 1038–1044.

Zhang, Y., Zhao, X. Y., Fu, H., Liang, Z., Chi, Z. R., Zhao, X. B., & Feng, D. G. (2011). A Time Delay Neural Network model for simulating eye gaze data. *Journal of Experimental & Theoretical Artificial Intelligence*, *23*(1), 111–126.