Robust face tracking using colour Dempster-Shafer fusion and particle filter
Francis Faux, Franck Luthon

To cite this version:
Francis Faux, Franck Luthon. Robust face tracking using colour Dempster-Shafer fusion and particle filter. 9th International Conference on Information Fusion, Jul 2006, Firenze, Italy. pp.;, 2006. <hal-00405766>

HAL Id: hal-00405766
https://hal.archives-ouvertes.fr/hal-00405766
Submitted on 18 Mar 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Robust face tracking using colour Dempster-Shafer fusion and particle filter

F. Faux  
F. Luthon

1 Laboratoire LIUPPA (EA 3000)

Computer Science Dpt, University of Pau, France  
faux@iutbayonne.univ-pau.fr

Abstract

This paper describes a real time face detection and tracking system. The method consists in modelling the skin face by a pixel fusion process of three colour sources within the framework of the Demster-Shafer theory. The algorithm is composed of two phases. In a simple and fast initialising stage, the user selects successively in an image, a shadowy, an overexposed and a zone of mean intensity of the face. Then the fusion process models the face skin colour. Next, on the video sequence, a tracking phase uses the key idea that the face exterior edges are well approximated as an ellipse including the skin colour blob resulting from the fusion process. As ellipse detection gets easily disturbed in cluttered environments by edges caused by non-face objects, a simple and fast efficient least squares method for ellipse fitting is used. The ellipse parameters are taken into account by a stochastic algorithm using a particle filter in order to realise a robust face tracking in position, size and pose.

The originality of the method consists in modelling the face skin by a pixel fusion process of three independant cognitive colour sources. Moreover, mass sets are determined from a priori models taking into account contextual variables specific to the face under study. Hence, the face specificity which is to present shadowy (neck) and overexposed zones (nose, front) is considered, so that sensitivity to lighting conditions decreases.

Results of face skin modelling, fusion, ellipse fitting and tracking are illustrated and discussed in this paper. The limits of the method and future work are also commented in conclusion.

Keywords

Face detection, Dempster-Shafer, colour fusion, face tracking, condensation, skin hue.

1 Introduction

Face detection and tracking in a video sequence is useful in many applications such as videoconference, HMI, telesurveillance or robotics.

Above 150 detection methods [1, 2], as low level methods using cues such as texture, color [3, 4], edges, up to high level methods such as appearance models, neural networks or Support Vector Machines are proposed in the literature [5, 6, 7, 8].

In these approaches, skin color is often used as a first localisation and segmentation primitive in order to decrease the research zone. Because of its specificity, skin colour cue is very pertinent [9] and allows rapid size and orientation invariant algorithms.

However it remains an important difficulty to accommodate unsupervised varying lighting conditions. Normalised RGB, HLS colour spaces are the most used. But these colour transforms remain quite sensitive to lighting conditions and very noisy in shadowy zones. In order to overcome these drawbacks, increase robustness and contrast, skin colour cues are mapped into the logarithmic colour space LUX [10]. Moreover, face is a non-rigid object which peculiarity is to present shadowy (neck) and overexposed zones (nose, front). Their time varying localisation because of movement in the video sequence yields complex modelisation. So contextual variables are included in the process to deal with this aspect.

Pixel-based skin colour detection techniques can be gathered in explicit methods, which use empirical rules of decision, and statistics [11]. However, information to model is never perfect because of image acquisition principle (3D space towards 2D plan transformation). There exists different forms of artefact including ambiguity, imprecision or inadequacy. We ought to take into account these difficulties to improve modelisation. Classical probabilistic methods (Bayesian inference rule) loose performance when early vision learning stage is not representative of real measurements. That is the reason why a pixel fusion process of skin colour cues within the framework of the Dempster Shafer theory is used in this paper [12, 13].

A simple and fast initialising stage takes into account ground truth as the user selects successively on one image, a shadowy, an overexposed and a zone of mean intensity. For each zone, mass sets are determined from a priori models of three colour sources (in the LUX colour space), contextual variables and source confidence degrees. Then mass sets are fused and blobs are computed.

