Face Detection Using the Theory of Evidence

Franck Luthon

CHAPTER 9

Face Detection Using the Theory of Evidence*

Franck Luthon†
Computer Science Laboratory LIUPPA
University of Pau Pays Adour UPPA, Anglet, France

Abstract

Abstract: Face detection and tracking by computer vision is widely used for multimedia applications, video surveillance or human computer interaction. Unlike current techniques that are based on huge training datasets and complex algorithms to get generic face models (*e.g.* active appearance models), the proposed approach using evidence theory handles simple contextual knowledge representative of the application background, thanks to a quick semi-supervised initialization. The transferable belief model is used to counteract the incompleteness of the prior model due to a lack of exhaustiveness in the learning stage.

The method consists of two main successive steps in a loop: detection, then tracking. In the detection phase, an evidential face model is built by merging basic beliefs carried by a Viola-Jones face detector and a skin color detector. The mass functions are assigned to information sources computed from a specific nonlinear color space. In order to deal with color information dependence in the fusion process, a cautious combination rule is used. The pignistic probabilities of the face model guarantee the compatibility between the belief framework and the probabilistic framework. They are the inputs of a bootstrap particle filter which yields face tracking at video rate. The proper tuning of the few evidential model parameters leads to tracking performance in real-time. Quantitative evaluation of the proposed method gives a detection rate reaching 80%, comparable to what can be found in the literature. Nevertheless, the proposed method requires a scanty initialization only (brief training) and allows a fast processing.

Keywords: Belief function, Cautious rule, Classification, Computer vision, Conjunctive rule, Dempster-Shafer, Face tracking, Fusion of information, LUX color space, Mass set, Particle filter, Pattern recognition, Pignistic probability, Region of interest, Skin hue, Source of information, Transferable belief model, Uncertainty management, Viola-Jones detector, Visual servoing.

1 Introduction

Real-time face detection and tracking in video sequences has been studied for more than twenty years by the computer vision and pattern recognition community, owing to the multiplicity

*Book chapter in "Advances in Face Image Analysis: Theory and Applications" (F. Dornaika Ed.), 2015 Bentham Science Publishers, pp.169-200

†Address: IUT GIM, 2 allée du parc Montaury, 64600 Anglet, France; E-mail: Franck.Luthon@univ-pau.fr

of applications: teleconferencing, closed-circuit television (CCTV), human machine interface, robotics. Despite the ongoing progress in image processing and the increase in computation speed of digital processors, the design of generic and robust algorithms is still the object of active research. Indeed, face image analysis (either detection, recognition or tracking) is made difficult by the variability of appearance of this deformable moving object due to many factors: individual morphological differences (nose shape, eye color, skin color, beard), presence of visual artifacts (glasses, occlusions, make-up), illumination variations (shadow, highlight), facial expression changes depending on context (social, cultural, emotional). Those are difficult to model and do not easily cope with real-time implementations. Moreover, the scene background might disturb detection, in case of foreground-background similarity or background clutter.

To handle the face specificity, a semi-supervised learning method is presented here, where the user selects manually a zone of the face in the first image of the video. This rapid initializing step constitutes the learning stage which yields simply a prior model for face class and background class. It is however dependent on the user subjectivity while selecting the face zone and it suffers from incompleteness because of a lack of exhaustiveness of this short training. In this context, a probabilistic modeling is not relevant. Therefore the proposed approach is based on belief functions: indeed the transferable belief model (TBM) is well suited to model partial knowledge in a complex system [1]. It was successfully applied to classification of emotions and facial expressions, or to human activity recognition [2].

The goal of the application is to automatically track the face of a person placed in the field of view of a motorized pan-tilt-zoom camera (or simply a webcam). The tracking technique should be as robust as possible to occlusions, pose, scale, background and illumination changes. It should take control of the camera to perform an automatic centering of the face in the image plane during the whole video sequence. The algorithm consists of two main steps: face detection, then tracking (Fig. 1). An elliptical region of interest (ROI) including the face is

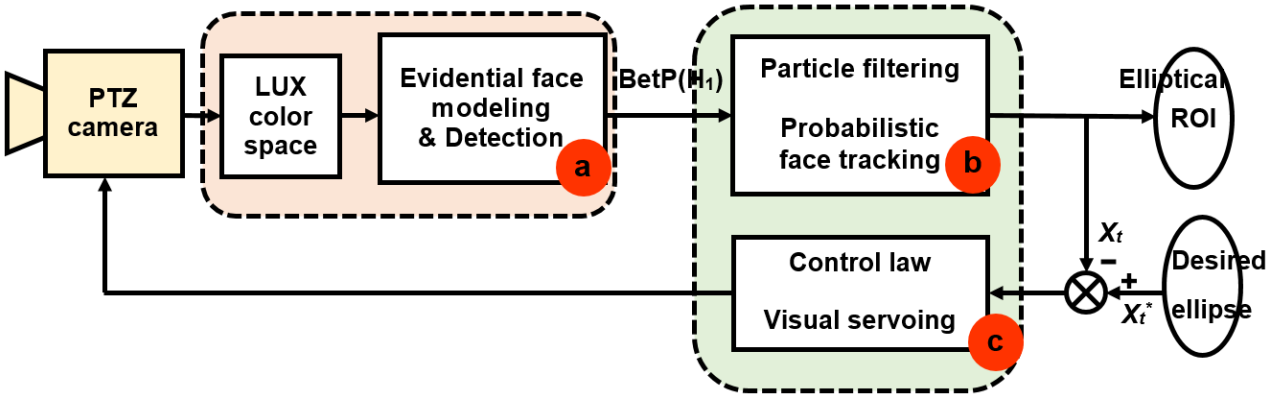


Figure 1: Overview of the processing with feedback loop: a) face detection by evidential modeling; b) face tracking by particle filtering; c) camera control by visual servoing

computed by particle filtering, and held at the center of the image by visual servoing. The context of application is indoor environments, typically a laboratory or an office. As regards acquisition conditions, the distance between user and sensor ranges from about 50 cm to a few meters. Ordinary lighting conditions prevail (uncontrolled illumination context), possibly in the presence of additional light sources, like a desk lamp or the influence of outside light entering through a window.

After a state of the art about face detection, the theory of belief functions is briefly exposed.

The proposed evidential model for face detection is then detailed in the application section. The tracking with particle filter and visual servoing of the camera are described. Performance analysis, both qualitative and quantitative, is presented. The chapter ends with a discussion.

2 State of the Art

Face detection methods may be grouped into two categories differing in the way of processing prior information [3]. It is also worth making a difference between detection methods dedicated to still images, where complex algorithms can be used, and methods dedicated to video sequences where the computation cost is of major concern for real-time processing.

Feature-based methods use as primitives local properties of the face. The so-called low-level analysis (or early vision) handles the information obtained directly from the pixels such as luminance or color, or indirectly after computation of edges, motion or texture from pixels neighborhood. Color is a key feature because of its invariance with respect to translation, rotation or scale. Nevertheless skin color is made of a great variety of hues depending both on the person and on illumination conditions (shadowy, pale, overexposed skin). Therefore the design of a robust hue detector requires the choice of a proper colorimetric space [4, 5]. Anyway, the primitives and estimates induced from low level analysis remain ambiguous (ill-posed problem). To validate the detection, additional information is required. The feature analysis is based both on the knowledge of an adequate prior model (a priori constraints, contextual information) and on measurements of normalized distances and angles derived from the individual description of face parts (eyes, nose, mouth). With this first family of methods, the processing is potentially fast, as few training is necessary. The methods for parameter extraction are often specific to the context at hand, and are designed empirically based on color, edge or motion cues. The parameter tuning relies on heuristics.

Holistic approaches, by contrast, address the detection problem as a global identification problem (high level analysis). The key-point is to compare a test image with a generic face model and to deduce if there is resemblance or not. Priors about geometrical or physiological specificities are discarded to limit the modeling errors due to incomplete and imprecise knowledge of the face. These methods rely on the learning of a generic face model from a database of samples as much complete as possible. Linear methods of subspaces, statistical approaches (Monte-Carlo methods), support vector machines (SVM) or neural networks can be used. An important step forward was made when the first holistic face detector with real-time capability was proposed by Viola and Jones [6]. It is based on an automatic selection of *2D* Haar wavelet filters applied to monochrome images and it uses a cascade of boosted classifiers with increasing complexity. The active shape models (ASM) introduced by Cootes and Taylor [7], are deformable models which depict the highest level of appearance of face features. Once initialized near a facial part, the model modifies its local characteristics (outline, contrast) and evolves gradually in order to take the shape of the target feature. The active appearance models (AAM) are an extension of the ASM by Cootes *et al.* [8]. The use of the third dimension, namely the temporal one, can lead to real-time *3D* deformable face models varying according to morphological parameters. Therefore, this second family of methods provides flexibility with respect to different contexts such as number of faces in the scene or type of lighting. However, these methods are strongly dependent on the choice and quality of the face models: they require a huge training dataset to be sufficiently representative. Whatever the face database used, it is of course rarely exhaustive and its choice remains a full problem. In addition, algorithms are

complex and induce heavy computation cost.

Here, two complementary face detection methods will collaborate in a fusion process [9]. One of them refers to induction, the other one refers to deduction, as illustrated in Fig. 2, and as explained in the audioslide available online [10]. First, among the feature-based methods, a

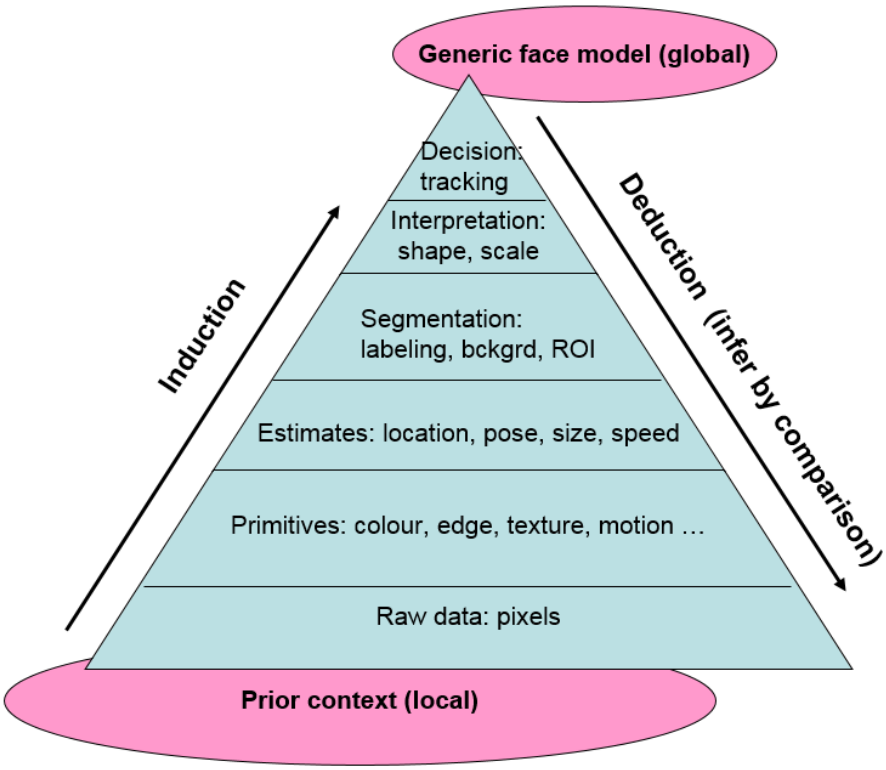


Figure 2: Image processing pyramidal framework.

skin color discriminating detector is chosen. Indeed, its properties of invariance with respect to motion allow to track a face whatever its pose during the video sequence. Second, among holistic approaches, the Viola-Jones (VJ) face detector is chosen due to its real-time ability and the availability of an open source implementation. It provides a target container (rectangular bounding box surrounding the face) highly reliable in the case of front-view faces. However as the authors [6] have made their classifier public but not their training, the classifier used here was not trained on our data. We will see that the proposed method, which applies evidence theory, circumvents this point. The key point, then, is the proper fusion of information delivered by the two detectors.

