Théorie de l'évidence pour suivi de visage

Francis Faux, Franck Luthon

Théorie de l'évidence pour suivi de visage

Francis Faux, Franck Luthon

Université de Pau et des Pays de l'Adour, Laboratoire d'Informatique EA3000
IUT de Bayonne-Pays Basque, 2 allée du parc Montaury, F-64600 Anglet
Francis.Faux, Franck.Luthon@univ-pau.fr

RÉSUMÉ. *Le suivi de visage par caméra vidéo est abordé ici sous l'angle de la fusion évidentielle. La méthode proposée repose sur un apprentissage sommaire basé sur une initialisation supervisée. Le formalisme du modèle de croyances transférables est utilisé pour pallier l'incomplétude du modèle* a priori *de visage due au manque d'exhaustivité de la base d'apprentissage. L'algorithme se décompose en deux étapes. La phase de détection de visage synthétise un modèle évidentiel où les attributs du détecteur de Viola et Jones sont convertis en fonctions de croyance, et fusionnés avec des fonctions de masse couleur modélisant un détecteur de teinte chair, opérant dans un espace chromatique original obtenu par transformation logarithmique. Pour fusionner les sources couleur dépendantes, nous proposons un opérateur de compromis inspiré de la règle prudente de Denœux. Pour la phase de suivi, les probabilités pignistiques issues du modèle de visage garantissent la compatibilité entre les cadres crédibiliste et probabiliste. Elles alimentent un filtre particulaire classique qui permet le suivi du visage en temps réel. Nous analysons l'influence des paramètres du modèle évidentiel sur la qualité du suivi.*

ABSTRACT. *This paper deals with real time face detection and tracking by a video camera. The method is based on a simple and fast initializing stage for learning. The transferable belief model is used to deal with the prior model incompleteness due to the lack of exhaustiveness of the learning stage. The algorithm works in two steps. The detection phase synthesizes an evidential face model by merging basic beliefs elaborated from the Viola and Jones face detector and from colour mass functions. These functions are computed from information sources in a logarithmic colour space. To deal with the colour information dependence in the fusion process, we propose a compromise operator close to the Denœux cautious rule. As regards the tracking phase, the pignistic probabilities from the face model guarantee the compatibility between the believes and the probability formalism. They are the inputs of a particle filter which ensures face tracking at video rate. The optimal parameter tuning of the evidential model is discussed.*

MOTS-CLÉS : *détection de visage, espace couleur LUX, filtrage particulaire, Dempster-Shafer, modèle de croyances transférables, règle prudente de Denœux, reconnaissance des formes.*

KEYWORDS: *face detection, tracking, LUX colour space, particle filter, evidence theory, Dempster-Shafer, transferable belief model, cautious rule, computer vision, pattern recognition.*

DOI:10.3166/TS.28.515-545

Extended abstract

This paper presents an original method both for face detection based on an evidential face model and for face tracking with a classical bootstrap particle filter technique. The aim of the application is to automatically track the face of a person located in the field of view of a motorized pan-tilt-zoom camera. The proposed method takes control of the servo-camera to perform a dynamic centering of the face in the image plane during the whole video sequence. The application context is limited to indoor environments, typically a laboratory or an office.

Face detection by computer vision is made difficult by the variability of appearance of this deformable moving object due not only to the lighting variations on the face zone such as shadows or highlights, and to the background clutter that disturbs the detection, but also to individual morphological differences (nose shape, eye color, skin color, beard), to face expression changes, or to the presence of visual artifacts like glasses or occlusions. In this paper, to address the face complexity, a supervised method is proposed, where the user selects manually a zone of the face on the first image of the video sequence. This fast initializing step constitutes the learning stage which yields the prior model. The proposed method handles simple contextual knowledge representative of the application background during a quick initializing stage, contrary to current techniques that are based on huge learning databases and complex algorithms to get generic face models (*e.g.* active appearance models). The transferable belief model is used to counteract the incompleteness of the prior model due to the lack of exhaustiveness of the learning stage and to the subjectivity on the face appearance.

The algorithm works in two steps. The detection phase is a pre-processing step consisting in the fusion of information to get an evidential face model. It merges complementary basic beliefs (or mass functions) elaborated from the Viola and Jones (VJ) face detector and from a skin colour detector. The VJ detector generates a target container (or bounding box) very reliable when the face is in front-view or slightly from profile. However it fails in the case of important rotations or occlusions, or when it recognizes falsely a face-like artifact in the background. In order to model the VJ attribute by a belief function, we assign to each pixel a simple mass according to its position with respect to the bounding box and proportionally to a parameter of reliability γ, that may be tuned online from the data available. The colour mass functions are computed from information sources in a logarithmic colour space (Logarithmic hUe eXtension LUX). A classification approach using the Appriou model is formalized to build simple mass functions integrating the colour information. Colour sources are obviously not independent since they are computed from the same raw data (RGB color pixels), whereas independence is granted only if two pieces of evidence have been obtained by different means. To deal with this colour information dependence in the fusion process, we propose a compromise operator close to the Denœux cautious rule. In order to limit the risk of artifact detection, the algorithm dynamically discounts the reliability of the VJ mass functions by estimating the conflict K inside the bounding box. Indeed, when the VJ detector recognizes falsely a shape-like artifact of a face

in the background with a high reliability degree, skin colour and VJ mass functions are discording, so that an important conflict is generated inside the VJ face container. Finally, in order to synthetize a discriminatory face model, the colour mass sets and the VJ mass sets are fused via the Florea rule.

As regards the tracking phase, the evidential face model constitutes the entry to a tracking filter. Indeed, the pignistic probabilities from the face model serve as inputs for computing the weights of a particle filter which ensures face tracking at video rate. The pignistic probabilities guarantee the compatibility between the belief formalism and the probability formalism. Probabilistic tracking by a particle filter is well suited here since the face, positioned relatively close to the camera, has unpredictable egomotion and frequent direction changes. The goal is to estimate the parameters of a state vector which represents the cinematic characteristics of the target-object, *i.e.* the face. The outer contour of the face is approximated by an ellipse whose parameters are stored in the state vector. The tracking algorithm begins classically with an initialization step. The zone of the face selected manually by the user during the learning stage is used to initialize the parameters of the state vector. Then the algorithm is organized according to two main successive stages: (i) first, the coordinates of the centre of the state vector are estimated by computing the quadratic sum of the pignistic probabilities contained inside each ellipse; (ii) then, the ellipse size and orientation are estimated using an elliptic measure based on least squares fitting method; (iii) if necessary, one resampling operation is performed, when the informative content associated with the particle estimating the state vector is lower than a predefined threshold value.

We illustrate the algorithm behaviour on various sequences registered in our laboratory in the presence of total or partial occlusion or pose variation, and we make also a performance comparison on a benchmark sequence often used in the literature. In order to quantify the tracking performances, the ROC curves (Receiver Operating Characteristics) are drawn for various values of the influence parameters: namely the VJ face detector reliability parameter γ and the color compromise parameter η. The optimal and adaptive parameter tuning of the evidential model is discussed. By setting jointly the adaptive parameter values of the evidential model and the particle filter, it is shown that a noticeable improvement of the tracking behaviour is achieved.

1. Introduction

La détection et le suivi de visage en temps réel dans une séquence vidéo constituent une problématique largement étudiée depuis plus d'une dizaine d'années par la communauté du traitement d'image et de la vision assistée par ordinateur, en raison de la variété des applications : vidéosurveillance, télémédecine, systèmes de téléconférence, interaction homme machine destinée à de nouvelles interfaces, robotique. Cependant, malgré les progrès notables en traitement d'image et en technologie des ordinateurs (réduction du temps de calcul), le développement d'algorithmes génériques et robustes de détection et de suivi de visage fait toujours l'objet de recherches actives.

En plus des variations d'éclairage et du contenu du fond de l'image qui peuvent perturber la détection, les autres facteurs contextuels qui complexifient la détection par ordinateur d'un visage sont de deux types. D'abord les différences d'apparence dues à la morphologie de l'individu (forme du nez, couleur des yeux, de la peau), ou à la présence d'artefacts visuels et occultations (présence de lunettes, main devant le visage). Ensuite les changements d'expression dictés par les codes socio-culturels qui sont fonctions de la société, du contexte historique, de facteurs émotionnels et psychologiques, du rapport à autrui. Abondamment étudiés en sciences humaines (sciences cognitives, psychologie, sociologie) ces points sont partiellement pris en compte en traitement d'image pour la reconnaissance du visage ou l'analyse d'expressions. Ils sont paradoxalement peu examinés en détection du visage proprement dite. En effet, ils ne s'intègrent pas aisément dans un algorithme de détection temps réel car leur modélisation est complexe.

Dans cet article, pour appréhender la complexité du visage, nous faisons appel à l'utilisateur qui sélectionne manuellement une zone de visage sur la première image de la séquence vidéo. Cette initialisation rapide constitue l'apprentissage. Celui-ci fournit un modèle *a priori* intéressant par sa simplicité mais qui souffre de la subjectivité de la sélection de la zone par l'usager et de l'incomplétude à cause du manque d'exhaustivité de l'apprentissage. Une modélisation probabiliste n'est donc pas suffisante dans ce contexte. C'est pourquoi la modélisation du visage se situe dans le cadre de la théorie de l'évidence et considère le modèle de croyances transférables (MCT) (Smets, 1990). Celui-ci s'avère bien adapté pour modéliser l'absence de connaissance d'un système complexe. Ainsi, Hammal (Hammal *et al.*, 2007) a démontré l'efficacité du MCT pour classifier des émotions et des expressions faciales. Ramasso (Ramasso *et al.*, 2010) a utilisé ce formalisme dans le contexte de la reconnaissance des activités humaines.

L'application a pour finalité le suivi automatique du visage d'une personne placée dans le champ de vision d'une caméra motorisée *pan-tilt-zoom*. La caméra asservie réalise un cadrage dynamique du visage afin de le maintenir au centre du plan image au cours de la séquence vidéo. Le suivi doit être robuste aux occultations, aux variations de pose, aux changements d'échelle, d'éclairage et de fond. Concernant les conditions d'acquisition, le contexte de l'application se limite à un environnement d'intérieur, typiquement un laboratoire ou un bureau. La distance séparant le visage du capteur est comprise entre 50 cm et quelques mètres. Un éclairage ordinaire est supposé (pas d'éclairage contrôlé), éventuellement en présence d'une source lumineuse d'appoint (lampe de bureau) ou de l'influence de la lumière extérieure. L'algorithme se décompose en deux phases (Faux, 2009) : la modélisation du visage puis la procédure de suivi qui génère une ellipse englobant le visage (figure 1).

Le chapitre 2 présente un état de l'art des techniques de détection du visage. La section 3 énonce les concepts de la théorie de l'évidence de Dempster-Shafer, formalisme sur lequel se base notre modèle évidentiel de visage détaillé au paragraphe 4. La question du suivi est développée au chapitre 5. Enfin une évaluation et une analyse des performances de l'algorithme sont fournies au paragraphe 6.

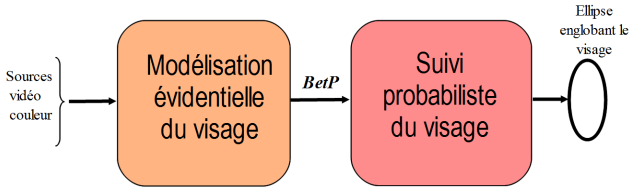


Figure 1. Synoptique de l'algorithme : détection par modélisation évidentielle et suivi probabiliste par filtrage particulaire

2. Détection informatique du visage : état de l'art

Les premiers travaux des années 1960 (Sakai *et al.*, 1969), liés à l'essor de l'intelligence artificielle, emploient des techniques heuristiques et anthropométriques avec diverses hypothèses simplificatrices telles qu'un fond uni et un visage de face. Cependant ces techniques trop simplistes souffrent d'un manque de flexibilité, de robustesse et ne répondent que partiellement au défi de la détection de visage. Aujourd'hui, les méthodes de détection se classent en deux catégories qui diffèrent selon la façon de considérer l'information *a priori* du visage (Hjelmås, Low, 2001 ; Yang *et al.*, 2002).

