Theory of evidence for face detection and tracking
Francis Faux, Franck Luthon
University of Pau and Adour River UPPA, Computer Science Laboratory LIUPPA, IUT GIM Anglet, France

ARTICLE INFO

Article history:
Received 6 April 2011
Received in revised form 16 February 2012
Accepted 16 February 2012
Available online xxx

Keywords:
Face tracking
Transferable belief model
Cautious rule
Particle filter
LUX colour space
Computer vision

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.

© 2012 Elsevier Inc. All rights reserved.

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......
Face detection methods can be grouped into two categories differing as to the processing of prior information [4–6]. 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 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 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 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 choice of the data patterns extracted from the image.

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. Whathoever, 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 $\Omega$, is the finite set of answers to this question. A mass
function is an application of the $2^\Omega$ parts of $\Omega$ towards the interval $[0, 1]$ which satisfies:

$$\sum_{A \subseteq \Omega} m(A) = 1.$$  \hspace{1cm} (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 $A \subseteq \Omega$ is set along with a weight function $w \in [0, 1]$ so that:

$$m(A) = 1 - w,$$

$$m(\Omega) = w.$$  \hspace{1cm} (2)

Denoted by $m = A^w$, it represents the belief put in $\Omega$ 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 $\Omega$ is a focal set) may be expressed, by canonical decomposition [20], as the conjunctive combination (see definition
below) of simple mass functions: $m(A) = \bigcap_{A \subseteq \Omega} A^w(A)$.

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

$$m_{\cap}(\Omega) = \sum_{A_1 \cap \ldots \cap A_f = A} \left( \prod_{j=1}^{f} m_j(A) \right), \quad \forall A \subseteq \Omega.$$  \hspace{1cm} (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 ($m_{\cap}(\emptyset) \neq 0$). Dempster proposed
a normalization version of this law better known as the Dempster combination rule or orthogonal sum [14]:
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 [23–25]. 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 $\otimes$, 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 $m_1$ 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:

$$w_{1 \otimes 2}(A) = w_1(A) \wedge w_2(A) \quad \forall A \subseteq \Omega,$$

$$m_{1 \otimes 2}(A) = \bigcap \omega A_{\wedge}^{w_1}(A) \wedge w_2(A).$$

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

$$m_{1 \oplus 2}(A) = \frac{m_{1 \otimes 2}(A)}{1 - m_{1 \otimes 2}(\emptyset)}, \quad \forall A \subseteq \Omega, A \neq \emptyset,$$

$$m_{1 \oplus 2}(\emptyset) = 0.$$  

The bold disjunctive rule, denoted by $\odot$, 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]:

$$m_{ij}(\{H_i\}) = 0,$$

$$m_{ij}(\{\overline{H_i}\}) = d_{ij} [1 - R_j p(x_j|H_i)],$$

$$m_{ij}(\Omega) = 1 - m_{ij}(\{\overline{H_i}\}).$$

is a discounting coefficient which characterizes the a priori degree of confidence in the knowledge 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 irrelevant. $R_j$ 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 [1,16]. 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:

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

(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.

![Fig. 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.](http://dx.doi.org/10.1016/j.ijar.2012.02.002)
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 $\Omega$, 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 $\Omega$, $\{H_1\}$ 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 $M = 256$):

$$
L = (R + 1)^{0.3}(G + 1)^{0.6}(B + 1)^{0.1} - 1
$$

$$
U = \begin{cases}
  \frac{M}{2} \left( \frac{R+1}{L+1} \right) & \text{for } R < L \\
  M - \frac{M}{2} \left( \frac{L+1}{R+1} \right) & \text{otherwise}
\end{cases}
$$

$$
X = \begin{cases}
  \frac{M}{2} \left( \frac{B+1}{L+1} \right) & \text{for } B < L \\
  M - \frac{M}{2} \left( \frac{L+1}{B+1} \right) & \text{otherwise}.
\end{cases}
$$

$L$ 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).

In order to model the VJ attribute by a belief function, a simple mass denoted by $m_v(\cdot)$ 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:

$$
m_v = \{H_1\}^{1-\gamma}, \forall p \in \text{BB}, \quad (13)
$$

$$
m_v = \{H_2\}^{1-\gamma}, \forall p \notin \text{BB}. \quad (14)
$$

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 $\Omega$. 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.