3 Theory of Belief Functions

3.1 Mass Sets

The theory of belief functions, also called Dempster-Shafer theory or evidence theory, dates back to the 1970s. Inspired by the upper and lower probabilities first studied by Dempster [11], then by Shafer [12], it may be interpreted as a formal quantitative model of degrees of belief. This theory increases modeling flexibility and allows to solve complex problems since: (i) it does not require complete prior knowledge about the problem at hand, and (ii) it offers

the possibility to distribute the belief in compound hypotheses (and not only on singletons as is the case in the probability modeling). It was successfully applied to image fusion for medical application in magnetic resonance imaging [13].

The basic concept of the evidence theory is the mass function which characterizes the opinion of an agent about a question or the state of a system. The frame of discernment Ω is the finite set of answers (called focal elements) to this question (typ. $\Omega = \{H_1, H_2\}$ in the simple binary case with two hypotheses only). A mass function $m(.)$ is an application of the set 2^Ω (typ. $2^\Omega = \{\emptyset, H_1, H_2, \Omega\}$) towards the real interval $[0, 1]$ which satisfies:

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

This constraint guarantees a commensurability between several mass sets. The mass $m(A)$ is the part of belief placed strictly in A.

Belief $Bel(.)$, plausibility $Pl(.)$ and commonality $q(.)$ are three common measures derived from the mass function. They are defined, $\forall A \subseteq \Omega, A \neq \emptyset$:

$$Bel(A) = \sum_{B \subseteq A, B \neq \emptyset} m(B) \quad ; \quad Pl(A) = \sum_{B \cap A \neq \emptyset} m(B) \quad \text{and} \quad q(A) = \sum_{B \supseteq A} m(B) \tag{2}$$

For the empty set: $Bel(\emptyset) = Pl(\emptyset) = 0$ and $q(\emptyset) = 1$. The interval $[Bel(A), Pl(A)]$ is the confidence interval that represents the lower and upper bounds of the likelihood of the subset A. The maximum of plausibility is often used as decision criterion.

A simple mass set, or elementary state of belief, is defined by a mass function m so that $A \subset \Omega$ is set along with a weight $w \in [0, 1]$:

$$\begin{aligned} m(A) &= 1 - w, \\ m(\Omega) &= w, \\ m(B) &= 0, \quad \forall B, B \neq A, B \neq \Omega. \end{aligned} \tag{3}$$

Denoted as shortcut $m = A^w$, it represents the belief put in Ω instead of A. For any A, A^1 ($w = 1$) is the vacuous simple mass function ($m(A) = 0$), whereas A^0 ($w = 0$) is the categorical simple mass function ($m(A) = 1$).

A complex state of belief may be modeled with a set of independently weighted hypotheses. This is called canonical decomposition: any non categorical mass function m (*i.e.* when Ω is one of the a focal elements, that is when $m(\Omega) \neq 0$) may be expressed as the conjunctive combination (defined in Eq. 5) of simple mass sets: $m = \textcircled{\cap}_{A \subset \Omega} A^{w(A)}$, where the weights are computed from commonalities: $\log w(A) = -\sum_{B \supseteq A} (-1)^{|B|-|A|} \log q(B)$.

3.2 Modeling of Mass Functions

The mass function modeling is a non trivial problem. Difficulty grows if one wants to assign beliefs to compound hypotheses (*e.g.* $H_1 \cup H_2 \cup H_3$). One may distinguish models based on distance computations, stemming from pattern recognition [14] where mass functions are built from available learning vectors, and models using likelihood computations, stemming form Bayesian probabilistic approach. These last ones decompose into global [15] and separable methods.

Separable methods build a belief function for each hypothesis H_i of the frame of discernment. They rely on an initial learning for estimating conditional probabilities $P(s_j|H_i)$ where s_j

represents an observation of the source j and H_i is one of the hypotheses. This approach was first proposed by Smets [16] then used by Appriou for multisensor signal fusion [17]. Appriou's model #1 is derived from the generalized Bayesian theorem:

$$\begin{cases} m_{ij}(H_i) &= 0, \\ m_{ij}(\overline{H_i}) &= d_{ij}[1 - R_j.P(s_j|H_i)], \\ m_{ij}(\Omega) &= 1 - m_{ij}(\overline{H_i}). \end{cases} \tag{4}$$

where $\overline{H_i}$ is the hypothesis opposite of H_i. The discounting coefficient d_{ij} characterizes the *a priori* degree of confidence in the knowldege of the distribution $P(s_j|H_i)$. It stands for the metaknowledge about the representativeness of the learning of each class H_i with each source j. This parameter tends to 1 when the learning is perfectly representative of the actual distribution, whereas $d_{ij} \to 0$ when the distribution of probabilities is poorly estimated (*e.g.* in case of a too small training dataset). R_j is a coefficient weighting the probabilities. It acts as a normalization factor bounding the dynamic range: $R_j \in [0; 1/\max\{P(s_j|H_i)\}]$. For $R_j = 0$, only the *a priori* reliability of the source is taken into account, otherwise the actual data are also considered.

The two types of approaches (distance *i.e.,* model-based, and likelihood *i.e.,* case-based) yield similar performances when applied to classification problems [18]. Here, a separable likelihood approach is chosen. Indeed, as our method uses a simple and hence incomplete learning stage, it is safer to estimate conditional probabilities and to fix a priori reliability degrees, rather than mass sets directly. Furthermore, Appriou's model #1 turns out to be well suited for facial analysis as one learns easily the face class against all the other classes (here the background class only), since a specific detector may be tuned on this class.

3.3 Combination of Beliefs

The belief combination, also called revision, is involved when one has new information, coded in the form of a belief function, to merge with existing mass functions, in order to make up a synthesis of knowledge in a multi-source environment. Two constraints must be fulfilled: every source of information belongs to the same frame of discernment Ω, and all sources are independent. Conjunctive and disjunctive rules are the two basic operators for combination. For J independent and totally reliable information sources, whose hypotheses are defined in Ω, the result of the conjunctive combination, denoted by $m_{\textcircled{\cap}}$, is:

$$m_{\textcircled{\cap}}(A) = \textcircled{\cap}_{A\subset\Omega}\ m_j(A) = \sum_{A_1\cap\ldots\cap A_J=A} \left(\prod_{j=1}^{J} m_j(A_j)\right), \quad \forall A \subseteq \Omega. \tag{5}$$

This rule is commutative, associative, with the total ignorance as neutral element and the total certainty as absorbing element. It is however not idempotent. This rule leads generally to an unnormalized mass of conflict ($m_{\textcircled{\cap}}(\emptyset) \neq 0$). Dempster proposed a normalization version of this law better known as the Dempster combination rule, or orthogonal sum [11]:

$$\begin{aligned} m_{\oplus}(A) &= \frac{m_{\textcircled{\cap}}(A)}{1-K}, \quad \forall A \subseteq \Omega, A \neq \emptyset, \\ m_{\oplus}(\emptyset) &= 0, \\ \text{where} \quad K &= m_{\textcircled{\cap}}(\emptyset) \quad \text{reflects the conflicting mass that belongs to } [0,1[. \end{aligned} \tag{6}$$

The disjunctive rule [16] replaces the intersection by the union in Eq. 5 and yields a mass denoted $m_{\textcircled{\cup}}(A)$. The disjunctive rule is used when at least one source of information is unreliable. This rule does not generate conflict but yields less precise fusion as the focal elements of the resulting mass function are widened. On the contrary, the conjunctive rule is used when all information sources are reliable. It yields a more precise fusion but might generate conflict.

3.4 Management of Conflict

When using the conjunctive combination, some information sources might be discordant and give incompatible propositions. The mass value $m(\emptyset)$ assigned to the empty set quantifies this conflict. Numerous combination rules are proposed to solve this problem [19]. For example, Florea advocates an intermediate solution between conjunction and disjunction, yielding a family of robust adaptive rules [20].

3.5 Combination Rules for Dependent Sources

Conjunctive and disjunctive rules rely on the assumption that the combined mass functions come from independent sources. In real-world situations however, this is not always true. To address this problem, Denœux *et al.* introduced two new rules: the cautious conjunctive rule and the bold disjunctive rule [21, 22].

The cautious conjunctive rule relies on the least commitment principle which states that, when several belief functions are compatible with a set of constraints, one should choose the least informative one. This principle means that one should not give more belief than required to an information source. It is similar to the maximum entropy principle in the theory of probabilities. Under the constraint that the combined mass be richer than m_1 and m_2, the least informative mass exists, is unique and is defined with the minimum (denoted by $\wedge$) of the weight functions associated with m_1 and m_2. If A^{w_1} and A^{w_2} are two simple mass sets, their combination by the cautious rule is the simple mass function denoted by $A^{w_1 \wedge w_2}$:

$$\begin{aligned} w_{min}(A) &= w_1(A) \wedge w_2(A), \quad \forall A \subset \Omega, \\ m_{1\wedge 2} &= \textcircled{\cap}_{A\subset\Omega} A^{w_{min}(A)}. \end{aligned} \tag{7}$$

The normalized version of this cautious rule denoted by $\textcircled{\wedge}$ is defined by replacing the conjunctive rule $\textcircled{\cap}$ by the Dempster rule $\oplus$ in Eq.7 so that:

$$\begin{aligned} m_{1\textcircled{\wedge}2}(A) &= \frac{m_{1\wedge 2}(A)}{1 - m_{1\wedge 2}(\emptyset)}, \quad \forall A \subseteq \Omega, A \neq \emptyset, \\ m_{1\textcircled{\wedge}2}(\emptyset) &= 0 \end{aligned} \tag{8}$$

The bold disjunctive rule, denoted by $\textcircled{\vee}$, is the operator opposite of the cautious rule: it takes the maximum of weights instead of their minimum. In [23], these new rules were extended to become adaptive. The properties of the cautious and bold rules result from those of the minimum and maximum: commutative, associative and idempotent.

