MOdèle de saillaNce Spatio-Temporel par Rareté sur hyperhistogrammE: MONSTRE

IOANNIS CASSAGNE[1], NICOLAS RICHE[1], MARC DECOMBAS[1], MATEI MANCAS[1]

[1] Université de Mons (UMONS) – Faculté Polytechnique (FPMs)
Mons, Belgique

[1]{Ioannis.Cassagne ; Marc.Decombas}@gmail.com, {Nicolas.Riche ; Matei.Mancas}@umons.ac.be

Résumé – Les modèles de saillance permettent de mettre en exergue les zones qui attirent le regard humain. La plupart de ces modèles ont été conçu pour des applications aux images, et cherche à s'étendre aux vidéos. Dans ce papier, nous proposons une méthode prenant nativement en compte les données spatio-temporelles. Pour cela les données statiques et dynamiques sont représentées grâce à des surfaces caractéristiques, les hyperhistogrammes.

Abstract - Saliency models highlight areas that attract the human eye. Most of these models have been designed for applications to static images, and seeks to extend to videos. In this paper, we propose a method natively taking into account the spatial and temporal data. For this, the static and dynamic data are represented with features surfaces, the hyperhistograms.

1 Introduction

Les modèles de saillance cherchent à prédire les lieux de l'attention humaine. Dans [1][3], l'attention humaine est définie comme le processus permettant de se concentrer sur certaines informations tout en écartant les autres. L'attention humaine fonctionne selon deux schémas différents, d'une part l'attention descendante, où le sujet est attiré par quelque chose qu'il cherche. D'autre part, l'attention ascendante, où le sujet répond aux stimuli les plus importants. Ces stimuli sont définis à l'aide des caractéristiques intrinsèques à l'image tandis que des connaissances apriori sont utilisées pour les modèles descendants. L'idée majeure derrière la majorité des nombreux modèles est sensiblement la même : identifier des caractéristiques inhabituelles dans un contexte donné en cherchant les informations rares, nouvelles, ou encore importantes. Ces différents modèles ont de nombreuses applications telles que la prédiction de regard [3], la compression par analyse de contenu [4], le recalage vidéo [5] ou encore le résumé vidéo [6].

Itti et al. [7] ont proposé un modèle statique basé sur trois éléments: la couleur, la luminance et l'orientation. Harel et al. [8] ont amélioré ce modèle en créant des cartes de caractéristiques à différentes échelles spatiales et proposé un modèle de saillance visuelle basé sur un graphique (GBVS). Cette approche s'appuie un graphique connecté à tous les pixels de chaque carte de caractéristiques. L'importance de chaque pixel est pondéré de manière inversement proportionnelle à la similitude par rapport aux autres valeurs et ainsi que de leur distance dans l'espace. Dans [9], Marat et al. , un modèle bio-inspiré est proposé, il décompose chaque trame d'une vidéo en trois cartes: 1) une carte de saillance statique illustrant les régions qui diffèrent de leur contexte, 2) une carte dynamique de saillance soulignant les régions mobiles et 3) une carte de saillance du visage accentuant les zones où des visages sont détectés. Enfin, toutes ces cartes sont fusionnées dans une carte principale de saillance. Le modèle de saillance de Rahtu et al. [10] possède l'avantage d'être multi-échelle, ne nécessite pas d'apprentissage. Il est calculé dans l'espace couleur perceptuel CIE Lab. Afin de prendre en compte le mouvement dans la scène, l'intensité de mouvement est ajoutée comme caractéristique d'entrée.

Basé sur [11] [12], un modèle Spatio-Temporel RARE (ST-RARE) a été proposé dans [13] et celui-ci intègre des caractéristiques dynamiques comme l'amplitude et la direction du mouvement. Un filtrage temporel est également utilisé pour apporter une robustesse temporelle.

Dans cet article, nous proposons un nouveau modèle de prévision de la rareté basé sur des histogrammes tridimensionnels, les hyperhistogrammes, MONSTRE. Les contributions de cet article sont: 1) une nouvelle méthode d'extraction des informations temporelles, 2) un nouveau processus pour sélectionner les caractéristiques importantes basées sur une surface de la rareté, 3) une carte de rehaussement finale en utilisant un algorithme SLIC [20], une gaussienne centrée et ainsi qu'un traceur. Ces contributions conduisent à un meilleur modèle du point de vue des points de fixation des yeux et ainsi que dans la prédiction des objets importants dans une scène. Le modèle MONSTRE est donc plus stable temporellement et plus orienté objet.

