SCANPNT : UNE MÉTHODE DE PRÉDICTION DU CHEMIN DE PARCOURS OCULAIRE PAR APPRENTISSAGE PROFOND

Mohamed Amine KERKOURI[1], Marouane TLIBA[1], Aladine CHETOUANI[1], Alessandro BRUNO[2]

[1]Laboratoire PRISME, 45000, Orléans,France

[2]Bournemouth university, Bournemouth, United Kingdom

mohamed-amine.kerkouri, marouane.tliba, aladine.chetouani@univ-orleans.fr
abruno@bournmouth.co.uk

Résumé – Dans cet article, nous proposons une nouvelle approche pour prédire des parcours oculaires de longueur variable dans les images en utilisant une architecture légère. L'architecture est entièrement convolutive et utilise des priors apprenables appartenant au domaine de tâche. Cela pour affiner les caractéristiques des stimuli et introduire des biais spécifiques du domaine. Elle incorpore également des branches de prédiction parallèles pour prédire les coordonnées séparément pour chaque dimension en basant sur des parties du modèle 3D PointNet. Nous corrigeons également le problème introduit par la propriété d'invariance de permutation de PointNet en ajoutant un module d'encodage positionnel pour assurer l'impact de l'ordre des caractéristiques. Les résultats obtenus sur un jeu de données sont compétitifs par rapport aux modèles de l'état de l'art. Projet TIC-ART financé par la Région Centre-Val de Loire).

Abstract – In this paper, we propose a new approach for predicting variable-length scanpath in images using a lightweight architecture. The architecture is fully convolutional and employs learnable task domain priors to refine the stimuli features and introduces domain-specific biases. It also incorporates parallel prediction branches to predict the coordinates separately for each dimension while basing their architecture on parts of the PointNet 3D Model, which excels at learning geometrical representations. We also fix the problem introduced by the permutation invariance property of PointNet by adding a positional encoding module to ensure the impact of the order of the features. The results obtained on well-known dataset are competitive with state-of-the-art models.

1 Introduction

Le mécanisme visuel, appelé "attention visuelle", oblige l'observateur à ne prêter attention qu'à certaines régions spécifiques de la scène. Ce phénomène se produit lors des mouvements oculaires saccadés qui représentent le déplacement du regard d'une région à une autre pour un stimulus visuel. Lorsque les mouvements oculaires se concentrent sur une région, le regard se fixe sur des points spécifiques, à savoir les "points de fixation". Ces derniers peuvent être collectés à l'aide d'eye-trackers permettant la projection des points de fixation de plusieurs observateurs sur une carte binaire appelée "carte de fixation". Par-dessus, une "carte de saillance" est généralement obtenue à l'aide de filtres de lissage pour donner une distribution spatiale des points de fixation sur les stimuli visuels. Les cartes de saillance sont généralement représentées sous forme de cartes de chaleur normalisées où chaque valeur de pixel représente la probabilité que le pixel attire l'attention des observateurs.

Le mécanisme décrit ci-dessus confère au système visuel humain une efficacité exceptionnelle. La prédiction de la saillance permet d'améliorer de nombreuses applications de vision par ordinateur comme la prédiction de qualité des images [1], la compression d'image [2], la recherche et la récupération d'images [3], l'amélioration de l'image pour les personnes atteintes de DVC (déficience de la vision des couleurs) [4].

Depuis la publication des travaux fondamentaux de Koch [5] et sa mise en œuvre dans [6] avec un modèle multi-échelle reposant sur l'extraction de caractéristiques de bas niveau (c'est-à-dire la couleur, l'intensité et l'orientation), de nombreux articles scientifiques ont été publiés pour la prédiction de la saillance comme [7], [8], [9] et [10]. Dans [11], les auteurs ont généré des scanpaths à partir d'une carte de saillance et les caractéristiques statistiques dérivées de plusieurs jeux de données. Dans la [12], la carte de saillance a été modélisée comme un champ de gravité où la masse du regard se déplace selon une loi physique.

A l'avènement de l'apprentissage profond, les taux de précision de plusieurs tâches de vision par ordinateur ont augmentés. Cela a permis d'améliorer considérablement la prédiction de la saillance et des chemins oculaires. Dans [13], des couches LSTM et un modèle VGG ont été utilisés avec un apprentissage contradictoire. Une carte de saillance fovéale a été utilisée avec des cartes d'inhibition du retour pour prédire les trajets de scanpath dans le modèle [14]. Les auteurs de [15] ont présenté un modèle de bout en bout permettant de prédire simultanément le trajet oculaire et la carte de saillance d'une image.