Tab. 1 illustrates the computation of the three combination rules $\textcircled{\cap}$, $\oplus$, and $\textcircled{\wedge}$ in the case of two non separable sources, given by their respective mass sets m_1 and m_2. Note that one obtains here generalized simple mass functions yielding weights $w(\emptyset) > 1$. Indeed, weights w are no longer constrained to belong to the interval $[0; 1]$ for generalized simple mass sets.

Table 1: Combination of 2 masses m_1, m_2: conjunction $\textcircled{\cap}$, Dempster-Shafer $\oplus$, cautious $\textcircled{\wedge}$

A	m_1	m_2	$m_{\textcircled{\cap}}$	$m_{\oplus}$	q_1	q_2	w_1	w_2	w_{min}	$m_{1\wedge 2}$	$m_{\textcircled{\wedge}}$
$\emptyset$	0	0	0.19	0	1	1	1.333	1.35	1.333	0.081	0
H_1	0.5	0.7	0.66	0.815	0.8	0.9	0.375	0.222	0.222	0.622	0.677
H_2	0.2	0.1	0.09	0.111	0.5	0.3	0.6	0.667	0.6	0.119	0.129
Ω	0.3	0.2	0.06	0.074	0.3	0.2				0.178	0.194

3.6 Decision with Transferable Belief Model

The TBM is a subjectivist interpretation of a mass function that models the partial knowledge of the value of a variable [24]. The TBM is a mental model with two levels: the credal level and the pignistic one. The credal level includes the static part of the model representing the knowledge in the form of mass functions, plus the dynamic part of the model which corresponds to the combination of beliefs. Decision is done at the pignistic level that transforms the masses into probabilities by equally sharing the conflict among every normalized mass function. For all $A \in 2^{\Omega}$ with $A \neq \emptyset$, the pignistic probability $BetP$ is defined as:

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

where $|B|$ denotes the cardinal of set B. Typically, in the binary case of two disjoint hypotheses without conflict ($m(\emptyset) = 0$), one gets: $BetP(A) = m(A) + m(\Omega)/2$, since $|\Omega| = 2$ and $[A| = 1$. Note that the computation of the pignistic probability implies a loss of information, since the degree of ignorance $m(\Omega)$ is dispatched among all the various hypotheses. The decision consists simply in choosing the hypothesis that gives the maximum of $BetP$, which is similar to the maximum plausibility criterion [17].

4 Application to Evidential Face Model

The face modeling strategy consists of an evidential fusion process using two complementary information sources: a VJ face shape detector (Fig. 3a) and a skin color detector (Fig. 3b). In order to account for the dependence between color sources, the fusion process uses the cautious rule to merge color mass sets. The fusion of color mass sets and VJ mass sets (Fig. 3c) gives a robust face model (applied here to indoor environment). For skin hue modeling, the learning stage consists of a quick initialization (Fig. 3i). This learning step is interesting for its simplicity, but it is obviously not exhaustive since it suffers from incompleteness as only the first video frame is taken into account. A classic Bayesian probabilistic approach is inefficient in this case. Therefore, the TBM framework (Fig. 3d) is used instead, since it is adequate to model partial knowledge of prior models that are incomplete, ambiguous, imprecise or unreliable. It is efficient for uncertainty management.

To each pixel p, a frame of discernment is associated with two mutually exclusive classes: $\Omega_p = \{\{H_{1p}\}, \{H_{2p}\}\}$, where $\{H_{1p}\}$ represents the face hypothesis and $\{H_{2p}\}$ represents the complementary set called the background. Dealing with only two hypotheses limits the complexity and thus the processing time, which is important for real-time tracking. Moreover, it is enough for the face/non-face binary decision. To simplify the notations in the following, we will skip the index p, and only write Ω, $\{H_1\}$ and $\{H_2\}$ for all those quantities that relate to pixel p.

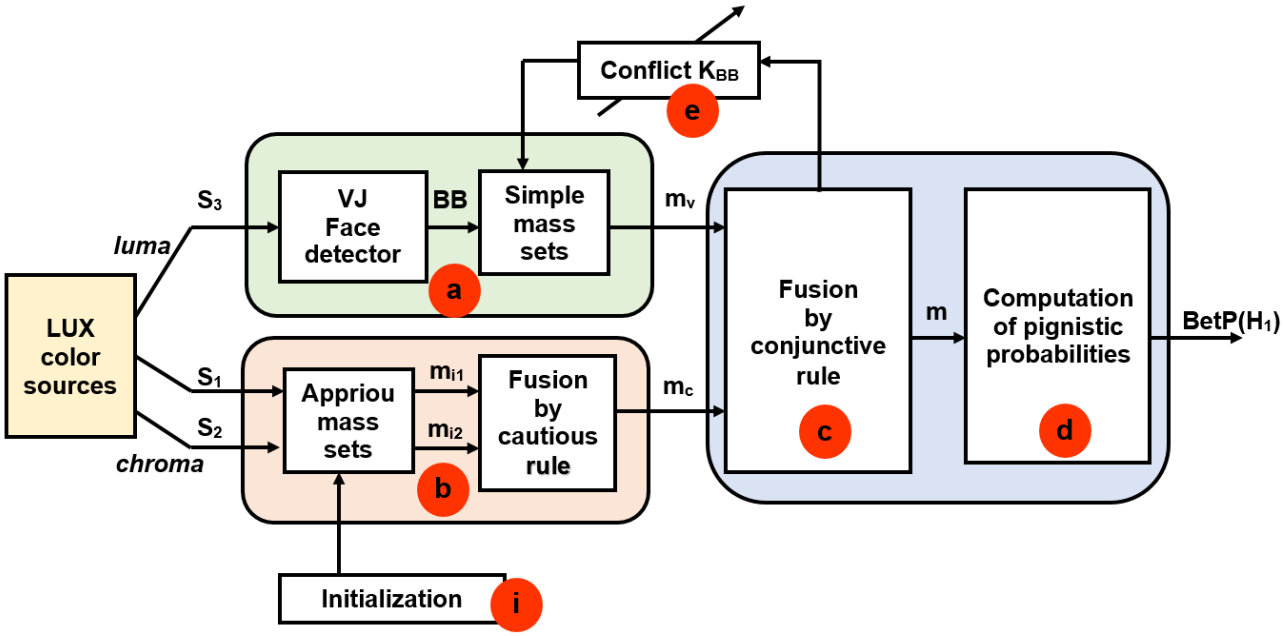


Figure 3: Block-diagram of the evidential face model: a) mass sets of the VJ face detector; b) color mass sets; c) fusion of VJ and color mass sets; d) computation of pignistic probabilities in the TBM; e) conflict management feedback; i) initialization.

4.1 Information Sources

Skin color is a relevant information since it allows to implement fast algorithms that are invariant to orientation and scale. However, skin color distribution strongly depends on lighting conditions and on the color space chosen [5]. To improve robustness to light changes, the logarithmic LUX color space [25] may be used instead of linear color spaces like RGB, YUV or other nonlinear spaces like HSV which is sensitive to noise. The three components of LUX space are computed from RGB components (with $M = 256$):

$$\begin{aligned} L &= (R+1)^{0.3}(G+1)^{0.6}(B+1)^{0.1} - 1 \\ U &= \begin{cases} \frac{M}{2}\left(\frac{R+1}{L+1}\right) & \text{for } R < L \\ M - \frac{M}{2}\left(\frac{L+1}{R+1}\right) & \text{otherwise} \end{cases} \\ X &= \begin{cases} \frac{M}{2}\left(\frac{B+1}{L+1}\right) & \text{for } B < L \\ M - \frac{M}{2}\left(\frac{L+1}{B+1}\right) & \text{otherwise.} \end{cases} \end{aligned} \tag{10}$$

L stands for the logarithmic luminance, whereas U and X are the two logarithmic chrominances (resp. red and blue). This nonlinear color space based on logarithmic image processing is known for rendering good contrast in low luminance. Besides, since it is inspired by biology (logarithmic response of retina cells), it ensures an efficient description of hues, it is little sensitive to noise and has proved its efficiency in color segmentation, compression or rendering [26]. Fig.4 illustrates the adaptive property of LUX in bright or dark context.

Hereafter, the three information sources s_j $(j = 1, 2, 3)$ used for face modeling are: $(s_1 = U,\ s_2 = X)$ for the skin hue detector, and $s_3 = L$ for the VJ detector respectively.

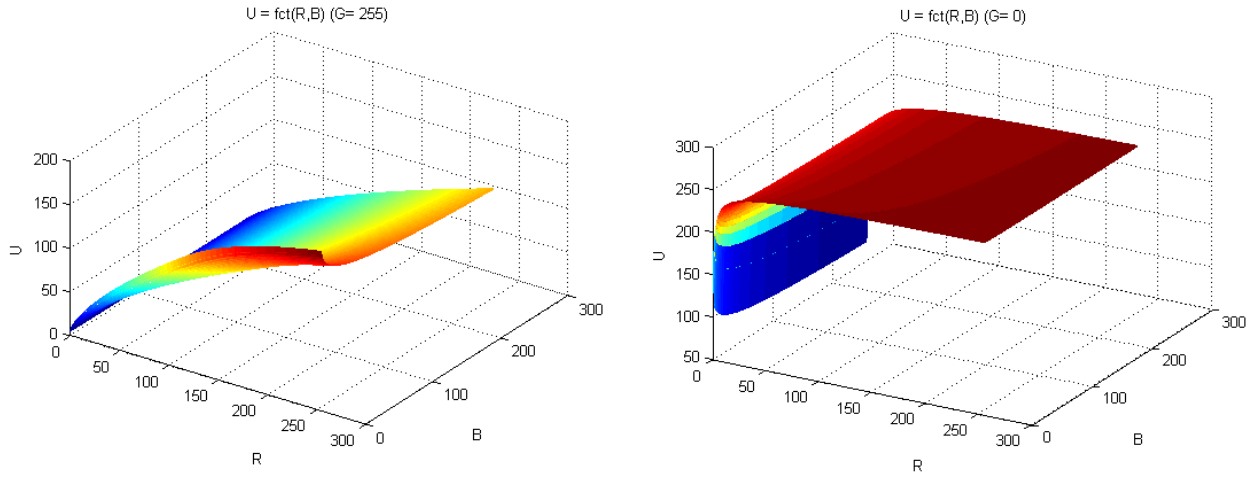


Figure 4: Variation of U as a function of (R, B) in two contexts: a) $G = 255$ (bright); b) $G = 0$ (dark)

4.2 Mass Sets for VJ Face Detector

This section explains how to obtain the mass m_v from the luma component L (Fig. 3a). The VJ face detector works on gray levels (source $s_3 = L$). It gives a target container (rectangular bounding box around the face denoted by BB) highly reliable when the face is in front-view or slightly from profile (Fig. 5a, 5b, 5c). However it fails in case of important rotation or occlusion or when it recognizes a shape-like face-artifact in the cluttered background (Fig. 5d).

a) sequence #1 b) sequence #2 c) sequence #3 d) sequence #4

Figure 5: Rectanguler bounding box BB produced by the VJ face detector in various sequences: a), b), c) correct detections; d) false detection.