Le document est structuré de la manière suivante. Dans Sec. 2, MONSTRE est décrit en détail. Sec. 3 fournit une évaluation du modèle proposé sur une grande variété de vidéos comparée à la vérité terrain des données des fixations des regards et des objets manuellement segmentés. Enfin, Sec. 4 présente une discussion et conclut.

2 MONSTRE

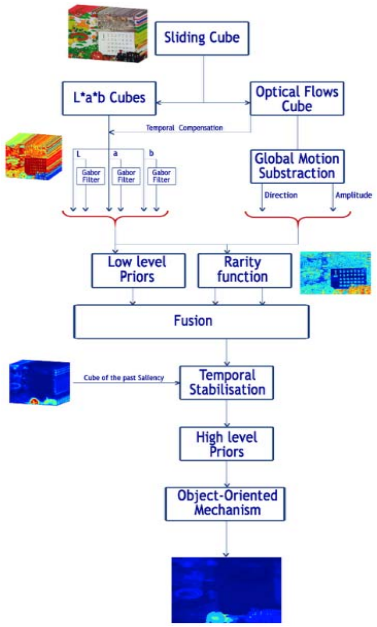

Figure 1 : Vue d'ensemble de MONSTRE. De haut en bas: Caractéristiques extraites sur une fenêtre glissante, A priori de bas niveau et mécanisme de rareté, Etape de fusion, Post-traitement

Le Figure 1 représente le schéma global de MONSTRE. Un "cube glissant " (2D + t) est utilisé pour extraire les caractéristiques spatiales et temporelles. Après prétraitement de ces caractéristiques, les éléments rares de la vidéo sont extraits en se basant sur des hyperhistogrammes. Ces hyperhistogrammes sont une concaténation dans le temps de tous les histogrammes des caractéristiques extraits pour chaque image spatiale composant le cube. Des informations a priori de bas niveau, tel que le comportement spécifique vis-à-vis des couleurs, sont également ajoutées. Une fusion des différentes cartes est réalisée puis stabilisée dans le temps. Enfin des aprioris de haut niveau (comme la détection de visage) et un algorithme de superpixel sont appliqués au modèle pour fournir une approche basée sur les objets. Dans les paragraphes suivants, chacune des étapes de l'algorithme est détaillée.

2.1 Extraction des caractéristiques

On construit un cube vidéo (x, y, t) grâce à une fenêtre glissante temporelle afin d'avoir les informations statiques et dynamiques de la trame courante, mais aussi des précédentes.

Six caractéristiques statiques sont extraites de cette vidéo; Trois cubes de couleur (un de luminance et deux de chrominance) sont définis dans l'espace de couleur CIE Lab. Enfin huit plans d'orientation sont obtenus par un filtrage de Gabor puis combinés ensemble à trois échelles différentes permettant d'avoir 3 cubes de textures avec trois échelles différentes.

Le flux optique, défini en [14] calculé sur la composante de luminance est utilisé pour créer deux cubes de caractéristiques dynamiques (l'un pour l'amplitude du mouvement et l'autre pour la direction du mouvement). Pour compenser le mouvement de la caméra, une soustraction du mouvement global est réalisée.

2.2 A priori et rareté multi-échelle

Deux cartes basées sur des aprioris de bas niveau sont calculées à partir de [15]: 1) la première est liée à la fréquence. En effet, le comportement humain peut être modélisé par un filtrage passe-bande. 2) le second est sur les couleurs. Certaines études [15] montrent que les couleurs chaudes, comme le rouge et le jaune, sont plus attirantes dans le système visuel humain que les couleurs froides.

Un mécanisme de rareté est ensuite appliqué sur chaque carte de de caractéristiques (spatiale et temporelle). L'idée principale vient de [12] et est basée sur le fait qu'un stimulus n'est pas nécessairement saillant seul, mais seulement dans un contexte spécifique. L'idée est ici étendue à un contexte spatio-temporel en lieu et place d'un contexte seulement spatial. En effet, l'histogramme monodimensionnel utilisé devient un hyperhistogramme qui est une surface 2D (figure 2, l'image b).

