Theory of Evidence for Face Detection and Tracking

Francis Faux, Franck Luthon[1]

University of Pau and Adour River UPPA,
Computer Science Laboratory LIUPPA, IUT GIM Anglet, France

Abstract

This paper deals with face detection and tracking by computer vision for multimedia applications. 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 proposed method handles simple contextual knowledge representative of the application background thanks to a quick supervised initialization. The transferable belief model is used to counteract the incompleteness of the prior model due first to a lack of exhaustiveness of the learning stage and secondly to the subjectivity of the task of face segmentation. The algorithm contains two main steps: detection and tracking. In the detection phase, an evidential face model is estimated by merging basic beliefs elaborated from Viola and Jones face detector and from a skin colour detector, for the assignment of mass functions. These functions are computed as the merging of sources in a specific nonlinear colour space. In order to deal with colour information dependence in the fusion process, the Denœux cautious rule is used. The pignistic probabilities stemming from the face model guarantee the compatibility between the belief framework and the probabilistic framework. They are the entries of a bootstrap particle filter which yields face tracking at video rate. We show that the proper tuning of the evidential model parameters improves the 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. However the proposed method requires only a weak initialization.

Keywords: Face Tracking, Evidence Theory, Dempster-Shafer, Particle Filter, Transferable Belief Model, Cautious Rule, LUX Colour Space, Image

[1] Email address: Franck.Luthon@univ-pau.fr

Processing, Pattern Recognition, Computer Vision.

1. Introduction

Real time face detection and tracking in video sequences has been studied for more than ten years by the image processing and computer vision communities, owing to the multiplicity of applications: teleconferencing, CCTV, human machine interaction, robotics. However, despite the ongoing progress in image processing and the increasing computation speed of digital processors, the design of generic and robust algorithms is still the object of active research. Indeed, face detection by computer is made difficult by the variability of appearance of this deformable moving object due to individual morphological differences (nose shape, eye colour, skin colour, beard), to the presence of visual artifacts (glasses) or occlusions, to illumination variations on the face zone (shadow, highlight), to face expression changes that depend on contextual (social, cultural, emotional) factors. Widely studied in human sciences (cognitive sciences, psychology, sociology) these last points are only partially taken into account in computer vision for face recognition or expression analysis, if not at all for face detection. Indeed, they are difficult to model and do not easily cope with real time implementation. Moreover, the scene background content can also disturb detection (foreground-background similarity or background clutter).

In this paper, to handle the face specificity, a supervised learning method is proposed, where the user selects manually a zone of the face on the first image of the video sequence. This rapid initializing step constitutes the learning stage which yields very simply to the prior model. It is however related to the user subjectivity while selecting the face zone and it suffers from incompleteness because of a lack of exhaustivity in the learning stage. In this context, a probabilistic modelling is not relevant. Therefore the proposed method for face modelling is based on belief functions: indeed the transferable belief model (TBM) [1] is well suited to model partial knowledge in a complex system. Hammal demonstrated the efficiency of TBM for the classification of emotions and facial expressions [2], and Ramasso used this framework successfully for human activity recognition [3].

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 just a webcam). The tracking should be as robust as possible to occlusions, pose, scale, background and illumination changes. The proposed algorithm takes control of the servo-camera to perform a dynamic centering of the face location in the image plane during the whole video sequence. The algorithm is made of two main steps: the face detection, then the tracking procedure (Fig. 1).

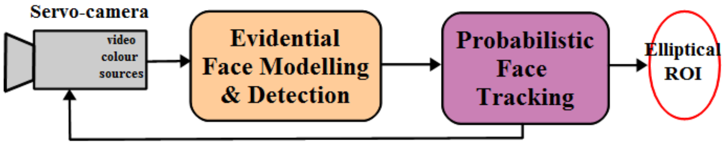

Servo-camera
video colour sources
Evidential Face Modelling & Detection
Probabilistic Face Tracking
Elliptical ROI

Figure 1: Overview of the two-step algorithm: a) face detection by evidential modelling; b) face tracking by particle filtering and visual servoing.

An elliptical region of interest (ROI) including the face is computed by particle filtering, and held at the center of the image plane by visual servoing. The context of application is limited to indoor environment, typically a laboratory or an office. As regards acquisition conditions, the distance between user and sensor ranges from approximately 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.

Section 2 presents a state of the art about face detection. Section 3 recalls briefly the theory of belief functions. The proposed evidential face model for face detection is detailed in section 4. The tracking with particle filter and visual servoing of the camera are described in chapter 5. Performance analysis of the algorithm, both qualitative and quantitative, is presented in section 6. Finally, a discussion in section 7 concludes the paper.

2. Related works

Face detection methods can be grouped into two categories differing as to the processing of prior information [4, 5, 6]. Nevertheless, this classification is not exhaustive since numerous methods use mixed approaches. It is also important to make a difference between detection methods dedicated to still images, where complex algorithms can be used, and those dedicated to video sequences where the computation cost is of major concern for real-time processing.

Feature-based methods use as primitives physical properties of the face. They rely on numerous heuristics for the proper choice of the data patterns extracted from the image. The so-called low-level analysis (or early vision) handles the information obtained directly from the properties of the pixels such as luminance or colour [7, 8], or indirectly by mathematical computation of edges, motion or texture from pixel neighborhoods. For example, the wavelet transform is efficient to extract face features. Colour is a key feature because of its specific properties and its invariance w.r.t. rotation and translation. Nevertheless skin colour is made of a large variety of hue shades (shadowy, pale, overexposed skin) depending both on the subject and on illumination conditions. Therefore the construction of a robust hue detector requires the choice of a proper colorimetric space [9]. Anyway, the primitives produced by low level analysis remain ambiguous. To validate the detection, it is necessary to use additional information. The feature analysis is based both on the knowledge of an adequate face model (prior model) 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, processings are potentially fast as no learning base is necessary. The methods for parameter extraction are often specific to the context at hand, and are constructed empirically on colour, edge or motion cues.

Holistic approaches, by contrast, address the detection problem as a general identification problem. The key-point is to compare an image with a generic face model and to decide if there is resemblance or not. Priors about geometrical or physiological specificities are discarded to limit the modelling errors due to incomplete and imprecise knowledge of the face. These methods are based on the learning of a face model from a base of examples as much complete as possible. Linear methods of subspaces, statistical approaches (Monte-Carlo methods), support vector machines or neural networks can be used. An important step was done when the first holistic face detector with real-time capacities was proposed by Viola and Jones [10]. It is based on an automatic selection of *2D* Haar filters applied to the monochrome image and it uses a cascade of boosted classifiers with increasing complexity. Some variants of this algorithm are adapted to faces with variable pose [11]. The active shape models (ASM), introduced by Cootes and Taylor [12], are deformable models which depict the highest level of appearance of the face features. Once initialized near a facial component, 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. [13]. The use of the third dimension, namely the temporal one, can lead to a real-time *3D* deformable face model varying according to morphological parameters during a video sequence. Therefore, this second family of methods provides some flexibility to the different contexts such as the number of faces in the scene or the type of lighting. Nevertheless these methods are strongly dependent on the choice and quality of the face models, and they require an important mass of data that is sufficiently representative. Whatsoever, the learning database is of course never exhaustive and its construction remains a full problem.

In this paper, we help collaborate two complementary face detection methods in a fusion process. First, among the feature-based methods, our choice focuses on a skin colour discriminating detector. Indeed, its properties of invariance w.r.t. motion allow to track the face whatever its pose during the video sequence. Second, the Viola and Jones (VJ) face detector is preferred among the holistic approaches, due to its real time properties 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 [10] have made their classifier public but not their training, the classifier used here has not been trained on our data. We will see that the proposed method circumvents this point.

3. Theory of belief functions

3.1. Belief functions

The theory of belief functions also called Dempster-Shafer theory or evidence theory, dates back to the 1970s. Inspired by the upper and lower probability notions studied first by Dempster [14], then by Shafer [15], it can be interpreted in a subjective way as a formal quantitative model of degrees of belief [16]. This theory increases modelling flexibility and allows to solve complex problems since: (i) it does not necessarily 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 modelling). It was successfully applied to multisensor signals [17] and to image fusion [18, 19].

The first concept in the evidence theory is the mass function which characterizes the opinion of an agent on a question or on the state of a system. The frame of discernment, denoted by Ω, is the finite set of answers to this question. A mass function is an application of the 2^{Ω} parts of Ω towards the 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 function $m(A)$ is interpreted as the part of belief placed strictly in A. A simple mass function, or elementary state of belief, is defined as a belief function with a mass m so that $A \subset \Omega$ is set along with a weight function $w \in [0,1]$ so that:

$$\begin{aligned} m(A) &= 1 - w, \\ m(\Omega) &= w. \end{aligned} \tag{2}$$

Denoted by $m = A^{w}$, it represents the belief put in Ω but not in A. For any A, A^{1} $(w = 1)$ is the empty simple mass function whereas A^{0} $(w = 0)$ is the categorical simple mass function.

In order to represent a complex state of belief, it is possible to build a set formed by these independently weighted propositions. Indeed, under two conditions recalled hereafter in section 3.2, any non dogmatic mass function $m(A)$ (*i.e.* when Ω is a focal set) may be expressed, by canonical decomposition [20], as the conjunctive combination (see definition below) of simple mass functions: $m(A) = \textcircled{\scriptsize $\cap$}_{A \subset \Omega} A^{w(A)}$.

3.2. 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 [21]. Conjunctive and disjunctive rules are the two main operators for combination. For J independent and totally reliable information sources, whose hypotheses are defined on Ω, the result of the conjunctive combination denoted by $m_{\textcircled{\scriptsize $\cap$}}$ is:

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

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 ($m_{\text{\textcircled{$\cap$}}}(\emptyset) \neq 0$). Dempster proposed a normalization version of this law better known as the Dempster combination rule or orthogonal sum [14]:

$$\begin{aligned} m_{\oplus}(A) &= \frac{m_{\text{\textcircled{$\cap$}}}(A)}{1-K}, \quad \forall A \subseteq \Omega, A \neq \emptyset, & (4)\\ m_{\oplus}(\emptyset) &= 0, & \\ \text{with} \quad K &= m_{\text{\textcircled{$\cap$}}}(\emptyset). & (5) \end{aligned}$$

K reflects the conflicting mass that varies within $[0, 1]$.

The disjunctive rule [22] replaces the intersection by the union in Eq. 3 and yields a mass denoted $m_{\text{\textcircled{$\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 functions are widened. On the contrary, the conjunctive rule is used when all the information sources are reliable. It yields a more precise fusion but may generate conflict.

3.3. Management of conflict

During the conjunctive combination, some combined sources may be discordant and show incompatible propositions. The mass function affected to the empty set quantifies this conflict. Numerous combination rules were proposed to solve this problem [23, 24, 25]. Florea proposed a family of adaptive rules which advocate an intermediate solution between conjunction and disjunction [26]. In [27], the Florea family was extended under the name of mixed rules family.

