# Robust Tracking for Processing of Videos of Communication's Gestures

Frédérick Gianni, Christophe Collet, and Patrice Dalle

Institut de Recherche en Informatique de Toulouse,
Université Paul Sabatier, Toulouse, France
{gianni, collet, dalle}@irit.fr

**Abstract.** This paper presents a method of image processing used in a mono-vision system in order to study semiotic gestures. We present a robust method to track the hands and face of a person performing gestural communication and the Signs' language communication. A model of skin is used to compute the observation density as a skin colour distribution in the image. Three particle filter trackers are implemented, with re-sampling and annealed update steps to increase their robustness to occultation and high acceleration variations of body parts'. Evaluations of the trackers with and without these enhancements, show the improvement that they bring.

**keywords** Hands and Head Tracking, Skin Colour Segmentation, Particle Filter, Sign Languages, Video Analysis.

## 1 Introduction

In the context of the research undertaken on gestural man machine communication and on signs language's (SL), we are interested in the study of image processing tools able to automate part of the video annotation and then to build gestures recognition systems. In these contexts, the gestures should be performed naturally, without of any constraints. Hands' movements are thus very fast in particular in SL, and one of the major problems is to find a robust tracking method. In this paper, we present an enhanced tracking method using particle filtering. First we present the realisation context of the gestures studied and the parts of the body involved : hands and face. Next, we detail the method used to model and to track body parts. In the last section we present results of robustness of the tracking method.

## 2 Communication gestures

During a communication such as a dialogue, work presentation, lot of gestures are emitted by numerous body parts. Each ones having their own meaning. According to Cadoz [1] the functionalities of the gestures are epistemic, ergotic and

semiotic. We are here focus on the semiotic function, the semantic information convey by the gesture. McNeil [2] gives a classification of those kinds of gestures : iconic, metaphoric, deictic or relative to the beat of the given information. If one wants to analyse gestures according to this classification, there is a need to retrieve, qualify and quantify the information given by the body parts to interpret the whole meaning of a communication. Those informations are useful in order to build interactive gesture systems for man-machine communication [3] and for linguistic's analysis in the studies of SL.

## 3 Tracking of body parts

Human motion tracking needs accurate features detection and features correspondence between frames using position, velocity and intensity information. In our approach, the feature correspondence is achieved using statistical estimators via a particle filter for the head and the two hands. As the particle filter models the uncertainty, it will provide a robuste framework for the tracking of the hands of a person communicating in french sign language.

### Particle Filter

The particle filter (PF) aims at estimating a sequence of hidden parameters $x_t$ from only the observed data $z_t$. The idea is to approximate the probability distribution by a weighted sample set :

$$\{(s_t^{(0)}, \pi_t^{(0)}) \dots (s_t^{(n)}, \pi_t^{(n)})\}$$

with $n = 1, \dots, N$ numbers of samples used. Each sample $s$ represents one state of the tracked object with a corresponding discrete sampling probability $\pi$.

The state is modelled as :

$$s_t = [x, y, \dot{x}, \dot{y}, \ddot{x}, \ddot{y}]^t$$

the position, velocity and acceleration of the sample $s$ in the observation at time $t$. Three states are maintained during the tracking, one for each body part tracked. We track the head and the hands separately, each of those areas is represented by one sample set. In the prediction phase, the samples are propagated throught a dynamic model : a first order auto-regressive process model $x_t = Sx_{t-1} + \eta$, where $\eta$ is a multivariate Gausian random variable and $S$ a transition matrix.

We use the particle filter defined in [4] applied in a color based context to achieve robustness against non rigidity and rotation. The observation density $p(z_t|x_t)$ is modelled as a skin colour distribution using the histogram backprojection method (fig.1).