Next, on the video sequence, a tracking phase uses the key idea that the face exterior edges are well approximated as an ellipse including the skin colour blob resulting from the fusion process [14, 15]. As ellipse detection gets easily disturbed in cluttered environments by edges caused by non-face objects, a simple and fast efficient least squares method for ellipse fitting is used [16]. The ellipse parameters (center, minor axis, major axis and orientation) are taken into account by a stochastic algorithm using a particle filter in order to realise a robust face tracking in position, size and pose.

The remainder of the paper is organised as follows. Section 2 develops the modelisation and fusion processes. Modelisation results are illustrated and commented in section
3. Section 4 explains the tracking algorithm and shows results. Finally, section 5 summarises the contribution and opens the suggestions for future work.

# Method description

## 2.1 Sources and discernment field

Skin colour cues are mapped into the logarithmic colour space LUX [10]. The expression of the LUX components from the RGB colour space coded on 3 x 8 bits are:

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

\[
U = \begin{cases} 
128 \frac{L+1}{R+1} & \text{for } R > L \\
256 - 128 \frac{B+1}{B+1} & \text{otherwise}
\end{cases}
\]

\[
X = \begin{cases} 
128 \frac{L+1}{R+1} & \text{for } B > L \\
256 - 128 \frac{B+1}{B+1} & \text{otherwise}
\end{cases}
\]

The fusion process uses 3 sources (colour “sensors”) called \( S_j \), \( j = 1, 2, 3 \) such as:

\[
S_1 = U, \quad S_2 = X \quad \text{et} \quad S_3 = 0.5(L + U) = W
\]

For the source \( S_3 \), instead of \( L \) we add to \( S_1 \) the luminance component \( L \), very rich from the semantic point of view, in order to better characterise the skin colour variations due to lighting conditions. Each source \( S_j \) provides a measurement noted \( M_j \).

The discernment field \( \Omega \) is defined by 2 assumptions such as:

\[
\Omega = \{ \omega_1, \omega_2 \}, \quad \omega_1 \text{ represents the face assumption and } \omega_2, \text{ complement of } \omega_1, \text{ symbolises the background } (\omega_2 = \bar{\omega}_1).
\]

## 2.2 A priori model

In order to determine the a priori model, in an initialisation stage, the user selects successively on an image three characteristic zones of the face: a shadowy zone, a zone of mean intensity and an overexposed zone. Three contextual variables \( z_i \), \( i = 1, 2, 3 \) are taken into account: shadowy zone \( z_1 \); mean zone \( z_2 \); overexposed zone \( z_3 \).

The histograms calculated on each selected zone and for each measurement \( M_j \) allows to determine the conditional probability densities \( p(M_j | \omega_1, z_i) \). These are approximated by Gaussian distributions \( N_{ij}(\mu_{ij}, \sigma_{ij}) \) (Fig. 3) where \( \mu_{ij} \) and \( \sigma_{ij} \) are respectively the mean and the standard deviation of the measurement \( M_j \) on the zone \( z_i \). In addition to statistical information on the zone, spatial data distribution (\( M_1, M_2, M_3 \)) in the colour space \((U, X, W)\) (Eq. 1), that is to say the colour domain \( D_i \), contributes to modelisation. For each context \( z_i \), 3 pairs of coefficients \( \alpha_{ij} \) and \( \beta_{ij} \) maps 6 straight-line segments \( M_{ij} \) which synthesise a parallelepiped. These coefficients are adjusted such as the parallelepiped encompasses at best the colour domain \( D_i \) (Fig. 4). Then, the function skin_{ij} defines the a priori model such as:

\[
\text{skin}_{ij} = \begin{cases} 
\max(N_{ij}(\mu_{ij}, \sigma_{ij})) & \text{if } M_{ij} \leq M_j \leq Mx_{ij} \\
0 & \text{else}
\end{cases}
\]

## 2.3 Mass sets

Appriou [17] suggested to introduce each a priori density of probability \( p(M_j | \omega_1, z_i) \) and its corresponding confidence degree \( d_{ij} \) in a mass set \( m_{ij}(\cdot) \). This set is defined by an axiomatic approach in the discernment field \( \Omega \). In the method developed here the densities of probability are replaced by the functions skin_{ij} which implicitly take into account the spatial data distribution in the colour space \((U, X, W)\).