In order to model the VJ detector by a belief function, a simple mass set $m_v(.)$ is assigned to each pixel p, according to its position with respect to BB. The mass is defined by a weight parameter (reliability) $\nu \in [0, 1]$:

$$\begin{aligned} m_v &= \{H_1\}^{1-\nu}, \forall p \in BB, \\ m_v &= \{H_2\}^{1-\nu}, \forall p \notin BB. \end{aligned} \tag{11}$$

The value $1 - \nu$ stands for the uncertainty in the belief about $\{H_1\}$ inside BB (resp. $\{H_2\}$ outside BB). For $\nu = 0$, the information source is not reliable at all, and the maximal belief is put on the tautology $\Omega = H_1 \cup H_2$. For $\nu = 1$, the source is reliable, the mass is maximal for the face class $\{H_1\}$ inside BB, and for the background class $\{H_2\}$ outside of BB.

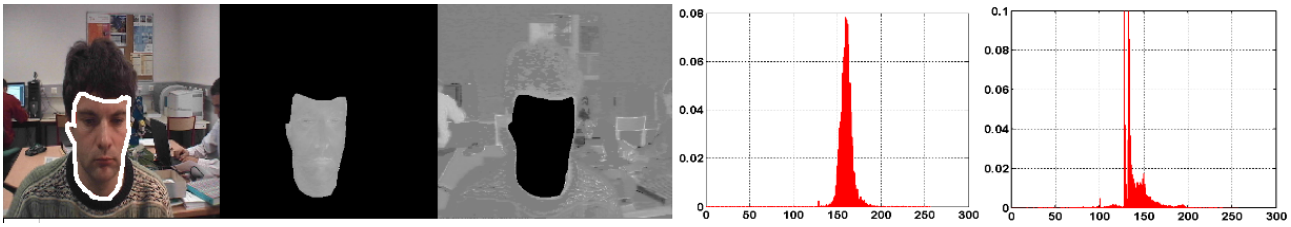
4.3 Color Mass Functions

This section explains how the color masses m_c are computed from the chroma components (Fig. 3b). For the current image, let use the following notations:

- S is the set of source vectors of size $Z \times J$, where Z is the image size (typically 400×400), and J is the dimension of the color space. Here, $J = 2$ since only two chromatic information sources s_1 and s_2 are used to build the color masses. s_j represents the color plane j of S,
- s_{jp} is one elementary observation data. It is the jth component of the color vector associated with pixel p,
- c_p is the class of pixel p (hidden primitive corresponding to one of the two hypotheses: face H_1 or non-face $H_2 = \overline{H_1}$).

Given a pixel p with known observation s_{jp} but of unknown class c_p, the classification problem consists in producing a belief about the current value of its class c_p without using any learning dataset apart from a quick initialization on the first image.

Appriou's model #1 (Eq. 4) requires the conditional likelihoods of the classes, that characterize the relationship between color components s_j and hypotheses H_1 or H_2. These priors are obtained during a semi-supervised learning step when the user selects manually in the first image of the video a free-shape zone of the face including mainly skin (Fig 6a). Hair should be discarded. This selection exhibits both: (i) a prior model of the face zone including mainly skin hue (Fig. 6b), (ii) a prior model of the background by considering all pixels outside of the selected zone (Fig. 6c). Histograms are built by considering all the color observations s_{jp} inside the face zone, resp. outside (background). Then, four conditional probabilities $P(s_j|H_i)$ (for source $j = 1$ or 2, and hypothesis $i = 1$ or 2) are deduced by simple normalization of the histograms as exemplified in Fig. 6d, 6e.



a) b) c) d) e)

Figure 6: Initialization on sequence #2: a) selected area of the face on the first image of the video; b) source s_1: face zone; c) source s_1: background; d) distribution $P(s_1|H_1)$; e) distribution $P(s_1|H_2)$.

Four Appriou mass sets $m_{ij}(H_i)$ are assigned to each pixel p having color value s_{jp} (one for each source s_j, $j \in \{1;2\}$ and for each class H_i, $i \in \{1;2\}$):

$$\begin{aligned} m_{ij}(H_i) &= 0, \\ m_{ij}(\overline{H_i}) &= d_{ij}[1 - R_j.P(s_{jp}|H_i)], \\ m_{ij}(\Omega) &= 1 - m_{ij}(\overline{H_i}). \end{aligned} \tag{12}$$

Given a pixel p, its probability $P(s_{jp}|H_i)$ is quickly retrieved from the tabulated histograms by a look-up table (L.U.T.) addressing operation. Parameter R_j, that weights the conditional

likelihoods, is set to its maximal value. For simplicity, all parameters d_{ij} are initialized to the same value $d_0 = 0.9$ (we mention in the conclusion some hints to implement a more sophisticated model). One takes $d_0 < 1$ in order to force non categorical mass sets (i.e. $m_{ij}(\Omega) \neq 0$). This Appriou model gives two complementary mass functions, one for each color source s_j, $j \in \{1;2\}$ (Eq. 3 with weights denoted by w_{ij}) so that for any hypothesis $A = \{H_i\}$ and any observation s_{jp}, one can compute the weight:

$$w_{ij}(\overline{A}) = 1 - m_{ij}(\overline{A}) \text{ computed from the prior model } P(s_{jp}|A). \tag{13}$$

Altogether, this yields four simple mass sets per pixel (two sources j, two hypotheses i).

4.4 Color Fusion by Cautious Rule

The concept of independence means that two pieces of evidence are obtained by different ways. Color sources s_1 and s_2 (the two logarithmic hues in LUX space), and hence their mass functions m_{ij} are obviously not independent as they are computed from the same raw data R, G, B (Eq. 10). Indeed, when red component R varies, both values of U and X change. To deal with the fusion of information from dependent sources, a conservative combination rule like the Denœux cautious conjunctive rule is well suited. Because of its conjunctive property, it strengthens the certainty in the information fusion. Nevertheless it ensures that the recursive combination of information with itself always gives the same result (idempotence). In that case, independence of information sources is not mandatory: idempotence authorizes dependence. So, it offers a compromise between reinforcement and idempotence. Here, this fusion operator with idempotence property is preferred.

For two distinct weights belonging to the interval $[0, 1]$, the cautious rule is defined by Eq. 7. Here, we have: $w_1 = w_{i1}$ (for red chrominance U), $w_2 = w_{i2}$ (for blue chrominance X), and $A \in 2^\Omega = \{\emptyset, \{H_1\}, \{H_2\}, \Omega\}$. The combined weights w are computed as: $w(A) = \min\{w_{ij}(A)\}$. Finally, the color masses $m_{i1 \wedge i2}(A)$ assigned to each pixel p are computed by Eq. 7, and then normalized by Eq. 8, giving the final masses $m_c(A)$ that yield, for each pixel, the belief in each class H_i (Tab. 2).

Table 2: Fusion by cautious rule: computation of color mass sets for pixel p

A	$w(.)$	$m_{i1 \wedge i2}(.)$	$m_c(.)$
$\emptyset$		$[1 - w(H_1)][1 - w(H_2)]$	0
$\{H_1\}$	$\min\{w_{ij}(H_1)\}$	$[1 - w(H_1)]w(H_2)$	$m_c(H_1)$
$\{H_2\}$	$\min\{w_{ij}(H_2)\}$	$w(H_1)[1 - w(H_2)]$	$m_c(H_2)$
Ω		$w(H_1)w(H_2)$	$m_c(\Omega)$

Typical results of this color fusion are shown in Fig. 7. The evidential model classifies correctly the image regions whose color corresponds to skin hue (face, arms). The red tee-shirt in seq. #4 is correctly detected as background by the cautious rule. The model fails however in certain background areas whose color is too close to skin hue.

4.4.1 Illustrative Example

Let us illustrate the processing with a sample case study. Table 3 shows the weights w_{ij} obtained from the following conditional probabilities:

$P(s_1|H_1) = 0.05$, $P(s_1|H_2) = 0.04$, $P(s_2|H_1) = 0.07$ and $P(s_2|H_2) = 0.01$.

Figure 7: Fusion results of color sources s_1 and s_2 by the cautious rule (display of pignistic probability $BetP(H_1)$) for the four sequences of Fig. 5, with $R_j = R_{max}$ and $d_{ij} = d_0 = 0.9$.

The discounting coefficient is set to $d_0 = 0.9$ and R_j is set to its maximal value: $R_1 = R_2 = 1/0.1 = 10$ (by taking as reference the sample histograms in Fig.6). The combined weights $w(A)$ are simply the minimal values among the $w_{ij}(A)$. The color mass set m_c resulting from the combination of weight function w is given in Tab. 3. The decision is cleary for H_1 (heaviest mass).

Table 3: Example of cautious color fusion: weights w_{ij}, combined weights w and masses m_c.

	red source U		blue source X		combination		
A	w_{11}	w_{21}	w_{12}	w_{22}	$w(.)$	$m_{i1\wedge i2}(.)$	$m_c(.)$
$\emptyset$						0.3645	0
$\{H_1\}$		0.46		0.19	0.19	0.4455	0.701
$\{H_2\}$	0.55		0.73		0.55	0.0855	0.1345
Ω						0.1045	0.1644

Let us compare with the classic Bayesian approach. The a posteriori probability is given by:

$$P(H_1|s_1, s_2) = \frac{P(H_1) \prod_j P(s_j|H_1)}{\sum_i P(H_i) \prod_j P(s_j|H_i)}. \tag{14}$$

First, let us suppose equiprobability: $P(H_1) = P(H_2) = 0.5$; then one obtains: $P(H_1|s_1, s_2) = 0.897$. The decision is clearly for H_1.