![Fig. 3. Bounding box produced by the VJ face detector in various sequences: (a)–(c) correct detection and (d) false detection.](http://dx.doi.org/10.1016/j.ijar.2012.02.002)
4.4. Colour masses

This section together 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:

- $\{p\}^P$ is the set of pixels in the image, where $P$ is the image size (typically $400 \times 400$),
- $S$ 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 source. It is the $j$th 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).

Two mass sets $m_{ijp}(H_1)$ (one for each class $H_i$, $i \in \{1; 2\}$) are assigned to every pixel $p$ with colour attribute $s_{jp}$ so that:

$$
\begin{align*}
    m_{ijp}(\{H_1\}) &= 0, \\
    m_{ijp}(\{H_1\}) &= d_j [1 - R_j \cdot p(s_{jp} | H_1)], \\
    m_{ijp}(\Omega) &= 1 - m_{ijp}(\{H_i\}).
\end{align*}
$$

$p(s_{jp} | H_1)$ is simply obtained by a look-up table (L.U.T.) addressing operation. Parameter $R_j$, that weights the data model defined by the conditional likelihoods, is set to its maximal value. For simplicity, all parameters $d_j$ 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 $s_{jp}$:

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

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 $s_{jp}, w_{ijp}(\cdot), m_{ijp}(\cdot)$.

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.
4.5.2. Illustrative example

The cautious rule. The model fails however in certain background zones whose colour is close to skin hue. whose colour corresponds to the skin hue (face, harms). The red tee-shirt in seq. #4 is correctly detected as background by mc 0 for red chrominance (Fig. 5). surface thanks to the proper action of visual tracking, then it always gives the same result. In that case, the independence of information sources is not mandatory and idempotence property, it strengthens the certainty during the information fusion, so that the resulting mass is more committed than the sample histograms in Fig. 4). The combined weight function denoted by w(·) decreases (Fig. 8). If we suppose equiprobability (9). Typical results of this procedure are shown in Fig. 5. The evidential model classifies correctly the zones of the image H 07, or equivalently BetP(H 1 |Ω) = 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).

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 1 for red chrominance (U-component), w 2 = w 2 for blue chrominance (X-component) and A ∈ {∅, {H 1 }, {H 2 }}. Then, the combined weight function denoted by w is computed such as w(A) = \bigwedge w j (A). Finally, the colour masses m c (A) = m 1(· ∩ 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 j } (see Table 1).

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.

4.5.2. Illustrative example

Let us illustrate the processing with a sample case study. Table 2 shows the weight functions w j 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 0 is set to its maximal value: R 0 = R 1 = 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 j (A). The colour mass sets resulting from the combination of weight function w are also given in Table 2.

Let us compare with the classic Bayesian approach. The a posteriori probability is: p(H 1 |s 1 , s 2 ) = [p(H 1 ) \bigwedge p(s 1 |H i )] / [\bigwedge p(H i ) \bigwedge p(s 1 |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 × 0.07)/[0.2(0.05 × 0.07) + 0.8(0.04 × 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 staked in indecision (p(H 1 |s 1 , s 2 ) ≈ 0.5). Similarly to the maximum a posteriori criterion, the evidential decision consists in choosing the hypothesis {H j } 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 ) = 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).

Table 1

<table>
<thead>
<tr>
<th>A</th>
<th>w(·)</th>
<th>m c · (·)</th>
</tr>
</thead>
<tbody>
<tr>
<td>∅</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>{H 1 }</td>
<td>min[w 1</td>
<td>min[w 1</td>
</tr>
<tr>
<td>{H 2 }</td>
<td>min[w 2</td>
<td>min[w 2</td>
</tr>
<tr>
<td>Ω</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2

<table>
<thead>
<tr>
<th>A</th>
<th>w 1(·)</th>
<th>w 2(·)</th>
<th>w 21(·)</th>
<th>w 22(·)</th>
<th>w(·)</th>
<th>m c · (·)</th>
</tr>
</thead>
<tbody>
<tr>
<td>∅</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.3645</td>
</tr>
<tr>
<td>{H 1 }</td>
<td>1</td>
<td>0.46</td>
<td>1</td>
<td>0.19</td>
<td>0.19</td>
<td>0.4455</td>
</tr>
<tr>
<td>{H 2 }</td>
<td>0.55</td>
<td>1</td>
<td>0.73</td>
<td>1</td>
<td>0.55</td>
<td>0.0855</td>
</tr>
<tr>
<td>Ω</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.1045</td>
</tr>
</tbody>
</table>