Focal elements associated to \( m_{ij}(\cdot) \) are \( \omega_1, \omega_2 \) and \( \Omega \). Mass sets are defined by:

\[
m_{ij}(\omega_1) = \frac{d_{ij} R_{s} \text{skin}_{ij}}{1 + R_{s} \text{skin}_{ij}} \quad \text{(3)}
\]

\[
m_{ij}(\omega_2) = \frac{d_{ij}}{1 + R_{s} \text{skin}_{ij}} \quad \text{(3)}
\]

\[
m_{ij}(\Omega) = 1 - d_{ij}
\]

The weighting coefficients \( R_i \) take into account the data skin_{ij} for each zone \( i \).

## 2.4 Confidence degrees

For \( S_1 = U \) and \( S_2 = X \) sources, the ambiguity between classes is weak. Consequently a probabilistic modelling is used \( d_{ij} = d_{i} \approx 1 \) (\( m(\Omega) = 0 \)). On the other hand the \( S_3 \) source depends on brightness. The reliability of this source for the face class is maximum \( (d_{ij} = 1) \) only for the average grey level \( (\mu_{s}) \) of the modelled zone. Ambiguity between classes grows \( (m(\Omega) > 0) \) when \( M_3 \) deviates from \( \mu_s \). Therefore a fuzzy function (Eq. 4) is used to characterise the reliability of the \( S_3 \) source in the context \( z_i \) (Fig. 1).

\[
\text{as } \mu_{13} \leq \mu_{23} \leq \mu_{33} \text{ we obtain :}
\]

\[
\begin{align*}
\text{for } i &= 1 : \\
&\frac{M_3 + 2\mu_{23} - 3\mu_{13}}{2\mu_{23} - 2\mu_{13}} \quad \text{if } 3\mu_{13} - 2\mu_{23} \leq M_3 \leq \mu_{13} \\
&0 \quad \text{else}
\end{align*}
\]

\[
\begin{align*}
\text{for } i &= 2 : \\
&\frac{M_3 - \mu_{13}}{\mu_{23} - \mu_{13}} \quad \text{if } \mu_{13} \leq M_3 \leq \mu_{23} \\
&0 \quad \text{else}
\end{align*}
\]

\[
\begin{align*}
\text{for } i &= 3 : \\
&\frac{M_3 - 2\mu_{13}}{2\mu_{23} - 2\mu_{13}} \quad \text{if } \mu_{23} \leq M_3 \leq 3\mu_{33} - 2\mu_{23} \\
&0 \quad \text{else}
\end{align*}
\]

## 2.5 Decision

Three contextual masses \( m_i(s) \) are associated at each pixel \( s(x, y) \) of colour components \( (M_1; M_2; M_3) \) by the orthogonal normalised combination of Dempster-Shafer:

\[
m_i(s) = \oplus m_{ij}(s) \quad \text{(5)}
\]
3 Modelisation results

3.1 Shadowed zone a priori model

As the initialisation stage is the same for each zone, only the modelling of the shadowed zone $z_1$ is presented in detail hereafter. First, the user selects on an image a shadowed zone of the face (Fig. 2). The probability densities are approximated for each measurement $M_j$ by a Gaussian distribution $N_{1j}(\mu_{1j}, \sigma_{1j})$ (Fig. 3).

$$m(s) = m_1(s) \cup m_2(s) \cup m_3(s)$$  \hspace{1cm} (6)

3.2 Shadowed zone detection

The mass sets $m_{1j}(\omega_1) = \frac{d_{1j} R_1 \text{skin}_{1j}}{1 + R_1 \text{skin}_{1j}}$ (Eq. 3) depend on the parameter $R_1$ which weights the data importance of $\text{skin}_{1j}$ characterising the face class.