Les principales contributions de ce document sont résumées ci-dessous :

- Proposer un nouveau modèle d'apprentissage profond efficace pour la prédiction de scanpaths.
- Améliorer les capacités de prédiction du modèle en introduisant des priors adaptatifs au domaine de la tâche liés au biais des données de la tâche.
- Utilisez une architecture basée sur PointNet pour explorer sa capacité à apprendre les formes géométriques du chemin de oculaire.

Le reste de l'article est structuré comme suit : Dans la section 2, nous décrivons en détail la méthode proposée. Dans la section 3, nous présentons le protocole expérimental ainsi que les résultats quantitatifs et qualitatifs obtenus. La section 4 est consacrée aux conclusions.

2 MÉTHODE PROPOSÉE

Dans ce travail, nous proposons un modèle de réseau neuronal profond pour la prédiction de scanpath à partir d'images. Le modèle utilise des composants légers pour prédire les scanpaths avec des longueurs variables. La Fig. 1 illustre l'architecture globale du modèle. Tout d'abord, un réseau MobileNet pré-entraîné est utilisé comme extracteur de caractéristiques léger, codant l'entrée dans un espace de représentation différent. Ensuite, des cartes de biais apprises sont concaténées à la sortie de l'extracteur de caractéristiques, représentant le biais lié à l'attention visuelle. Les cartes de caractéristiques obtenues sont ensuite transmises à une couche convolutive de fusion qui combine les distributions de deux caractéristiques concaténées. Chacune des cartes de caractéristiques 2D résultantes est vectorisée et fusionnée avec un vecteur d'encodage positionnel. Les caractéristiques codées résultantes sont finalement passées par deux branches pour prédire les coordonnées verticales et horizontales de l'extracteur de caractéristiques.

2.1 Extracteur de caractéristiques

.

MobileNet [16] est un modèle qui a été conçu pour être déployé sur des plateformes mobiles et embarquées. Il est considéré comme léger car il a introduit l'utilisation de convolutions séparables par point et par profondeur. De ce fait, le nombre de paramètres nécessaires à la convolution est réduits. Ce modèle est ici employé pour l'extraction de caractéristiques. Il peut trouver des informations sémantiquement pertinentes pour notre tâche sans avoir besoin d'un long apprentissage.

2.2 Priors gaussiennes de domaine apprenable

Plusieurs études et résultats de recherche [17] montrent que les spectateurs humains ont tendance à concentrer leur attention sur les zones centrales des images. D'un point de vue mathématique, le " biais central " [18] suit généralement une distribution spatiale gaussienne dont la position moyenne est le centre de l'image ($\mu_{xy} = (image_w/2, image_h/2)$), tandis que son écart-type (σ) dépend de l'ensemble de données. Nous pouvons généraliser ce phénomène en représentant tous les biais de l'ensemble de données comme un ensemble de distributions gaussiennes de biais avec des moyennes et des écarts types différents Eq.2. Chaque distribution exprimée par l'équation 1 et représentée par une carte de chaleur en 2D, appelée "carte des priorités" :

$$G(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp^{-((\frac{x-\mu_x}{2\sigma_x})+(\frac{y-\mu_y}{2\sigma_y}))} \tag{1}$$

$$S = \{G_1(x,y), G_2(x,y), ..., G_16(x,y)\} \tag{2}$$

où S est l'ensemble des distributions gaussiennes, (x,y) représentent les coordonnées spatiales d'un point dans la carte. (μ_x, μ_y) et (σ_x, σ_y) sont les moyennes et les écart-types correspondants à la distribution, respectivement.

Ici, le modèle apprend à modéliser 16 distributions de cartes des priorités et les fusionne avec les caractéristiques extraites de l'encodeur. Ces caractéristiques incluent celles spécifiques à notre tâche. Il dispense l'encodeur d'apprendre ces biais et concentre ses performances sur les caractéristiques spécifiques aux stimuli.