Les méthodes à base de primitives utilisent les propriétés physiques du visage. Elles se basent sur de nombreuses heuristiques dans le choix des modèles ou des données extraites de l'image. L'analyse dite de bas niveau considère les informations obtenues, soit directement à partir des pixels telles que la luminance ou la couleur (Sigal *et al.*, 2004 ; Soriano *et al.*, 2003), soit indirectement par des opérations mathématiques sur ces pixels et leur voisinage dans le cas de la détection de contours, de mouvement ou de texture. Ainsi, la transformation en ondelettes s'avère efficace pour extraire les traits du visage (Huang *et al.*, 2005). La couleur du visage est certes une primitive pertinente en raison de ses propriétés spécifiques et de ses qualités d'invariance face aux mouvements de rotation et de translation. Cependant, la teinte chair se décline selon une grande variété de nuances (ombrée, pâle, surexposée) en fonction du sujet et de l'éclairage. C'est pourquoi, la construction d'un détecteur robuste de teinte doit considérer le problème délicat du choix de l'espace colorimétrique (Phung *et al.*, 2005 ; Vezhnevets *et al.*, 2003). Toutefois, ces primitives produites par l'analyse bas niveau demeurent ambigües. Pour confirmer la détection il s'avère indispensable d'employer une analyse supplémentaire. L'analyse de traits est basée sur la connaissance d'un modèle adéquat de visage, ou modèle *a priori*, et sur des mesures de distances normalisées et d'angles issues de la description individuelle des parties du visage (yeux, nez, bouche). Le regroupement des composantes faciales dans des constellations robustifie la détection dans le contexte de poses diverses ou de fonds complexes. Pour cette première famille de méthodes, les traitements sont potentiellement rapides car une base d'apprentissage n'est pas nécessaire. En revanche, les méthodes d'extraction des paramètres sont souvent basées sur des choix heuristiques spécifiques au contexte, et construites empiriquement sur des indices de couleur, de contour et de mouvement.

A contrario, les approches holistiques voient la détection comme un problème général d'identification, et traitent le visage comme un tout. Il s'agit de comparer une image à un modèle générique de visage et d'indiquer s'il y a ou non ressemblance. Les *a priori* sur les spécificités géométriques et physiologiques du visage sont évacués afin de limiter les erreurs de modélisation dues à la connaissance incomplète et imprécise du visage. Ces méthodes s'appuient sur l'apprentissage d'un modèle de visage à partir d'une base d'exemples aussi complète que possible. Les méthodes linéaires de sous-espaces (ACP), les approches statistiques (MMC), les machines à vecteurs de support ou séparateurs à vaste marge (SVM), les réseaux de neurones ou approches connexionnistes, peuvent s'envisager pour détecter le visage. Un pas important dans ce domaine a été fait avec le premier détecteur holistique présentant des capacités de temps réel proposé par Viola et Jones (Viola, Jones, 2001). Celui-ci se base sur une sélection automatique de filtres de Haar $2D$ appliqués à l'image monochrome. Le détecteur est composé de classifieurs *boostés* de complexité croissante mis en cascade. Des variantes de cet algorithme ont été implantées pour s'adapter à des visages avec des poses variables (Viola, Jones, 2003). Les modèles à forme active (*active shape models* - ASM) introduits par Cootes (Cootes, Taylor, 1992) sont des modèles déformables qui dépeignent le plus haut niveau d'apparence des parties du visage. Une fois initialisé à proximité d'une composante faciale, ce modèle modifie ses caractéristiques locales (contour, contraste) et se déforme graduellement pour prendre la forme de la composante. Les modèles à apparence active (AAM) sont une extension des modèles à forme active (Cootes *et al.*, 2001). L'exploitation de la dimension temporelle conduit à un modèle de visage $3D$ déformable variant selon des paramètres morphologiques ou d'expression et capable de s'adapter en déformation au cours d'une séquence vidéo (Knothe *et al.*, 2011). Par conséquent, cette deuxième famille de méthodes de détection procure une certaine souplesse vis-à-vis des différents contextes tels que le nombre de visages à traiter, le type d'éclairage. Pourtant ces méthodes, fortement conditionnées par le choix du modèle de visage, nécessitent une masse de données importante et suffisamment représentative. Or, la base d'apprentissage n'est jamais exhaustive et sa construction est un problème à part entière.

Pour finir, notons que cette classification n'est toutefois pas exclusive car de nombreuses méthodes utilisent des approches mixtes. Il est aussi important de différencier les méthodes de détection dédiées aux images fixes pour lesquelles des algorithmes complexes peuvent être mis en œuvre, de celles destinées aux séquences vidéo où le temps de calcul est une contrainte essentielle.

3. Théorie de l'évidence

3.1. Fonctions de croyance

La théorie de l'évidence, appelée aussi théorie des fonctions de croyance ou théorie de Dempster-Shafer, date des années 1970. Inspirée des notions de probabilités supérieures et inférieures étudiées d'abord par Dempster (Dempster, 1967), puis par Shafer (Shafer, 1976), elle peut être interprétée de manière plus générale d'un point de

vue subjectiviste comme un modèle formel quantitatif de degrés de confiance (Smets, 1990). Ne nécessitant pas forcément de connaissances *a priori* sur le problème à traiter, et offrant la possibilité de répartir la croyance sur des compositions d'hypothèses (et non uniquement sur les singletons comme dans la modélisation probabiliste), cette théorie procure une grande souplesse de modélisation et permet de résoudre des problèmes complexes. Ainsi, elle a été appliquée avec succès dans le domaine de la fusion multicapteurs, des signaux et des images.

Le concept premier dans la théorie de l'évidence est la notion de fonction de masse qui caractérise l'opinion d'un agent sur une question ou sur l'état d'un système. L'ensemble fini des réponses à cette question se nomme le cadre de discernement, noté Ω. Une fonction de masse est une application de l'ensemble des 2^{Ω} parties A de Ω vers l'intervalle $[0;1]$ qui vérifie :

$$\sum_{A \subseteq \Omega} m(A) = 1. \tag{1}$$

Cette normalisation garantit une commensurabilité entre plusieurs jeux de masses. La masse $m(A)$ s'interprète comme la part de croyance placée strictement sur A.

Une fonction de masse simple, ou état de croyance élémentaire, possède deux ensembles focaux de la forme :

$$\begin{aligned} m(A) &= 1 - w, \\ m(\Omega) &= w, \quad \text{avec} \quad w \in [0;1] \quad \text{et} \quad A \subseteq \Omega. \end{aligned} \tag{2}$$

Notée A^w, cette fonction de masse simple représente la masse mise non pas sur A mais sur Ω. Pour A quelconque, A^1 ($w = 1$) représente la fonction de masse vide tandis que la fonction de masse catégorique s'écrit A^0 ($w = 0$). Afin de représenter un état de croyance complexe, il est envisageable de construire un ensemble formé de ces propositions pondérées indépendantes. Ainsi la décomposition canonique décompose, sous certaines conditions, une fonction de croyance quelconque en un ensemble de masses simples combinées (Smets, 1995).

3.2. Combinaison des croyances

La combinaison ou révision des croyances intervient lorsqu'on dispose de nouvelles informations, codées sous forme de fonctions de croyance, à fusionner avec les fonctions de croyance existantes, afin de réaliser une synthèse de connaissance dans un environnement multi-sources. Deux contraintes sont à respecter : chaque source d'information s'exprime sur le même cadre de discernement, et les sources doivent être indépendantes (Yaghlane *et al.*, 2000). La règle conjonctive et la somme disjonctive sont les deux opérateurs principaux de combinaison.

Pour J sources d'information indépendantes et totalement fiables, dont les hypothèses sont définies sur le cadre de discernement Ω, la fonction de masse résultant de la combinaison conjonctive, notée $m_{\textcircled{\cap}}$, s'écrit :

$$m_{\textcircled{\cap}}(A) = \sum_{A_1 \cap \ldots \cap A_J = A} \left(\prod_{j=1}^{J} m_j(A_j) \right), \quad \forall A \subseteq \Omega. \tag{3}$$

Cette règle commutative et associative, possède pour élément neutre l'ignorance totale et pour élément absorbant la fonction certitude totale. Elle n'est cependant pas idempotente. Cette règle conduit généralement à une fonction de masse non normalisée ($m_{\textcircled{\cap}}(\emptyset) \neq 0$). Dempster a introduit une étape de normalisation afin d'aboutir à la loi de combinaison conjonctive normalisée plus connue sous le nom de règle de combinaison de Dempster ou somme orthogonale (Dempster, 1967).

La règle disjonctive (Smets, 1993) remplace l'intersection par l'union dans l'équation (3) :

$$m_{\textcircled{\cup}}(A) = \sum_{A_1 \cup A_2 \ldots \cup A_J = A} \left(\prod_{j=1}^{J} m_j(A_j) \right). \tag{4}$$

3.2.1. Nouvelles règles de combinaison

Les règles conjonctives et disjonctives supposent que les fonctions de croyance combinées sont issues de sources indépendantes. Or dans les applications, les sources ne le sont pas toujours. Pour répondre à ce problème, Denœux a proposé la règle conjonctive prudente (Denœux, 2008). Cette règle, notée $\textcircled{\wedge}$, s'appuie sur le principe d'engagement minimal qui indique que lorsque plusieurs fonctions de croyance sont compatibles avec un ensemble de contraintes, il faut choisir la moins informative (principe analogue au maximum d'entropie en théorie des probabilités). Ce principe stipule que l'on ne devrait pas donner plus de croyance que justifié par la source d'information. Sous la contrainte que m_{12} soit plus riche que m_1 et m_2, la fonction de masse la moins informative existe, est unique et est définie par le minimum (noté $\wedge$) des fonctions de pondération associées à m_1 et m_2. Ainsi, si A^{w_1} et A^{w_2} sont deux fonctions de masse simple, leur combinaison par la règle conjonctive prudente est la fonction de masse simple $A^{w_1 \wedge w_2}$ telle que :

$$\begin{aligned} w_{1 \textcircled{\wedge} 2}(A) &= w_1(A) \wedge w_2(A) \quad \forall A \subset \Omega, \\ m_{1 \textcircled{\wedge} 2}(A) &= \textcircled{\cap}_{A \subset \Omega} A^{w_1(A) \wedge w_2(A)}. \end{aligned} \tag{5}$$

Les propriétés de la règle prudente résultent de celles du minimum : commutativité, associativité, idempotence.

3.2.2. Gestion du conflit

Lors de la combinaison conjonctive, il est possible que certaines sources combinées soient discordantes et affirment des propositions incompatibles. La masse affectée à l'ensemble vide quantifie ce conflit. De nombreuses règles de combinaison ont

été proposées pour résoudre ce problème (Martin, Osswald, 2007 ; Smarandache, Dezert, 2009 ; Pichon, Denœux, 2009). Florea (Florea *et al.*, 2009) a proposé la famille des règles adaptatives qui fournit une solution intermédiaire entre la conjonction et la disjonction. Elle est définie pour tout $A \subseteq \Omega$ par :

$$m_f(A) = \alpha(K).m_{\textcircled{\cap}}(A) + \beta(K).m_{\textcircled{\cup}}(A) \quad \text{pour } A \neq \emptyset, \tag{6}$$

$$m_f(\emptyset) = 0, \tag{7}$$

où $\alpha(K)$ et $\beta(K) \in [0;1]$ sont deux fonctions du paramètre de conflit $K = m_{\textcircled{\cap}}(\emptyset)$.