Then, three masses $(m_{11}(s), m_{12}(s), m_{13}(s))$ are associated with each pixel in the site $s(x,y)$ in function of its colour components $(M_1, M_2, M_3)$.

Finally a single mass $m_1(s)$ is affected to each pixel by the conjunctive fusion rule (Eq. 5).

The 3 top images of Fig. 6 show the experimental results for various values of $R_1$. For $R_1 = 1$ modelling is satisfactory as the shadowed skin colour (in white) present under the chin, the eyes contours and under the hair is detected. For $R_1 = 5$ or $R_1 = 10$ the modelling zone widens and false detections appear.

3.3 Modelisation synthesis

Fig. 6 (middle and bottom) shows the results of the conjunctive fusion for the mean intensity and overexposed zones, whose process is similar to those presented in sections 3.1 and 3.2 for the shadowed zone. Modelling detects accurately the shadowed and overexposed skin face zones (Fig. 6 top and bottom). Although more representative of the face, the average model (Fig. 6 middle), presents artefacts in the shadowy or overexposed zones. This is the reason why the disjunctive fusion (Eq. 6) is used to combine the various models and optimise the detection result (Fig. 7 left and center).

When two masses are combined, modelling is insufficient and presents defaults on mean intensity parts of the face (Fig. 7 a, b), on the shadowy parts of the face (Fig. 7 d, e) or on the overexposed parts of the face (Fig. 7 g, h).
Modelling is optimal when the 3 masses are combined ($R_1 = 1$, $R_2 = 1$, $R_3 = 1$) (Fig. 7 j). However a too important weighting of parameter $R_i$ deteriorates slightly the modelling quality (Fig. 7 k).

Within the framework of the Dempster-Shafer theory, in spite of a complex background made up of colour elements close to skin colour (table, sweater, poster), the fusion process ensures a good detection and proves to be robust to occlusions and pose variations (Fig. 8 bottom).

On the other hand, Fig. 7 right and Fig. 8 middle represent the result of modelling when ($d_{ij} = 1$) i.e. when the reliability of the sensors is not taken into account any more. Consequently, $m_{ij} = 0$ (cf Eq. 3) and only the face ($\omega_1$) and non-face ($\omega_2$) classes are considered. Within this probabilistic framework, the detection result is of worse quality in the zones where the data inaccuracy is not negligible because of the less reliable learning stage model.

4 Tracking by condensation algorithm

4.1 Particle filter formalism

The particle filter algorithm initially designed for signal processing problems [18], was applied in computer vision under the name of ”condensation ” algorithm [19]. In the case of a single object, the vector $X_t$ representing the hidden state of the object of interest follows the law of evolution (Eq. 7) and is observed by the vector $Y_t$ at discrete times according to Eq. 8:

$$X_t = F_t(X_{t-1}, V_t)$$

$$Y_t = H_t(X_t, W_t)$$

No assumption is made on $F_t$ and $H_t$ functions and the processes $V_t$ and $W_t$ are two white noises not necessarily Gaussian, mutually independent and independent of the initial condition $X_0$.

4.2 Application to face tracking

In the framework of face tracking, where movement is not foreseeable and frequently changes in direction, particle filtering finds all its justification.

Moreover, the key idea that the face exterior edges are well approximated as an ellipse of center noted $(x_{c_1}, y_{c_1})$, of mi-
The coefficient $\beta$ing model [20]:

$$X_t = [X_{1t}, X_{2t}]$$

where $X_{1t} = [x_{ct}, y_{ct}]$ and $X_{2t} = [\ell_t, L_t, \theta_t]$. After a phase of initialisation, the tracking method proceeds in two stages. First the algorithm determines the center then estimates the size and the orientation of the ellipse.

**Initialization.** In order to select the face whatever its initial position in the image, the particles are distributed in position according to a uniform probability law whereas the parameters of size and orientation are fixed at a constant value ($\ell_0 = 20; L_0 = 25; \theta_0 = 0$) (Fig. 9). The weight of the particles is $q_n^1 = 1/N$ where N is the number of particles (typ. $N = 150$).