2.3 Encodage positionnel

Comme mentionné ci-dessus, les deux branches de notre modèle sont inspirées de PointNet. Cette conception introduit un biais inductif pour les caractéristiques, et par extension, les coordonnées de fixation générées sont invariantes aux permutations. Cette propriété présente une erreur hypothétique car les scanpaths sont séquentiels par nature. Par conséquent, nous avons introduit un module de codage positionnel dans notre architecture pour résoudre ce problème et fusionner ces informations. Inspiré par le travail de [19], nous configurons le vecteur d'encodage comme suit :

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}}) \tag{3}$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}}) \tag{4}$$

où pos est la position, i est la dimension et d_{model} représente la dimensionnalité des vecteurs encodés.

Ces fonctions permettent d'encoder des positions normalisées sans écraser l'information contenue dans le vecteur sémantique.

Les vecteurs de position et sémantique sont combinés comme suit :

$$EV = PE + SV \tag{5}$$

Où PE est le vecteur contenant les valeurs d'encodage positionnel. SV est le vecteur sémantique résultant du processus d'aplatissement après la convolution de fusion. Tandis que EV représente le vecteur transmis aux deux branches.

2.4 Branches de prédiction de la trajectoire visuelle

Afin de prédire les points de fixation composant le chemin de balayage, nous avons employé 2 branches, chacune d'entre elles prédit la séquence de coordonnées sur une dimension du plan 2D. Cela permet de démêler la prédiction des coordonnées sans les dé-corréler complètement.

Inspirée de l'architecture PointNet [20], chaque branche est composée de deux blocs constitués d'une couche convolutionnelle 1D suivie d'une couche Max-pooling. La couche de convolution agrège les caractéristiques sur le même canal tandis que la couche de Max-Pooling regroupe les informations entre les canaux transmettant ainsi les valeurs d'activation les plus élevées d'un patch de canaux. Cette dernière permet d'incorporer les caractéristiques locales par le biais de la convolution avec celles globales provenant du Max-pooling.

2.5 Apprentissage et perte

Le modèle a été entraîné, validé et testé sur 9000, 1000 et 5000 images de l'ensemble de données Salicon [21], respectivement. Les coordonnées des scanpaths ont été normalisées avant l'entraînement en fonction des dimensions de l'image puis complétées pour une forme spécifique de 16 fixations car 97,34% des images ne dépassent pas ce nombre de points. Nous avons entraîné notre modèle en utilisant la fonction de perte suivante :

$$L(y,\hat{y}) = \frac{1}{N}\sum_{i=0}^{N}\sqrt{{y_x}^2 - \hat{{y_x}}^2} + \frac{1}{N}\sum_{i=0}^{N}\sqrt{{y_y}^2 - \hat{{y_y}}^2} \tag{6}$$

où N représente le nombre de points de fixation de la trajectoire de balayage, y_x et y_y sont les coordonnées de la trajectoire de balayage de la vérité de base, tandis que $\hat{y_x}$ et $\hat{y_y}$ sont les coordonnées de la trajectoire de balayage prédite, respectivement.

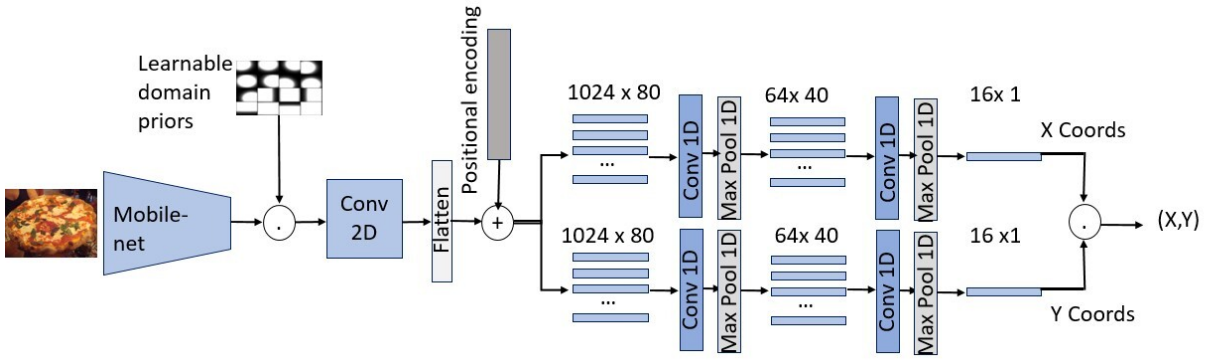

FIGURE 1 – **ScanPNT** : Scanpath PoiNTs.