3.3. Modélisation des fonctions de masse

L'estimation des fonctions de masse est un problème difficile qui n'a pas de solution universelle. La difficulté est augmentée si l'on veut affecter des masses aux hypothèses composées. On distingue les modélisations fondées sur le calcul d'une distance (Denœux, 1995 ; Zouhal, Denoeux, 1998), et celles basées sur le calcul d'une vraisemblance. Ces dernières se décomposent en méthodes globales (Shafer, 1976) et méthodes séparables. Une comparaison de ces deux approches a été réalisée par Denœux et Smets (Denœux, Smet, 2006). Les méthodes séparables construisent une fonction de croyance pour chacune des hypothèses H_i du cadre de discernement. Ce type d'approche a été proposé indépendamment par Smets (Smets, 1986) et par Appriou (Appriou, 1999). Ces modèles d'inspiration probabiliste supposent une estimation supervisée des probabilités conditionnelles $p(x_j|H_i)$ par apprentissage (où x_j représente une observation de la source j). Appriou préconise l'utilisation du modèle obtenu à partir du théorème de Bayes généralisé proposé par Smets (Smets, 1986) :

$$\begin{cases} m_j(\{H_i\}) = 0, \\ m_j(\{\overline{H_i}\}) = d_{ij}[1 - R_j.p(x_j|H_i)], \\ m_j(\Omega) = 1 - m_j(\{\overline{H_i}\}). \end{cases} \tag{8}$$

R_j, coefficient de pondération des probabilités, est un facteur de normalisation dont le domaine de définition est contraint par : $R_j \in [0, (\max_i \{p(x_j|H_i)\})^{-1}]$. Il garantit l'obtention d'une fonction de croyance normalisée. Pour $R_j = 0$, seule la fiabilité de la source est prise en compte, sinon les données sont également considérées. Le coefficient d_{ij} est un coefficient d'affaiblissement lié à la classe H_i qui caractérise le degré de représentativité de l'apprentissage. Ce coefficient vaut 1 lorsque les densités sont parfaitement représentatives de l'apprentissage et 0 lorsque la distribution de probabilité est complètement méconnue.

Ce modèle s'avère bien adapté lorsque : (i) l'on apprend facilement une classe contre toutes les autres, ce qui est fréquent en reconnaissance des formes dans les images, en particulier pour la détection de visage ; (ii) chaque classe est déterminée à partir d'un détecteur adapté (par exemple pour une application de vidéoconférence, un détecteur de teinte chair de forme elliptique dans une image de locuteur permet de définir le probabilité d'appartenance à la classe visage, mais n'est pas capable de distinguer entre d'autres classes).

3.4. Modèle des croyances transférables (MCT)

Le modèle MCT est une interprétation subjectiviste, dans laquelle une fonction de croyance modélise la connaissance partielle de la valeur d'une variable (Smets, Kennes, 1994). Le MCT comporte deux niveaux : le niveau crédal et le niveau pignistique. Le niveau crédal comprend principalement l'étape de modélisation, c'est-à-dire la partie statique du modèle représentant les connaissances sous forme de fonctions de croyance, ainsi que la combinaison de ces croyances appelée révision qui correspond à la partie dynamique du modèle.

La décision s'effectue au niveau pignistique qui impose une transformation des fonctions de croyances en distributions de probabilité. La transformation pignistique consiste à partager équitablement chaque masse de croyance normalisée en se basant sur le principe de la raison suffisante, qui stipule qu'en l'absence de raison de privilégier une hypothèse plutôt qu'une autre, on doit supposer les hypothèses équiprobables (Denœux, 1997). On obtient ainsi une distribution de probabilité pignistique, notée $BetP$, définie pour tout $A \in 2^{\Omega}$ avec $A \neq \emptyset$ par :

$$BetP(A) = \sum_{B \in 2^{\Omega}\ ;\ B \neq \emptyset} \frac{|A \cap B|}{|B|} \frac{m(B)}{1 - m(\emptyset)}, \quad \text{avec } m(\emptyset) \neq 1. \tag{9}$$

Ainsi, on associe une seule probabilité pignistique à une fonction de masse mais inversement, une infinité de fonctions de masse à une distribution pignistique. Cela reflète la perte d'information occasionnée lors du passage entre le niveau crédal et le niveau pignistique.

4. Modélisation évidentielle du visage

4.1. Introduction

Notre approche consiste à restreindre la phase d'apprentissage à une initialisation sommaire où l'utilisateur sélectionne manuellement sur la première image de la séquence vidéo la zone de visage à suivre. La forte incomplétude de cet apprentissage simple ne permet pas d'envisager une modélisation probabiliste efficace qui supposerait une base d'apprentissage exhaustive des différentes classes. C'est pourquoi nos développements se situent dans le cadre du MCT, formalisme permettant la représentation d'informations partielles. Notre modélisation du visage repose sur un processus de fusion évidentielle faisant collaborer deux familles d'attributs ou sources complémentaires (figure 2) : le détecteur de visages de Viola et Jones (VJ) connu pour ses bonnes performances et un détecteur de teinte chair qui est un bon critère discriminant.

À chaque pixel p de l'image est associé un cadre de discernement comprenant deux classes mutuellement exclusives : $\Omega_p = \{H_{1p}, H_{2p}\}$, où $\{H_{1p}\}$ représente l'hypothèse visage et $\{H_{2p}\}$ symbolise ce qui n'est pas le visage (*i.e.,* le complémentaire). La limitation du cadre de discernement à deux hypothèses réduit la complexité et donc le temps de calcul, ce qui s'avère important en vue d'une procédure de suivi. Dans la

suite, pour simplifier l'écriture, nous omettrons l'indice p et noterons simplement Ω, H_1 et H_2 ces quantités relatives au pixel.

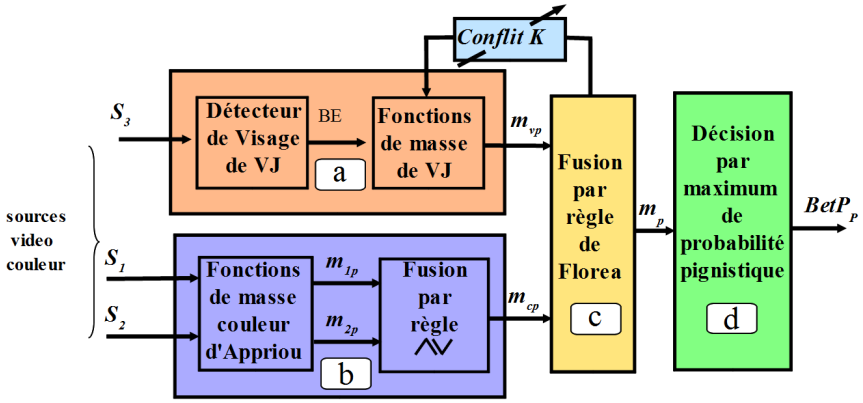


Figure 2. Schéma-bloc du modèle évidentiel de visage pour la synthèse des probabilités pignistiques : a) jeux de masse des attributs de VJ ; b) jeux de masse de la couleur ; c) fusion des masses ; d) décision

4.1.1. Sources d'information couleur

La couleur de la peau est une information importante dans notre application car elle permet l'implantation d'algorithmes rapides, invariants aux changements d'orientation et d'échelle. Cependant, la distribution de la couleur chair du visage varie sensiblement en fonction de l'éclairage et de l'espace couleur considéré. Afin d'améliorer la robustesse aux variations d'éclairage, notre choix s'est orienté vers l'espace logarithmique couleur *LUX* au lieu des espaces couleur linéaires comme RGB ou $YCrCb$. Les trois composantes sont calculées à partir des composantes RGB comme suit :

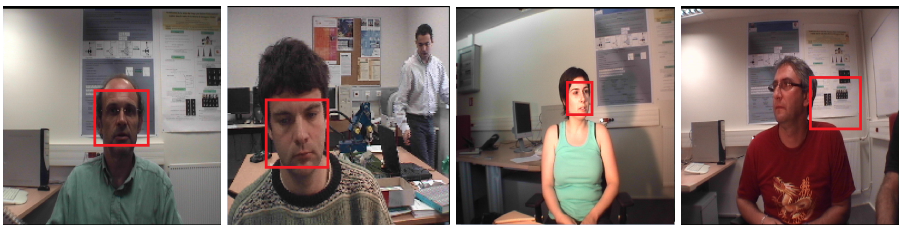
$$\begin{aligned} L &= (R+1)^{0,3}(G+1)^{0,6}(B+1)^{0,1} - 1 \\ U &= \begin{cases} \frac{M}{2}\left(\frac{R+1}{L+1}\right) \text{ pour } R < L \\ M - \frac{M}{2}\left(\frac{L+1}{R+1}\right) \text{ sinon} \end{cases} \quad \text{avec :} \quad M = 256 \\ X &= \begin{cases} \frac{M}{2}\left(\frac{B+1}{L+1}\right) \text{ pour } B < L \\ M - \frac{M}{2}\left(\frac{L+1}{B+1}\right) \text{ sinon.} \end{cases} \end{aligned} \tag{10}$$

L représente la luminance logarithmique, tandis que U et X sont les chrominances logarithmiques, resp. le rouge et le bleu. Cet espace couleur non-linéaire basé sur une transformation logarithmique offre un rendu de contraste performant pour les zones de faible luminance. Il assure une description efficace des teintes, il est peu sensible au bruit et a montré son efficacité en segmentation couleur (Liévin, Luthon, 2004), en compression couleur (Luthon, Beaumesnil, 2004) et en rendu couleur (Luthon *et al.*, 2010). Les trois sources d'information couleur utilisées ici pour la modélisation du

visage, notées s_j $(j = 1, 2, 3)$, sont respectivement : $s_1 = U; s_2 = X$ pour la teinte chair, et $s_3 = L$ pour le détecteur de VJ.

4.2. Modéle évidentiel du détecteur de Viola et Jones

Le détecteur de VJ opère sur l'information de luminance uniquement (source s_3). Il génère une boîte englobante rectangulaire du visage très fiable lorsque le visage est de face ou légèrement de profil (figure 3a, b, c). Cependant il est défaillant en présence de fortes occultations, de rotations ou lorsqu'il reconnaît un ersatz de visage dans le fond de l'image (figure 3d).



a) séquence No. 1 b) séquence No. 5 c) séquence No. 6 d) séquence No. 2

Figure 3. Exemples de boîte englobante fournie par le détecteur de VJ : a), b), c) détections correctes ; d) détection incorrecte

Afin de modéliser cet attribut de Viola et Jones par une fonction de croyance (figure 2a), nous affectons au pixel p une fonction de masse notée $m_v(.)$, en fonction de sa position relativement à la boîte englobante notée BE et proportionnellement à un paramètre de fiabilité noté $\gamma \in [0; 1]$:

$$\left.\begin{array}{rcl} m_v(\{H_1\}) & = & \gamma \\ m_v(\{H_2\}) & = & 0 \\ m_v(\Omega) & = & 1-\gamma \end{array}\right\} \forall p \in BE \qquad \left.\begin{array}{rcl} m_v(\{H_1\}) & = & 0 \\ m_v(\{H_2\}) & = & \gamma \\ m_v(\Omega) & = & 1-\gamma \end{array}\right\} \forall p \notin BE$$

Pour $\gamma = 0$, aucune hypothèse n'est privilégiée car la source d'information n'est pas fiable. Toute la masse est associée à la tautologie Ω et il n'y a donc pas de discrimination selon l'appartenance du pixel à BE. Pour $\gamma = 1$, la masse est maximale pour la classe visage $\{H_1\}$ à l'intérieur de BE, et pour le fond $\{H_2\}$ à l'extérieur de celle-ci. Pour des valeurs de γ intermédiaires, cette modélisation simple est avantageuse car elle module l'information à la fois en fonction de γ et de la position du pixel. Nous verrons plus loin (section 4.6) que γ peut être ajusté en fonction du conflit.