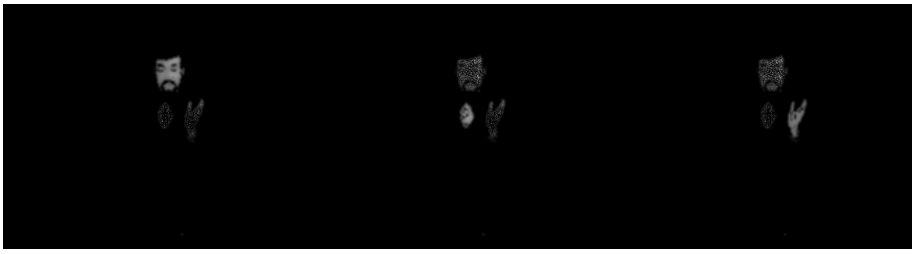


**Fig. 1.** Observation densities as a skin colour for head, right hand and left hand.

In the particle filter, a resampling step is used to avoid the problem of degeneracy of the algorithm, that is, avoiding the situation that all but one of the importance weights are close to zero. The stratified resampling proposed by Kitagawa [5] is used, because it is optimal in terms of variance.

**Particle filter with annealed update step (APF).** To maintain a good representation of the posterior probability, one may iterate the algorithm a certain amount of time. But this leads to an over-representation of the possible local maximum. This is caused by the weighting function applied at each iteration. The annealed effect proposed by [6] and with a more generic formulation [4] provides a mean to apply the weighting function to the sample set smoothly.

Update with simulated annealing

For $m$=$M, \ldots, 1$

For $i = 0, \ldots, N$, $\pi_{t,m}^{(i)} \leftarrow p(z_t|\tilde{x}_{t,m} = \tilde{s}_{t,m}^{(i)})^{\beta_m}$

For $i = 0, \ldots, N$, $\pi_{t,m}^{(i)} \leftarrow \dfrac{\pi_{t,m}^{(i)}}{\sum_{j=1}^{n} \pi_{t,m}^{(j)}}$

For $i = 0, \ldots, N$, $x_{t,m}^{(i)} \leftarrow \tilde{x}_{t,m}^{(j)}$ with likelihood $\pi_{t,m}^{(j)}$

For $i = 0, \ldots, N$, $\tilde{x}_{t,m-1}^{(i)} \leftarrow \tilde{s}_{t,m-1}^{(i)} \leftarrow \tilde{s}_{t,m}^{(i)} + \mathbf{B}_m$

For $i = 0, \ldots, N$ $\pi_{t,0}^{(i)} \leftarrow p(z_t|x_t = \tilde{s}_t^{(i)})$

For $i = 0, \ldots, N$ $\pi_{t,0}^{(i)} \leftarrow \dfrac{\pi_{t,0}^{(i)}}{\sum_{j=1}^{n} \pi_{t,0}^{(j)}}$

The value of $\beta_m$ will determine the rate of annealling. The value of this parameter is chosen following the recommandations of Deutscher *et al.* [6]. From the survival diagnostic :

$$D = \left( \sum_{n=1}^{N} (\pi^{(n)})^2 \right)^{-1}$$

MacCormick [7] provides the particle survival rate $\alpha = \frac{D}{N}$. The annealing rate can be computed as follow: for each iteration an annealing rate $\alpha_m$ is given and using a gradient descent we can find the corresponding $\beta^m$ :

$$\alpha_{i-1} = \left( N \sum_{i=0}^{N} ((\pi^{(i)})^{\beta_{m-1}})^2 \right)^{-1}$$

$$\alpha_i = \left( N \sum_{i=0}^{N} ((\pi^{(i)})^{\beta_m})^2 \right)^{-1} \text{ with } \beta_m = \tfrac{1}{2}\beta_{m-1}$$

while $\alpha_i - \alpha_{desired} > \varepsilon$

$$\Delta_\alpha = \frac{\alpha_i - \alpha_{i-1}}{\beta_m - \beta_{m-1}}$$

$$\beta_{m-1} = \beta_m$$

$$\beta_m = \beta_m + \frac{\alpha_{desired} - \alpha_i}{\Delta_\alpha}$$

$$\alpha_{i-1} = \alpha_i$$

$$\alpha_i = \left( N \sum_{i=0}^{N} ((\pi^{(i)})^{\beta_m})^2 \right)^{-1}$$

$\mathbf{B}_m$ is a multi-variate gaussian random variable with mean $\mathbf{0}$ and variance $\mathbf{P}_m$. The diffusion variance vector has been set as $\mathbf{P}_m = (\alpha_M \alpha_{M-1} \dots \alpha_m)$ [6].