Le mécanisme de rareté est illustré sur la figure 2 à la composante de luminance en trois étapes: a) une décomposition de pyramide gaussienne fournit des fonctionnalités des cartes cube à différentes échelles, b) pour chaque cube, une surface d'histogramme (un hyperhistogramme) est traitée, c) La rareté est calculée sur l'ensemble du hyperhistogramme, mais seule la trame courante est extraite.

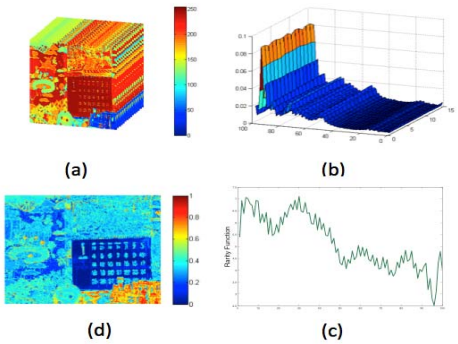

Figure 2 : Illustration du mécanisme de Rareté

2.3 Fusion

Le processus de fusion a deux étapes principales: 1) les caractéristiques spatiales sont combinées avec les aprioris de bas niveau avec une fusion maximale. La valeur maximale entre deux cartes est prise pour chaque pixel. 2) Ces cartes spatiales de rareté sont ensuite combinées avec les cartes temporelles. Les cartes qui ont des pics importants par rapport à leur moyenne ont un poids plus élevé. Une carte de saillance unique est finalement obtenue.

2.4 Post-traitement et amélioration

La carte de saillance obtenue dans la section précédente est encore améliorée en utilisant trois techniques différentes. Tout d'abord, un post-traitement effectue la stabilisation temporelle en utilisant un bref historique des cartes de saillance des images précédentes. Deuxièmement, les aprioris de haut niveau sont ajoutés. Des études antérieures [20] ont montré que l'information saillante est principalement située au centre de l'image pour des images naturelles. Pour modéliser cet apriori, une gaussienne centrée est appliquée à la carte.

Enfin, un algorithme SLIC [20] est utilisé avec DBSCAN [21] pour extraire des superpixels dans l'image. Ces superpixels sont des groupes de pixels ayant des niveaux de couleurs similaires. Ils fournissent des informations sur la forme des objets dans l'image. La carte de saillance est moyennée pour chaque superpixel. De cette façon, la carte finale possèdera une approche orientée objet.

3 Evaluation des résultats

3.1 Base de données et métriques

La base de données STRAP vidéo, basée sur [15] puis étendue dans [17], comprend 12 vidéos avec les données de suivi de l'œil et les masques binaires segmentés manuellement. La figure 3 montre une séquence vidéo extraite de la base de données avec la trame d'origine sur la colonne de gauche, le masque binaire sur la colonne du milieu et une carte de chaleur des données de suivi de l'œil.

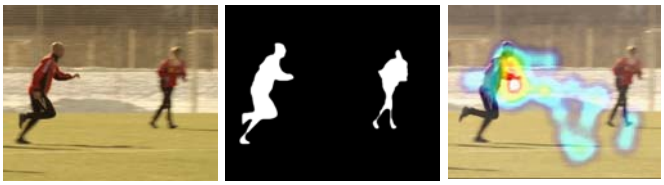

Figure 3: Base de données

Pour comparer les résultats de MONSTRE avec d'autres modèles vidéo de saillance, trois mesures différentes sont utilisées. Basé sur les données de suivi de l'œil, l'aire sous la courbe ROC (AUROC) [18] se concentre sur l'emplacement de saillance à des positions du regard. Le NSS [19] met l'accent sur les valeurs de saillance à des positions du regard. Pour AUROC et NSS, des scores élevés indiquent une meilleure performance quant aux données de suivi de l'œil. Pour la comparaison aux masques binaires, la F-Measure est utilisée. Cette mesure est basée sur de vrais positifs (TP), vrais négatifs (TN), les faux positifs (FP) et de faux négatifs (FN) qui comparent les résultats prévus avec les résultats de référence. Il est défini comme une combinaison de la précision et le rappel où la précision est le nombre de points pertinents par rapport au nombre total de points trouvé et le rappel est le nombre de points pertinents par rapport au nombre total de points importants dans la référence. Cette mesure montre la capacité de la méthode à prédire l'objet saillant et non seulement le regard des yeux.