3.4. New combination rules

Conjunctive and disjunctive rules rely on the assumption that the combined mass functions come from independent sources. However in real-world applications, this is not always the case. To address this problem, Denœux introduced two new rules: the cautious conjunctive rule and the bold disjunctive one [28, 29].

The cautious conjunctive rule, denoted by $\text{\textcircled{$\wedge$}}$, 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 m_{12} is richer than m_1 and m_2, the least informative mass exists, is unique and is defined as 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 masses, their combination by the cautious rule is the simple mass function denoted by $A^{w_1 \wedge w_2}$ so that:

$$\begin{aligned} w_{1 \owedge 2}(A) &= w_1(A) \wedge w_2(A) \quad \forall A \subset \Omega, & (6)\\ m_{1 \owedge 2}(A) &= \bigcirc\!\!\!\!\!\cap_{A \subset \Omega} A^{w_1(A) \wedge w_2(A)}. & (7) \end{aligned}$$

A normalized version of this cautious rule denoted by $\owedge^*$ is defined by replacing the conjunctive rule $\bigcirc\!\!\!\!\!\cap$ by the Dempster rule $\oplus$:

$$\begin{aligned} m_{1 \owedge^* 2}(A) &= \frac{m_{1 \owedge 2}(A)}{1 - m_{1 \owedge 2}(\emptyset)}, \quad \forall A \subseteq \Omega, A \neq \emptyset, & (8)\\ m_{1 \owedge^* 2}(\emptyset) &= 0 & (9) \end{aligned}$$

The bold disjunctive rule, denoted by $\ovee$, is the dual operator of the cautious rule. In [30], 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.

3.5. Modelling of mass functions

The mass function modelling is a difficult problem with no universal solution. Difficulty is increased if one wants to assign beliefs in compound hypotheses. One can distinguish models based on distance, stemming from pattern recognition [31, 32] where mass functions are built only from learning vectors, and the models using likelihood computation. These last ones decompose in global methods [15, 33] and separable ones.

Separable methods build a belief function for each hypothesis H_i of the frame of discernment. This kind of approach was first proposed by Smets [34] then used by Appriou [35]. These models, stemming form a probabilistic inspiration, rely on an initial learning for the estimation of conditional probabilities $p(x_j|H_i)$ where x_j represents an observation of the source j and H_i is one of the hypotheses. Appriou recommends to use the model obtained from the generalized Bayes theorem (GBT) proposed by Smets [22]:

$$\left\{ \begin{array}{lcl} m_{ij}(\{H_i\}) &=& 0, \\ m_{ij}(\{\overline{H_i}\}) &=& d_{ij}[1 - R_j.p(x_j|H_i)], \\ m_{ij}(\Omega) &=& 1 - m_{ij}(\{\overline{H_i}\}). \end{array} \right. \qquad (10)$$

d_{ij} is a discounting coefficient which characterizes the *a priori* degree of confidence in the knowldege of each distribution $p(x_j|H_i)$. It represents some kind of metaknowledge about the representativeness degree of the learning of each class H_i with each source j. This parameter is equal to 1 when the densities are perfectly representative of the learning, whereas $d_{ij} = 0$ when the distribution of probabilities is completely underestimated. R_j is a coefficient weighting the probabilities. It acts as a normalization factor bounding the dynamic range: $R_j \in [0; (\max_i \{p(x_j|H_i)\})^{-1}]$. For $R_j = 0$, only the *a priori* source reliability is taken into account, otherwise the data are also considered.

A comparative study of these two types of approaches (distance *i.e.,* model-based and likelihood *i.e.,* case-based [36]) shows that the performance of these methods applied to classification problems does not differ drastically. Thus the choice of the model remains a delicate topic. In our application, as the method is based on a very simple, and hence incomplete learning stage, it is relatively easier to estimate the conditional probabilities and the *a priori* reliability degrees, rather than the mass sets directly. Furthermore, this model 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.

3.6. Transferable Belief Model

The TBM is a subjectivist interpretation, where a mass function models the partial knowledge of the value of a variable [16, 1]. The TBM is a mental model with two levels: the credal level and the pignistic one. The credal level mainly includes the static part of the model representing the knowledge in the form of mass functions, and the belief combination called revision which corresponds to the model's dynamic part. Decision is done at the pignistic level that transforms the mass into probability distributions by fairly sharing every normalized mass function. The pignistic probability denoted by $BetP$ is defined for all $A \in 2^\Omega$ with $A \neq \emptyset$ as:

$$BetP(A) = \sum_{B \in 2^\Omega \ ; \ B \neq \emptyset} \frac{|A \cap B|}{|B|} \frac{m(B)}{1 - m(\emptyset)}, \quad \text{with } m(\emptyset) \neq 1. \tag{11}$$

Note that the computation of the pignistic probability implies a loss of information at the transition between credal and pignistic levels, since the conflict is dispatched among the various hypotheses.

4. Evidential Face Model

4.1. Proposed strategy

Let us recall that the proposed algorithm consists of a face detection stage, which serves as input for the face tracking procedure (Fig. 1). The face modelling is based on an evidential fusion process using two families of complementary sources: the VJ face detector (Fig. 2a) and a skin colour discriminating detector (Fig. 2b). The fusion of colour mass sets and VJ mass sets (Fig. 2c) gives a model representative of the face in the various contexts of application (restricted to indoor environment).

As regards the model for skin hue, the learning stage reduces to a quick initialization (Fig. 2e). This learning step is interesting for its simplicity, but it is obviously not exhaustive and it suffers from incompleteness as only the first video frame is taken into account. A classic probabilistic approach is inefficient in this case. Therefore, the proposed method takes place within the TBM framework, that is adequate to model the incompleteness (partial knowledge) of a prior model. Moreover, in order to account for the dependence between colour sources, we propose a variant of the Appriou fusion process (Eq. 10) using the Denœux cautious conjunctive rule to merge the colour mass sets.

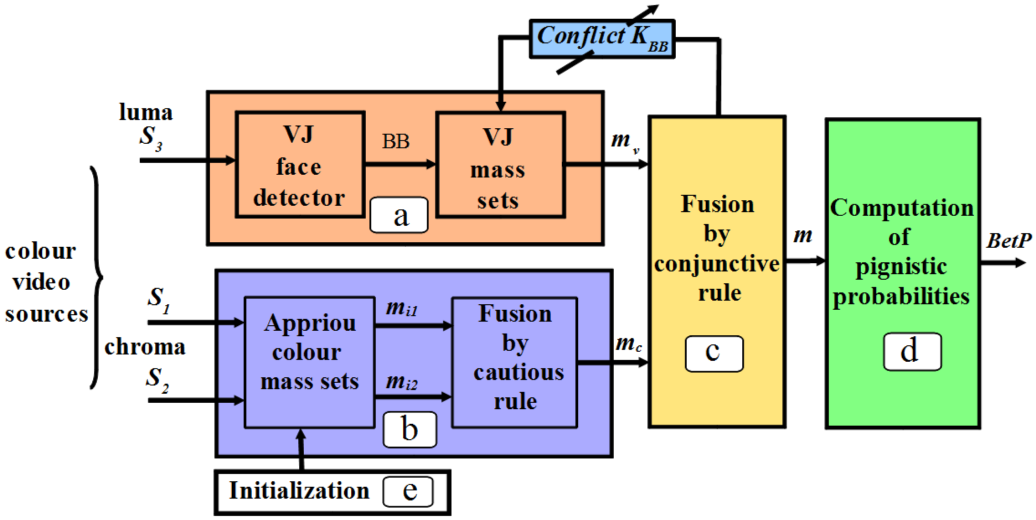


Figure 2: Block-diagram of evidential face modelling: a) mass sets of VJ face detector attributes; b) colour mass sets; c) fusion of VJ and colour mass sets; d) computation of pignistic probabilities; e) initialization.

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 what is not a face (*i.e.*, the complementary set called the background). This limitation put on Ω, with only these two hypotheses, reduces the complexity and thus the processing time, which is important for real-time tracking. To simplify the notations in the following, we will skip the index p and only write Ω, $\{H_1\}$ and $\{H_2\}$ for all the quantities related to pixel p.

4.2. Information sources

Face skin colour is a relevant information since it allows to implement fast algorithms that are invariant to orientation or scale changes. However, skin colour distribution strongly depends on the lighting conditions and on the colour space chosen [9]. To improve robustness to light changes, we choose the LUX logarithmic colour space instead of linear colour spaces like RGB, YCrCb or other nonlinear spaces like HSV which is more sensitive to noise [37]. The three components of LUX space are computed from the RGB components as follows (with $M = 256$):

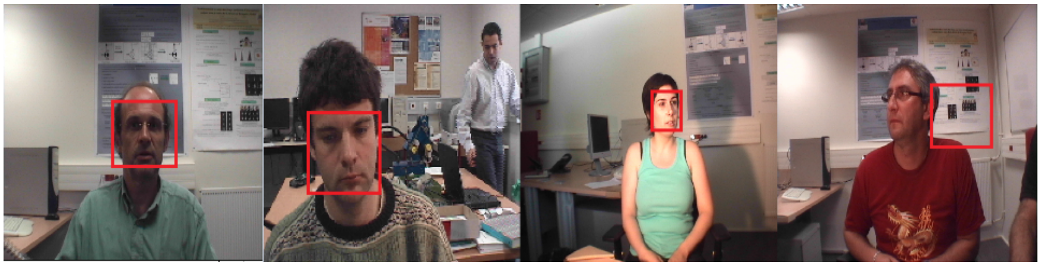
$$\begin{array}{l} L = (R+1)^{0.3}(G+1)^{0.6}(B+1)^{0.1} - 1 \\ U = \begin{cases} \frac{M}{2}\left(\frac{R+1}{L+1}\right) & \text{for } R < L \\ M - \frac{M}{2}\left(\frac{L+1}{R+1}\right) & \text{otherwise} \end{cases} \\ X = \begin{cases} \frac{M}{2}\left(\frac{B+1}{L+1}\right) & \text{for } B < L \\ M - \frac{M}{2}\left(\frac{L+1}{B+1}\right) & \text{otherwise.} \end{cases} \end{array} \tag{12}$$

L stands for the logarithmic luminance, whereas U and X are the two logarithmic chrominances (resp. red and blue). This nonlinear colour space based on the logarithmic image processing transform is known for rendering a good contrast even for low luminance [38]. Besides, since it is inspired by biology (*cf.* logarithmic response of retina cells) [39], it ensures an efficient description of hues, it is little sensitive to noise and has proved its efficiency in colour segmentation, colour compression or colour rendering [40]. Hereafter, the three information sources denoted by s_j $(j = 1, 2, 3)$, that will be used to model the face, are: $(s_1 = U,\ s_2 = X)$ for the skin hue, and $s_3 = L$ for the VJ detector.