**Multiple-object tracking.** Here arise the problem of data association which makes the problem harder than single object tracking. Multiple object tracking and data association techniques have been extensively studied in [8] and a number of statistical data association techniques such as probabilistic data association filter, joint probabilistic data association filter, multiple hypothesis tracking filter have been developed. These generally employ a combination of "blob" identification and some assumption on the target motion. One may think of avoiding this problem in a way of interpreting the target as "blob" which merge and split again [9]. A "blob" interpretation does not maintain the identity of the targets, and it's difficult to implement for target which are not easily separable.
What is needed is an "exclusion principle" such as the one provide by MacCormick [10]. This way we do not allow two targets to merge when their configuration become similar. First we compute the posterior density for each tracker, then we use it to penalize the measurement of the other trackers in a second computation of the posterior density.

## 4 Evaluations

Evaluations were conduct on a sequence of a person signing a story in the french sign language. We are here interested in testing the robustness of our approach

against high dynamics motion variation and body part occlusions. Such a language is a tracking challenge as the tracked targets are very similar, the performed gestures have got a lot of dynamic variations (fig. 2), the targets are relatively often occluded and it's a long sequence (3 000 frames).

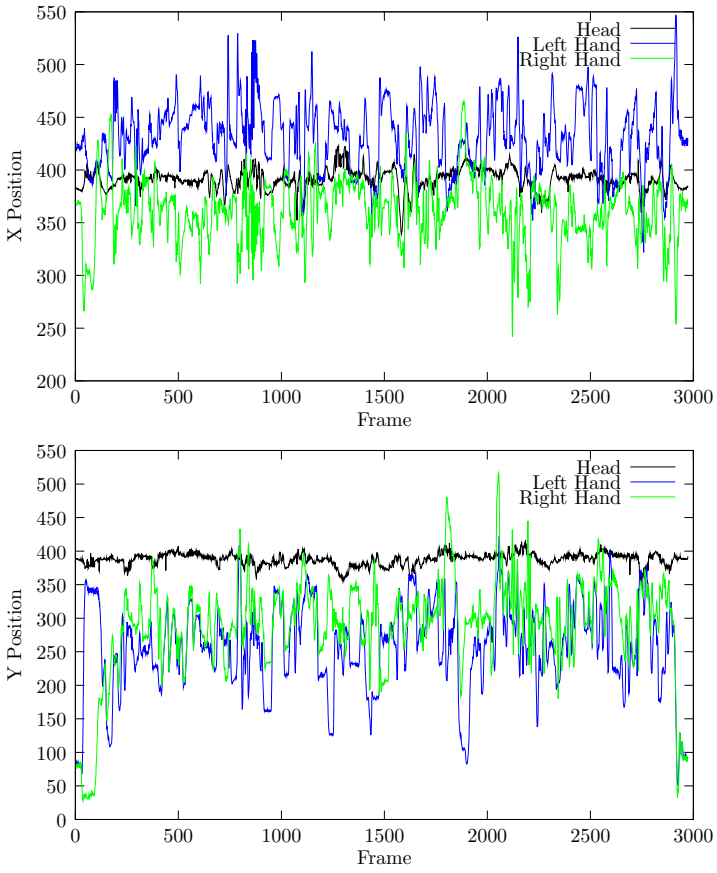
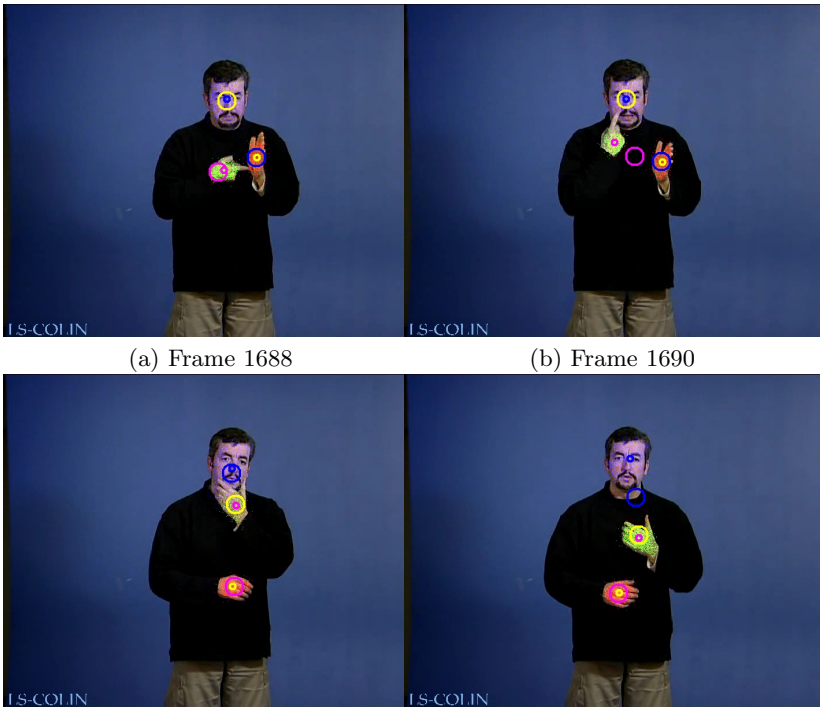