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:

\[
\begin{align*}
    w_1 \wedge \vee w_2(A) &= (1 - \eta) \min\{w_1(A), w_2(A)\} + \eta \max\{w_1(A), w_2(A)\} \\
    m_c(A) &= m_1 \wedge \vee m_2 = \bigwedge_{A \in \Omega} \bigvee w_1(A) \wedge \vee w_2(A).
\end{align*}
\]

Typical results of this adaptive procedure are shown in Fig. 6. The fusion quality varies as a function of parameter \(\eta\). By raising the value of \(\eta\), the weight \(w_1 \wedge \vee w_2(. ) = w_1(.) \wedge \vee w_2(.)\) 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.

Note that the value of parameter \(\eta\) 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 \(\gamma\), 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:

\[ m(A) = m_c \cap m_v(A). \]  

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 \( m_c((H_1)) \leq 0.5 \) 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 \( \gamma_0 \) of parameter \( \gamma \) by considering the global conflict inside the bounding box so that (cf. feedback loop in Fig. 2):

\[
\begin{align*}
\gamma_t &= \gamma_0 & \text{for } t = 0, \\
\gamma_t &= \gamma_0 (1 - K_{BB}) & \text{for } t > 0, \text{ with } K_{BB} = \frac{1}{N_{BB}} \sum_{p \in BB} K_p
\end{align*}
\]

\( K_p = m_c((H_2)) \times \gamma_t, \forall p \in BB \) is the conflict between colour and VJ elementary masses at pixel level, \( N_{BB} \) is the number of pixels inside the bounding box and \( K_p \) 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 \( \gamma \) is detailed in Table 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 filter (section 5.1). The pignistic probability \( \text{BetP}_p(\cdot) \) associated with the face class is:

\[
\begin{align*}
\text{BetP}_p((H_1)) &= m((H_1)) + m(\Omega)/2, \\
\text{BetP}_p((H_1)) &= m_c((H_1)) + \left(1 + \frac{\gamma_t}{2}\right)m_c(\Omega), \quad \forall p \in BB \\
\text{BetP}_p((H_1)) &= (1 - \gamma_t)\left[m_c((H_1)) + \frac{1}{2}m_c(\Omega)\right], \quad \forall p \notin BB
\end{align*}
\]

Since \( \text{BetP}_p \) 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).

Table 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 \( \gamma \) and to the colour uncertainty \( m_c(\Omega) \).

\begin{table}[h]
\centering
\caption{Fusion of \( m_c \) and \( m_v \) by the conjunctive rule as a function of the pixel position (inside or outside the bounding box).}
\begin{tabular}{ |c|c|c| }
\hline
\( A \) & \( m \) for \( p \in BB \) & \( m \) for \( p \notin BB \) \\
\hline
\( \emptyset \) & \( m_c((H_2)) \cdot \gamma_t \) & \( m_c((H_1)) \cdot \gamma_t \) \\
\( \{H_1\} \) & \( m_c((H_1)) + m_c(\Omega) \cdot \gamma_t \) & \( m_c((H_1)) \cdot (1 - \gamma_t) \) \\
\( \{H_2\} \) & \( m_c((H_2)) \cdot (1 - \gamma_t) \) & \( m_c((H_2)) + m_c(\Omega) \cdot \gamma_t \) \\
\( \Omega \) & \( m_c(\Omega) \cdot (1 - \gamma_t) \) & \( m_c(\Omega) \cdot (1 - \gamma_t) \) \\
\hline
\end{tabular}
\end{table}
unpredictable with frequent direction changes. The scene is tracking by predicting the future state (localization) of a dynamic system (the target) using past measurements. Kalman Dempster–Shafer theory is proposed in [50]. The algorithm presents an original solution to the problem of multi-camera disrupted sources using contextual information brought by a particle filter. An extension of Bayesian particle filters to the $X_t$ e.g. mentioned in Section 2 about detection: (i) first, low level methods achieve tracking by performing colour segmentation, tracked object) in the video stream [44]. Tracking techniques can be grouped into three categories, some of them already orientation parameters of one (or more) face along time. The goal is to obtain in real-time the trajectory of the target (or after the face model detection. Face tracking is defined as the process of estimation of the shape, appearance, position and estimation of the posterior conditional probability distribution $p$.