4.3. Mass functions of the VJ face detector

In this section, we explain how to obtain the mass m_v from the luma component L (cf. Fig. 2a). The VJ face detector works on grey levels (source

$s_3 = L$). It generates a target container (*i.e.* a rectangular bounding box around the face denoted by BB) highly reliable when the face is in front-view or slightly from profile (Fig. 3a, b, c). However it fails in the case of important rotation or occlusions or when it recognizes a shape-like face-artifact in the background (Fig. 3d).



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

Figure 3: Bounding box produced by the VJ face detector in various sequences: a), b), c) correct detection; d) false detection.

In order to model the VJ attribute by a belief function, a simple mass denoted by $m_v(.)$ is assigned to each pixel p, according to its position with respect to the bounding box and proportionally to a parameter of reliability $\gamma \in [0, 1]$ so that:

$$m_v = \{H_1\}^{1-\gamma}, \forall p \in BB, \tag{13}$$

$$m_v = \{H_2\}^{1-\gamma}, \forall p \notin BB. \tag{14}$$

The value $1-\gamma$ stands for the uncertainty in the belief about $\{H_1\}$ in Eq. 13 (resp. $\{H_2\}$ in Eq. 14). For $\gamma = 0$ the information source is not reliable and the maximal belief is associated to the tautology Ω. For $\gamma = 1$ the source is reliable, the mass is maximal for the face class $\{H_1\}$ inside the bounding box, and for the background class $\{H_2\}$ outside of BB.

4.4. Colour masses

This section togheter with the next one (section 4.5) explain how the colour masses m_c are computed from the chroma components (cf. Fig. 2b). A classification approach is taken to build the mass sets coding the colour information. For the current image, the following notations are used:

- $\{p\}_1^P$ is the set of pixels in the image, where P is the image size (typically 400×400),

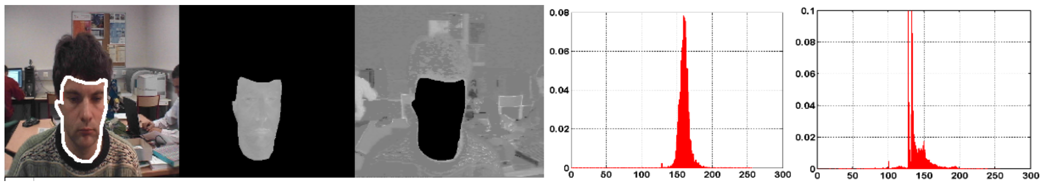
- S is the set of source vectors of size $P \times J$, where J is the dimension of the colour space. Here, one takes $J = 2$ since only the two chromatic information sources s_1 and s_2 are used for the definition of colour masses (cf. Fig. 2b). s_j represents the colour plane j of S,

- s_{jp} is an observation data. It is the jth component of the colour vector associated with pixel p,

- c_p is the class of pixel p (hidden primitive corresponding to one of the two hypotheses: face or non-face).

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

The Appriou model (Eq. 10) requires the conditional likelihood of the classes, *i.e.,* prior models which characterize the relationship between the component s_j and the hypotheses H_1 and H_2. These prior models are generated during the supervised learning step when the user selects manually on the first image of the video sequence a free-shape zone of the face including mainly skin (Fig 4a). Hair is not considered. This selection allows to exhibit both: (i) a prior model of the face zone including mainly skin hue (Fig. 4b), (ii) a prior model representative of the background by considering the pixels outside of the selected zone (Fig. 4c). Histograms are built by considering all the colour attributes s_{jp} inside the face zone, or outside (background). The conditional probability densities $p(s_j|H_1)$, respectively $p(s_j|H_2)$, are deduced from histograms by a simple normalization procedure (Fig. 4d, e).



a) b) c) d) e)

Figure 4: Initialization on sequence #2: a) selected area of the face on the first image of the video sequence; b) source s_1: face zone; c) source s_1: background; d) density of probability $p(s_1|H_1)$; e) density of probability $p(s_1|H_2)$.

Two mass sets $m_{ijp}(\{H_i\})$ (one for each class $\{H_i\}, i \in \{1;2\}$) are assigned to every pixel p with colour attribute s_{jp} so that:

$$\begin{array}{rcl} m_{ijp}(\{H_i\}) & = & 0, \\ m_{ijp}(\{\overline{H_i}\}) & = & d_{ij}[1 - R_j.p(s_{jp}|H_i)], \\ m_{ijp}(\Omega) & = & 1 - m_{ijp}(\{\overline{H_i}\}). \end{array} \tag{15}$$

$p(s_{jp}|H_i)$ is simply obtained by a look-up table (L.U.T.) addressing operation. Parameter R_j, that weighs the data model defined by the conditional likelihoods, is set to its maximal value. For simplicity, all parameters d_{ij} are initialized to the same value $d_0 = 0.9$ (but we mention in the conclusion some hints to implement a more sophisticated model). Note that one takes $d_0 < 1$ in order to guarantee the non dogmatic character of the mass sets ($m_{ijp}(\Omega) > 0$). This Appriou model exhibits two sets of complementary simple mass functions, one for each hypothesis H_i, $i \in \{1;2\}$ (Eq. 2 with weights denoted by w_{ijp}) so that for any $A = \{H_i\}$ and any source s_{jp}:

$$w_{ijp}(\overline{A}) = 1 - m_{ijp}(\overline{A}) \;\text{ from prior model }\; p(s_{jp}|A). \tag{16}$$

Keeping in mind that all elementary quantities refer to pixel p, we will omit the index p to simplify the notations in the rest of the paper, and write s_j, w_{ij}, m_{ij} etc. instead of s_{jp}, $w_{ijp}(.)$, $m_{ijp}(.)$.

4.5. Colour fusion

4.5.1. Fusion by cautious conjunctive rule

The concept of independence means intuitively that two pieces of evidence have been obtained in some sense by different ways [41]. Colour sources s_1 and s_2 (the two logarithmic hues computed from LUX space, resp. red and blue) and hence the mass functions m_{ij} are obviously not independent as they are computed from the same raw data (R, G, B in Eq. 12). Indeed, when R component varies, both values of sources U and X change. To deal with dependent sources, a solution for the fusion of beliefs consists in adopting a conservative attitude by applying the Denœux cautious conjunctive rule.

In the case of two distinct weights belonging to the interval $[0, 1]$, this rule is defined by Eq. 6, here with: $w_1 = w_{i1}$ for red chrominance (U-component), $w_2 = w_{i2}$ for blue chrominance (X-component) and $A \in \{\emptyset, \{H_1\}, \{H_2\}\}$. Then, the combined weight function denoted by w is computed such as $w(A) = \bigwedge w_{ij}(A)$. Finally, the colour masses $m_c(A) = m_{i1 \wedge i2}(A)$ assigned to each pixel p are given by Eq. 7. These masses are normalized (Eq. 8) and denoted by $m_{c^*}(.)$ yielding for each pixel significative beliefs in each class $\{H_i\}$ (see Tab. 1).

Table 1: Denœux cautious rule : computation of colour mass sets for pixel p

A	$w(.)$	$m_c(.)$	$m_{c^*}(.)$
$\emptyset$	1	$[1 - w(\{H_1\})][1 - w(\{H_2\})]$	0
$\{H_1\}$	$\min\{w_{1j}\} = \bigwedge w_{1j}(.)$	$[1 - w(\{H_1\})]w(\{H_2\})$	$m_{c^*}(\{H_1\})$
$\{H_2\}$	$\min\{w_{2j}\} = \bigwedge w_{2j}(.)$	$w(\{H_1\})[1 - w(\{H_2\})]$	$m_{c^*}(\{H_2\})$
Ω		$w(\{H_1\})w(\{H_2\})$	$m_{c^*}(\Omega)$

The idempotent combination rule constitutes an alternative to the classic conjunctive rule. Because of its conjunctive property, it strengthens the certainty during the information fusion, so that the resulting mass is more committed than the mass functions from which it is originated. Moreover it ensures that the recursive combination of an information with itself always gives the same result. In that case, the independence of information sources is not mandatory and idempotence authorizes dependence. So, a dilemma appears between reinforcement and idempotence. In our face colour model, a fusion operator with this idempotence property is favored.

Typical results of this procedure are shown in Fig. 5. The evidential model classifies correctly the zones of the image whose colour corresponds to the skin hue (face, harms). The red tee-shirt in seq. #4 is correctly detected as background by the cautious rule. The model fails however in certain background zones whose colour is close to skin hue.

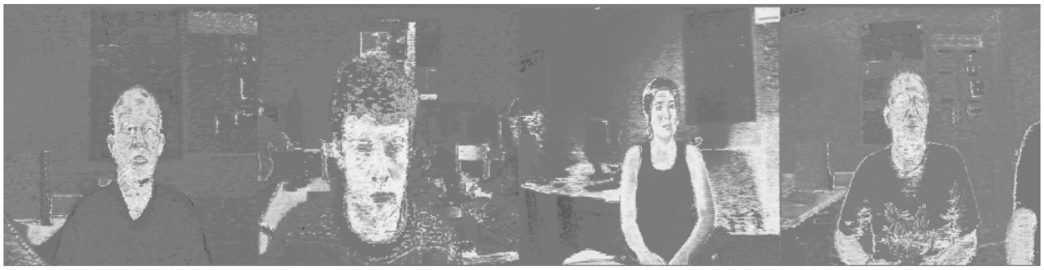


Figure 5: Fusion results (pignistic probability $BetP_p(H_1)$) of colour sources s_1 and s_2 by the cautious conjunctive rule for the four sequences of Fig. 3, with $R_j = R_{max}$ and $d_{ij} = d_0 = 0.9$.

4.5.2. Illustrative Example

Let us illustrate the processing with a sample case study. Table 2 shows the weight functions w_{ij} obtained from the following conditional probabilities $p(s_j|H_i)$: $p(s_1|H_1) = 0.05$, $p(s_1|H_2) = 0.04$, $p(s_2|H_1) = 0.07$ and $p(s_2|H_2) = 0.01$. 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.4). The combined weight $w(A)$ is simply the minimal value among the four $w_{ij}(A)$. The colour mass sets resulting from the combination of weight function w are also given in Tab. 2.

Table 2: Example of cautious colour fusion: weights w_{ij}, combined weights w and masses m_c.