**Fig. 2.** The ground truth trajectories of the head and the two hands, (x,y) coordinates.

As expected the APF achieves a better robustness against strong dynamic variations than the PF and in the same way against local maxima (fig. 3).

We have estimated the frame to frame tracking that is to say if a link between two physical objects detected at two consecutive time instants is correctly computed or not. The metric uses the following comparison information:

1. Detected object at time $t$ and $t+1$ are related to the same reference data using the euclidean distance and a threshold.
2. A link exists between detected objects at time $t$ and $t+1$ and a link exists also in reference data



(a) Frame 1688 (b) Frame 1690

(c) Frame 116 (d) Frame 125

**Fig. 3.** The APF (small circles) achieve a better robustness against strong dynamic variations (a,b) and against local maxima (c,d) than the PF (large circle)

If there are several links between detected objects related to the same reference data, the one which maximize the overlap with reference data is kept as the good link and is removed for further association. Using this we compute four metrics (fig.4) :

- **Good tracking** $GT$, reference data link matching a link between two physical objects.
- **False tracking** $FT$, a link between two physical objects not matching any reference data.
- **Miss tracking 1** $MT1$, reference data link not found due to frame-to-frame tracking shortcomings, reject of case (1)
- **Miss tracking 2** $MT2$, reference data link not found due to frame-to-frame tracking shortcomings, reject of case (1) and (2).

and then :

Precision $\frac{GT}{GT+FT}$, Sensitivity1 $\frac{GT}{GT+MT1}$ and Sensitivity2 $\frac{GT}{GT+MT2}$

We have performed evaluations with a different particle number 2000, 3000 and 6000 to outline the behavior of the filter (fig. 5). The optimal number of