Then, if one takes: $P(H_1) = 0.2$, $P(H_2) = 0.8$, by supposing that the face size is kept to about 20% of the image surface thanks to the proper action of visual servoing, then: $P(H_1|s_1, s_2) = \frac{0.2(0.05\times0.07)}{0.2(0.05\times0.07)+0.8(0.04\times0.01)} = 0.686$. Decision is still for H_1.

In contrary, if one has: $P(H_1) = 0.1$, $P(H_2) = 0.9$ (*i.e.* when the face size decreases), then one gets stuck in indecision since $P(H_1|s_1, s_2) \approx 0.5$. Similarly to the maximum *a posteriori* criterion, the evidential decision consists in choosing the hypothesis H_i that has the maximum mass, and thus the maximum plausibility Pl or the maximum pignistic probability $BetP$. In this case, we get: $Pl(H_1) = m_c(H_1) + m_c(\Omega) = 0.8655$, and $Pl(H_2) = 0.299$, or equivalently: $BetP(H_1) = 0.783$, and $BetP(H_2) = 0.217$; the decision is still easy to take. So, the proposed method outperforms the Bayesian approach when the prior probability decreases (i.e. when $P(H_1) << 0.5$).

4.5 Global Fusion of Color and VJ Mass Sets by Conjunctive Rule

In this section, we describe the fusion of color masses m_c with VJ masses m_v (Fig. 3c). On one hand, the color model faithfully shows the skin hue but is not able to differentiate the face color from that of an arm or a hand. On the other hand, the VJ face detector detects a front-view face with a high reliability as it validates the presence of eyes, nose and mouth in the bounding box, but it might fail in case of rotated faces or background artifacts. As the information of these two sources is complementary, it is interesting to make them collaborate in order to synthetize a more robust face model. Since these two pieces of information are elaborated from the same image raw data, the question to address before implementing a proper fusion is to know whether they are dependent or not. For that purpose, a simple test is presented here: the merging of these two sources is compared using resp. the cautious rule (Fig. 8a) and the classic conjunctive rule (Fig. 8b).

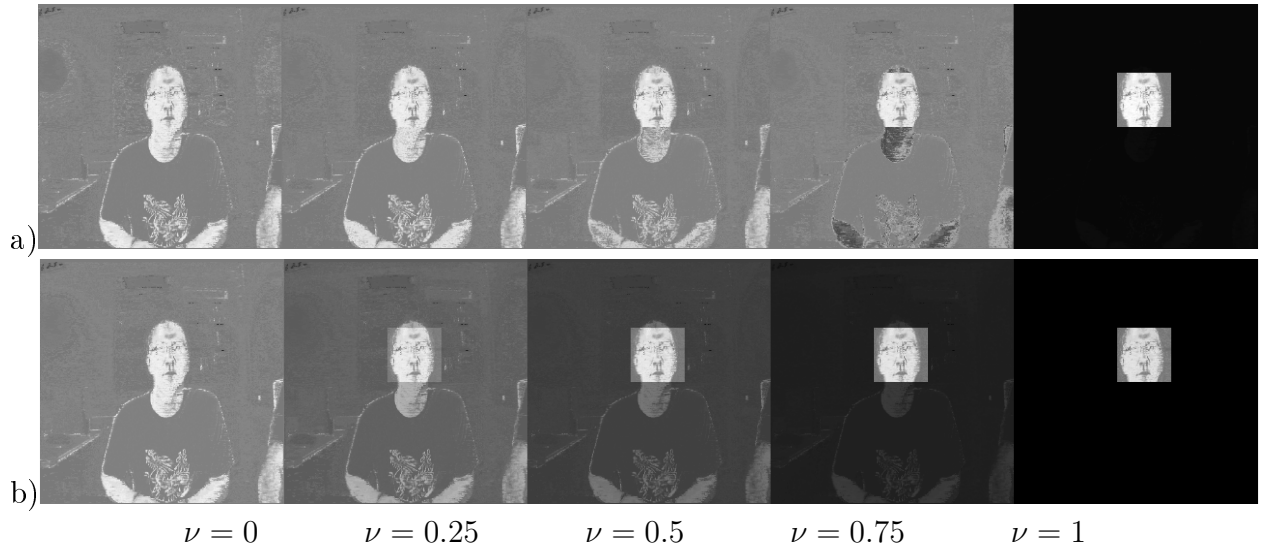


Figure 8: Fusion results of color and VJ mass functions on sequence #4 for five values of ν: a) by the cautious conjunctive rule; b) by the classic conjunctive rule.

For $\nu < 0.75$ the cautious rule favours the color masses as their weights are lower (hence the masses are heavier) than the VJ ones. The VJ information has little influence for low values of ν, and the fusion process is inefficient in that case. On the contrary, using the classic conjunctive rule, the VJ information is taken into account as soon as $\nu > 0$. The background is toned down proportionally to parameter ν, and the effect of the bounding box is more visible. The certainty on the face class is more strengthened with the classic conjunctive rule. One can induce from this simplistic test that VJ information is relatively independent from the color sources (even if this is not a formal proof of independence). This seems coherent as the VJ bounding box is computed using *2D* Haar filters applied on the L component, whereas color cues are computed from U and X components. Therefore, color and VJ mass functions are combined using the classic conjunctive rule (Eq.5):

$$m(A) = m_c(A) \textcircled{\cap} m_v(A). \tag{15}$$

A problem occurs when the VJ detector recognizes a face-like artifact in the background (Fig. 5d) with a high reliability ($\nu \geq 0.5$). In this case, skin color ($m_c(H_1) < 0.5$) and VJ mass functions disagree. This yields an important conflict inside the bounding box BB. In

order to limit false detection, we dynamically discount the initial value ν_0 of parameter ν by considering the global conflict inside BB (cf. feedback loop, Fig. 3e):

$$\begin{aligned} \nu_t &= \nu_0 && \text{for } t = 0, \\ \nu_t &= \nu_0(1 - K_{BB}) && \text{for } t > 0, \quad \text{with } K_{BB} = \frac{1}{N_{BB}} \sum_{p \in BB} K_p \end{aligned} \tag{16}$$

N_{BB} is the number of pixels inside the bounding box, and $\forall p \in BB,\ K_p = m_c(H_2) \times \nu_t$ is the conflict $m(\emptyset)$ between color and VJ masses at pixel level, thus K_{BB} denotes the average conflict. The mass m resulting from the conjunctive combination of m_c and m_v with the implementation of this discounting strategy on ν is detailed in Tab. 4.

Table 4: Fusion of m_c and m_v by the conjunctive rule, depending on pixel position

A	$m(.)$ for $p \in BB$	$m(.)$ for $p \notin BB$
$\emptyset$	$m_c(H_2) \cdot \nu_t$	$m_c(H_1) \cdot \nu_t$
$\{H_1\}$	$m_c(H_1) + m_c(\Omega) \cdot \nu_t$	$m_c(H_1) \cdot (1 - \nu_t)$
$\{H_2\}$	$m_c(H_2) \cdot (1 - \nu_t)$	$m_c(H_2) + m_c(\Omega) \cdot \nu_t$
Ω	$m_c(\Omega) \cdot (1 - \nu_t)$	$m_c(\Omega) \cdot (1 - \nu_t)$

Table 5: Output of evidential model: decision depending on color masses m_c and VJ detector reliability ν.

$m_c(.)$		VJ	$m_c(\Omega)$	$BetP(H_1)$		decision	
$\{H_1\}$	$\{H_2\}$	ν		$p \in BB$	$p \notin BB$	$p \in BB$	$p \notin BB$
0	1	0	0	0	0	$\{H_2\}$	$\{H_2\}$
		0.5					
		1					
0.5	0.5	0	0	0.5	0.5	indecisive	indecisive
		0.5		0.67	0.33	$\{H_1\}$	$\{H_2\}$
		1		1	0	$\{H_1\}$	$\{H_2\}$
1	0	0	0	1	1	$\{H_1\}$	$\{H_1\}$
		0.5			1		$\{H_1\}$
		1			0		$\{H_2\}$
0	0	0	1	0.5	0.5	indecisive	indecisive
		0.5		0.75	0.25	$\{H_1\}$	$\{H_2\}$
		1		1	0	$\{H_1\}$	$\{H_2\}$
0	0.5	0	0.5	0.25	0.25	$\{H_2\}$	$\{H_2\}$
		0.5		0.5	0.125	indecisive	
		1		1	0	$\{H1\}$	
0.5	0	0	0.5	0.75	0.75	$\{H_1\}$	$\{H_1\}$
		0.5		0.875	0.5		indecisive
		1		1	0		$\{H_2\}$

4.6 Computation of Pignistic Probabilities

This section describes the final step of the face modeling to get pignistic probabilities (Fig. 3d). The transformation of the mass functions $m(.)$ into the probabilistic framework is necessary for taking the decision and for the tracking operated by particle filter described in next section. The pignistic probability attributed to the face class $\{H_1\}$ is, for each pixel p:

$$BetP(H_1) = [m(H_1) + m(\Omega)/2]/[1 - m(\emptyset)], \tag{17}$$

Since $BetP$ belongs to $[0, 1]$, it is multiplied by 255 in order to display legible gray level images of this probability (like in Fig. 7).

Tab. 5 summarizes the behaviour of the evidential face model when the pixel hue is either close to face hue ($m_c(H_1) \to 1$), really different ($m_c(H_2) \to 1$) or in between ($m_c(H_2) \to 0.5$), and according to VJ detector reliability parameter ν. Note that color uncertainty is of course: $m_c(\Omega) = 1 - m_c(H_1) - m_c(H_2)$. The performance of the evidential model depends both on color masses and on the VJ face detector reliability (Fig. 8). Face is correctly detected if both $\nu \geq 0.5$ and $m_c(H_1) + m_c(\Omega) \geq 0.5$. A too low value of ν ($\nu < 0.5$) limits the influence of the VJ face detector and finally reduces the evidential model to a simple skin color detector. A too high value of ν ($\nu > 0.9$) can be counter-productive when the VJ detector fails and focuses on an artifact with color close to skin hue. Therefore we recommend to initialize the ν value such as $0.7 \leq \nu_0 \leq 0.9$. When the VJ face detector fails, *i.e.* when it does not deliver any bounding box, ν is temporarily set to zero.