A	$w_{11}(.)$	$w_{12}(.)$	$w_{21}(.)$	$w_{22}(.)$	$w(.)$	$m_c(.)$	$m_{c^*}(.)$
$\emptyset$	1	1	1	1	1	0.3645	0
$\{H_1\}$	1	0.46	1	0.19	0.19	0.4455	0.701
$\{H_2\}$	0.55	1	0.73	1	0.55	0.0855	0.1345
Ω						0.1045	0.1644

Let us compare with the classic Bayesian approach. The *a posteriori* probability is: $p(H_1|s_1, s_2) = [p(H_1)\prod_j p(s_j|H_1)]/[\sum_i p(H_i)\prod_j p(s_j|H_i)]$. If we take $p(H_1) = 0.2$, $p(H_2) = 0.8$ by supposing that the face size is kept to about 20% of the image surface thanks to the proper action of visual tracking, then $p(H_1|s_1, s_2) = 0.2(0.05 \times 0.07)/[0.2(0.05 \times 0.07) + 0.8(0.04 \times 0.01)] = 0.686$. If we suppose equiprobability $p(H_1) = 0.5$, then one obtains: $p(H_1|s_1, s_2) = 0.897$. In contrary if we have $p(H_1) = 0.1$, $p(H_2) = 0.9$ (*i.e.* the face size decreases), we get stacked in indecision ($p(H_1|s_1, s_2) \approx 0.5$). 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 example we get $Pl(\{H_1\}) = m_{c^*}(\{H_1\}) + m_{c^*}(\Omega) = 0.8655$, and $Pl(\{H_2\}) = 0.299$, or equivalently $BetP(\{H_1\}) = 0.783$, $BetP(\{H_2\}) = 0.217$): the decision is still easy to take. So that the proposed method outperforms the Bayesian approach when the prior probability decreases ($p(H_1) < 0.5$).

4.5.3. Colour fusion by compromise rule

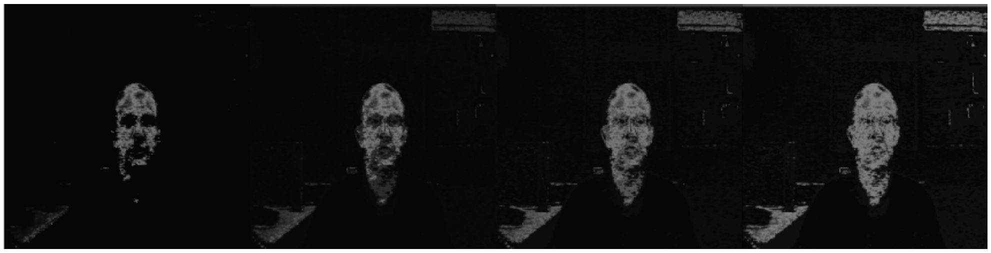
In a previous work [42], another version of Appriou's model is used allowing to obtain colour masses from only one conditional probability ($p(s_j|H_1)$), coupled with a fusion strategy which consists in an adaptive compromise rule varying between the min ($\wedge$) and the max ($\vee$) and denoted by $\wedge\vee$. In the case of two weights w_1 and w_2 belonging to the interval $[0, 1]$, this rule is defined by:

$$w_{1\ \wedge\vee\ 2}(A) = (1-\eta)\min\{w_1(A), w_2(A)\} + \eta\max\{w_1(A), w_2(A)\} \qquad (17)$$

where $A \in \{\emptyset, \{H_1\}, \{H_2\}\}$. For $\eta = 0$ we get the min used in the cautious conjunctive rule, whereas for $\eta = 1$ we get the max operator close to the disjunctive rule. Then the masses $m_c(A)$ are given by (instead of Eq. 7):

$$m_c(A) = m_{1\wedge\vee 2}(A) = \mathop{\bigcirc\hspace{-0.75em}\cap}_{A\subset\Omega} A^{w_1(A)\ \wedge\vee\ w_2(A)}. \qquad (18)$$

Typical results of this adaptive procedure are shown in Fig. 6. The fusion quality varies as a function of parameter η. By raising the value of η, the weight $w_{1\wedge\vee 2}(.) = w_1(.)\wedge\vee w_2(.)$ assigned to pixel p is increased according to the difference between w_1 and w_2, except of course when $w_1 = w_2$. On Fig. 6, the neck is poorly detected for $\eta = 0$ (*i.e.* with cautious rule) whereas for $\eta = 1$ the face is correctly detected. The counterpart of this improvement is a highlighting of certain background zones whose colour is close to the skin hue. An empirical optimum of the modelling is reached for a setting such as $\eta \approx 0.5$. So, the use of the compromise operator influences the colour model quality. In the present paper, only the limiting case: $\eta = 0$ will be considered.



$\eta = 0$ $\eta = 0.33$ $\eta = 0.66$ $\eta = 1$

Figure 6: Fusion results (pignistic probability $BetP_p(H_1)$) of colour sources s_1 and s_2 by the compromise operator for four values of η on sequence #1

Note that the value of parameter η can be learned or estimated accurately online through the computation of the covariance or coherence of the colour sources. Note also that the compromise operator $\wedge\vee$ is commutative but not associative (which could be a handicap if one wants to fuse three sources or more, and make colour-ordered fusion). Nevertheless it gives good results when one learns only the face class H_1 (i.e. with a simplistic Appriou model) and when the Florea adaptive rule is used instead of the conjunctive rule used here for final fusion of VJ and colours (Fig. 2c). Another variant of contextual fusion with three zones was also proposed in [43]. Some results about these variants will be compared in section 6.2.

4.6. Global fusion of colour and VJ mass sets by conjunctive rule

In this section, we describe the fusion of colour masses m_c with VJ-masses m_v (cf. Fig. 2c). On one hand, the colour model faithfully shows the skin hue but is not able to differentiate the face colour from that of an arm or a hand for example. 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 may fail in the case of rotated faces or background artifacts. As the informative content of these two sources is complementary, it seems 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. 7a) and the classic conjunctive rule (Fig. 7b, Eq. 3).

For $\gamma < 0.75$ the cautious rule favours the colour masses as colour weights are lower than the VJ ones. The VJ information has little influence for low values of γ, 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 $\gamma > 0$. The background is toned down proportionally to this parameter, and the effect of the bounding box is more apparent. The certainty on the face class is more strengthened with the classic conjunctive rule. One can induce from this simplistic test that the VJ information is relatively independent from the colour sources (even if this is not a formal proof of independence). This seems coherent as the VJ bounding box has been computed using *2D* Haar filters on the L component, *i.e.*, a means really different from the one used to computed colour cues U and X.

Therefore, the colour and the VJ mass functions are combined using the classic conjunctive rule (Eq.3) so that:

$$m(A) = m_{c^* \bigcap\!\!\!\!\bigcirc v}(A). \tag{19}$$

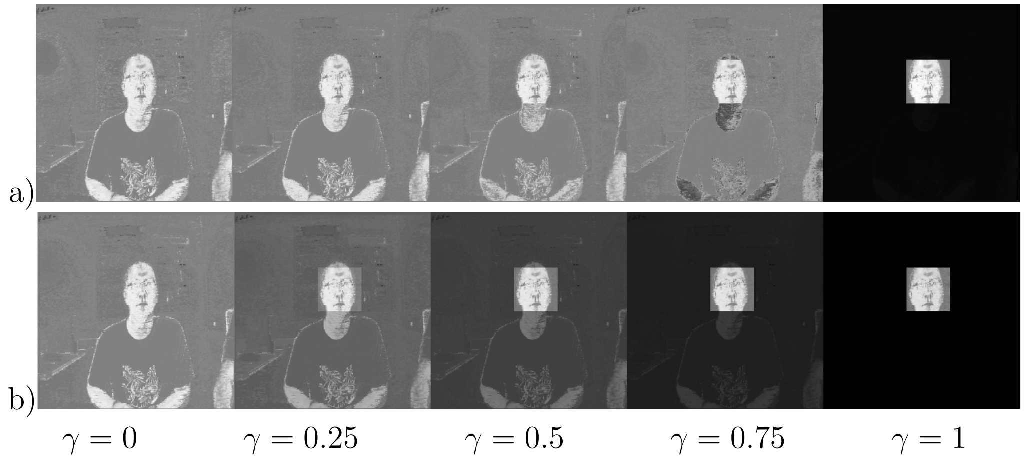


Figure 7: Fusion results of colours and VJ mass functions on sequence #4 for five values of γ: a) by the cautious conjunctive rule (with $\eta = 0$); b) by the classic conjunctive rule.

A problem occurs when the VJ detector recognizes a face-like artifact in the background (Fig. 3d) with a high reliability ($\gamma \geq 0.5$). In this case, skin colour ($m_{c^*}(\{H_1\}) < 0.5$) and VJ mass functions disagree inside the bounding box BB. This yields an important conflict inside BB. In order to limit this risk of false detection, we dynamically discount the initial value γ_0 of parameter γ by considering the global conflict inside the bounding box so that (cf. feedback loop in Fig. 2):

$$\gamma_t = \gamma_0 \quad \text{for } t = 0, \tag{20}$$

$$\gamma_t = \gamma_0(1 - K_{BB}) \quad \text{for } t > 0, \quad \text{with } K_{BB} = \frac{1}{N_{BB}} \sum_{p \in BB} K_p \tag{21}$$

$K_p = m_{c^*}(\{H_2\}) \times \gamma_t, \forall p \in BB$ is the conflict between colour and VJ elementary masses at pixel level, N_{BB} is the number of pixels inside the bounding box and 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 γ is detailed in Tab. 3.

4.7. Computation of the pignistic probabilities

This section describes the final step of the face modelling (Fig. 2d) to get pignistic probabilities. The transformation of the mass functions m into the probabilistic framework is necessary for the tracking operated by particle

Table 3: Fusion of m_{c^*} and m_v by the conjunctive rule as a function of the pixel position (inside or outside the bounding box)

A	m for $p \in BB$	m for $p \notin BB$
$\emptyset$	$m_{c^*}(\{H_2\}) \cdot \gamma_t$	$m_{c^*}(\{H_1\}) \cdot \gamma_t$
$\{H_1\}$	$m_{c^*}(\{H_1\}) + m_{c^*}(\Omega) \cdot \gamma_t$	$m_{c^*}(\{H_1\}) \cdot (1 - \gamma_t)$
$\{H_2\}$	$m_{c^*}(\{H_2\}) \cdot (1 - \gamma_t)$	$m_{c^*}(\{H_2\}) + m_{c^*}(\Omega) \cdot \gamma_t$
Ω	$m_{c^*}(\Omega) \cdot (1 - \gamma_t)$	$m_{c^*}(\Omega) \cdot (1 - \gamma_t)$

filter (section 5.1). The pignistic probability $BetP_p(.)$ associated with the face class is:

$$BetP_p(\{H_1\}) = m(\{H_1\}) + m(\Omega)/2, \quad (22)$$

$$\text{so} \begin{cases} BetP_p(\{H_1\}) = m_{c^*}(\{H_1\}) + \left(\frac{1+\gamma_t}{2}\right) m_{c^*}(\Omega), & \forall p \in BB \\ BetP_p(\{H_1\}) = (1 - \gamma_t)\left[m_{c^*}(\{H_1\}) + \frac{1}{2} m_{c^*}(\Omega)\right], & \forall p \notin BB \end{cases} \quad (23)$$

Since $BetP_p$ belongs to $[0, 1]$, it is multiplied by 255 in order to display a legible grey level image of this probability (like in Fig. 5).

Tab. 4 summarizes the evidential face model behaviour when the pixel hue is either close to that of the face ($m_{c^*}(\{H_1\}) \rightarrow 1$), really different ($m_{c^*}(\{H_2\}) \rightarrow 1$) or in between ($m_{c^*}(\{H_2\}) \rightarrow 0.5$), according to the VJ detector reliability parameter γ and to the colour uncertainty $m_{c^*}(\Omega)$.