Model	MM Shape	MM Dir	MM Len	MM Pos	MM Mean	NSS	Congruency
PathGan[13]	0.9608	0.5698	**0.9530**	0.8172	0.8252	-0.2904	0.0825
Le Meur[11]	0.9505	0.6231	0.9488	0.8605	0.8457	0.8780	0.4784
G-Eymol[12]	0.9338	0.6271	0.9521	**0.8967**	0,8524	0.8727	0.3449
SALYPATH [15]	**0.9659**	0.6275	0.9521	0.8965	**0,8605**	0.3472	0.4572
Our model	0.9552	**0.6466**	0.9509	0.8873	0.8600	**1.0062**	**0.5170**

TABLE 1 – Résultats de la prédiction de scanpath sur Salicon.

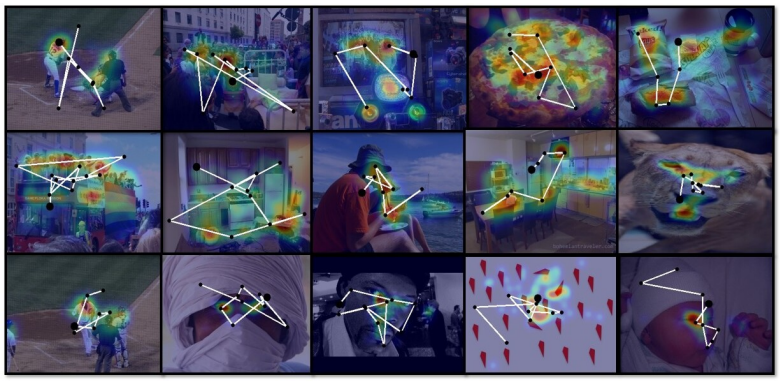

FIGURE 2 – Exemples de scanpaths prédis sur des images de Salicon.

3 Résultats expérimentaux

3.1 Jeux de données

Nous avons évalué notre méthode sur l'ensemble de données **Salicon** [21]. Il se compose de 9000 images pour la formation, 1000 images pour la validation et 5000 images pour le test avec les cartes de saillance et les données de points de fixation correspondantes. Dans notre travail, nous avons validé environ 250 000 scanpaths sur ce jeu de données. Le grand nombre de scanpaths permet de garantir la validité empirique des résultats.

-	Mean	Std	Min	Median	Max
Salicon GT	8.073658	1.516080	2.500000	8.209316	12.666667
Salicon Pred	9.051000	1.577117	5.0000	9.000000	15.0000

TABLE 2 – Descriptions statistiques sur les longueurs des scanpaths.

3.2 Protocole expérimental

Pour évaluer les performances de notre méthode, nous avons utilisé trois métriques : MultiMatch, NSS et Congruence. La métrique MultiMatch (MM) permet de comparer deux vecteurs de scanpaths et calcule la similarité à l'aide de cinq caractéristiques (Forme, Direction, Longueur, Position et Durée). Comme le modèle ne prédit que les coordonnées spatiales, nous n'utilisons que les quatre premières caractéristiques et mesurons la performance globale avec leur valeur moyenne. Les deux autres mesures comparent les scanpaths prédites avec une carte de saillance de l'image. Le NSS calcule la valeur de saillance moyenne des emplacements de fixation du scanpath. La $Congruence$ calcule le rapport de nombre des points de fixation prédits qui se trouvent dans les régions saillantes après seuillage de la carte de saillance.

3.3 Résultats

Les résultats de l'évaluation sur le jeu de données Salicon sont présentés dans le tableau. 1. Nous comparons également les résultats obtenus avec les méthodes de l'état de l'art. Les meilleurs résultats sont mis en évidence en gras. On peut constater que notre modèle a obtenu des résultats compétitifs pour de nombreuses métriques. Nous avons obtenu des résultats proches de Salypath et Pathgan sur la $MMShape$, ce qui indique que notre modèle apprend la

forme vectorielle des scanpaths de manière acceptable. Nous avons obtenu des résultats similaires sur la métrique $MMLength$ où seule la méthode de Le Meur dépasse la nôtre d'une marge négligeable. Nous avons également obtenu de bonnes performances sur la métrique $MMPosition$. Sur la métrique $MMDirection$, nous avons surpassé les méthodes considérées par une bonne marge, ceci concerne la capacité des branches inspirées de PointNet à apprendre des formes géométriques. Sur la métrique globale de $MMMean$, les résultats de notre modèle sont similaires à ceux du modèle Salypath. Sur les métriques hybrides de $congruence$ et NSS, notre modèle surpasse le modèle de Le Meur, ce qui révèle la capacité de notre modèle à discerner les régions saillantes dans les images.