4.3. Modéle évidentiel de la couleur du visage

4.3.1. Fonctions de masse couleur

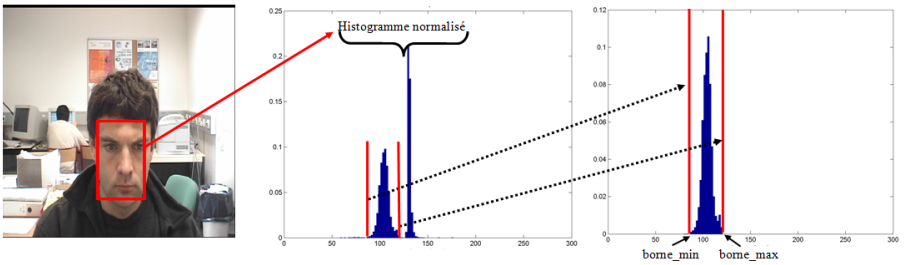
Une approche de type classification est formalisée pour construire des jeux de fonctions de masse intégrant l'information couleur. Soient :

– $\{p\}_1^P$ l'ensemble des P pixels présents dans l'image, où P est la taille de l'image (typiquement $P = 400 \times 400$) ;

– s_j une source d'information (plan couleur $j = 1, 2$). Une observation est notée s_{jp} : il s'agit de la composante j du vecteur d'attribut couleur associé au pixel p ;

– c_p la classe du pixel p (primitive cachée correspondant à une hypothèse).

Étant donné un pixel p avec une observation connue s_{jp}, mais de classe inconnue c_p, le problème consiste à construire une croyance sur la valeur actuelle de la classe c_p sans utiliser aucune base d'apprentissage sauf celle issue de l'initialisation sommaire sur la première image.



a) b) c)

Figure 4. Initialisation sur la séquence No. 3 :
a) zone visage sélectionnée sur la première image de la vidéo ;
b) densité de probabilité $p(s_2|H_1)$; c) fonction $skin_2$ bornée

Le modèle d'Appriou (équation (8)) est bien adapté au problème car il prend en compte la fiabilité de la source d'information. Ce modèle nécessite la connaissance des vraisemblances conditionnellement aux classes, *i.e.,* un modèle *a priori*, que nous noterons $skin_j$ (équation (11)), qui caractérise la relation entre l'observation s_j et l'hypothèse visage $\{H_1\}$. Ce modèle est généré à partir de l'apprentissage supervisé où l'opérateur sélectionne manuellement une zone visage sur la première image de la séquence vidéo (figure 4a). Les histogrammes sont calculés en considérant tous les attributs couleur s_{jp} dans cette zone. Les densités de probabilité conditionnelles $p(s_j|H_1)$ sont alors déduites des histogrammes par une simple opération de normalisation (figure 4b). Afin de maximiser le résultat de la classification, nous proposons une stratégie où l'utilisateur a la possibilité de limiter ces densités de probabilité. Les bornes choisies sont tracées en rouge sur la figure 4b. Alors, la fonction $skin_j$ est définie par (figure 4c) :

$$skin_j(\xi) = \begin{cases} p([s_j = \xi]|H_1) & \text{si } borne_{min} < \xi < borne_{max} \\ 0 & \text{sinon,} \end{cases} \tag{11}$$

où $\xi \in [0; M[$ est le niveau d'intensité de la composante s_j. Notons que $skin_j(s_{jp})$ s'obtient simplement par une opération de lecture de table (*Look Up Table*).

Dès lors, durant la séquence vidéo, le jeu de fonctions de masse couleur affecté à chaque pixel p de l'image à l'instant t est tel que :

$$\begin{array}{rcl} m_{jp}(\{H_1\}) & = & 0, \\ m_{jp}(\{H_2\}) & = & d_j[1 - R_j.skin_j(s_{jp})], \\ m_{jp}(\Omega) & = & 1 - m_{jp}(\{H_2\}). \end{array} \quad (12)$$

$m_{jp}(.)$ est la masse affectée à s_{jp}. R_j pondère les données $skin_j$ et est fixé à sa valeur maximale. d_j est initialisé de manière *ad hoc* par l'opérateur sur la première image de la séquence vidéo. Nous garantissons le caractère non dogmatique des fonctions de masse en imposant un coefficient de fiabilité $d_j < 1$. La masse d'ignorance $m_{jp}(\Omega)$ demeure par conséquent strictement positive. Ces fonctions sont des fonctions de masse simples (équation (2)) avec des pondérations notées w_{jp} telles que :

$$w_{jp}(\{H_1\}) = 1 \quad \text{et} \quad w_{jp}(\{H_2\}) = m_{jp}(\Omega).$$

Finalement, pour générer un jeu de fonctions de masse couleur global plus informatif noté $m_c(.)$, les masses $m_{jp}(.)$ sont fusionnées comme décrit ci-après (figure 2b).

4.3.2. Fusion par l'opérateur de compromis

Les sources couleur sont dépendantes. En effet, si la composante R varie, les valeurs des trois sources L, U et X sont modifiées (équation (10)). Dans ce contexte de non-indépendance, une solution consiste à adopter une attitude conservative en appliquant la règle conjonctive prudente de Denœux (cf. section 3.2.1) pour fusionner les fonctions de masse de couleur. Deux variantes de la règle prudente peuvent être envisagées. Une règle conjonctive adaptative considère que les sources ne sont ni distinctes ni très redondantes. Les poids w_{jp} sont agrégés avec une t-norme paramétrique *i.e.,* avec un opérateur de combinaison variant entre le produit et le $\min$ (Denœux, 2008 ; Kallel, Le Hégarat-Mascle, 2009). Dans notre contexte applicatif cette fusion adaptative n'améliore pas la qualité de la modélisation couleur de façon significative, par rapport à la règle prudente. La complexité de l'opérateur accroît inutilement le temps de calcul sans contrepartie. De plus, une limitation de la t-norme provient de son absence d'effet compensatoire. Une stratégie de fusion avec une propriété de compensation semble pertinente. Nous proposons une solution qui consiste à utiliser une règle de compromis adaptative très simple, notée $\wedge\vee$, variant entre le $\min$ et le $\max$. Cette règle est définie dans le cas de deux poids distincts w_{jp} et w_{kp} avec $j \neq k$, $0 \leq w_{jp} \leq 1$ et $0 \leq w_{kp} \leq 1$ par :

$$w_{jp \wedge\vee kp}(A) = (1-\eta) \times \min(w_{jp}(A), w_{kp}(A)) + \eta \times \max(w_{jp}(A), w_{kp}(A)) \quad (13)$$

avec $A \in \{\emptyset, \{H_1\}, \{H_2\}\}$. Pour $\eta = 0$ nous retrouvons le $\min$ utilisé dans la règle conjonctive prudente, tandis que pour $\eta = 1$ nous obtenons l'opérateur $\max$ proche de la règle disjonctive. Ainsi, pour deux attributs de couleur s_{1p} et s_{2p}, les masses combinées $m_{1p \wedge\vee 2p}$ sont déterminées par la règle suivante :

$$m_{1p \wedge\vee 2p}(A) = \mathbb{O}_{A \subset \Omega} A^{w_{1p}(A) \ \wedge\vee \ w_{2p}(A)}. \quad (14)$$

Les résultats de cette procédure adaptative sont illustrés sur la figure 5. On perçoit nettement la variation de la qualité de la fusion en fonction du réglage du paramètre η. En augmentant la valeur de η, le poids $w_{1p\wedge\vee 2p}(.) = w_{1p}(.) \wedge \vee w_{2p}(.)$ affecté au pixel p est augmenté en fonction de la différence entre w_{1p} et w_{2p}, sauf bien sûr si $w_{1p} = w_{2p}$. Sur la figure 5a, le cou s'avère mal détecté pour $\eta = 0$ tandis que pour $\eta = 1$ tout le visage est parfaitement classifié. La contrepartie de cette amélioration de la modélisation du visage est la mise en évidence de certaines zones du fond de couleur proche de la teinte chair. Ceci est nettement visible sur la figure 5b. Le tee-shirt de couleur rouge, correctement classifié par la règle prudente ($\eta = 0$) est détecté à tort avec $\eta = 1$. L'emploi d'un opérateur de compromis influence donc la qualité du modèle de couleur.

Figure 5. Résultats (probabilités pignistiques) de la fusion des sources couleur s_1 et s_2 par l'opérateur de compromis pour quatre valeurs de η : a) séquence No. 1 ; b) séquence No. 2

REMARQUE 1. — Nous avons proposé par ailleurs (Faux, Luthon, 2006) une variante dite contextuelle, qui combine deux ou trois sources couleur (s_1; s_2 et s_3) et ceci pour trois contextes caractéristiques du visage correspondant respectivement à une zone ombrée (cou), une zone saturée (front, nez) et une zone globale du visage. Ici, le modèle proposé est simplifié par l'utilisation d'un apprentissage simple sur une seule zone globale du visage et par l'utilisation de deux sources de couleur uniquement ($s_1 = U$ et $s_2 = X$) en raison d'un éclairage relativement homogène dans l'application. □

REMARQUE 2. — Les règles de combinaison idempotentes constituent une alternative aux règles de combinaisons conjonctives classiques. Elles ne renforcent pas la certitude lors des combinaisons d'informations dépendantes mais assurent que la combinaison récursive d'une information avec elle-même donne toujours le même résultat. Dans ce cas l'indépendance des sources est moins cruciale et l'idempotence autorise

la dépendance. Il apparaît donc un dilemne renforcement/idempotence. Dans notre modèle de couleur nous avons favorisé un opérateur de fusion idempotent et générant un certain « flou » dans les images. □

4.4. Fusion des deux jeux de masses par la règle de Florea

Le modèle couleur modélise fidèlement la teinte chair mais ne différencie évidemment pas une teinte du visage de celle d'un bras ou d'une main par exemple. C'est pourquoi, afin d'élaborer un modèle discriminant de visage nous fusionnons les jeux de masses de couleur avec le jeu de masses de VJ par la règle de Florea (Florea *et al.*, 2009). Florea préconise logiquement l'emploi d'une règle disjonctive en présence de fort conflit tandis qu'une approche conjonctive est employée dans le cas contraire (équation (6)). Cette règle paraît pertinente dans notre application car elle est robuste dans le sens où elle s'adapte à des sources dont le degré de fiabilité est inconnu. En effet, ces réglages s'avèrent difficiles à quantifier en raison du caractère non stationnaire du contexte. Ainsi nous combinons les informations couleur (m_c) et les attributs de VJ (m_v) et nous élaborons un jeu de masse unique m_p (figure 2c) tel que :

$$m_p(A) = \alpha_p . m_c \textcircled{\cup}_v(A) + \beta_p . m_c \textcircled{\cap}_v(A) \quad \forall A \neq \emptyset. \tag{15}$$

avec $\beta_p = \frac{1-K_p}{1-K_p+K_p^2}$ et $\alpha_p = \frac{K_p}{1-K_p+K_p^2}$ et où le conflit K_p entre les sources vaut :

$$K_p = m_c(\{H_1\}) \times m_v(\{H_2\}) + m_c(\{H_2\}) \times m_v(\{H_1\}) \tag{16}$$

$$\text{d'où } \begin{cases} K_p = m_c(\{H_2\}) \times \gamma, & \forall p \in BE, \\ K_p = 0, & \forall p \notin BE. \end{cases} \tag{17}$$

Le tableau 1 détaille le calcul de m_p obtenu par l'opérateur de Florea à partir des combinaisons conjonctive et disjonctive des masses m_c et m_v.