The performance of the evidential model depends both on the colour model quality and on the VJ face detector reliability (Fig. 7). Face is correctly detected if both $\gamma \geq 0.5$ and $m_{c^*}(\{H_1\}) + m_{c^*}(\Omega) \geq 0.5$. A too low value of γ ($\gamma < 0.5$) limits the influence of the VJ face detector and finally reduces the evidential model to a simple skin colour detector. A too high value of γ ($\gamma > 0.9$) can be counter-productive when the VJ detector fails and focuses on an artifact with colour close to skin hue. Therefore we recommend to initialize the γ value such as $0.7 \leq \gamma_0 \leq 0.9$. Note that when the VJ face detector is in default, *i.e.* when it does not deliver any bounding box, γ is set to zero.

5. Probabilistic face tracking

This section describes the second part of the processing, namely the face tracking procedure (cf. Fig. 1b) that takes place after the face model detection. Face tracking is defined as the process of estimation of the shape, appearance, position and orientation parameters of one (or more) face along

Table 4: Outputs of the evidential model: fusion of colour masses $m_{c^*}(.)$ and VJ face detector reliability γ.

$m_{c^*}(.)$		VJ	$m_{c^*}(\Omega)$	$BetP_p(\{H_1\})$		decision	
$\{H_1\}$	$\{H_2\}$	γ		$p \in BB$	$p \notin BB$	$p \in BB$	$p \notin BB$
0	1	0	0	0	0	$\{H_2\}$	$\{H_2\}$
		0.5		0	0	$\{H_2\}$	$\{H_2\}$
		1		0	0	$\{H_2\}$	$\{H_2\}$
0.5	0.5	0	0	0.5	0.5	indecisive	indecisive
		0.5		0.5	0.25	indecisive	$\{H_2\}$
		1		0.5	0	indecisive	$\{H_2\}$
1	0	0	0	1	1	$\{H_1\}$	$\{H_1\}$
		0.5		1	0.5	$\{H_1\}$	indecisive
		1		1	0	$\{H_1\}$	$\{H_2\}$
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		0.375	0.125	$\{H_2\}$	$\{H_2\}$
		1		0.5	0	indecisive	$\{H_2\}$
0.5	0	0	0.5	0.75	0.75	$\{H_1\}$	$\{H_1\}$
		0.5		0.875	0.375	$\{H_1\}$	$\{H_2\}$
		1		1	0	$\{H_1\}$	$\{H_2\}$

time. The goal is to obtain in real-time the trajectory of the target (or tracked object) in the video stream [44]. Tracking techniques can be grouped into three categories, some of them already mentioned in section 2 about detection: (i) first, low level methods achieve tracking by performing colour segmentation, *e.g.*, mean-shift [45], background substraction in the case of uniform or stationary background, or optical flow estimation; (ii) second, snakes or AAM track the face by template matching [46, 47]; (iii) finally, filtering methods perform a temporal tracking by predicting the future state (localization) of a dynamic system (the target) using past measurements. Kalman filtering is employed for Gaussian uni-modal models whereas particle filter is widely used for nonlinear models, non-Gaussian processes [48]. Klein [49] implements an efficient approach for several visual tracking situations which combines disrupted sources using contextual information brought by a particle filter. An extension of Bayesian particle filters to the Dempster-Shafer theory is proposed in [50]. The algorithm presents an original solution to the problem of multi-camera people tracking in indoor environments.

In our application context, the face is a deformable object placed relatively close to the camera, whose egomotion is unpredictable with frequent direction changes. The scene is *a priori* cluttered with a varying background due to camera mobility. Therefore we have chosen a probabilistic tracking method by a bootstrap particle filter as this technique is efficient for objects with nonlinear trajectory and as it takes into account the temporal redundancy between frames. The goal is to estimate the parameters of a state vector denoted by X_t 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 centre (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 $X_t = [x_{c_t}, y_{c_t}, h_t, l_t, \theta_t]$. The particle filtering 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 $Y_t = BetP$, *i.e.,* an image whose high-value pixels indicate the presence of the face at time t. Knowing 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. 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 *a posteriori* law is approximated by:

$$p(X_t|Y_{1:t}) \approx \sum_{n=1}^{N} \omega_t^{(n)} \delta_{\lambda_t^{(n)}}. \tag{24}$$

Let us recall that the evidential face model constitutes the entry to the tracking filter (Fig. 1). The tracking algorithm begins classically with an initialization step (Fig. 8e). The zone of the face selected manually by the user during the learning stage is used to intialize the parameters of X_t. Then the algorithm is organized according to two main successive stages depicted in Fig. 8f and 8g: (i) first, the coordinates of the centre of the state vector (x_{c_t}, y_{c_t}) are estimated by particle filtering (section 5.1); (ii) then, the ellipse size and orientation (h_t, l_t, θ_t) are estimated by a second particle filtering (section 5.2). If necessary, a resampling operation [51] is triggered inbetween (Fig. 8h): it occurs when the informative content associated with the particle estimating the 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 $50 \leq N \leq 100$ is the number of particles. Then, one draws randomly new positions of the face by propagating particles following a uniform law $\mathcal{U}_X$. When a particle finds a face zone again, the filter converges after a few iterations, which ensures tracking to resume.

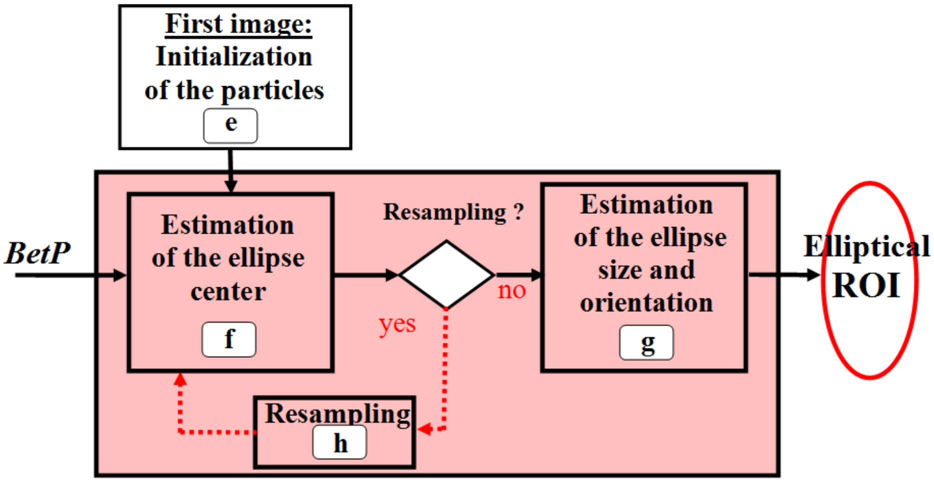

First image:
Initialization of the particles
e
BetP
Estimation of the ellipse center
f
Resampling ?
no
yes
Estimation of the ellipse size and orientation
g
Resampling
h
Elliptical ROI

Figure 8: Block-diagram of the tracking algorithm by particle filtering.

5.1. Estimation of the ellipse centre

The state vector is reduced here to $X_t = [x_{c_t}, y_{c_t}]$. A simple dynamic model inspired by the work of Pérez [52] randomly distributes the centers of the particles in the image so that:

$$p(\tilde{X}_t|X_{t-1}) = (1-\nu)\mathcal{N}(\tilde{X}_t|X_{t-1}, \Sigma) + \nu\mathcal{U}_{\tilde{X}}(\tilde{X}_t) \tag{25}$$

where $\mathcal{N}(.|\mu, \Sigma)$ is a normal Gaussian law with average μ and covariance Σ. The diagonal matrix $\Sigma = \text{diag}(\sigma_{x_{c_t}}, \sigma_{y_{c_t}}) = \text{diag}(5, 5)$ sets the *a priori* constraints: it gives the variances imposed to the position components of the state vector. The coefficient ν weights the uniform distribution: $0 \leq \nu \leq 1$. 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 component is heuristically set to $\nu = 0.1$ so that the majority of particles (90%) remains around the centre predicted at time $t-1$. It ensures some inertia in the particle distribution along time. A too high value of ν 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. 9a, the influence of the Gaussian distribution is characterized by the concentration of most particles around the centre estimated from the previous image. We see the influence of parameter ν as a few isolated particles spread over other regions in the image background.

After the particle prediction, the filter evaluates the adequacy 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 level of adequacy is expressed by the quadratic sum of pignistic probabilities $BetP_p(\{H_1\})$ contained inside the ellipse. Hence the estimated weight of each particle $\Lambda_t^{(n)}$ is given by:

$$\tilde{\omega}_t^{(n)} = \sum_{p \in \tilde{X}_t^{(n)}} [BetP_p(\{H_1\})]^2. \tag{26}$$

The adequacy criterion is the maximum likelihood. It selects the most significant ellipse and its centre defines the position components of the state vector (Fig. 9b).

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. 23) ensures the compatibility with the probabilistic framework of the particle filter (the compound hypothesis Ω does not appear any more). The mutual exclusion principle which stipulates that two hypotheses must be antagonist is respected. This is the justification of the pignistic probability choice as the output of the face model.

5.2. Estimation of size and pose

The size and pose components at time t are predicted by running again the particle filter with the same dynamic model as given by Eq. 25, but with the state vector reduced to $X_t = [h_t, l_t, \theta_t]$ (as particles are now propagated according to size and pose only, around the center (x_{c_t}, y_{c_t}) estimated previously in section 5.1) and with the parameter setting $\nu = 0$. Indeed it is not relevant to take erratic variations of the size and pose parameters of the state vector into account. Fig. 9c illustrates the distribution of the different predicted ellipses around the centre x_{c_t}, y_{c_t}.

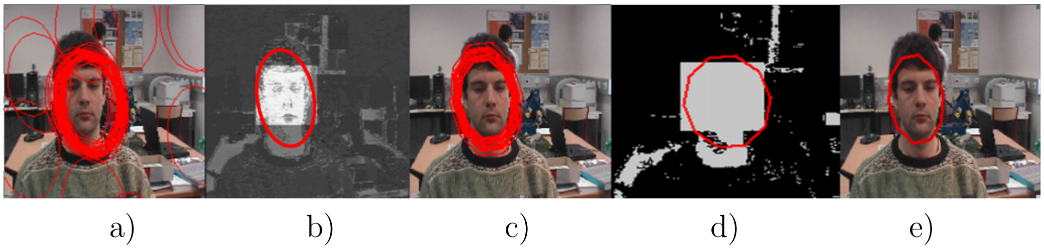


Figure 9: Sequence #2: a) particles during the centre estimation stage ($N = 50$); b) position filtering result; c) particles during the size and pose estimation step; d) measured ellipse (observation); e) ellipse filtering result.