The performance of the evidential model depends both on the colour model quality and on the VJ face detector reliability $\gamma$. 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 $\gamma$ ($\gamma < 0.5$) 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 $\gamma$ ($\gamma > 0.9$) 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 $\gamma$ value such as 0.7 ≤ $\gamma_0$ ≤ 0.9. Note that when the VJ face detector is in default, i.e., when it does not deliver any bounding box, $\gamma$ is set to zero.

<table>
<thead>
<tr>
<th>$m_{c^*}([H_1])$</th>
<th>$m_{c^*}([H_2])$</th>
<th>$\gamma$</th>
<th>$m_{c^*}([\Omega])$</th>
<th>$\text{BetP}_\gamma([H_1])$</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$p \in \text{BB}$</td>
<td>$p \notin \text{BB}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$[H_2]$</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>$[H_2]$</td>
</tr>
<tr>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>Indecisive</td>
<td>$[H_1]$</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Indecisive</td>
<td>$[H_1]$</td>
</tr>
<tr>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td>Indecisive</td>
<td>$[H_1]$</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.75</td>
<td>0.25</td>
<td>Indecisive</td>
<td>$[H_1]$</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Indecisive</td>
<td>$[H_1]$</td>
</tr>
<tr>
<td>0</td>
<td>0.5</td>
<td>0.25</td>
<td>0.25</td>
<td>Indecisive</td>
<td>$[H_1]$</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0.75</td>
<td>0.75</td>
<td>Indecisive</td>
<td>$[H_1]$</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.75</td>
<td>0.75</td>
<td>Indecisive</td>
<td>$[H_1]$</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Indecisive</td>
<td>$[H_1]$</td>
</tr>
</tbody>
</table>

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. Tracking is defined as the process of estimation of the shape, appearance, position and orientation parameters of one (or more) face along 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 subtraction 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 $(x_c, y_c)$, main axis $h_t$, minor axis $l_t$ and orientation $\theta_t$. These parameters are grouped into the state vector $X_t = [x_c, y_c, h_t, l_t, \theta_t]$. 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 $Y_t = \text{BetP}$, 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 $n$th particle at time $t$. The \textit{a posteriori} law is approximated by:

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

---

**5.1. Estimation of the ellipse centre**

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

$$p(\tilde{X}_t | X_{t-1}) = (1 - \nu) N(\tilde{X}_t | X_{t-1}, \Sigma) + \nu U_{\tilde{X}}(\tilde{X}_t)$$ \hfill (25)

where $N(.,|\mu, \Sigma)$ is a normal Gaussian law with average $\mu$ and covariance $\Sigma$. The diagonal matrix $\Sigma = \text{diag}(\sigma_{x_{ct}}, \sigma_{y_{ct}}) = \text{diag}(5, 5)$ sets the \textit{a priori} constraints: it gives the variances imposed to the position components of the state vector. The coefficient $\nu$ 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 $\nu$ 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 $\nu$ 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 $\text{BetP}_P(\{H_1\})$ contained inside the ellipse. Hence the estimated weight of each particle $\Lambda_t^{(n)}$ is given by:

$$\tilde{\omega}_t^{(n)} = \sum_{p \in \tilde{X}_t^{(n)}} [\text{BetP}_P(\{H_1\})]^2.$$ \hfill (26)
Fig. 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.

Fig. 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.

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 \(\Omega_1\) 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 \((x_{ct}, y_{ct})\) 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 \((x_{ct}, y_{ct})\).

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 morphological operation of image filling is applied to this image in order to exhibit a shape (in grey on Fig. 9d) around the center \((x_{ct}, y_{ct})\), whose contour is extracted. Finally an elliptic approximation of this contour based on a least squares fitting method \([53]\) yields a measured ellipse, which constitutes the new observation (in red on Figure 9d) whose parameters are denoted by \(\hat{Y}_t = \begin{bmatrix} \hat{h}_t, \hat{l}_t, \hat{\theta}_t \end{bmatrix}\). 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)).

\[
\tilde{\omega}_t^{(n)} = p \left( \hat{Y}_t | \hat{X}_t^{(n)} \right) \propto \frac{1}{\left( \hat{h}_t - \hat{h}_t^{(n)} \right)^2 + \left( \hat{l}_t - \hat{l}_t^{(n)} \right)^2 + \left( \hat{\theta}_t - \hat{\theta}_t^{(n)} \right)^2}
\] (27)

At last, the maximum likelihood criterion selects the most significant particle: among all the predicted ellipses around the previously 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 \(\Sigma_{max} = \text{diag}(\sigma_{ht}, \sigma_{lt}, \sigma_{\theta t}) = \text{diag}(5, 5, 0.1)\) 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].