Dans la Fig. 2, nous présentons les scanpaths prédits de différentes longueurs superposés aux cartes de saillance. Nous pouvons observer que pour les exemples donnés, les points de fixation des trajectoires de balayage capturent bien les emplacements saillants des images. Ils couvrent les caractéristiques distinctives des images représentant des visages et capturent les objets importants dans les images montrant plusieurs personnes ou objets. Nous vérifions également la prédiction correcte des distributions de longueur pour les scanpaths. Le tableau 2 montrent que les distributions de longueurs entre le $Salicon_{GT}$ et le $Salicon_{Pred}$ sont assez similaires.

4 Conclusion

Dans cet article, nous avons introduit une nouvelle architecture qui exploite le réseau MobileNet léger et améliore les caractéristiques prédites à l'aide des cartes de priorités gaussiennes apprises par rapport au biais du domaine de la tâche. Les caractéristiques ont été employées dans la prédiction des coordonnées des points de fixation en utilisant 2 branches qui utilisent une architecture similaire au réseau 3D PointNet. Nous avons comparé notre modèle avec des modèles de l'état de l'art et il a obtenu des résultats quantitatifs compétitifs sur plusieurs métriques. En même temps, il a obtenu des résultats qualitatifs exceptionnels par rapport à d'autres modèles et a vérifié la capacité de notre modèle à prédire correctement un nombre correct de fixations pour une image.

Références

[1] Aladine Chetouani and Leida Li, "On the use of a scanpath predictor and convolutional neural network for blind image quality assessment," *Signal Processing : Image Communication*, vol. 89, pp. 115963, 2020.

[2] Yash Patel, Srikar Appalaraju, and R Manmatha, "Saliency driven perceptual image compression," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 227–236.

[3] Haoxiang Wang, Zhihui Li, Yang Li, Brij B Gupta, and Chang Choi, "Visual saliency guided complex image retrieval," *Pattern Recognition Letters*, vol. 130, pp. 64–72, 2020.

[4] Alessandro Bruno, Francesco Gugliuzza, Edoardo Ardizzone, Calogero Carlo Giunta, and Roberto Pirrone, "Image content enhancement through salient regions segmentation for people with color vision deficiencies," *i-Perception*, vol. 10, no. 3, pp. 2041669519841073, 2019.

[5] Christof Koch and Shimon Ullman, "Shifts in selective visual attention : towards the underlying neural circuitry," in *Matters of intelligence*, pp. 115–141. Springer, 1987.

[6] Laurent Itti and Christof Koch, "Computational modelling of visual attention," *Nature reviews neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[7] Jonathan Harel, Christof Koch, and Pietro Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, 2007, pp. 545–552.

[8] Robert Peters and Laurent Itti, "The role of fourier phase information in predicting saliency," *Journal of Vision*, vol. 8, no. 6, pp. 879–879, 2008.

[9] Chenlei Guo, Qi Ma, and Liming Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[10] Neil Bruce and John Tsotsos, "Saliency based on information maximization," in *Advances in neural information processing systems*, 2006, pp. 155–162.

[11] Olivier Le Meur and Zhi Liu, "Saccadic model of eye movements for free-viewing condition," *Vision research*, vol. 116, pp. 152–164, 2015.

[12] Dario Zanca, Stefano Melacci, and Marco Gori, "Gravitational laws of focus of attention," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[13] Marc Assens, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor, "Pathgan : visual scanpath prediction with generative adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

[14] Wentao Bao and Zhenzhong Chen, "Human scanpath prediction based on deep convolutional saccadic model," *Neurocomputing*, 2020.

[15] Mohamed A. Kerkouri, Marouane Tliba, Aladine Chetouani, and Rachid Harba, "Salypath : A deep-based architecture for visual attention prediction," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 1464–1468.

[16] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, "Mobilenets : Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv :1704.04861*, 2017.

[17] Po-He Tseng, Ran Carmi, Ian GM Cameron, Douglas P Munoz, and Laurent Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes," *Journal of vision*, vol. 9, no. 7, pp. 4–4, 2009.

[18] Sabira K Mannan, Keith H Ruddock, and David S Wooding, "The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images.," *Spatial vision*, 1996.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[20] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas, "Pointnet : Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[21] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao, "Salicon : Saliency in context.," in *CVPR*. 2015, pp. 1072–1080, IEEE Computer Society.