For the correction step, the following observation is used: the pignistic probabilities stemming from the evidential model are first binarised using a simple thresholding technique. Then a morphomathematical operation of image filling is applied to this image in order to exhibit a shape (in grey on Fig. 9d) around the center (x_{c_t}, y_{c_t}), whose contour is extracted. Finally an elliptic approximation of this contour based on a least squares fitting method [53] yields a measured ellipse, which constitues the new observation (in red on Figure 9d) whose parameters are denoted by $\hat{X}_t = \left[\hat{h}_t, \hat{l}_t, \hat{\theta}_t\right]$. The correction step evaluates the importance weight $\tilde{\omega}_t^{(n)}$ as inversely proportional to the Euclidian distance between the predicted ellipse and the measured one

(Eq. 27).

$$\tilde{\omega}_t^{(n)} = p\left(Y_t^{(n)} | \tilde{X}_t^{(n)}\right) \propto \frac{1}{(\hat{h}_t - h_t^n)^2 + (\hat{l}_t - l_t^n)^2 + (\hat{\theta}_t - \theta_t^n)^2} \qquad (27)$$

At last, the maximum likelihood criterion selects the most significant particle: among all the predicted ellipses around the previoulsy estimated center (Fig 9c) the algorithm selects the one (Fig. 9e) whose size and pose are closest to the observation (measured ellipse in red on Fig. 9d). Note that maximal values of variance $\Sigma_{max} = \text{diag}(\sigma_{h_t}, \sigma_{l_t}, \sigma_{\theta_t}) = \text{diag}(5, 5, 0.1)$ imposed in the model ensure that particles deviate little from the state vector components estimated at time $t-1$.

5.3. Visual servoing

The visual servoing controls the three degrees of freedom (panoramic, tilt, zoom) of the PTZ camera (Fig. 10). The purpose is to keep the face in the center of the image plane, and this with a reasonable size (approximately 10% of the image size). The tracking (task of centering) and the zoom control (scaling) strategies are elaborated by a classic approach [54].

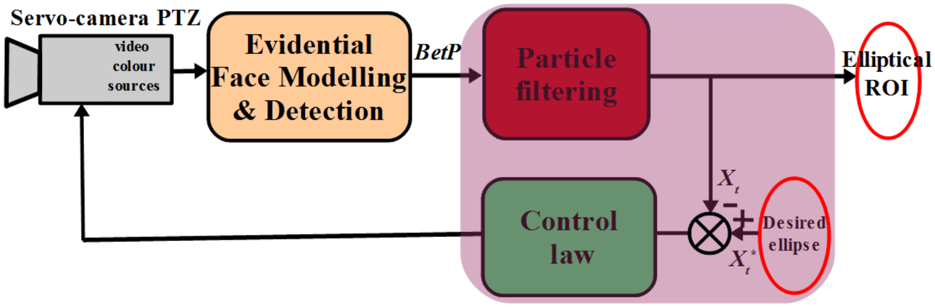
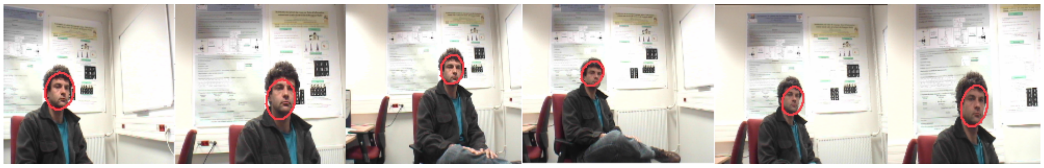


Figure 10: Visual servoing scheme: $X_t^* = [0, 0, 120, 100, 0]$ is the servoing command and X_t is the state vector coming from the particle filter.

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 operator moves backward (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 with visual servoing of the camera in position (pan and tilt) and control of the zoom (for sequence #8).

6. Performance analysis

Performance evaluation of tracking systems is mandatory. However this requires both the definition of quantitative criteria like precision, robustness or execution time, and the availability of a ground truth (GT), that is, a set of data coding the real positions of the face image by image. However the task of obtaining the GT by a human expertise is relatively subjective and tedious. 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 (Fig. 3). During a total occlusion, the face is supposed to be missing.

6.1. Qualitative evaluation

The algorithm behaviour is illustrated with two sequences: (i) in the presence of total or partial occlusion and pose variations with sequence #1 registered in our laboratory, (ii) in the presence of pose changes, lighting and background variations, disruptive elements (the operator removes then puts his glasses back again) with complex sequence David Indoor used in numerous recent articles [55].

In sequence #1 (Fig. 12), the Viola and Jones masses increase the informative content in the face zone on images im_{57} and im_{73}: the pignistic probabilities are most significant (white pixels in Fig. 12b) on the face zone where colour and VJ attributes are fused, but not on other skin colour regions (arms, hands, or neck). No bounding box is delivered by the VJ detector in the case of images $im_{60}, im_{66}, im_{69}$, so that $\gamma_t = 0$ is set in the evidential model since only colour information is valid. Therefore, in the presence of total occlusion (im_{66}), the resulting ellipse lies on the hand of the user. The uniform component ($\nu = 0.1$) of the particle filter dynamic model (Eq. 25) ensures a correct repositioning when a candidate particle locates again on

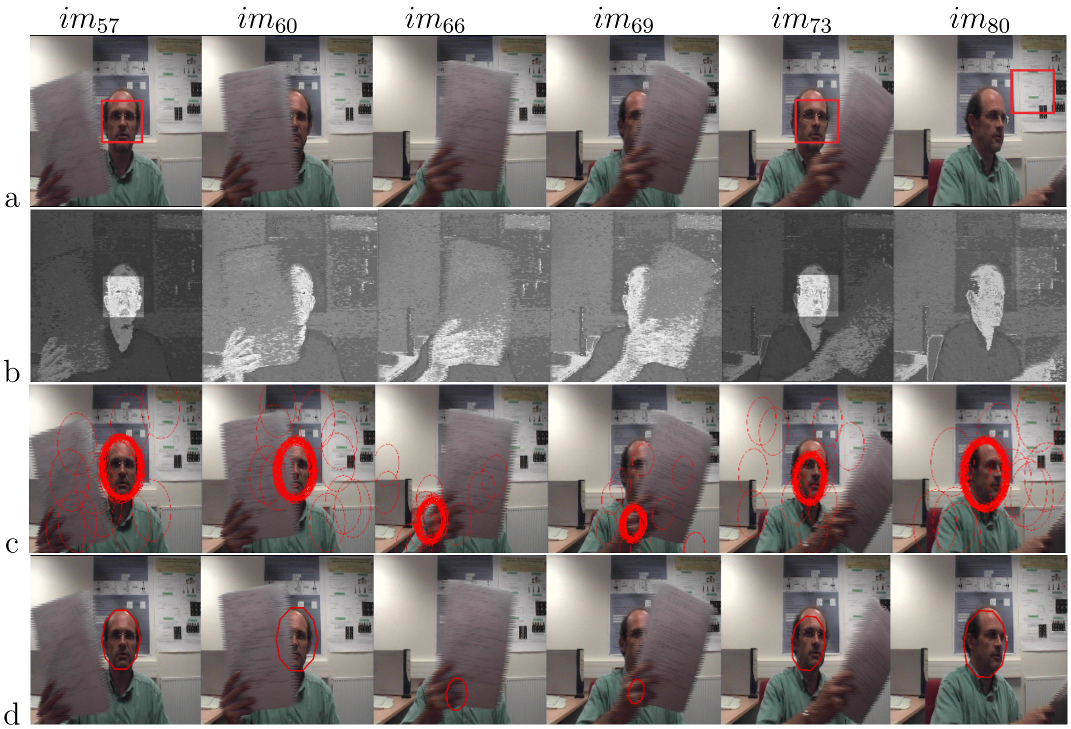


Figure 12: Face tracking for sequence #1: a) Bounding box (in red) supplied by the VJ face detector; b) Pignistic probability stemming from the face model; c) Particle filter particles during the centre estimation step of X_t; d) Ellipse resulting from particle filter.

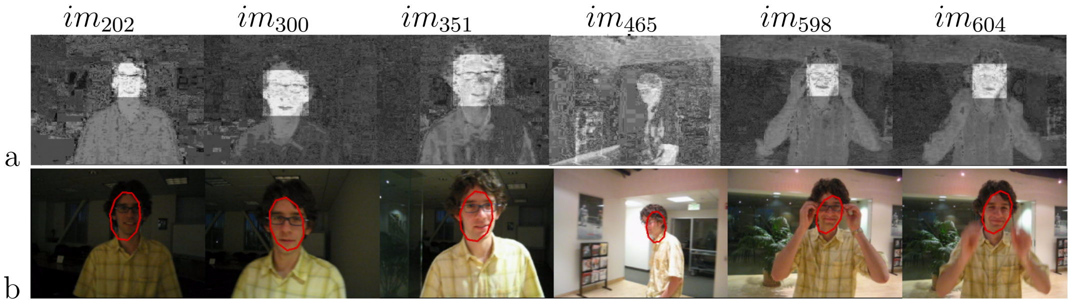


Figure 13: Results of tracking on sequence David Indoor: a) evidential fusion (pignistic probability $BetP_p(H1)$); b) ellipse positioning.

the face zone (im_{73}). The VJ information may degrade the tracking quality when the VJ detector focuses on a face-like artifact (frame im_{80}). An important conflict ($K_{BB} = 0.7$) is measured inside the VJ bounding box. Then the adjustment of parameter γ_t ($\gamma_t \rightarrow 0.24$ as $\gamma_0 = 0.8$) favours more the colour 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}), but not on the first frame of the video sequence as usually. Indeed the complete absence of lighting on this first frame precludes the exhibition of a prior model representative of the skin colour.

On frames im_{202} and im_{300}, the pignistic probabilities are most significant in the face zone where colour 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 the 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 colour information is considered ($\gamma_t = 0$). When hands are in contact with the face in frame im_{598}, the estimations of center and pose remain correct. When the hands go away from the face, they are not tracked any longer (frame im_{604}). This shows the robustness of the proposed method: the presence of disruptive elements alters weakly pose and size estimation and only slighty perturbs the tracking in position.

6.2. Quantitative evaluation

In order to quantify the tracking performances in various contexts on statistically significant data, we have manually segmented (*i.e.* cut-out) the face in 1,400 images of 7 video sequences registered in our laboratory (giving the ground truth GT at a rate of 1 image per second, which is appropriate for the application), and in 500 images of the David Indoor benchmark sequence [55].

Additional sequences complement those already presented in Fig. 12 and Fig. 13: in sequence #2 another person moves in the background (Fig. 3b); in sequence #3 a woman with red make-up rotates on her chair (Fig. 3c); in sequence #4 the person dressed with a reddish tee-shirt is in front-view and removes his glasses (Fig. 3d); in sequence #5 a black person goes away from the camera; in sequence #6 a person moves his head near the camera and in sequence #7 a person rotates on his chair and removes his glasses.