4.5. Décision évidentielle

La règle de Florea n'affecte pas de masse sur le conflit ($m_p(\emptyset) = 0$) (équation (6)). Dès lors la probabilité pignistique $BetP_p(.)$ associée à la classe visage n'est pas surévaluée en fonction du conflit et a pour expression :

$$BetP_p(\{H_1\}) = m_p(\{H_1\}) + m_p(\Omega)/2, \tag{18}$$

$$\text{d'où } \begin{cases} BetP_p(\{H_1\}) = \dfrac{\alpha_p}{2} + \dfrac{\beta_p}{2}(1+\gamma)(1-m_c(\{H_2\})), & \forall p \in BE \\ BetP_p(\{H_1\}) = \dfrac{1}{2}(1-\gamma)(1-m_c(\{H_2\})) & \forall p \notin BE. \end{cases} \tag{19}$$

Comme $BetP_p \in [0;1]$, elle est multipliée par 255 afin de créer une image lisible en niveaux de gris de cette probabilité.

Tableau 1. Fusion par la règle de Florea de m_c et m_v en fonction de la position du pixel p (à l'intérieur ou hors de la boîte englobante BE)

$masses$	$p \in BE$	$p \notin BE$
$m_{(c \textcircled{\cap} v)}(\emptyset)$	$m_c(\{H_2\})\gamma$	0
$m_{(c \textcircled{\cap} v)}(\{H_1\})$	$\gamma(1 - m_c(\{H_2\}))$	0
$m_{(c \textcircled{\cap} v)}(\{H_2\})$	$m_c(\{H_2\})(1-\gamma)$	$m_c(\{H_2\})(1-\gamma)+\gamma$
$m_{(c \textcircled{\cap} v)}(\Omega)$	$(1-\gamma)(1-m_c(\{H_2\}))$	$(1-\gamma)(1-m_c(\{H_2\}))$
$m_{(c \textcircled{\cup} v)}(\emptyset)$	0	0
$m_{(c \textcircled{\cup} v)}(\{H_1\})$	0	0
$m_{(c \textcircled{\cup} v)}(\{H_2\})$	0	$m_c(\{H_2\})\gamma$
$m_{(c \textcircled{\cup} v)}(\Omega)$	1	$1 - m_c(\{H_2\})\gamma$
$m_p(\emptyset)$	0	0
$m_p(\{H_1\})$	$\beta_p(1 - m_c(\{H_2\}))\gamma$	0
$m_p(\{H_2\})$	$\beta_p(1-\gamma)m_c(\{H_2\})$	$m_c(\{H_2\})(1-\gamma)+\gamma$
$m_p(\Omega)$	$\alpha_p + \beta_p(1-\gamma)(1-m_c(\{H_2\}))$	$(1-\gamma)(1-m_c(\{H_2\}))$

Le tableau 2 résume le comportement du modèle évidentiel de visage lorsque le pixel a une teinte proche ($m_c(\{H_2\}) \rightarrow 0$) ou différente ($m_c(\{H_2\}) \rightarrow 1$) de celle du visage, et ceci en fonction de la fiabilité γ du détecteur de VJ. La performance du modèle évidentiel dépend à la fois de la qualité du modèle couleur et de la fiabilité du détecteur de VJ. C'est pourquoi nous préconisons d'initialiser la valeur de γ telle que $0,5 \leq \gamma \leq 0,7$. Une valeur de γ faible ($\gamma < 0,2$) conduit à limiter l'influence du détecteur de VJ et réduit le modèle évidentiel à un simple détecteur de teinte chair. Lorsque le détecteur de VJ est en défaut (quand il ne délivre aucune boîte englobante) nous fixons $\gamma = 0$.

Tableau 2. Résultats du modèle évidentiel en fonction des informations couleur $m_c(\{H_2\})$ et de la fiabilité γ du détecteur de Viola et Jones

$m_c(\{H_2\})$	γ	$BetP_p(\{H_1\})$ $\forall\, p \in BE$	$BetP_p(\{H_1\})$ $\forall\, p \notin BE$	**Décision** $\forall\, p \in BE$	**Décision** $\forall\, p \notin BE$
$\rightarrow 0$	0	$0,5$	$0,5$	Indécision	Indécision
$\rightarrow 0$	$0,5$	$0,75$	$0,25$	$\{H_1\}$	$\{H_2\}$
$\rightarrow 0$	1	1	0	$\{H_1\}$	$\{H_2\}$
$0,5$	0	$0,25$	$0,25$	$\{H_2\}$	$\{H_2\}$
$0,5$	$0,5$	$0,5$	$0,125$	Indécision	$\{H_2\}$
$0,5$	1	$0,5$	0	Indécision	$\{H_2\}$
$\rightarrow 1$	0	0	0	$\{H_2\}$	$\{H_2\}$
$\rightarrow 1$	$0,5$	$0,33$	0	$\{H_2\}$	$\{H_2\}$
$\rightarrow 1$	1	$0,5$	0	Indécision	$\{H_2\}$

4.6. Affaiblissement de γ par rétroaction

Lorsque le détecteur de VJ se positionne sur un erzatz de visage (figure 3 d) avec une fiabilité élevée ($\gamma \geq 0,5$), les informations de teinte ($m_c(\{H_2\}) > 0,5$) et de VJ sont discordantes et le conflit est important dans cette zone.

Afin de limiter ce risque de détection d'un leurre, nous calculons le conflit global noté K. Il correspond à la somme du conflit contenu dans BE normalisée en fonction de la taille de BE telle que :

$$K = \frac{1}{N_B} \sum_{p \in BE} K_p \tag{20}$$

où N_B est le nombre de pixels appartenant à BE. K est alors utilisé pour affaiblir dynamiquement γ par rétroaction tel que : $\gamma_t = \gamma_0 \times (1 - K_{t-1})$ où K_{t-1} est le conflit calculé à l'image précédente (cf. boucle de rétroaction sur la figure 2).

5. Suivi de visage

Le suivi de visage consiste à estimer la position, la taille et la pose du visage pour obtenir sa trajectoire en temps réel durant la séquence video (Yilmaz *et al.*, 2006). Les nombreuses techniques de suivi peuvent être classées en trois catégories. Les méthodes bas niveau assurent le suivi en considèrant l'information de couleur (*mean-shift*) (Comaniciu *et al.*, 2003), par soustraction du fond dans le cas d'un fond uniforme ou stationnaire, ou en estimant le flux optique. Les *snakes*, les modèles de forme ou d'apparence (AAM) assurent le suivi par une mise en correspondance de modèles (*template matching*) (Isard, MacCormick, 2001 ; Rathi *et al.*, 2007). Ces méthodes utilisent des algorithmes semblables à ceux évoqués en détection de visage (section 2) mais en respectant la contrainte du temps réel. Enfin les méthodes de filtrage assurent un suivi temporel par une prédiction de l'état (la localisation) d'un système dynamique (l'objet à suivre) à l'image suivante en se basant sur des mesures antérieures. Le filtre de Kalman s'emploie pour des modèles uni-modaux gaussiens tandis que les filtres particulaires ont été largement utilisés pour modéliser des processus non linéaires non gaussiens (Arulampalam *et al.*, 2002). Ainsi, pour diverses situations de suivi, Klein (Klein *et al.*, 2010) propose une approche efficace qui fusionne des sources défaillantes en utilisant des informations contextuelles issues d'un filtre particulaire. Muñoz (Muñoz-Salinas *et al.*, 2009) propose une extension du filtre particulaire bayésien à la théorie de l'évidence. L'algorithme présente une solution originale au suivi multi-caméras de personnes dans le contexte d'un environnement d'intérieur.

Dans notre application, le visage est un objet déformable relativement proche de la caméra, dont le mouvement propre est peu prévisible avec des changements de direction. La scène est $a\ priori$ encombrée et non stationnaire en raison de la mobilité de la caméra. Dans ce contexte, nous avons choisi une méthode de suivi par filtrage particulaire de type *bootstrap* car cette technique s'avère efficace lorsque l'objet a

une trajectoire non linéaire et elle prend en compte la redondance temporelle entre les images. Le but est d'estimer de manière robuste les paramètres d'un vecteur d'état qui contient les caractéristiques cinématiques de l'objet à suivre (le visage). Les contours extérieurs du visage sont approximés à l'instant t par une ellipse de centre (x_{c_t}, y_{c_t}), d'axe mineur l_t, de grand axe h_t et d'orientation θ_t. Ces paramètres sont regroupés dans le vecteur d'état $X_t = [x_{c_t}, y_{c_t}, l_t, h_t, \theta_t]$.

L'algorithme de filtrage particulaire applique un filtre bayésien récursif à plusieurs localisations hypothétiques du visage, et fusionne ces hypothèses en fonction de leur vraisemblance, conditionnellement à l'état prédit. Sachant que Y_t est l'observation issue de l'image à l'instant t, le filtre particulaire approxime la distribution de probabilité conditionnelle *a posteriori* $p(X_t|Y_{1:t})$ sous la forme d'une combinaison linéaire pondérée de masses de Dirac appelées particules. Une particule $\Lambda_t^{(n)} = \{\lambda_t^{(n)}, \omega_t^{(n)}\}$ représente une hypothèse sur l'état de la cible. $\lambda_t^{(n)}$ correspond à la position et $\omega_t^{(n)}$ au poids affecté à la particule n à l'instant t. La loi *a posteriori* est alors approximée par :

$$p(X_t|Y_{1:t}) \approx \sum_{n=1}^{N} \omega_t^{(n)} \delta_{\lambda_t^{(n)}}. \tag{21}$$

Le modèle évidentiel de visage constitue l'entrée de la procédure de suivi (cf. figure 1). L'algorithme de suivi (figure 6) débute par une phase d'initialisation. La zone

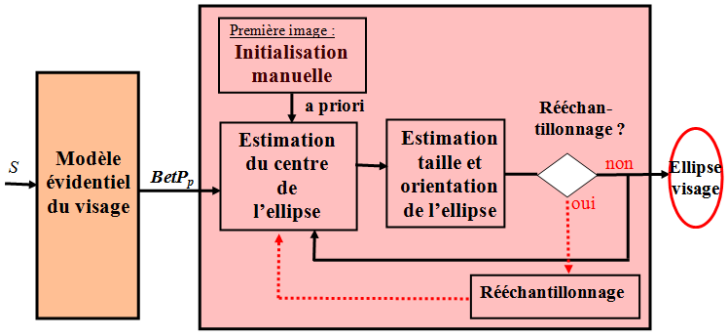


Figure 6. Synoptique de l'algorithme de suivi par filtrage particulaire

de visage rectangulaire sélectionnée manuellement par l'opérateur lors de l'apprentissage supervisé est utilisée pour initialiser les paramètres de X_t. Puis l'algorithme opère en deux étapes successives (décrites aux paragraphes 5.1 et 5.2) :

– dans un premier temps, les coordonnées du centre (x_{c_t}, y_{c_t}) sont déterminées,

– ensuite, la taille et l'orientation (pose) de l'ellipse (l_t, h_t, θ_t) sont estimées.