5 Probabilistic Face Tracking

This section describes the second part of the processing, namely the face tracking procedure (Fig. 1b). The goal is to obtain in real-time the trajectory of the target (tracked object) in the video stream [27]. Tracking techniques can be grouped into three categories: (i) low level methods achieve tracking by performing color segmentation, background substraction (in case of stationary background), or optical flow estimation; (ii) active contours, snakes or AAM track the face by template matching; (iii) filtering methods perform temporal tracking by predicting the future state (localization) of a dynamic system (the target) using past measurements. Kalman filtering is used for Gaussian uni-modal models, whereas particle filtering is widely used for nonlinear models, non-Gaussian processes [28]. An extension of Bayesian particle filters to Dempster-Shafer theory is proposed in [29] for multi-camera people tracking in indoor environments. Evidential particle filtering is also used in [30] for robust multiple-source object tracking.

In the application context here, the face is a deformable object moving close to the camera, whose egomotion is unpredictable with frequent direction changes. The scene is a priori cluttered, with changes in background due to camera motion. Therefore a probabilistic tracking method based on a bootstrap particle filter is chosen, as this technique is efficient for objects with nonlinear trajectory and as it takes the temporal redundancy between frames into account. The goal is to estimate the parameters of a state vector which represents the cinematics of the target, *i.e.* the face at time t. The outer contour of the face is approximated by an ellipse with center (x_{c_t}, y_{c_t}), main axis h_t, minor axis l_t and orientation θ_t. These parameters are grouped into the state vector $X_t = [x_{c_t}, y_{c_t}, h_t, l_t, \theta_t]$ to be estimated. The particle filtering technique applies a recursive Bayesian filter to several hypothetical face locations, and merges these hypotheses according to their likelihood, conditionally to the predicted state.

The observation used as input for the particle filter is $Y_t = BetP(H_1)$, *i.e.*, an image whose high-valued pixels indicate the presence of the face at time t (cf. Fig. 8). The knowledge of these observations Y_t allows to recover the *a posteriori* probabilities: the particle filter estimates the posterior conditional probability distribution $P(X_t|Y_{1:t})$ under the form of a linear combination of weighted Dirac masses called particles:

$$P(X_t|Y_{1:t}) = \sum_{n=1}^{N} \omega_t^{(n)} \delta_{\lambda_t^{(n)}}. \tag{18}$$

A particle $\Lambda_t^{(n)} = \{\lambda_t^{(n)}, \omega_t^{(n)}\}$ represents an hypothesis on the state of the target. $\lambda_t^{(n)}$ denotes position and $\omega_t^{(n)}$ denotes weight assigned to the nth particle at time t.

The tracking algorithm begins with an initialization step (Fig. 9i). The zone of the face selected manually by the user during the learning stage is used to intialize X_t. Then the algorithm consists of two main successive stages: (i) first, the coordinates of the center of the state vector (x_{c_t}, y_{c_t}) are estimated by particle filtering (Fig. 9f); (ii) then, the ellipse size and orientation (h_t, l_t, θ_t) are estimated by a second particle filter (Fig. 9g).

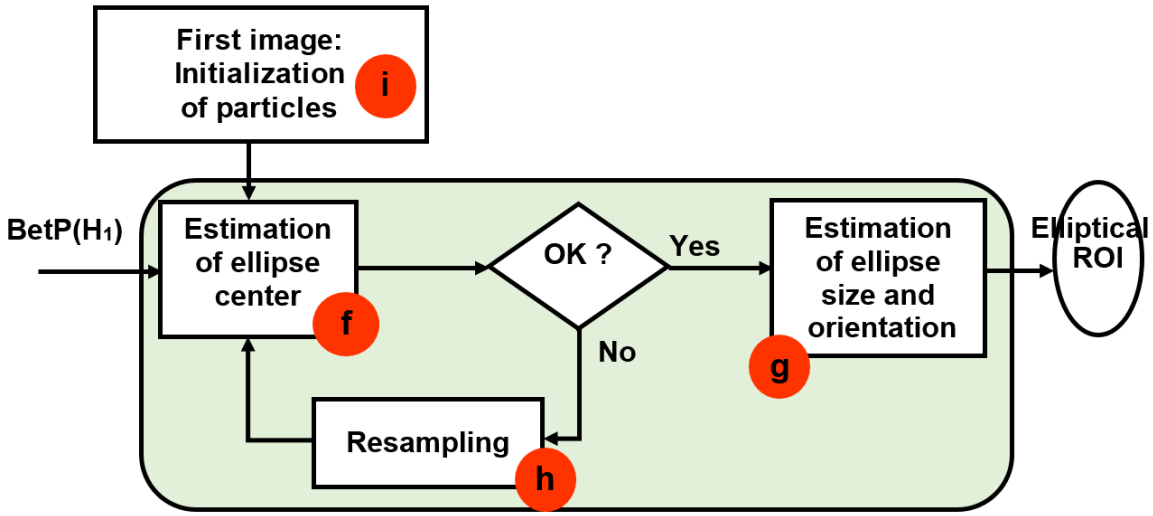


Figure 9: Block-diagram of the tracking algorithm by particle filtering.

If necessary, a resampling operation [31] is triggered inbetween (Fig. 9h): it occurs when the informative content associated with the particle estimating the state vector position is lower than a preset threshold value NR_{thresh} (typ. set to 10000 for an image size of 400×400, which is about 5% of image size). In that case, all the weights are equally reset to: $\omega_t^{(n)} = 1/N$, where N is the number of particles (typ. $N = 50$). Then, one draws randomly new positions of the face by generating particles from a uniform law $\mathcal{U}_X$ (see Eq. 19). When a particle finds a face zone again, the filter converges after a few iterations, which ensures tracking to resume.

5.1 Estimation of Ellipse Center

The state vector reduces here to $X_t = [x_{c_t}, y_{c_t}]$. A simple dynamic model [32] randomly distributes the centers of the particles in the image:

$$P(\tilde{X}_t|X_{t-1}) = (1-\alpha)\mathcal{N}(\tilde{X}_t|X_{t-1}, \Sigma) + \alpha\mathcal{U}_{\tilde{X}}(\tilde{X}_t) \tag{19}$$

where $\mathcal{N}(.|\mu, \Sigma)$ is a normal Gaussian law with average μ and covariance Σ. The diagonal matrix $\Sigma = \text{diag}(\sigma_{x_{c_t}}, \sigma_{y_{c_t}})$ sets the *a priori* constraints: it imposes the variances to the position components of the state vector (typ. $\Sigma = \text{diag}(5, 5)$). The coefficient α weights the uniform distribution: $0 \leq \alpha \leq 1$. It accounts for the rare erratic face movements acting as jumps in the video sequence. It also helps the algorithm resume tracking after a momentary period of partial or total occlusion. This uniform factor is heuristically set to $\alpha = 0.1$ so that the majority of particles (90%) remains around the center predicted at time $t-1$. It ensures some inertia in the particle distribution along time. A too high value of α is counter-productive in presence of multiple or erratic blobs in the frame. Indeed the risk of multiple jumps is increased, that can cause filter instability.

In Fig. 10a, the influence of the Gaussian distribution is characterized by the concentration of most particles around the center estimated from the previous image. The influence of parameter α can be seen, as a few isolated particles spread over other regions in the image background.

After the particle prediction, the filter evaluates the fitting of Y_t measured in the predicted ellipse $\tilde{X}_t^{(n)}$ with the face model data to compute the likelihood $P(Y_t|\tilde{X}_t)$. The fitting criterion is the quadratic sum of pignistic probabilities $BetP(H_1)$ contained inside the ellipse. Hence the estimated weight of each particle is given by:

$$\tilde{\omega}_t^{(n)} = \sum_{p \in \tilde{X}_t^{(n)}} [BetP(H_1)]^2. \tag{20}$$

The fitting criterion is the maximum likelihood, to select the most significant ellipse whose center position gives the state vector (Fig. 10b).

The nonlinearity (quadratic sum) used to compute the weight $\tilde{\omega}_t^{(n)}$ favours particles containing pignistic probabilities of high values. The transformation of the mass set into pignistic probabilities (Eq. 17) ensures the compatibility with the probabilistic framework of particle filtering (the compound hypothesis Ω does not appear any longer). The mutual exclusion principle, which states that two hypotheses must be antagonist is fulfilled. This justifies the choice of pignistic probabilities as output of the face model.

a) b) c) d)

Figure 10: Sequence #2: a) particles generated for the center estimation stage ($N = 50$); b) best position result; c) particles generated for the size and pose estimation step; d) final best ellipse in size and pose.

5.2 Estimation of Size and Pose

The size and pose at time t are predicted by running the particle filter again, with a dynamic model similar to Eq. 19, but with a state vector reduced to $X_t = [h_t, l_t, \theta_t]$, as particles are

now propagated around the fixed center (x_{c_t}, y_{c_t}) already estimated, and with a parameter setting $\alpha = 0$. Indeed it is not relevant to take erratic variations of size and pose into account. The covariance matrix $\Sigma = \text{diag}(\sigma_{h_t}, \sigma_{l_t}, \sigma_{\theta_t})$ constrains the model so that particles deviate little from those estimated at time $t-1$ (typ. $\Sigma = \text{diag}(5, 5, 0.1)$). Fig. 10c illustrates the distribution of the different predicted ellipses around the center x_{c_t}, y_{c_t}.

For the final correction step, the following observation is used: the pignistic probabilities from the evidential model are filtered (by nonlinear morphological image filling), and then thresholded to exhibit a binary shape whose contour is approximated by least squares fitting to an ellipse (measured ellipse) that serves as new observation Y_t for the second particle filter. The weights are simply the inverse of the MSE between the predicted ellipse and the measured ellipse. At last, the maximum likelihood criterion selects the most significant particle: among all the predicted ellipses around the previoulsy estimated center (Fig 10c), the algorithm selects the one (Fig. 10d) whose size and pose are closest to the observation (*i.e.* measured ellipse).

5.3 Visual Servoing

The purpose is to keep the face in the center of the image plane, with an almost constant size (approximately 10% of the image size). The tracking (task of centering) and the zoom control strategy (task of scaling) are done with a classic regulation approach (Fig. 1c). The visual servoing controls the three degrees of freedom of the PTZ camera (panoramic, tilt, zoom) . In Fig. 1, X_t^* stands for the servoing command (desired ellipse center, size and pose, typ. $X_t^* = [0, 0, 120, 100, 0]$) and X_t is the state vector measured from the particle filter. Fig. 11 shows the visual servoing behavior. On image im_{15} the face is located on the left side of the field of view. The joint action of panoramic motion and zoom focuses the face in the center of the image plane in image im_{18}. From image im_{20} to im_{24}, the user moves backward on his chair (and hence gets smaller). Then, the control of the zoom and the vertical movement of the camera (tilt) allow to refocus the face in the center of the image with the desired size (image 29).

im_{15} im_{18} im_{20} im_{24} im_{26} im_{29}

Figure 11: Tracking results for sequence #8, with visual servoing of the camera in position (pan and tilt) and control of the zoom.