Pixels located inside the cut-out face represent the ground truth (GT). The tracking algorithm delivers an ellipse denoted by ROI (region of interest) derived from the particle filter. True positive pixels (TP) belong to the intersection: $TP = GT \cap ROI$, whereas false positives (FP) lay outside of GT: $FP = ROI \cap \overline{GT}$.

Two measures are used to quantify the tracking performance, namely the Precision and the Recall defined as follows:

$$\text{Precision} = \frac{|TP|}{|ROI|} \qquad\qquad \text{Recall} = \frac{|TP|}{|GT|}. \tag{28}$$

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. They are 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 proposed method. From these measurements, the ROC curves (Receiver Operating Characteristics) are built with coordinates $x = (1-\text{Precision})$ and $y = \text{Recall}$, 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 setting of the parameter value. This study gives the sensibility of the method to the VJ detector reliability parameter γ and to the compromise parameter η (section 4.5.3).

Three ROC curves are displayed in Fig. 14. The curve “all data” shows the global tracking performances resulting from the whole dataset whereas the curve “seq.#1(cautious rule)” shows only results from data-subset of sequence #1. The curve “seq.#1(compromise rule)” displays the best performances obtained with the approach presented in section 4.5.3.

The curve “all data” indicates that when colour information only is used (*i.e.*, $\gamma = 0$), the global tracking quality is poor (Precision ≈ 0.57, Recall ≈ 0.82). Tracking is notably improved by a weak contribution of the VJ detector and the best performances are reached for $\gamma \approx 0.3$ (Precision ≈ 0.63, Recall ≈ 0.72). When γ increases ($\gamma > 0.6$), the VJ information plays a major role in the evidential face model. Precision is quasi-constant (≈ 0.73) while Recall is poor and little varies ($\in [0.53; 0.56]$).

The point drawn for the adaptive parameter $\gamma = \gamma_t$ shows the tracking performances obtained when the discounting factor by feedback is imple-

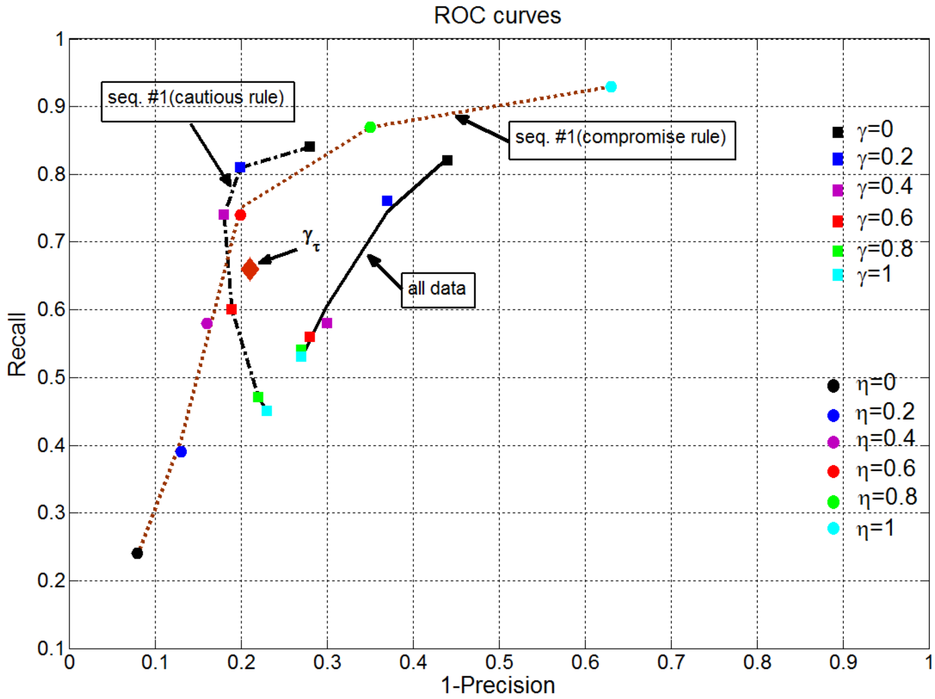


Figure 14: ROC curves: "all data" resulting from the whole dataset (with $\gamma \in [0, 1]$ and with $\gamma = \gamma_t$); "seq.#1(cautious rule)" resulting from the data-subset of sequence #1 with $\gamma \in [0, 1]$; "seq.#1(compromise rule)" resulting from the data-subset of sequence #1 for the approach presented in section 4.5.3 (with $\gamma = 0.1$ and $\eta \in [0, 1]$)

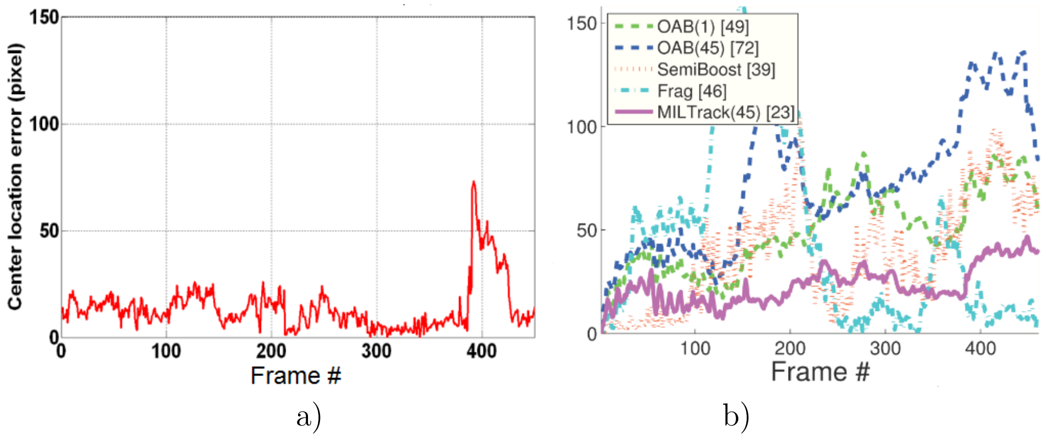


Figure 15: Tracking results (centre location error) on the sequence David Indoor, with: a) the proposed method b) various algorithms according to Babenko [55].

mented (Eq. 21). This dynamic setting of γ leads to a performance optimization (Precision ≈ 0.79, Recall ≈ 0.67).

For sequence #1 where the face is often in profile or hidden, the colour masses are more relevant than the VJ ones. Two approaches are compared: (i) the method (in section 4.5.3) that uses a scant Appriou model (only one prior model $p(s_j|H_1)$) and a compromise rule to merge the colour mass functions in the face evidential model, (ii) the more complex model in current work that uses a rigorous Appriou model and the cautious rule to fuse the colour mass functions. For the first model, the optimal performances are reached when $\gamma = 0.1$ and for a value of the compromise parameter correctly adjusted, *i.e.*, $\eta \approx 0.65$ (Fig 14, seq.#1 compromise rule). So, a joint action of compromise on the colour masses and moderation on the VJ masses leads to performance optimization (Precision and Recall $\approx$ 75%). The second model improves the optimal performances of the first one (Fig 14 seq. #1 cautious rule) for $0 < \gamma < 0.4$ (Precision and Recall $\approx$ 80%). Results are comparable to those of standard classifiers whose detection rate reaches 70 to 80% [56].

This comparative study shows that the best performances are reached with the complex evidential face model. It is also more robust w.r.t. context variations. However the searching of more simplicity can be profitable: (i) the settings are made easier because of the low number of parameters, (ii) the computation cost is reduced what is useful for real time applications. Satisfactory performances are reached with the simplest method for a particular sequence with a parameter setting adapted to the application context. So this simple algorithm could be preferred when the background and the lighting conditions vary in a limited way during the video sequence.

Another important evaluation criterion for the assessment of the algorithm performance is the center location error denoted by:

$$\varepsilon = \sqrt{(x_{GT_t} - x_{c_t})^2 + (y_{GT_t} - y_{c_t})^2},$$

where x_{GT_t}, y_{GT_t} are the coordinates of the face gravity centre 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 $\varepsilon_{max} = 30$ pixels during the majority of the sequence (Fig. 15a), our algorithm exceeds the performances of the best algorithm (MILTrack) evaluated in [55] (Fig. 15b). Our approach fails locally on the images 380 to 430, *i.e.,* when our algorithm positions on an artifact.

In Tab. 5, the mean location error ε_{mean} and the standard deviation Σ_{mean} are estimated on the whole set of images for: (i) each sequence, (ii) all sequences (all data). The average localisation error is of 23 pixels with a standard deviation of 36 pixels. These performances are of the same order as those presented in the literature about face tracking by particle filter. The average localisation error during the tracking is of 22.4 pixels with the Condensation algorithm and of 16.3 pixels with the adaptive particle filter APF [57].

Table 5: Mean location error and standard deviation of: sequences #1 to #7, the sequence David Indoor and the whole set of images

Seq	#1	#2	#3	#4	#5	#6	#7	Dav Ind	all data
Mean	24	17	56	17	11	13	32	15	23
Std	36	12	61	19	4	30	47	11	36

As regards the computation cost, the processing time is ≈ 1s/image for an image of size 400 × 400 with a Pentium 4, CPU 2.4 GHz and 500 Mo of RAM memory. The computation simplicity makes this method usable in a real-time video (even if it is not the case in our actual prototype developed with Matlab and LabVIEW to make simulations easier).

7. Discussion

This paper has presented an original method both for face detection based on an evidential modelling and for face tracking with a classical bootstrap particle filter technique. Our previous theorical contribution was to propose a compromise operator in the colour fusion process. Here we adopt another strategy which takes the background class H_2 in addition to face class H_1 into account. Concerning the face tracking application, Precision and Recall rates may reach 80% with an adequate parameter setting, but noteworthy without having to build a huge learning database, which is the originality of our approach. The computation simplicity makes this method usable in a real-time video. Our results show a robustness improvement of the dynamic fusion thanks to idempotent combination rules which limit the belief contraction. By setting jointly the adaptive parameter values of the evidential model and the particle filter, we show that it is possible to finely tune the tracking behaviour.

The statistical results of section 6.2 confirm the qualitative observations reported in section 6.1. In the current work, the optimal setting of parameter values ($\gamma = 0.3$) is deduced from the averaging of experimental results. Consequently, this study poorly estimates the setting of the parameter for a transient variation of context on a part of the video sequence (but it still works). A time-dynamic adjustment of parameters is required to improve the tracking robustness (as done for γ in Eq. 21).

As future works, the dynamic setting of the algorithm parameters is under investigation. Besides, distinct values for parameter γ could be canvassed ($\gamma_1 \neq \gamma_2$) and also various values for parameters d_{ij}. Indeed, *a priori* knowledge about the acquisition could be used for that 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 ($\Rightarrow \gamma_1 > \gamma_2$). The mass function modelling 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 maximize the beliefs.

Acknowledgements

We thank the anonymous reviewers for their fruitful remarks, that gave us new ideas to improve the method proposed in this paper, and we thank Gisèle Sigal, assistant professor in English at our institute, for her help in reviewing the lingusitic quality of the manuscript.

