Analyse de signaux sociaux non verbaux de vidéos d'entretiens en face à face *

S. Bovée[1,2] O. Lézoray [1] P. Hamel[2]
[1] Université de Caen, GREYC, CNRS UMR 6072, Caen, France
[2] Zero To One Technology, Campus Effiscience, Colombelles, France

1 Introduction

La communication non verbale est l'un des phénomènes les plus répandus de notre vie quotidienne. En effet, les êtres humains montrent constamment des signaux non verbaux comportementaux tels que les expressions faciales, le regard, les postures, etc. Si cela joue un rôle si important dans notre vie de tous les jours, il semble alors naturel de vouloir capturer, par une analyse faite par ordinateur, la signification de nos signaux sociaux non verbaux [6]. Dans cet article nous nous intéressons à l'analyse de tels signaux dans des entretiens en face à face.

2 Vidéos d'entretiens en face à face

Les entretiens en face à face sont particulièrement utiles dans le cadre de formations. Par exemple, cela peut permettre à un manager commercial d'acquérir une photographie précise des compétences de ses vendeurs dans le cadre d'un rendez-vous client. Pour cela, des participants sont filmés dans des mises en situation construites sur la base de la vente d'un produit (un participant dans le rôle du professionnel, l'autre dans le rôle du client). Les données audiovisuelles recueillies peuvent être ensuite analysées afin d'amener les participants à prendre conscience de leur savoir-faire et à augmenter leurs performances. Plusieurs signaux sociaux sont essentiels à capturer et analyser. Certains sont liés à une communication verbale (le temps d'écoute par rapport au temps de parole, les coupures dans l'entretien, etc.) ou à une communication non verbale (les temps de regard, hochements de tête, mimiques du visage, etc.). Si certains outils de l'état de l'art sont bien établis pour la détection des visages et leur modélisation à partir de points anatomiques détectés [3], leur utilisation dans le cadre de la production de résumés automatiques de signaux sociaux dans des vidéos a été très peu abordée. Dans cet article, nous présentons les premiers résultats d'une approche pour la quantification des temps de regard.

3 Estimation des temps de regard

Les données dont nous disposons sont les suivantes : les deux personnes en entretien sont filmées simultanément (tête et buste) mais indépendamment par deux caméras. À partir de ces deux enregistrements vidéos, nous cherchons à estimer quand une personne regarde l'autre. L'estimation proposée des temps de regard dans des vidéos se décompose en plusieurs étapes que nous présentons ci-après.

3.1 Points caractéristiques du visage

Les vidéos étant prises dans des environnements non contrôlés, la première étape consiste à localiser précisément la personne dans la vidéo. Nous procédons donc tout d'abord à une reconnaissance du visage par la méthode de Viola et Jones [7] (carré rouge de la Figure 1). Dans cette zone du visage sont extraits ensuite des points caractéristiques du visage à l'aide de l'approche présentée dans [4]. Cette approche est particulièrement efficace et fonctionne de la manière suivante. Elle estime la position de $M = 7$ points caractéristiques du visage (commissures des lèvres et des yeux, nez) dans une image I de taille $h \times w$. Les points caractéristiques $(s_1, \ldots, s_M)$ ont une position $s_i \in \mathcal{S}_i \subset \{1, \ldots, h\} \times \{1, \ldots, w\}$ et la qualité d'une configuration $(s_1, \ldots, s_M) \in \mathcal{S}_1 \times \cdots \times \mathcal{S}_M$ est estimée par la fonction objectif

$$f(I, \mathcal{S}) = \sum_{i=1}^{M} q_i(I, s_i) + \sum_{(s_i, s_j) \in E} g_{ij}(s_i, s_j) \qquad (1)$$

Le terme $q_i(I, s_i)$ correspond à un modèle d'apparence estimant la correspondance entre les positions des points s_i et l'image I. Le terme $g_{ij}(s_i, s_j)$ correspond à une mesure de coût de déformation évaluant les positions de points voisins selon des contraintes exprimées sous la forme d'un graphe dirigé acyclique dont les noeuds sont les points s_i et les arêtes les contraintes. Les paramètres des deux termes sont appris par un SVM structuré et l'optimisation de (1) est effectuée par programmation dynamique. À partir de cette détection, nous effectuons plusieurs post-traitements. Premièrement, à partir des points caractéristiques des yeux, des rectangles englobants sont extraits en utilisant des apriori anatomiques. Deuxièmement, à partir des points caractéristiques de la bouche, deux points sont détectés pour les centres des deux lèvres en combinant informations couleur et gradient avec un apriori sur la structure anatomique des lèvres [1]. Troisièmement, chaque sourcil est extrait en appliquant un seuillage adaptatif sur une zone située au dessus de lœil et trois points le représentant en sont déduits. Enfin, pour augmenter la robustesse de l'extraction des points caractéristiques, une étape de suivi de ceux-ci est effectuée. La figure 1 (partie gauche) présente l'ensemble de ces informations extraites.

*Travaux effectués dans le cadre d'une convention CIFRE ANRT.