![Figure 9: Particle initialisation](image)

**Model center estimation.** In this stage, in order to determine the state vector position at time $t$ size and orientation parameters are fixed. They are the components of $X_{2t}$ estimated at time $t - 1$.

The dynamics of the vector $X_{1t}$ is described by the following model [20]:

$$p(X_{1t}/X_{1t-1}) = (1 - \beta_u)N(X_{1t}/X_{1t-1}, \Sigma) + \beta_u U(X_{1t})$$

where $N(.)$ represents the Gaussian distribution of average $\mu$ and of covariance $\Sigma$.

$U(.)$ represents the uniform distribution on the unit $\chi$.

The coefficient $\beta_u$, $0 \leq \beta_u \leq 1$ weights the uniform distribution and $\Sigma = diag(\sigma_x^2, \sigma_y^2)$ is the diagonal matrix made up of the variances of the $X_{1t}$ state vector components.

The introduction of a uniform component ($\beta_u = 0.1$) manages the rare erratic movements like jumps in the video sequence. It helps also the algorithm to be locked after one period of partial or total occlusion.

The model of measurement $Y_{1n,t}$ weights more or less strongly a particle $n$ in function of the squared sum of masses inside the particle $n$. The probability of observation is consequently defined by:

$$p(Y_{1t}/X_{1t}) = \prod_n p(Y_{1n,t}/X_{1n,t})$$

The maximum of plausibility criterion allows selecting the most significant ellipse and its center defines the state vector position components. At the end of the initialisation stage, the estimated position (Fig. 10 b) does not correspond to face center because of incorrect face and orientation ellipse dimensioning. After some steps, these parameters being corrected (see hereafter), the algorithm estimates with a better accuracy the ellipse center (Fig. 10 d).

**Size and pose estimation.** From the ellipse center defined at the previous stage, a filling step then a binarisation is realised.

Moreover, the dynamics of the vector $X_{2t}$ is described by the following model:

$$p(X_{2t}/X_{2t-1}) = N(X_{2t}/X_{2t-1}, \Sigma)$$

where $\Sigma = diag(\ell_t^2, L_t^2, \theta_t^2)$ is the diagonal matrix made up of the variances of the $X_{2t}$ vector components.

Then face edges are extracted then approximated by an ellipse called measurement ellipse (Fig. 11 c,d). The Fitzgibbon al. [16] algorithm for ellipse fitting realises a compromise between speed and accuracy and is very robust to noise. It provides a measurement of ellipse parameters $\hat{X}_t = [\hat{x}_{ct}, \hat{y}_{ct}, \hat{\ell}_t, \hat{L}_t, \hat{\theta}_t]$.

Consequently the probability density is:

$$p(Y_{2,t}/X_{2t}) = \prod_n p(Y_{2n,t}/X_{2n,t})$$

where $Y_{2n,t} \propto d^2$

$d$ is the euclidean distance between the predicted ellipse $X_{2n,t}$ and the measured ellipse noted $\hat{X}_{2t} = [\hat{\ell}_t, \hat{L}_t, \hat{\theta}_t]$.

The minimum of plausibility criterium enables to isolate the particle having the most probable form.

The results on the sequence (Fig. 12) exhibit the behavior of the algorithm in the context of face viewed from aside (image 15) or partial occlusion (image 27).

5 Conclusion

The originality of the method consists in modelling the face skin by a Dempster-Shafer pixel fusion of three cognitive independant colour sources. Moreover, mass sets are determined from a priori models taking into account contextual
variables specific to the face under study. The shape criterion (ellipse) used in the particle filter algorithm removes modelisation artefacts and so improves segmentation and tracking.

However the fusion of only colour information is sometimes insufficient. The use of other independent cues such as the movement, texture or edges should improve detection. A feedback loop will be integrated in order to dynamically adapt the fusion process parameters and thus to optimise the information reliability (precision, uncertainty …).

Finally, the objective will be to elaborate an active vision system where the dynamic collaboration of information and a feedback loop should contribute to increase segmentation and tracking robustness.

References