Lorsque le contenu informatif (équation (23)) associé à la particule estimant le vecteur d'état est inférieur à un seuil fixé à l'initialisation, un rééchantillonnage simple est opéré. Les poids de l'ensemble des particules sont alors fixés à : $\omega_t^{(n)} = 1/N$ (où N est le nombre de particules). Nous tirons au hasard des positions éventuelles du visage

en propageant les particules suivant une loi uniforme $\mathcal{U}_{\mathcal{X}}$. Lorsque la position tirée retrouve une partie du visage, la convergence du filtre après quelques itérations assure la reprise complète du suivi.

5.1. Estimation du centre du modèle

Un modèle dynamique assez simple du vecteur d'état, réduit ici à $X_t = [x_{c_t}, y_{c_t}]$, inspiré des travaux de Pérez (Pérez *et al.*, 2004), distribue aléatoirement le centre des particules dans l'image selon la loi :

$$p(\tilde{X}_t|X_{t-1}) = (1-\nu)\mathcal{N}(\tilde{X}_t|X_{t-1},\Sigma) + \nu\mathcal{U}_{\tilde{X}}(\tilde{X}_t) \tag{22}$$

où $\mathcal{N}(.|\mu,\Sigma)$ est la distribution gaussienne de moyenne μ et de covariance Σ. La matrice diagonale $\Sigma = diag(\sigma^2_{xc_t}, \sigma^2_{yc_t})$ fixe les variances imposées aux composantes de position du vecteur d'état. Le coefficient ν pondère la distribution uniforme : $0 \leq \nu \leq 1$. Cette composante uniforme ($\nu = 0, 1$) gère les rares mouvements erratiques perçus comme des sauts dans la séquence vidéo. Elle aide aussi l'algorithme à se déverrouiller après une période d'occultation partielle ou totale momentanée. Ainsi, sur la figure 7a, l'influence de la distribution gaussienne se caractérise par la concentration de la plupart des particules autour du centre estimé à l'image précédente. On note aussi clairement le rôle du paramètre ν car une dizaine de particules isolées explorent d'autres zones de l'image.

Suite à la prédiction, le filtre particulaire évalue l'adéquation de l'observation $Y_t^{(n)}$ mesurée dans l'ellipse prédite $\tilde{X}_t^{(n)}$, avec les données du modèle de visage pour calculer la vraisemblance $p(Y_t/\tilde{X}_t)$. Pour quantifier cette adéquation, nous considérons la somme quadratique des probabilités pignistiques $BetP_p(\{H_1\})$ contenues à l'intérieur de l'ellipse. Dès lors, le poids estimé de chaque particule n est donné par :

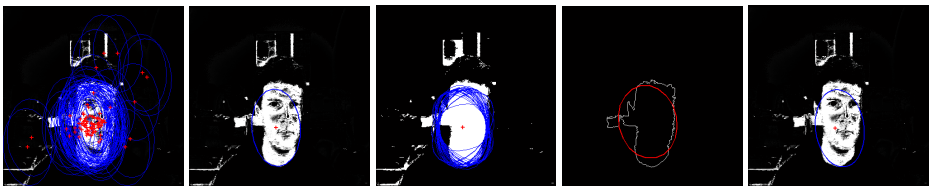
$$\tilde{\omega}_t^{(n)} = \sum_{p\in n} (BetP_p(\{H_1\}))^2. \tag{23}$$

Finalement, le critère du maximum de vraisemblance sélectionne l'ellipse la plus significative et son centre définit les composantes de position du vecteur d'état (figure 7b).

La transformation des masses du modèle évidentiel de visage en probabilités pignistiques (équation (19)), assure la compatibilité avec le cadre probabiliste du filtre particulaire (la disjonction d'hypothèses Ω n'apparaît plus).

5.2. Estimation de la taille et de la pose

Les composantes de taille et de pose à l'instant t du vecteur d'état réduit à $X_t = [l_t, h_t, \theta_t]$ sont prédites par le modèle dynamique de l'équation (22) mais en fixant le paramètre $\nu = 0$. En effet, il n'est pas pertinent de considérer des mouvements



a) b) c) d) e)

Figure 7. Séquence No. 3 : a) particules lors de la phase d'estimation du centre ($N = 80$) ; b) résultat du filtrage en position ; c) particules lors de la phase d'estimation de la taille ; d) mesure de l'ellipse ; e) résultat du filtre particulaire (les pixels blancs dénotent la teinte chair)

erratiques à propos des paramètres de taille ou de pose du vecteur d'état. La figure 7c illustre la répartition des différentes ellipses prédites autour du centre x_{c_t}, y_{c_t}.

Lors de la phase de correction, les probabilités pignistiques issues du modèle évidentiel sont binarisées. Une opération morphomathématique de remplissage s'applique alors sur cette image afin d'exhiber une forme autour du centre (x_{c_t}, y_{c_t}) dont le contour est extrait (en blanc sur la figure 7d). Ensuite, une mesure elliptique de ce contour basée sur une méthode d'ajustement par moindres carrés (Fitzgibbon *et al.*, 1999) (en rouge sur la figure 7d) définit l'ellipse dite de mesure dont les paramètres sont notés : $\left[\hat{l}_t, \hat{h}_t, \hat{\theta}_t\right]$. Dès lors, l'étape de correction évalue le poids d'importance $\tilde{\omega}_t^{(n)}$ comme inversement proportionnel à la distance euclidienne entre l'ellipse prédite et l'ellipse de mesure. Le critère du maximum de vraisemblance permet de sélectionner la particule la plus significative. Parmi l'ensemble des ellipses prédites (figure 7c) l'algorithme sélectionne celle (figure 7e) dont la taille et l'orientation sont les plus proches de l'ellipse de mesure (en rouge sur la figure 7d). Le résultat du filtrage (figure 7e) est correct bien que la mesure soit perturbée (figure 7d) par un artefact issu du fond. En effet ce résultat dépend aussi de la forme et de l'orientation des particules lors de la phase de prédiction, donc du paramétrage du modèle dynamique du filtre. La valeur de la variance Σ_{max} imposée dans le modèle assure que les particules s'écartent faiblement des composantes du vecteur d'état estimé à l'instant $t-1$.

6. Analyse des résultats du suivi vidéo

L'évaluation des performances des algorithmes de suivi nécessite à la fois la définition d'un critère quantitatif tel que la précision, la robustesse, ou la vitesse d'exécution et une vérité terrain, c'est-à-dire un ensemble de données codant image par image la position réelle du visage obtenu par une expertise humaine. L'obtention de cette vérité est relativement subjective et laborieuse. Ici, nous considérons le visage présent dans l'image lorsqu'une part suffisante de la peau de celui-ci est visible. Les cheveux ne sont pas pris en compte. Le visage peut se situer de face, mais aussi de profil (figure 3). Lors d'une occultation totale le visage est considéré comme absent.

6.1. Évaluation qualitative

Nous illustrons par deux séquences le comportement de l'algorithme en présence d'occultations totales ou partielles (figure 8), de variations de pose et d'éléments perturbateurs tels que les mains à proximité de la tête (figure 9).

Figure 8. Suivi de visage pour la séquence No. 1 : a) Boîte englobante (en rouge) délivrée par le détecteur de VJ ; b) Probabilité pignistique issue du modèle de visage pour $\gamma = 0,5$ et $\eta = 0,5$; c) Particules du filtre particulaire lors de la phase d'estimation du centre de X_t ; d) Ellipse résultat du filtrage particulaire

Sur les images im_{57} et im_{71} de la figure 8a, les masses de VJ renforcent le contenu informatif associé à la zone visage tout en affaiblissant celui extérieur à la boîte englobante. Les probabilités pignistiques sont les plus significatives (en blanc figure 8b) sur la zone visage et non sur les autres régions de couleur de peau (bras, main, cou).Le détecteur de VJ est mis en défaut sur l'image im_{80} en raison de la rotation du visage et de la présence d'un artefact ressemblant à un visage dans le fond (affiche en haut à droite). L'ajustement du paramètre de compromis ($\eta = 0,5$) ainsi que l'affaiblissement de γ par le conflit (rétroaction) élèvent le contenu informatif associé à l'information teinte et le visage demeure correctement détecté. Aucune boîte englobante n'est délivrée par le détecteur de VJ sur les images $im_{60}, im_{66}, im_{69}$. Dans ce cas, nous imposons $\gamma = 0$: seule l'information de couleur est considérée. C'est pourquoi, en présence d'occultation totale (im_{66}), l'ellipse solution se positionne sur la main de la personne. La faible composante uniforme ($\nu = 0,1$) dans le modèle dynamique du

filtre (équation (22)) assure un raccrochage du filtre dès qu'une particule candidate se place à nouveau sur une zone de visage (im_{71}).

Dans la séquence de la figure 9, l'utilisateur se situe de face et retire ses lunettes. Ses deux mains viennent se positionner près du visage à l'image im_{82} avant de commencer à s'en écarter à l'image im_{113}. La personne est vêtue d'un tee-shirt de couleur assez proche de la teinte chair. Le détecteur de VJ fournit une boîte englobante fiable durant toute cette séquence où le visage est de face. Le résultat du suivi est parfait lorsque les mains effleurent le visage sur l'image im_{82}. Lorsque les mains sont en contact franc avec celui-ci (im_{110} à im_{113}), l'ellipse est déviée en orientation : l'estimation du centre est correcte mais la pose est erronée. Lorsque les mains s'éloignent du visage, l'algorithme ne les suit pas et se repositionne correctement sur le visage à l'image im_{120}. La présence d'éléments perturbateurs altère donc partiellement l'estimation de la pose et de la taille mais perturbe peu le suivi en position.

Figure 9. Suivi de visage sur la séquence No. 2 : ellipse résultat du filtrage particulaire

6.2. Évaluation quantitative

Afin de quantifier les performances du suivi dans des contextes variés et sur un lot de données statistiquement significatif, nous avons détouré manuellement le visage dans 500 images de 6 séquences vidéo (ceci à raison de 1 image/sec afin d'explorer chaque vidéo avec une cadence adaptée à l'applicatif). Trois séquences supplémentaires complètent celles déjà présentées : dans la séquence No. 4, l'utilisateur de couleur de peau noire, s'éloigne de la caméra ; dans la séquence No. 5, une personne se déplace en arrière plan (fond complexe) ; enfin la séquence No. 6 présente un visage de jeune femme maquillée. Les pixels situés à l'intérieur du détourage représentent la vérité terrain VT. L'algorithme de suivi délivre une ellipse notée ROI (pour *Region Of Interest*) issue du filtre particulaire. Les pixels vrais positifs VP appartiennent à l'intersection $VT \cap ROI$ tandis que les faux positifs (FP) sont inscrits dans la ROI mais situés à l'extérieur de VT. Pour évaluer la performance du suivi nous utilisons deux mesures classiques, la précision et le rappel définis par :

$$\text{Précision} = \frac{VP}{ROI} \qquad \text{Rappel} = \frac{VP}{VT}. \tag{24}$$

La précision est donc le rapport des mesures correctes (VP) sur l'ensemble des mesures ($ROI = VP \cup FP$). Le rappel représente le rapport des mesures correctes

sur l'ensemble de la vérité terrain (car $VT = VP \cup FN$). Les faux négatifs (FN) sont extérieurs à la ROI mais appartiennent à VT. Précision et rappel sont calculés individuellement sur chaque image. Ils sont ensuite moyennés sur chaque séquence pour mettre en évidence précisément l'influence des paramètres dans chaque contexte, et enfin sur l'ensemble des données afin de cerner le comportement global de l'algorithme. Ces mesures conduisent à l'élaboration de courbes COR (Caractéristiques Opérationnelles de Réception) d'abscisse $x = (1-$Précision$)$ et d'ordonnée $y =$ Rappel, obtenues pour différentes valeurs des paramètres d'influence (γ et η). Le point de la courbe le plus proche du point idéal ($x = 0; y = 1$) correspond à la meilleure valeur du paramétrage.