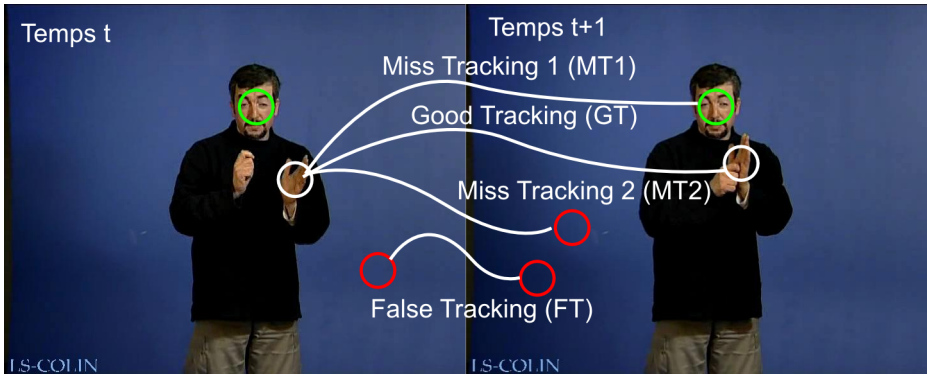



**Fig. 4.** Metrics used for frame-to-frame evaluation.

particles is around 3000 for the hands and around 6000 for the head, numbers that correspond to the size (in pixels) of those areas (table 1). For each filter the survival diagnostic has been evaluated and reflects those results.

**Table 1.** Precision and sensitivity results of the tracking with different particle numbers

| Number of particles | | Head | Right Hand | Left Hand |
|---|---|---|---|---|
| 2000 | P | 0.90393 | 0.85489 | 0.90024 |
| | S1 | 0.94091 | 0.98491 | 0.91875 |
| | S2 | 0.95833 | 0.86624 | 0.97810 |
| 3000 | P | 0.92912 | 0.96708 | 0.98589 |
| | S1 | 0.93890 | 0.97428 | 0.99323 |
| | S2 | 0.98892 | 0.99242 | 0.99256 |
| 6000 | P | 0.93013 | 0.95532 | 0.94995 |
| | S1 | 0.95024 | 0.96801 | 0.95929 |
| | S2 | 0.97775 | 0.98647 | 0.98985 |

The error in position has been computed from the euclidean distance between the computed position and the ground truth (fig. 5). Error peaks are caused by situations where the hands and head are very close to each others or merely occluding each other and move away under heavy acceleration. However the trackers do not miss their target a long time, they re-find them quickly. With a threshold of 50 pixels of distance (the tracker miss the target), one counts 224 errors for the head (7,4%), 25 and 79 for each hand (1,7%) on the 3000 images. The program is written in C++ and it can treat up to two frames per seconde without of any optimisation on a 1.86GHz Pentium M processor powered laptop.