3.2 Base de données et métriques

Pour valider MONSTRE, des expérimentations qualitatives et quantitatives ont été faites. La figure 4 montre les résultats qualitatifs de trois modèles différents à travers des cartes de chaleur. Le bleu illustre les zones qui ne sont pas importantes tandis que le rouge, les régions d'intérêt. Sur la première colonne de la figure 4, nous pouvons voir que notre approche définit ainsi les objets saillants. Pour STRARE (colonne du milieu), les objets saillants sont bien identifiés pour la séquence Football, mais également une partie de l'arrière-plan qui n'est pas ce qu'elle devrait être. Pour comparer les cartes thermiques avec la vérité terrain, il faut se référer à la figure 3.

Figure 4 : Cartes de chaleur résultats. De gauche à droite : HYPERAPTOR- STRARE - GBVS.

Pour la validation quantitative à la figure 5, les trois mesures décrites précédemment sont utilisées pour comparer quatre algorithmes de l'état de l'art et une gaussienne centrée. On voit qu'avec la métrique AUROC, notre modèle est le troisième. Cela est dû au fait que cette mesure est fortement influencée par une gaussienne centrée, qui peut être naturellement trouvées dans le modèle gaussien, mais aussi dans GBVS.
La métrique NSS est complémentaire à AUROC. Il peut être vu qu'à la suite de cette mesure, notre approche est statistiquement meilleure que l'état de l'art.
Lorsque nous comparons MONSTRE avec les autres modèles sur les masques binaires à l'aide de la F-mesure, MONSTRE surpasse statistiquement toutes les autres méthodes.
La figure 5 montre que MONSTRE est toujours mieux que les autres méthodes sur deux des mesures (NSS et F-mesure), la métrique AUROC étant très sensible vis-à-vis de la gaussienne centrée.

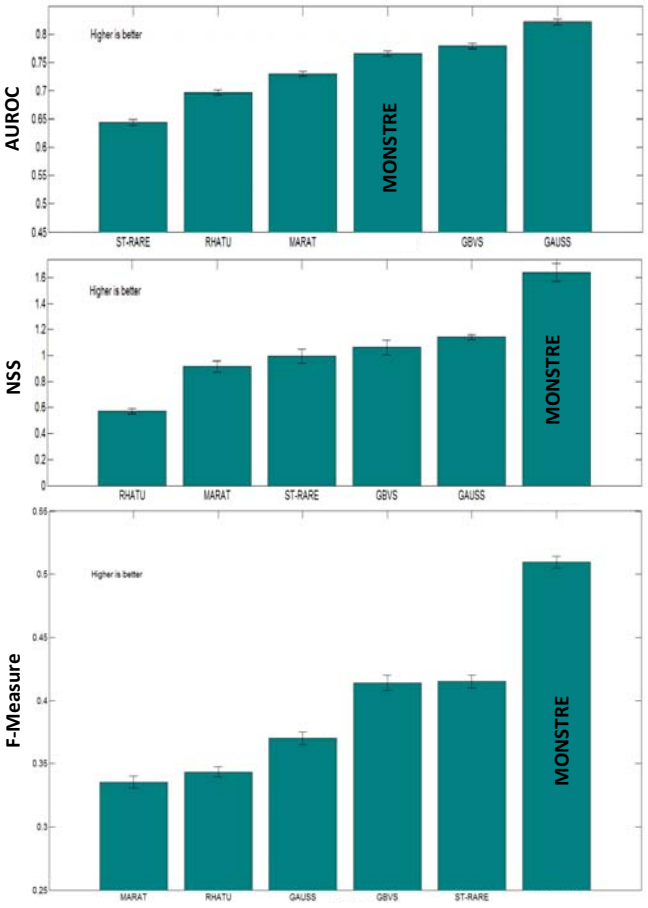

Figure 5 : Comparaison de MONSTRE

4 Conclusion

Dans cet article, nous proposons une nouvelle approche de saillance vidéo qui utilise des surfaces d'histogramme

lors du calcul de la rareté. Les deux caractéristiques statiques et dynamiques sont prises en compte. Des aprioris de bas et de haut niveau sont ajoutés ainsi qu'une pré-segmentation basée sur des superpixels. Cette nouvelle approche est évaluée sur base de données de 12 vidéos avec trois métriques différentes. Les vidéos sont très différents en termes de contenu et de mouvement (entre autres : fond encombré, déplacement de fond, passage devant la caméra, etc ...). La base de données est basée sur une œuvre existante qui est complétée avec l'ajout de masques binaires segmentés manuellement.
Si l'on excepte la métrique d'AUROC où MONSTRE est 3ème, sur NSS et F-mesure, notre modèle est bien meilleur que les approches concurrentes. L'utilisation des superpixels, fait que MONSTRE est également approprié pour la détection des objets saillant et non plus seulement pour la prédiction du regard.