Fig. 11 shows the visual servoing behavior. On image im15, 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 im18. From image im20 to im24, 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).

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 re-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 im57 and im73: 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 im60, im66, im69, so that γ1 = 0 is set in the evidential model since only colour information is valid. Therefore, in the presence of total occlusion (im66), the resulting ellipse lies on the hand of the user. The uniform component (ν = 0.1) of the particle filter dynamic model (Eq. (25)) ensures a correct repositioning when a candidate particle locates again on the face zone (im73). The VJ information may degrade the tracking quality when the VJ detector focuses on a face-like artifact (frame im80). An important conflict (K80 = 0.7) is measured inside the VJ bounding box. Then the adjustment of parameter γ1 (γ1 → 0.24 as 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 (im200), 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 im202 and im300, 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 im351) 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 im465, no bounding box is delivered by the VJ detector and only colour information is considered (γ1 = 0). When hands are in contact with the face in frame im598, the estimations of center and pose remain correct. When the hands go away from the face, they are not tracked any longer (frame im604). This shows the robustness of the proposed method: the presence of disruptive elements alters weakly pose and size estimation and only slightly 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 $GT$: $FP = ROI \cap GT$.

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

$$\text{Precision} = \frac{|TP|}{|ROI|}, \quad \text{Recall} = \frac{|TP|}{|GT|}. \quad (28)$$

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: $FN = ROI \cap GT$. 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.
Fig. 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 $\eta \in [0, 1]$).

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

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 $\gamma$ and to the compromise parameter $\eta$ (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 $\approx 0.57$, Recall $\approx 0.82$). Tracking is notably improved by a weak contribution of the VJ detector and the best performances are reached for $\gamma \approx 0.3$ (Precision $\approx 0.63$, Recall $\approx 0.72$). When $\gamma$ increases ($\gamma > 0.6$), the VJ information plays a major role in the evidential face model. Precision is quasi-constant ($\approx 0.73$) 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 implemented (Eq. (21)). This dynamic setting of $\gamma$ leads to a performance optimization (Precision $\approx 0.79$, Recall $\approx 0.67$).

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 $p(s_j|H_1)$). Please cite this article in press as: F. Faux, F. Luthon, Theory of evidence for face detection and tracking, Int. J. Approx. Reason (2012), http://dx.doi.org/10.1016/j.ijar.2012.02.002
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 $70–80\%$ [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:

$$
\varepsilon = \sqrt{(x_{GT} - x_c)^2 + (y_{GT} - y_c)^2},
$$

where $x_{GT}, y_{GT}$, and $x_c, y_c$ are the coordinates of the face gravity centre given by the ground truth (GT), whereas $x_c, y_c$ are the center location coordinates of the detected ellipse (ROI).

With a location error lower than $\varepsilon_{\text{max}} = 30$ 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 Table 5, the mean location error $\bar{\varepsilon}$ and the standard deviation $\Sigma_{\text{mean}}$ 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].

As regards the computation cost, the processing time is $\approx 1$ s/image for an image of size $400 \times 400$ 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).

### Table 5

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

<table>
<thead>
<tr>
<th>Seq</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
<th>#5</th>
<th>#6</th>
<th>#7</th>
<th>Dav Ind</th>
<th>All data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>24</td>
<td>17</td>
<td>56</td>
<td>17</td>
<td>11</td>
<td>13</td>
<td>32</td>
<td>15</td>
<td>23</td>
</tr>
<tr>
<td>Std</td>
<td>36</td>
<td>12</td>
<td>61</td>
<td>19</td>
<td>4</td>
<td>30</td>
<td>47</td>
<td>11</td>
<td>36</td>
</tr>
</tbody>
</table>

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 theoretical 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 $\gamma$ in Eq. (21)).

As future works, the dynamic setting of the algorithm parameters is under investigation. Besides, distinct values for parameter $\gamma$ 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 ($\Rightarrow d_{11} > d_{12}$). 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 linguistic quality of the manuscript.

References