6 Performance Analysis

Performance evaluation of tracking systems is mandatory. However this requires both the definition of quantitative criteria like precision, MSE, robustness, execution time, etc. and the availability of a ground truth (GT), that is, a dataset coding the exact position of the face image by image. However the task of obtaining the GT by a human expertise is tedious and subjective. Here, we consider the face present in the image when a sufficient part of its skin

is visible. Hair is not taken into account. Faces can be viewed full-frontal but also from aside. During a total occlusion, the face is supposed to be missing.

6.1 Qualitative Evaluation

The algorithm behaviour is illustrated with two sequences: (i) sequence #1 registered in our laboratory exhibits partial or total occlusions and pose variations; (ii) benchmark sequence David Indoor from the literature [33] contains pose changes, lighting and background variations, disruptive elements (the user removes his glasses, then puts them on again).

In sequence #1 (Fig. 12), the Viola-Jones masses increase the informative content in the face zone on images im_{57} and im_{73}: pignistic probabilities are most significant (white pixels in Fig. 12b) on the face zone where color and VJ mass sets are fused, but not on other skin color regions (arms, hands, or neck).

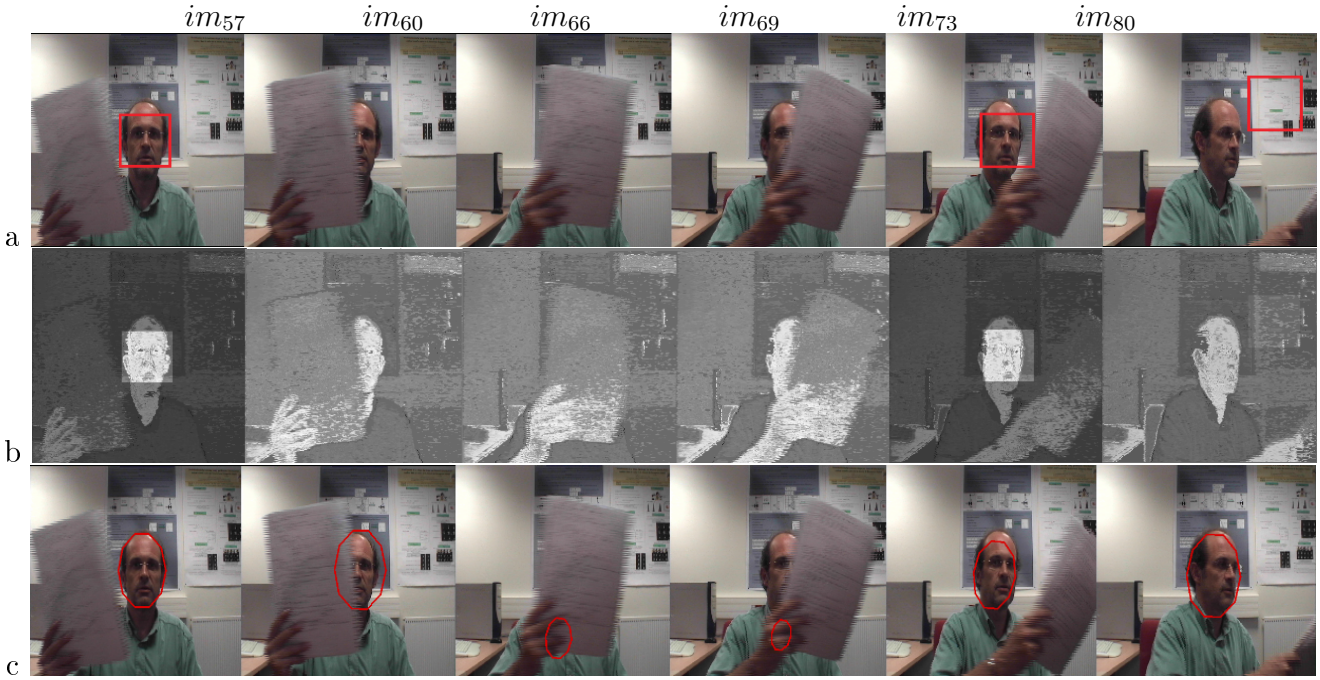


Figure 12: Face tracking for sequence #1: a) Bounding box supplied by the VJ face detector; b) Pignistic probability of the face model; c) Ellipse resulting from particle filter.

No bounding box is delivered by the VJ detector in images $im_{60}, im_{66}, im_{69}$, so that $\nu_t = 0$ is set in the evidential model since only color information is available. Therefore, in the presence of total occlusion (im_{66}), the resulting ellipse lies on the hand of the user. The uniform distribution in the filter dynamics (Eq. 19 with $\alpha = 0.1$) ensures a correct repositioning when a candidate particle locates on the face zone again (im_{73}). The VJ bounding box might degrade the tracking quality when the VJ detector focuses on a face-like artifact (frame im_{80}). An important conflict is measured inside the bounding box ($K_{BB} = 0.7$). Then the adjustment of parameter ν_t ($\nu_t \rightarrow 0.24$ since $\nu_0 = 0.8$) favours more the color information and the resulting ellipse correctly lies on the face.

In the sequence of Fig. 13, the learning stage is set up on an underexposed frame (im_{200}). On frames im_{202} and im_{300}, the pignistic probabilities are most significant in the face zone where color and VJ attributes are fused. As the person leaves the under-exposed hall (frame

im_{351}), tracking remains efficient: no updating of the evidential model is necessary even if illumination conditions have changed. As the face is in profile in frame im_{465}, no bounding box is delivered by the VJ detector and only color information is considered ($\nu_t = 0$). When hands are in contact with the face in frame im_{598}, the center and pose estimations deviate little. When the hands go away from the face, they are not tracked any longer (frame im_{604}). This shows the robustness of the method: the presence of disruptive elements alters weakly pose and size estimation and only slighty perturbs the tracking in position.

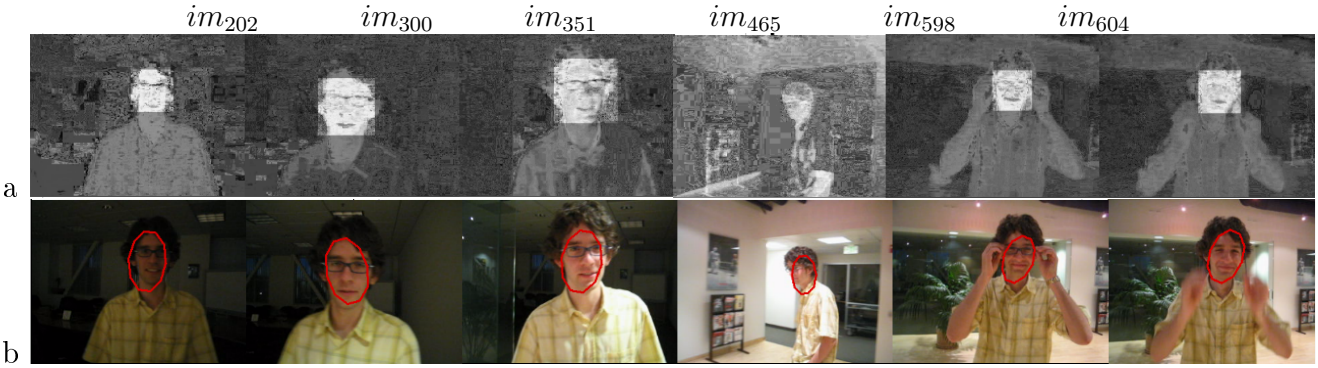


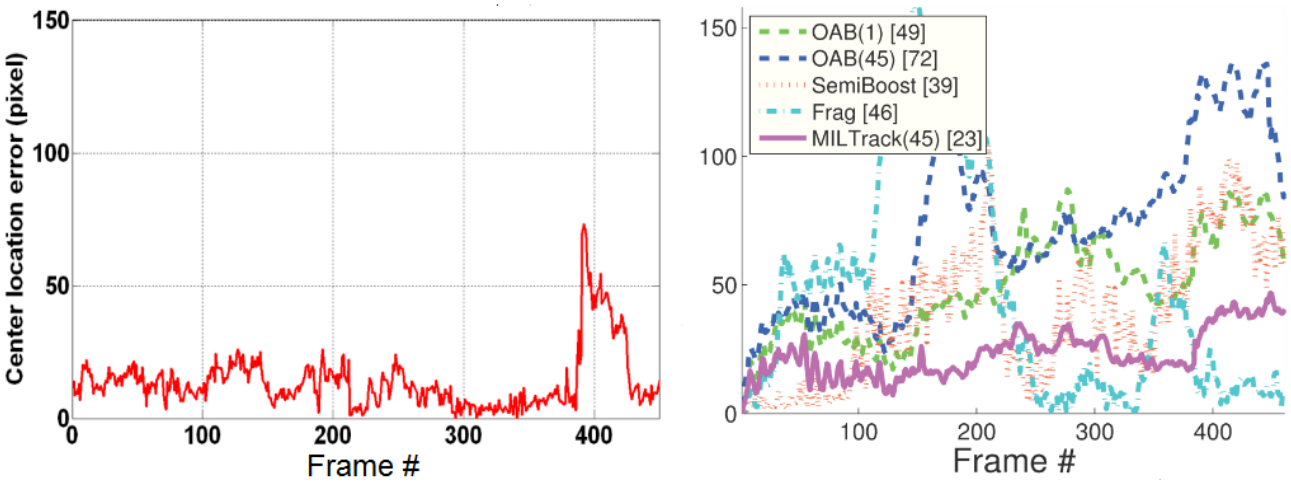
Figure 13: Tracking results on David Indoor sequence : a) evidential fusion (pignistic probability $BetP(H1)$); b) ellipse positioning.