5 Référence

[1] C. Koch, S. Ullman, "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry," Human Neurobiology, vol. 4, pp. 219-227, 1985.

[2] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y.H. Lai, N. Davis, F. Nuflo, "Modelling Visual Attention via Selective Tuning," Artificial Intelligence, vol. 78, no. 1-2, pp. 507-545, Oct. 1995.

[3] S. Lu, J.H. Lim "Saliency Modeling from Image Histograms", European Conference on Computer Vision (ECCV), pp. 321-332, Florence, Italy, 2012:

[4] M. Décombas, F. Dufaux, E. Renan, B. Pesquet-Popescu, F. Capman, "Improved seam carving for semantic video coding,", in Proc. IEEE International Workshop on Multimedia Signal Processing (MMSP 2012), Banff, Canada, Sept. 2012

[5] M. Rubinstein, A. Shamir, S. Avidan "Improved seam carving for video retargeting," ACM Trans. Graphics, vol. 27, no. 3 , pp. 1-16, 2008

[6] Z. Li, P. Ishwar, J. Konrad, "Video condensation by ribbon carving," IEEE Trans. Image Processing, vol. 18, no. 11, pp. 2572-2583, Nov. 2009

[7] L. Itti, C. Koch, E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. on pattern analysis and machine intelligence, vol. 20, no. 11, pp.1254-1259, 1998

[8] J. Harel, C. Koch, P. Perona, "Graph-based visual saliency," In Advances in neural information processing system, pp. 545-552, 2006

[9] S. Marat, T. H. Phuoc, L. Granjon, N. Guyader, D. Pellerin, A. Guérin-Dugué , "Modeling spatio-temporal saliency to predict gaze direction for short videos," International journal of computer vision, vol. 82, no. 3, pp. 231-243, 2009

[10] E. Rahtu, J. Kannala, M. Salo and J. Heikkilä, "Segmenting Salient Objects from Images and Videos", European Conference on Computer Vision (ECCV), Heraklion, Greece, Sept. 2010

[11] M. Mancas, N. Riche, J. Leroy, B. Gosselin, " Abnormal motion selection in crowds using bottom-up saliency," in Proc. IEEE International Conference on Image Processing (ICIP), Bruxelles, Belgium, 2011

[12] N. Riche, M. Mancas, B. Gosselin, T. Dutoit, " Rare: A new bottom-up saliency model" in Proc. IEEE International Conference on Image Processing (ICIP), Orlando, FL, Oct. 2012

[13] M. Décombas, N. Riche, F. Dufaux, B. Pesquet-Popescu, M. Mancas, B. Gosselin, T. Dutoit, Spatio-temporal saliency based on rare model, IEEE International Conference on Image Processing (ICIP), 2013.

[14] A. Chambolle, T. Pock, "A first-order primal-dual algorithm for convex problems with application to imaging," Technical Report, 2010

[15] Zhang, L., Gu, Z., & Li, H. (2013, September). SDSP: A novel saliency detection method by combining simple priors. In ICIP (pp. 171-175).

[16] H. Hadizadeh, M. J. Enriquez, and I. V. Bajić, "Eye-tracking database for a set of standard video sequences," IEEE Trans. Image Processing, vol. 21, no. 2, pp. 898-903, Feb. 2012

[17] M. Mancas, N. Riche, M. Décombas, Computational attention website http://tcts.fpms.ac.be/attention

[18] B. Lau, B.: Evaluation measures for saliency maps: AUROC,http://www.subcortex.net/research/code/area_under_roc_curve

[19] A. Borji, A.: Evaluation measures for saliency maps: CC_and_NSS,https://sites.google.com/site/saliencyevaluation/evaluation-measures

[20] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Susstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 34(11), 2274-2282.

[21] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, No. 34, pp. 226-231).

[22] V. Rijsbergen, C. Keith Joost, Information retrieval 1979, Butterworths, London