References

[1] P. Smets, R. Kennes, The Transferable Belief Model, Artificial Intelligence 66 (2) (1994) 191–243.

[2] Z. Hammal, L. Couvreur, A. Caplier, M. Rombaut, Facial Expression Classification: an Approach Based on the Fusion of Facial Deformations Using the Transferable Belief Model, International Journal of Approximate Reasoning 46 (3) (2007) 542–567.

[3] E. Ramasso, C. Panagiotakis, M. Rombaut, D. Pellerin, Belief Scheduler Based on Model Failure Detection in the TBM Framework. Application

to Human Activity Recognition, International Journal of Approximate Reasoning 51 (7) (2010) 846–865.

[4] E. Hjelmås, B. K. Low, Face Detection: a Survey, Computer Vision and Image Understanding 83 (2001) 236–274.

[5] M.-H. Yang, D. Kriegman, N. Ahuja, Detecting Faces in Images: a Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (1) (2002) 35–58.

[6] C. Zhang, Z. Zhang, A Survey of Recent Advances in Face Detection, Technical Report 66, Microsoft Research, Redmond, Washington, USA (June 2010).

[7] Y. Huang, X. Ao, Y. Li, Real Time Face Detection Based on Skin Tone Detector, International Journal of Computer Science and Network Security 9 (7) (2009) 71–77.

[8] J. Zhang, Q. Zhang, J. Hu, RGB Color Centroids Segmentation (CCS) for Face Detection, ICGST International Journal on Graphics, Vision and Image Processing, GVIP 9 (2) (2009) 1–9.

[9] P. Kakumanu, S. Makrogiannis, N. Bourbakis, A Survey of Skin-Color Modeling and Detection Methods, Pattern Recognition 40 (3) (2007) 1106–1122.

[10] P. Viola, M. Jones, Robust Real-Time Face Detection, International Journal of Computer Vision 57 (2) (2004) 137–154.

[11] Y. W. Wu, X. Y. Ai, An Improvement of Face Detection Using AdaBoost with Color Information, in: Proceedings of the 2008 ISECS International Colloquium on Computing, Communication, Control, and Management (CCCM'08), Guangzhou, China, 2008, pp. 317–321.

[12] T. F. Cootes, C. J. Taylor, D. H. Cooper, J. Graham, Active Shape Models - Their Training and Application, Computer Vision and Image Understanding 61 (1) (1995) 38–59.

[13] T. F. Cootes, G. J. Edwards, C. J. Taylor, Active Appearance Models, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (6) (2001) 681–685.

[14] A. P. Dempster, Upper and Lower Probabilities Induced by a Multivalued Mapping, Annals of Mathematical Statistics 38 (1967) 325–339.

[15] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, NJ, 1976.

[16] P. Smets, The Combination of Evidence in the Transferable Belief Model, IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (5) (1990) 447–458.

[17] W. Pieczynski, Multisensor Triplet Markov Chains and Theory of Evidence, International Journal of Approximate Reasoning 45 (1) (2007) 1–16.

[18] S. B. Chaabane, M. Sayadi, F. Fnaiech, E. Brassart, Colour Image Segmentation Using Homogeneity Method and Data Fusion Techniques, Eurasip Journal on Advances in Signal Processing (2010) 1–12.

[19] I. Bloch, Defining Belief Functions Using Mathematical Morphology. Application to Image Fusion under Imprecision, International Journal of Approximate Reasoning 48 (2) (2008) 437–465.

[20] P. Smets, The Canonical Decomposition of a Weighted Belief, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95), Morgan Kaufman, San Mateo, California, USA, 1995, pp. 1896–1901.

[21] B. B. Yaglane, P. Smets, K. Mellouli, Independence Concepts for Belief Functions, in: 8th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), Vol. 1, Madrid, Spain, 2000, pp. 357–364.

[22] P. Smets, Belief Functions: the Disjunctive Rule of Combination and the Generalized Bayesian Theorem, International Journal of Approximate Reasoning 9 (1993) 1–35.

[23] A. Martin, C. Osswald, Towards a Combinaison Rule to Deal with Partial Conflict and Specificity in Belief Functions Theory, in: International Conference on Information Fusion (FUSION'07), Quebec, Canada, 2007, pp. 9–12.

[24] F. Smarandache, J. Dezert, Advances and Applications of DSmT for Information Fusion, Collected Works, Vol. 3, American Research Press, 2009.

[25] F. Pichon, T. Denœux, Interpretation and Computation of Alpha-Junctions for Combining Belief Functions, in: 6th International Symposium on Imprecise Probability: Theories and Applications (ISIPTA'09), Durham, United Kingdom, 2009, pp. 1–10.

[26] M. C. Florea, A.-L. Jousselme, E. Bossé, D. Grenier, Robust Combination Rules for Evidence Theory, Information Fusion 10 (2009) 183–197.

[27] A. Martin, C. Osswald, J. Dezert, F. Smarandache, General Combination Rules for Qualitative and Quantitative Beliefs, Journal of Advances in Information Fusion 3 (2) (2008) 67–82.

[28] T. Denœux, Conjunctive and Disjunctive Combination of Belief Functions Induced by Non Distinct Bodies of Evidence, Artificial Intelligence 172 (2008) 234–264.

[29] B. Quost, M. H. Masson, T. Denœux, Classifier Fusion in the Dempster-Shafer Framework Using Optimized t-Norm Based Combination Rules, International Journal of Approximate Reasoning 52 (3) (2011) 353–374.

[30] A. Kallel, S. Le Hégarat-Mascle, Combination of Partially Non-Distinct Beliefs: the Cautious-Adaptive Rule, International Journal of Approximate Reasoning 50 (7) (2009) 1000–1021.

[31] T. Denœux, A k-Nearest Neighbour Classification Rule Based on Dempster-Shafer Theory, IEEE Transactions on Systems, Man and Cybernetics 25 (5) (1995) 804–813.

[32] L. M. Zouhal, T. Denoeux, An Evidence-Theoretic k-NN Rule Parameter Optimization, IEEE Transactions on Systems, Man and Cybernetics - Part C 28 (2) (1998) 263–271.

[33] P. Walley, S. Moral, Upper Probabilities Based Only on the Likelihood Function, Journal of Royal Statistical Society, Series B 61 (Part 4) (1999) 831–847.

[34] P. Smets, Bayes' Theorem Generalized for Belief Functions, in: European Conference on Artificial Intelligence (ECAI'86), Vol. 2, Brighton, UK, 1986, pp. 169–171.

[35] 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), 1999, pp. 5.1–5.31.

[36] T. Denœux, P. Smets, Classification Using Belief Functions: the Relationship between the Case-Based and Model-Based Approaches, IEEE Transactions on Systems, Man and Cybernetics B 36 (6) (2006) 1395–1406.

[37] M. Liévin, F. Luthon, Nonlinear Color Space and Spatiotemporal MRF for Hierarchical Segmentation of Face Features in Video, IEEE Transactions on Image Processing 13 (2004) 63–71.

[38] G. Deng, J.-C. Pinoli, Differentiation-Based Edge Detection Using the Logarithmic Image Processing Model, Journal of Mathematical Imaging and Vision 8 (1998) 161–180.

[39] D. Alleysson, J. Hérault, Variability in Color Discrimination Data Explained by a Generic Model with Nonlinear and Adaptive Processing, Color Research and Application 26 (2001) –, supplement.

[40] F. Luthon, B. Beaumesnil, N. Dubois, LUX Color Transform for Mosaic Image Rendering, in: Proceedings of the 17th IEEE Int. Conf. on Automation, Quality and Testing, Robotics (AQTR 2010), Vol. 3, Cluj-Napoca, Romania, 2010, pp. 93–98.

[41] R. R. Yager, Aggregating Non-Independent Dempster-Shafer Belief Structures, in: 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), Vol. 1, Màlaga, Spain, 2008, pp. 289–297.

[42] F. Faux, F. Luthon, Théorie de l'Evidence pour Suivi de Visage, Revue Traitement du Signal 28 (5) (2011) 517–547.

[43] F. Faux, F. Luthon, Robust Face Tracking Using Colour Dempster-Shafer Fusion and Particle Filter, in: The 9th International Conference on Information Fusion (FUSION'06), Firenze, Italy, 2006, pp. 1–7.

[44] A. Yilmaz, O. Javed, M. Shah, Object Tracking: a Survey, ACM Computing Surveys 38 (4) (2006) 1–45.

[45] D. Comaniciu, V. Ramesh, P. Meer, Kernel-Based Object Tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (5) (2003) 564–575.

[46] M. Isard, J. MacCormick, BraMBLe: a Bayesian Multiple-Blob Tracker, in: IEEE International Conference on Computer Vision (ICCV), 2001, pp. 34–41.

[47] Y. Rathi, N. Vaswani, A. Tannenbaum, A. Yezzi, Tracking Deforming Objects Using Particle Filtering for Geometric Active Contours, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (8) (2007) 1470–1475.

[48] M. S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking, IEEE Transactions on Signal Processing 50 (2) (2002) 174–188.

[49] J. Klein, C. Lecomte, P. Miché, Hierarchical and Conditional Combination of Belief Functions Induced by Visual Tracking, International Journal of Approximate Reasoning 51 (4) (2010) 410–428.

[50] R. Muñoz-Salinas, R. Medina-Carnicer, F. J. Madrid-Cuevas, A. Carmona-Poyato, Multi-Camera People Tracking Using Evidential Filters, International Journal of Approximate Reasoning 50 (5) (2009) 732–749.

[51] A. Doucet, S. J. Godsill, C. Andrieu, On Sequential Monte Carlo Sampling Methods for Bayesian Filtering, Statistics and Computing 10 (3) (2000) 197–208.

[52] P. Pérez, J. Vermaak, A. Blake, Data Fusion for Visual Tracking with Particles, Proceedings of IEEE 92 (3) (2004) 495–513.

[53] A. Fitzgibbon, M. Pilu, R. Fisher, Direct Least Squares Fitting of Ellipses, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (5) (1999) 476–480.

[54] F. C. A. Crétual, F. Chaumette, P. Bouthemy, Complex Object Tracking by Visual Servoing Based on 2D Image Motion, in: International Conference on Pattern Recognition (ICPR), Vol. 2, Brisbane, Australia, 1998, pp. 1251–1254.

[55] B. Babenko, M. H. Yang, S. Belongie, Robust Object Tracking with Online Multiple Instance Learning, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (8) (2011) 1619–1632.

[56] M. Castrillón, O. Déniz, D. Hernández, J. Lorenzo, A Comparison of Face and Facial Feature Detectors Based on the Viola-Jones General Object Detection Framework, Machine Vision and Applications 22 (2011) 481–494.

[57] W. Zheng, S. M. Bhandarkar, Face Detection and Tracking Using a Boosted Adaptive Particle Filter, Journal of Visual Communication and Image Representation 20 (2009) 9–27.