Par cette étude nous présentons la sensibilité de la méthode aux paramètres de compromis η et de fiabilité γ. La rétroaction par affaiblissement (section 4.6) n'est pas mise en œuvre ici afin d'exhiber clairement l'influence de γ sur l'algorithme.

La figure 10 indique que lorsque seule l'information couleur est considérée (*i.e.*, $\gamma = 0$), la qualité globale du suivi est médiocre même si le paramètre de compromis η permet une modulation large des performances (Précision $\in [0,42; 0,53]$, Rappel $\in [0,22; 0,55]$). Les performances optimales s'obtiennent pour $\gamma \approx 0,3$ associé à un ajustement de η tel que $0,2 \leq \eta \leq 0,6$. Pour $\gamma = 0,7$, l'information de VJ, supposée très fiable, joue le rôle majeur dans le modèle évidentiel du visage. Le paramètre de compromis η influence alors très faiblement les performances du suivi : la précision varie peu ($\in [0,63; 0,68]$) et le rappel est quasi-constant ($\approx 0,6$).

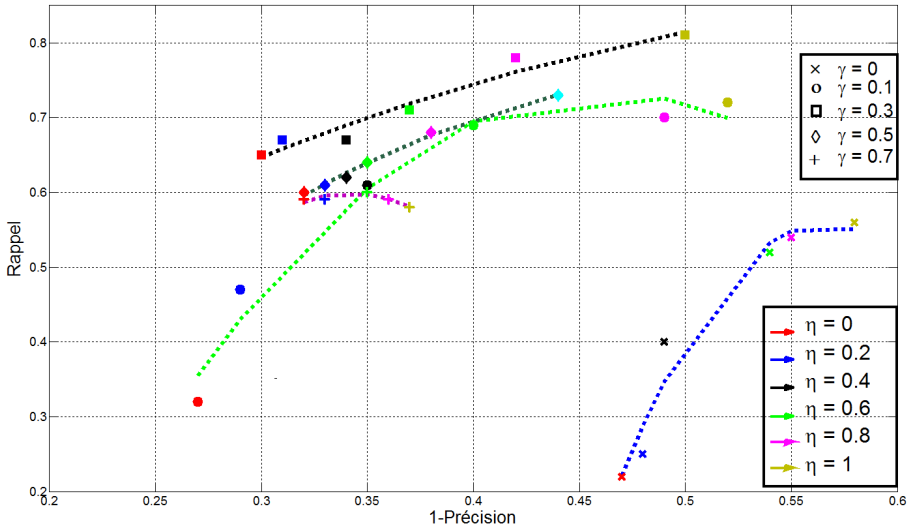


Figure 10. Courbes COR globales établies sur l'ensemble des données pour le jeu de valeurs des deux paramètres : $\gamma \in [0; 0,7]$ et $\eta \in [0; 1]$

Pour les séquences où le visage est souvent de profil ou occulté, les masses couleur sont plus significatives que celles de VJ. Les performances les plus faibles s'observent

pour $\gamma = 0,7$ (figure 11). Elles sont optimales en considérant partiellement l'information de VJ ($\gamma = 0,1$) et ceci pour une valeur de η correctement ajustée, *i.e.*, $\eta \approx 0,6$. Ainsi, l'action conjointe d'un compromis sur les masses couleur et la modération des informations de VJ conduit à une optimisation des performances (précision et rappel proches de 80 %). Ces résultats sont comparables à ceux des classifieurs standards qui présentent des taux de détection variant entre 70 et 80 % (Castrillón *et al.*, 2011).

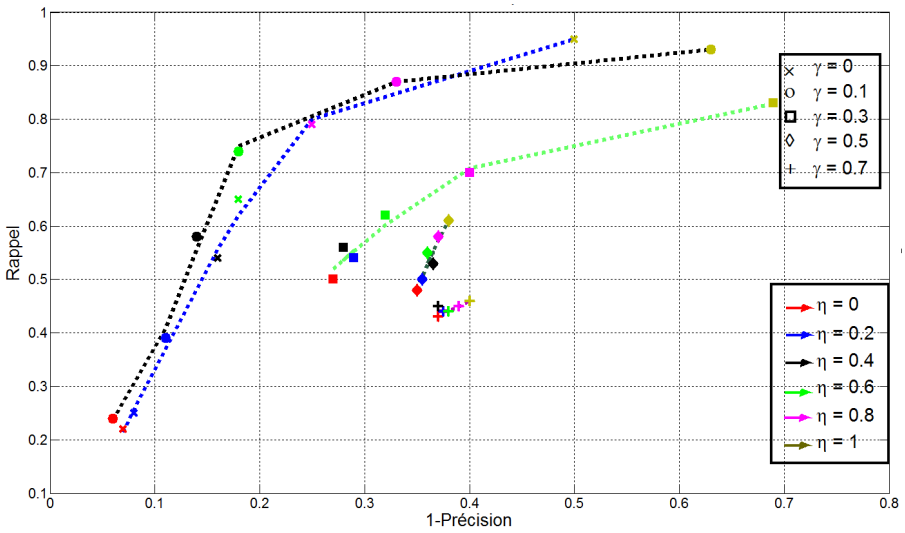


Figure 11. Courbes COR établies sur la séquence No. 1 pour les valeurs des paramètres : $\gamma \in [0; 0,7]$ *et* $\eta \in [0; 1]$

Un autre critère important dans l'évaluation des performances du suivi est l'erreur de localisation notée : $\varepsilon = \sqrt{(x_{VT_t} - x_{c_t})^2 + (y_{VT_t} - y_{c_t})^2}$.

x_{VT_t}, y_{VT_t} sont les coordonnées du centre de gravité du visage issu de la vérité terrain (VT), et x_{c_t}, y_{c_t} sont les paramètres de position de l'ellipse (ROI). L'erreur de localisation moyenne ainsi que l'écart type, calculés sur l'ensemble des images de la séquence, sont aussi des critéres d'évaluation couramment utilisés. A des fins de comparaison nous avons testé notre méthode sur la séquence *David Indoor* utilisée dans de nombreux articles récents (Babenko *et al.*, 2011). Le contexte de cette séquence est proche de celui rencontré dans notre application : changements de pose du visage, variations d'éclairage, présence d'éléments perturbateurs (l'opérateur enlève puis remet ses lunettes). Avec une erreur de position inférieure à 20 pixels sur la majorité de la séquence (figure 12), notre algorithme dépasse les performances du meilleur algorithme (MILTrack) évalué sur la figure 13 . Notre approche est défaillante localement sur les images 76 puis 181 à 187 *i.e.,* lorsque l'algorithme se positionne sur un artefact. L'erreur de localisation moyenne est de 16,7 pixels avec un écart type de 26,8 pixels. Ces performances sont du même ordre que celles présentées dans la littérature à propos du suivi de visage par filtrage particulaire où l'erreur moyenne de suivi est de 22,4

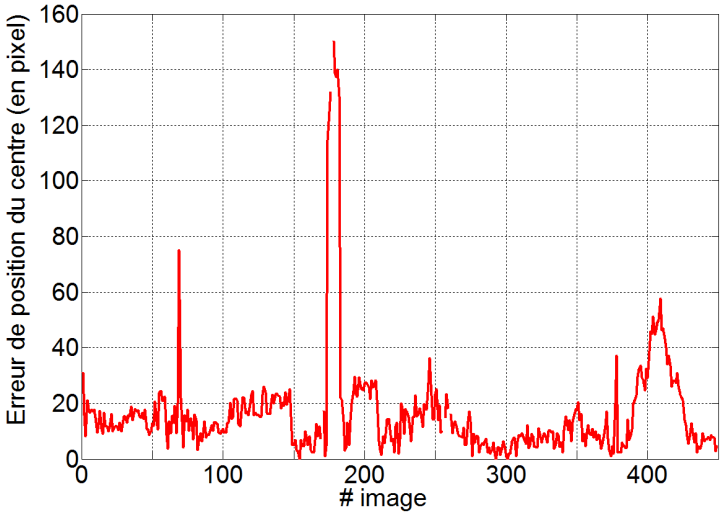


Figure 12. Résultat de suivi (erreur de position du centre) sur la séquence David Indoor, *avec* $\eta = 0,5$

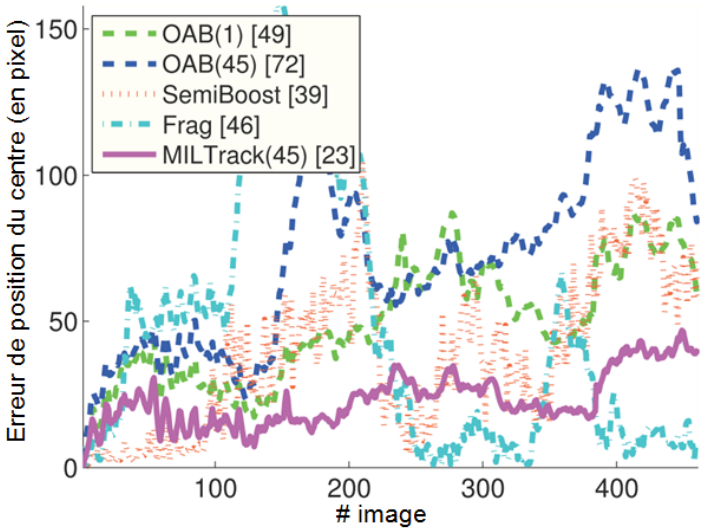


Figure 13. Résultats de suivi (erreur de position du centre) de différents algorithmes sur la séquence David Indoor *d'après Babenko*

pixels avec l'algorithme de Condensation et de 16,3 pixels pour un filtre particulaire adaptatif APF (Zheng, Bhandarkar, 2009).

6.3. Asservisement visuel

L'asservissement visuel (figure 14) contrôle les trois degrés de liberté (panoramique, inclinaison, zoom) de la caméra PTZ. Le but est de maintenir le visage au centre de l'image avec une taille raisonnable (environ 20 % de la taille de l'image). Les lois de commande de la poursuite (tâche de recentrage) et du zoom sont élaborées par une approche classique (Crétual, Bouthemy, 1998).

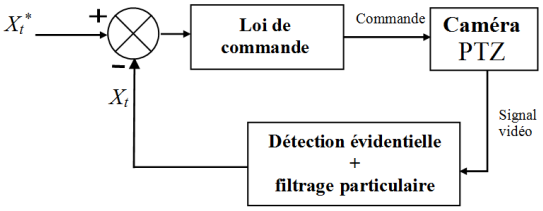


Figure 14. Synoptique de l'asservissement visuel avec : $X_t^* = [0, 0, 120, 100, 0]$ consigne de l'asservissement et X_t vecteur d'état issu du filtre particulaire

La figure 15 illustre le comportement de l'asservissement visuel. Sur l'image im_{14} le visage est situé sur la gauche du champ visuel. Sur l'im_{18}, l'action conjointe du déplacement panoramique et du zoom recentre le visage dans le plan image. De l'im_{20} à l'im_{24} l'opérateur se déplace vers le fond de la salle. Dès lors le zoom et le mouvement vertical de la caméra (*tilt*) recadrent le visage au centre de l'image avec une taille désirée.