3.2 Détection du centre de l'iris

Afin de quantifier l'information non verbale liée au regard et à sa direction, il est nécessaire d'extraire précisément le centre des yeux dans la zone rectangulaire délimitant l'œil d'une personne. Pour cela, nous avons adapté la méthode proposée par [5]. Cette méthode utilise des isophotes qui sont les courbes connectant des points de même intensité. Ceci repose sur le fait que les yeux sont des motifs d'intensités radiales symétriques et l'on cherche à localiser les centres des isophotes circulaires. En coordonnées cartésiennes, la courbure locale des isophotes est donnée par :

$$\kappa = -\frac{L_y^2 L_{xx} - 2L_{xx}L_y + L_x^2 L_{yy}}{(L_x^2 + L_y^2)^{\frac{3}{2}}} \quad (2)$$

avec L_x la dérivée première de la luminance selon la direction en x. La distance au centre de l'iris est déterminée par l'inverse de cette quantité. L'orientation est obtenue par une mesure de gradient, mais comme celui-ci indique toujours le plus fort changement de luminance, il est multiplié par l'inverse de κ pour lever l'ambiguité sur la direction du centre. Afin de privilégier les zones ou l'isophote suit les contours de l'oeil, une pondération est appliquée par un indicateur de forts contours [2]. Un vecteur de déplacement $\{D_x, D_y\}$ est ensuite estimé en chaque pixel (x, y) comme la position estimée des centres par :

$$\{D_x, D_y\} = -\frac{\{L_x, L_y\}\sqrt{L_{xx}^2 + 2L_{xy}^2 + L_{yy}^2}\,(L_x^2 + L_y^2)}{L_y^2 L_{xx} - 2L_x L_{xy} L_y + L_x^2 L_{yy}} \quad (3)$$

Pour en déduire la position du centre de l'iris, les $\{D_x, D_y\}$ sont stockés dans un accumulateur $M(x, y)$. Le maximum de cet accumulateur donne alors le centre de l'iris. Cependant, en conditions réelles, cet accumulateur devient moins robuste et le plus fort maximum ne correspond pas forcément à l'iris. Nous modifions tout d'abord l'accumulateur $M(x, y)$ en le multipliant par une carte couleur $C(x, y) = (255 - L(x, y)) * \frac{Cb(x,y)}{Cr(x,y)}$ afin de ne chercher à localiser le centre de l'iris que dans une zone colorée. Enfin, plutôt que de chercher à n'extraire qu'un seul maximum, nous en extrayons 3. Ce nombre a été fixé relativement à une base de référence d'iris et il permet de garantir qu'un des maxima correspond à l'iris. De ces 3 maxima ne sont conservés que ceux significatifs (gradient local supérieur à un seuil), voir figure 1, partie droite.

3.3 État ouvert/fermé de l'œil

Si la précédente méthode donne de très bons résultats pour des yeux ouverts, comme elle est appliquée de manière automatique, elle détectera également trois maxima pour des yeux fermés. Il nous faut donc déterminer l'état ouvert ou fermé de l'œil. La zone rectangulaire d'un œil est subdivisée en n blocs et pour chaque bloc est calculé un histogramme des descripteurs LBP de la luminance des pixels. La concaténation des histogrammes des blocs fournit un vecteur $\mathbf{x}_i$ décrivant l'œil. Un apprentissage d'un classifieur SVM non linéaire est ensuite effectué à partir d'une base de référence d'yeux ouverts et fermés. Ce classifieur permet d'obtenir un taux de classification de 94% (évalué en Leave One Out Cross-Validation). Ceci, combiné avec la détection du centre de l'iris et un suivi temporel, permet de déterminer le regard de la personne. Nous combinons actuellement cette information avec la pose de la tête afin d'estimer la direction du regard qui, mise en relation avec celle de l'autre interlocuteur, permettra d'estimer les temps de regard.

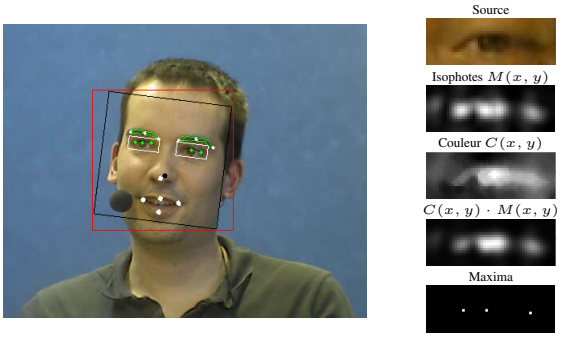

FIGURE 1 – Points caractéristiques du visage et de l'œil.

4 Conclusion

Nous avons présenté nos premiers résultats d'estimation des temps de regard dans des vidéos d'entretiens en face à face. Notre approche est automatique et repose sur l'extraction de points caractéristiques du visage et de l'iris à l'aide de descripteurs locaux et d'apprentissage. Les résultats préliminaires sont encourageants et devraient permettre de continuer vers l'extraction d'autres signaux sociaux non-verbaux exprimés par la bouche, les sourcils et de manière plus générale par les expressions faciales.

Références

[1] N. Eveno, A. Caplier, and P.-Y. Coulon. Key points based segmentation of lips. In *ICME (2)*, pages 125–128, 2002.

[2] Jan J Koenderink and Andrea J van Doorn. Surface shape and curvature scales. *Image and Vision Computing*, 10(8) :557 – 564, 1992.

[3] Albert Ali Salah and Theo Gevers, editors. *Computer Analysis of Human Behavior*. Springer, 2011.

[4] Michal Uřičář, Vojtěch Franc, and Václav Hlaváč. Detector of facial landmarks learned by the structured output SVM. In *VISAPP*, volume 1, pages 547–556, 2012.

[5] Roberto Valenti and Theo Gevers. Accurate eye center location through invariant isocentric patterns. *IEEE Trans. on PAMI*, 34(9) :1785–1798, 2012.

[6] A. Vinciarelli, H. Salamin, and M. Pantic. Social signal processing : Understanding social interactions through nonverbal behavior analysis. In *CVPR Workshops*, pages 42–49, 2009.

[7] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, volume 1, pages 511–518, 2001.