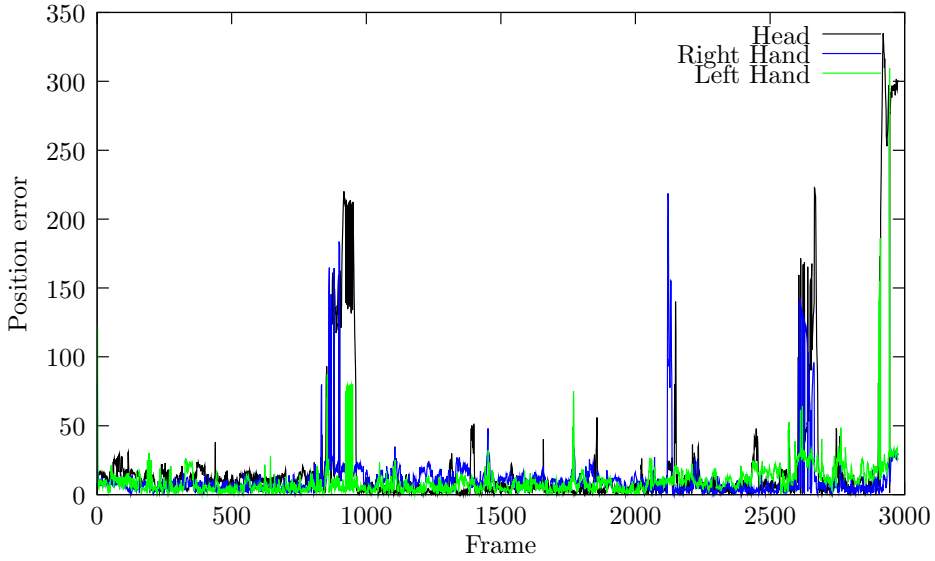



**Fig. 5.** Position error for the head, right hand and left hand

## 5 Conclusion and perspectives

We have presented a procedure to perform a visual tracking of very similar objects using particle filter. The particle filter has been provided with a simulated annealing update in order to improve the robustness against local maxima and high dynamics variations. An evaluation shows the improvement of this method compared to the original one. The results are promising, but a better multi object framework has to be developed to reduce labels mistakes.
Starting from these results, a study on the hands' shape changes is undertaken, being based on various measurements like, motion, Cartesian geometric moments [11] and features of texture. Thanks to these works, we wish to build tools based on computer vision techniques to help signs' langage analysis and to integrate these tools in our software of video's annotation's [12][13] to make the annotation's task easier than by hand.

## References

1. Cadoz, C.: Le geste canal de communication homme/machine. La communication "instrumentale". Techniques et Sciences Informatiques (13) (1994) 32–61
2. McNeill, D.: Hand and mind: What gestures reveal about thought. University of Chicago Press (1992)
3. Carbini, S., Viallet, J., Bernier, O., Bascle, B.: Tracking body parts of multiple people for multi-person multimodal interface. In: IEEE International Workshop on Human-Computer Interaction, Beijing, China (21 October 2005)

4. Gall, J., Potthoff, J., Schnoerr, C., Rosenhahn, B., Seidel, H.: Interacting and annealing particle filters: Mathematics and a recipe for applications. Technical Report MPI-I-2006-4-009, Max-Planck Institute for Computer Science (2006)
5. Kitagawa, G.: Monte carlo filter and smoother for non-gaussian nonlinear state space models. Journal of Computational and Graphical Statistics **5**(1) (1996) 1–25
6. Deutscher, J., Blake, A., I., R.: Articulated body motion capture by annealed particle filtering. Computer Vision and Pattern Recognition (2000)
7. MacCormick, J., Isard, M.: Partitioned sampling, articulated objects, and interface-quality hand tracking. In: Proc. of European Conf. on Computer Vision. Number 2 (2000) 3–19
8. Bar-Shalom, Y.: Tracking and data association. Academic Press Professional, Inc., San Diego, CA, USA (1987)
9. Haritaoglu, I., Harwood, D., Davis, L.: Ghost: a human body part labelling system using silhouette. In: Proc. of IEEE Conf. on Pattern Recognition. Number 1 (1998) 77–82
10. MacCormick, J., Blake, A.: A probabilistic exclusion principle for tracking multiple objects. Int. Journal on Computer Vision **39**(1) (2000)
11. Cassel, R., Collet, C., Gherbi, R.: Real-time acrobatic gesture analysis. In Gibet, S., Courty, N., Kamp, J.F., eds.: Gesture in Human-Computer Interaction and Simulation: $6^{th}$ International Gesture Workshop, GW 2005, Revised Selected Papers. Number 3881 in Lecture Notes in Computer Science, Berder Island, France, Springer-Verlag (18–20 May2005 2006) 88–99
12. Braffort, A., Choisier, A., Collet, C., Dalle, P., Gianni, F., Lenseigne, B., Segouat, J.: Toward an annotation software for video of sign language, including image processing tools and signing space modelling. In: Proc. of $4^{th}$ International Conference on Language Resources and Evaluation - LREC 2004. Volume 1., Lisbon, Portugal (26–28 May 2004) 201–203
13. Lenseigne, B., Dalle, P.: Using Signing Space as a Representation for Sign Language Processing . In Gibet, S., Courty, N., Kamp, J.F., eds.: Gesture in Human-Computer Interaction and Simulation: $6^{th}$ International Gesture Workshop, GW 2005, Revised Selected Papers. Number 3881 in Lecture Notes in Computer Science, Berder Island, France, Springer-Verlag (18–20 May2005 2006) 25–36