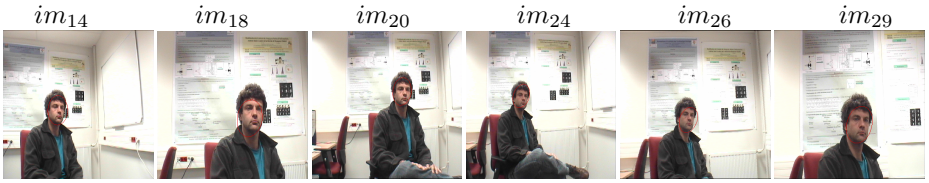


Figure 15. Résultat du suivi avec un asservissement visuel de la caméra en position (pan et tilt) et un contrôle du zoom

7. Discussion

Nous avons proposé une méthode originale de détection et suivi de visage basée sur une modélisation crédibiliste couplée à une technique de filtrage particulaire classique de type *bootstrap*. Notre contribution théorique porte sur un opérateur de compromis pour réaliser la fusion de sources d'information. Pour l'application au suivi

de visages, les taux de précision et de rappel peuvent atteindre 80 % avec un réglage adéquat des paramètres, et ceci sans avoir à construire de base d'apprentissage lourde, ce qui constitue l'originalité de notre approche. La simplicité des calculs rend l'approche utilisable en temps-réel vidéo (même si ce n'est pas le cas de notre prototype actuel développé avec Matlab et LabVIEW pour faciliter le prototypage et les simulations). Les résultats statistiques (section 6.2) confirment les constatations qualitatives évoquées en section 6.1. Cependant les valeurs optimales des paramètres ($\gamma = 0,3$ et $\eta = 0,6$) sont déduites d'un moyennage des résultats. Par conséquent, cette étude évalue mal le dimensionnement des paramètres pour une variation transitoire du contexte sur une partie de la séquence. Un ajustement dynamique des paramètres s'avère nécessaire pour améliorer la robustesse du suivi. C'est pourquoi nous avons proposé en section 4.6 un réglage de γ en fonction du conflit. De plus, le contrôle de l'interaction entre la fusion et le filtre particulaire demeure délicat. En effet, un apport trop important ou mal adapté de « flou » risque de dénaturer complètement le modèle de visage et conduire à une divergence du filtre. En tant que perspective, la mise à jour dynamique du paramètre η de compromis, en fonction d'une mesure de la dépendance des sources (intercorrélation ou cohérence) est une question à aborder. La modélisation des fonctions de masse pourrait aussi être améliorée en utilisant un apprentissage sur une vérité terrain sommaire pour estimer *a priori* les taux VP, FP, VN, FN et maximiser les croyances.

Bibliographie

Appriou A. (1999, novembre). Multisensor signal processing in the framework of the theory of evidence. In *Application of mathematical signal processing techniques to mission systems, Research and Technology Organisation (lecture series 216)*, p. 5.1-5.31.

Arulampalam M., S.Maskell, Gordon N., Clapp T. (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, vol. 50, n° 2, p. 174-188.

Babenko B., Yang M., Belongie S. (2011). Robust object tracking with online multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, n° 8, p. 1619-1632.

Castrillón M., Déniz O., Hernández D., Lorenzo J. (2011). A comparison of face and facial feature detectors based on the Viola-Jones general object detection framework. *Machine Vision and Applications*, vol. 22, p. 481-494.

Comaniciu D., Ramesh V., Meer P. (2003). Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, n° 5, p. 564-575.

Cootes T. F., Edwards G. J., Taylor C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, n° 6, p. 681-685.

Cootes T. F., Taylor C. J. (1992). Active shape models -'smart snakes'. In *Proceedings of British machine vision conference*, p. 266-275.

Crétual F. C. A., Bouthemy P. (1998). Complex object tracking by visual servoing based on 2D image motion. In *International conference on pattern recognition*, vol. 2, p. 1251-1254. Brisbane, Australie.

Dempster A. (1967). Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, vol. 38, p. 325-339.

Denœux T. (1995). A k-nearest neighbour classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics*, vol. 25, n° 5, p. 804-813.

Denœux T. (1997). Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition*, vol. 30, n° 7, p. 1095-1107.

Denœux T. (2008). Conjunctive and disjunctive combination of belief functions induced by non distinct bodies of evidence. *Artificial Intelligence*, vol. 172, p. 234-264.

Denœux T., Smet P. (2006). Classification using belief functions: the relationship between the case-based and model-based approaches. *IEEE Transactions on Systems, Man and Cybernetics B*, vol. 36, n° 6, p. 1395-1406.

Faux F. (2009). *Détection et suivi de visage par la théorie de l'évidence*. Thèse de doctorat non publiée, université de Pau et des Pays de l'Adour, Anglet, France.

Faux F., Luthon F. (2006). Robust face tracking using colour Dempster-Shafer fusion and particle filter. In *The 9th international conference on information fusion (FUSION'06)*, p. 1-7. Firenze, Italy.

Fitzgibbon A., Pilu M., Fisher R. (1999, mai). Direct least squares fitting of ellipses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, n° 5, p. 476-480.

Florea M. C., Jousselme A.-L., Bossé E., Grenier D. (2009). Robust combinaison rules for evidence theory. *Information Fusion*, vol. 10, p. 183-197.

Hammal Z., Couvreur L., Caplier A., Rombaut M. (2007, décembre). Facial expression classification: An approach based on the fusion of facial deformations using the transferable belief model. *International Journal of Approximate Reasoning*, vol. 46, n° 3, p. 542-567.

Hjelmås E., Low B. (2001, septembre). Face detection: A survey. *Computer Vision and Image Understanding*, vol. 83, p. 236-274.

Huang L.-L., Shimizu A., Kobatake H. (2005). Robust face detection using Gabor filter features. *Pattern Recognition Letters*, vol. 26, n° 11, p. 1641-1649.

Isard M., MacCormick J. (2001). BraMBLe: A Bayesian multiple-blob tracker. In *IEEE international conference on computer vision (ICCV)*, p. 34-41.

Kallel A., Le Hégarat-Mascle S. (2009). Combination of partially non-distinct beliefs: The cautious-adaptive rule. *International Journal of Approximate Reasoning*, vol. 50, n° 7, p. 1000-1021.

Klein J., Lecomte C., Miché P. (2010). Hierarchical and conditional combination of belief functions induced by visual tracking. *International Journal of Approximate Reasoning*, vol. 51, n° 4, p. 410-428.

Knothe R., Amberg B., Romdhani S., Blanz V., Vetter T. (2011). *Handbook of face recognition, Chapter Morphable models of faces*. Edited by Stan Li and Anil Jain, Springer-Verlag.

Liévin M., Luthon F. (2004). Nonlinear color space and spatiotemporal MRF for hierarchical segmentation of face features in video. *IEEE Transactions on Image Processing*, vol. 13, n° 1, p. 63-71.

Luthon F., Beaumesnil B. (2004). Color and R.O.I with JPEG2000 for wireless videosurveillance. In *International conference on image processing (ICIP' 04)*, p. 3205-3208.

Luthon F., Beaumesnil B., Dubois N. (2010). LUX color transform for mosaic image rendering. In *Proceedings of the 17th IEEE international conference on automation, quality and testing, robotics (AQTR 2010)*, vol. 3, p. 93-98. Cluj-Napoca, Romania.

Martin A., Osswald C. (2007). Towards a combinaison rule to deal with partial conflict and specificity in belief functions theory. In *International conference on information fusion (FUSION'07)*, p. 9-12. Québec, Canada.

Muñoz-Salinas R., Medina-Carnicer R., Madrid-Cuevas F. J., Carmona-Poyato A. (2009). Multi-camera people tracking using evidential filters. *International Journal of Approximate Reasoning*, vol. 50, n° 5, p. 732-749.

Phung S., Bouzerdoum A., Chai D. (2005). Skin segmentation using color pixel classification: Analysis and comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, n° 1, p. 148-154.

Pichon F., Denœux T. (2009). Interpretation and computation of alpha-junctions for combining belief functions. In *6th international symposium on imprecise probability: Theories and applications (ISIPTA '09)*, p. 1-10. Durham, United Kingdom.

Pérez P., Vermaak J., Blake A. (2004). Data fusion for visual tracking with particles. *Proceedings of IEEE*, vol. 92, n° 3, p. 495-513.

Ramasso E., Panagiotakis C., Rombaut M., Pellerin D. (2010). Belief scheduler based on model failure detection in the TBM framework. Application to human activity recognition. *International Journal of Approximate Reasoning*, vol. 51, n° 7, p. 846-865.

Rathi Y., Vaswani N., Tannenbaum A., Yezzi A. (2007). Tracking deforming objects using particle filtering for geometric active contours. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, n° 8, p. 1470-1475.

Sakai T., Nagao M., Fujibayashi S. (1969). Line extraction and pattern recognition in a photograph. *Pattern Recognition*, vol. 1, p. 233-248.

Shafer G. (1976). *A mathematical theory of evidence*. NJ, Princeton University Press.

Sigal L., Sclaroff S., Athitsos V. (2004). Skin color-based video segmentation under time-varying illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, n° 7, p. 862-877.

Smarandache F., Dezert J. (2009). *Advances and applications of DSmT for information fusion, collected works* (vol. 3). American Research Press.

Smets P. (1986). Bayes' theorem generalized for belief functions. In *European Conference on Artificial Intelligence (ECAI'86)*, vol. 2, p. 169-171. Brighton, UK.

Smets P. (1990). The combination of evidence in the Transferable Belief Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, n° 5, p. 447-458.

Smets P. (1993). Belief functions: the disjunctive rule of combinaison and the Generalized Bayesian Theorem. *International Journal of Approximate Reasoning*, vol. 9, p. 1-35.

Smets P. (1995). The canonical decomposition of a weighted belief. In *International joint conference on artificial inteligence*, p. 1896-1901. San Mateo, CA, USA, Morgan Kaufman.

Smets P., Kennes R. (1994). The Transferable Belief Model. *Artificial Intelligence*, vol. 66, n° 2, p. 191-234.

Soriano M., Martinkauppi B., Huovinen S., Laaksonen M. (2003). Adaptive skin color modeling using the skin locus for selecting training pixels. *Pattern Recognition*, vol. 36, n° 3, p. 681-690.

Vezhnevets V., Sazonov V., Andreeva A. (2003). A survey on pixel based skin color detection techniques. In *Proceedings of Graphicon*, p. 85-92. Moscow, Russia.

Viola P., Jones M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, vol. 1, p. 511-518.

Viola P., Jones M. (2003). *Fast multi-view face detection*. Rapport technique. Mitsubishi Electric Research Laboratories.

Yaghlane B. B., Smets P., Mellouli K. (2000). Independence concepts for belief functions. In *8th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU)*, vol. 1, p. 357-364. Madrid, Spain.

Yang M.-H., Kriegman D., Ahuja N. (2002). Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, n° 1, p. 35-58.

Yilmaz A., Javed O., Shah M. (2006). Object tracking: A survey. *ACM Computing Surveys*, vol. 38, n° 4, p. 1-45.

Zheng W., Bhandarkar S. M. (2009). Face detection and tracking using a boosted adaptive particle filter. *Journal of Visual Communication and Image Representation*, vol. 20, p. 9-27.

Zouhal L., Denoeux T. (1998). An evidence-theoric k-NN rule parameter optimization. *IEEE Transactions on Systems, Man and Cybernetics - Part C*, vol. 28, n° 2, p. 263-271.

Reçu le 28/02/2011
Accepté le 03/10/2011

Francis Faux *est docteur de l'université de Pau et des Pays de l'Adour depuis 2009. Il est professeur agrégé en Génie électrique. Il est membre du laboratoire d'informatique LIUPPA. Ses recherches portent sur les fonctions de croyance et leur application à la détection et au suivi de visage.*

Franck Luthon *est professeur de l'université de Pau et des Pays de l'Adour depuis 2000. Il est depuis 2007 le chef du département Génie industriel de l'IUT de Bayonne Pays basque à Anglet. Ses travaux de recherche portent sur le traitement d'images, les systèmes électroniques et l'enseignement technologique à distance.*