6.2 Quantitative Evaluation

In order to quantify the tracking performance in various contexts on statistically significant data, we have manually segmented (*i.e.* cut-out) the face in various video sequences registered in our laboratory to get the ground truth, and in 500 images of the David Indoor benchmark sequence [33]. Pixels located inside the cut-out face represent the ground truth (GT). The tracking algorithm delivers an ellipse denoted by ROI (region of interest). True positive pixels (TP) belong to the intersection: $TP = ROI \cap GT$, whereas false positives (FP) lay outside of GT: $FP = ROI \cap \overline{GT}$. Two measures are classicaly used to quantify tracking performance, namely **Precision**$= \frac{|TP|}{|ROI|}$ and **Recall**$= \frac{|TP|}{|GT|}$. Precision is the probability that a pixel detected as a face pixel is actually a face pixel: it is computed as the ratio of the correct measures (TP) on all measures taken ($ROI = TP \cup FP$). Recall is the probability that a face pixel is detected: it is computed as the ratio between correct measures and the whole ground truth (as $GT = TP \cup FN$). False negative pixels (FN) belong to the intersection: $FN = \overline{ROI} \cap GT$. Precision and Recall are computed individually on every image, then averaged on each sequence (to precisely exhibit the influence of the parameters in every context), and finally on all the data to assess the global performance of the method. From these measurements, the ROC curves (Receiver Operating Characteristics) are built with x-coordinate $x = (1 - \text{Precision})$ and y-coordinate $y = \text{Recall}$, and drawn for various values of the influence parameters. The point of the curves closest to the ideal point $(x = 0; y = 1)$ corresponds to the best tuning of parameter values. The study gives the sensibility of the method to the VJ detector reliability parameter ν. The point drawn for the adaptive parameter $\nu = \nu_t$ shows the tracking performance obtained when the discounting factor by feedback is implemented (Eq. 16). This dynamic setting of ν leads to a performance optimization (Precision and Recall $\approx 80\%$). ROC curves can be found

in [9]. Results are comparable to those of standard classifiers whose detection rate reaches 80% [34].

Another quantitative evaluation criterion for the assessment of tracking performance is the center location error: $\varepsilon = \sqrt{(x_{GT_t} - x_{c_t})^2 + (y_{GT_t} - y_{c_t})^2}$, where x_{GT_t}, y_{GT_t} are the coordinates of the face center given by the ground truth (GT), whereas x_{c_t}, y_{c_t} are the center location coordinates of the detected ellipse (ROI). With a location error lower than $\varepsilon_{max} = 25$ pixels during most of the sequence (Fig. 14a), the proposed algorithm exceeds the performances of the best algorithm (MILTrack) evaluated in [33] (Fig. 14b). Our approach fails locally on images 380 to 430, when the algorithm positions on an artifact. A mean location error $\varepsilon_{mean} = 15$ pixels and a standard deviation $\Sigma_{mean} = 11$ pixels on this benchmark sequence are performances similar to or even better than those presented in the literature about particle filter [35].



a) b)

Figure 14: Tracking results (center location error) on David Indoor sequence, with: a) the proposed method; b) various algorithms according to Babenko [33].

7 Discussion

This chapter has presented an original method both for face detection based on evidential modeling, and for face tracking with a classic particle filter technique. A strategy is adopted which takes the background class H_2 in addition to face class H_1 into account. Concerning the face tracking performance, Precision and Recall reach 80% with an adequate parameter setting, but noteworthy without having to train on a huge learning dataset, which is the originality of the approach. The computation simplicity makes the method usable in a real-time (tracking at video rate). The results show the robustness of the dynamic fusion thanks to idempotent combination rules which limit the belief contraction. By setting jointly the few adaptive parameters of the evidential model and of the particle filter, we show that it is possible to finely tune the tracking behaviour. It is also more robust with respect to context variation when background or lighting conditions change during the video sequence. The statistical results confirm the qualitative observations reported here.

In the current work, the optimal setting of parameter values is deduced from an averaging of few experimental data. Consequently, this study poorly estimates the setting of the parameters

to properly tackle transient variations of context in parts of a video sequence (but it still works). A time-dynamic adjustment of parameters is required to improve the tracking robustness (as done for ν in Eq. 16). Therefore, the dynamic setting of the algorithm parameters deserves further investigation: distinct values for parameter ν could be chosen, depending on the position with respect to BB ($\nu_1 \neq \nu_2$) and also various values for parameters d_{ij}. Indeed, *a priori* knowledge about the acquisition could be used for that purpose: for face tracking purpose, red is maybe more relevant than blue ($\Rightarrow d_{i1} > d_{i2}$). Moreover, the learning of the face class H_1 is certainly more accurate than the learning of the non-face class H_2 ($\Rightarrow d_{1j} > d_{2j}$). The bounding box may be more reliable for the face model than for non-face model ($\Rightarrow \nu_1 > \nu_2$). The mass function modeling could also be improved by using a rough learning on the ground truth in the first image at initialisation, to estimate the rates TP, FP, TN, FN and then modelize and maximize the beliefs as done in [36].

8 Conflict of Interest

The author declares that he has no conflict of interest to this work.

9 Acknowledgement

I am grateful to my former PhD students, Francis Faux and Marc Liévin, for their respective contribution to the evidential face model [37, 38] and to the nonlinear LUX color space.

References

[1] P. Smets and R. Kennes, "The transferable belief model," *Artificial Intelligence*, vol. 66, no. 2, pp. 191–243, 1994.

[2] E. Ramasso, C. Panagiotakis, M. Rombaut, and D. Pellerin, "Belief scheduler based on model failure detection in the TBM framework. Application to human activity recognition," *Int. Journal of Approximate Reasoning*, vol. 51, no. 7, pp. 846–865, Sep. 2010.

[3] M.-H. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 35–58, 2002.

[4] J. M. Chaves-Gonzalez, M. A. Vega-Rodriguez, J. Gomez-Pulido, and J. M. Sanchez-Perez, "Detecting skin in face recognition systems: A colour spaces study," *Digital Signal Processing*, vol. 20, no. 3, pp. 806–823, May 2010.

[5] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A survey of skin-color modeling and detection methods," *Pattern Recognition*, vol. 40, no. 3, pp. 1106–1122, 2007.

[6] P. Viola and M. Jones, "Robust real-time face detection," *Int. Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[7] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models - Their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[9] F. Faux and F. Luthon, "Theory of evidence for face detection and tracking," *Int. Journal of Approximate Reasoning*, vol. 53, no. 5, pp. 728–746, Jul. 2012.

[10] F. Luthon, "Audioslide ScienceDirect," https://www.youtube.com/watch?v=AaV51gz1GBU, 2013.

[11] A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *Annals of Mathematical Statistics*, vol. 38, pp. 325–339, 1967.

[12] G. Shafer, *A Mathematical Theory of Evidence*. New Jersey: Princeton University Press, 1976.

[13] I. Bloch, "Defining belief functions using mathematical morphology. Application to image fusion under imprecision," *Int. Journal of Approximate Reasoning*, vol. 48, no. 2, pp. 437–465, Jun. 2008.

[14] L. M. Zouhal and T. Denoeux, "An evidence-theoretic k-NN rule parameter optimization," *IEEE Transactions on Systems, Man and Cybernetics - Part C*, vol. 28, no. 2, pp. 263–271, 1998.

[15] P. Walley and S. Moral, "Upper probabilities based only on the likelihood function," *Journal of Royal Statistical Society, Series B*, vol. 61 (Part 4), pp. 831–847, 1999.

[16] P. Smets, "Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem," *Int. Journal of Approximate Reasoning*, vol. 9, pp. 1–35, 1993.

[17] A. Appriou, "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)*, Nov. 1999, pp. 5.1–5.31.

[18] T. Denœux and P. Smets, "Classification using belief functions: the relationship between the case-based and model-based approaches," *IEEE Transactions on Systems, Man and Cybernetics B*, vol. 36, no. 6, pp. 1395–1406, 2006.

[19] A. Martin, C. Osswald, J. Dezert, and F. Smarandache, "General combination rules for qualitative and quantitative beliefs," *Journal of Advances in Information Fusion*, vol. 3, no. 2, pp. 67–82, Dec. 2008.

[20] M. C. Florea, A.-L. Jousselme, E. Bossé, and D. Grenier, "Robust combination rules for evidence theory," *Information Fusion*, vol. 10, pp. 183–197, 2009.

[21] T. Denœux, "Conjunctive and disjunctive combination of belief functions induced by non distinct bodies of evidence," *Artificial Intelligence*, vol. 172, pp. 234–264, 2008.

[22] B. Quost, M. H. Masson, and T. Denœux, "Classifier fusion in the Dempster-Shafer framework using optimized t-norm based combination rules," *Int. Journal of Approximate Reasoning*, vol. 52, no. 3, pp. 353–374, 2011.

[23] A. Kallel and S. Le Hégarat-Mascle, "Combination of partially non-distinct beliefs: the cautious-adaptive rule," *Int. Journal of Approximate Reasoning*, vol. 50, no. 7, pp. 1000–1021, Jul. 2009.

[24] P. Smets, "Decision making in the TBM: the necessity of the pignisitic transformation," *Int. Journal of Approximate Reasoning*, vol. 38, no. 2, pp. 133–147, Feb. 2005.

[25] M. Liévin and F. Luthon, "Nonlinear color space and spatiotemporal MRF for hierarchical segmentation of face features in video," *IEEE Trans. on Image Processing*, vol. 13, no. 1, pp. 63–71, Jan. 2004.

[26] F. Luthon, B. Beaumesnil, and N.Dubois, "LUX color transform for mosaic image rendering," in *17th IEEE Int. Conf. on Automation, Quality and Testing, Robotics (AQTR 2010)*, Cluj-Napoca, Romania, May 28-30 2010, pp. 93–98, vol. III.

[27] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: a survey," *ACM Computing Surveys*, vol. 38, no. 4, pp. 1–45, 2006.

[28] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[29] R. Muñoz-Salinas, R. Medina-Carnicer, F. J. Madrid-Cuevas, and A. Carmona-Poyato, "Multi-camera people tracking using evidential filters," *Int. Journal of Approximate Reasoning*, vol. 50, no. 5, pp. 732–749, May 2009.

[30] J. Klein, C. Lecomte, and P. Miché, "Hierarchical and conditional combination of belief functions induced by visual tracking," *Int. Journal of Approximate Reasoning*, vol. 51, no. 4, pp. 410–428, Mar. 2010.

[31] A. Doucet, S. J. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statistics and Computing*, vol. 10, no. 3, pp. 197–208, 2000.

[32] P. Pérez, J. Vermaak, and A. Blake, "Data fusion for visual tracking with particles," *Proceedings of IEEE*, vol. 92, no. 3, pp. 495–513, 2004.

[33] B. Babenko, M. H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1619–1632, 2011.

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

[35] W. Zheng and S. M. Bhandarkar, "Face detection and tracking using a boosted adaptive particle filter," *Journal of Visual Communication and Image Representation*, vol. 20, pp. 9–27, 2009.

[36] M. Shoyaib, M. Abdullah-Al-Wadud, and O. Chae, "A skin detection approach based on the Dempster-Shafer theory of evidence," *Int. Journal of Approximate Reasoning*, vol. 53, no. 4, pp. 636–659, Jun. 2012.

[37] F. Faux and F. Luthon, "Théorie de l'évidence pour suivi de visage," *Traitement du Signal*, vol. 28, no. 5, pp. 515–545, Sept-Oct. 2011.

[38] F. Faux, "Détection et suivi de visage par la théorie de l'évidence," Ph.D. dissertation, Université de Pau et Pays de l'Adour, Anglet, France, Oct. 2009.