View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Infoscience - École polytechnique fédérale de Lausanne

Robust infants face tracking using active appearance models: a mixed-state CONDENSATION approach

Luigi Bagnato[1], Matteo Sorci[1], Gianluca Antonini[1], Giuseppe Baruffa[3], Andrea Maier[2], Peter Leathwood[2], and Jean-Philippe Thiran[1]

[1] Ecole Polytechnique Federale de Lausanne, Signal Processing Institute
Ecublens, 1015 Lausanne, Switzerland
{Luigi.Bagnato,Matteo.Sorci,Gianluca.Antonini,JP.Thiran}@epfl.ch
[2] Nestlé Research Center (CRN), Food-Consumer Interactions Department,
Lausanne, Switzerland
{Andrea.Maier,Peter.Leathwood}@rdls.nestle.com
[3] University of Perugia,Dept.of Electronic and Information Engineering
06125 Perugia, Italy
baruffa@diei.unipg.it

Abstract. In this paper a new extension of the CONDENSATION algorithm, with application to infants face tracking, will be introduced. In this work we address the problem of tracking a face and its features in baby video sequences. A mixed state particle filtering scheme is proposed, where the distribution of observations is derived from an active appearance model. The mixed state approach combines several dynamic models in order to account for different occlusion situations. Experiments on real video show that the proposed approach augments the tracker robustness to occlusions while maintaining the computational time competitive.

1 Introduction

The tracking of the face motion in a video sequence represents a challenging task in computer vision, because of the variability of facial appearance in real scenes, most notably due to changes in head pose, expressions, lighting or occlusions. This is especially challenging when an infant face is the tracking target. This task requires, by definition, the use of a model that describes the expected structure of the face. The Active Appearance Model (AAM) [1] is one of such techniques, which elegantly combines shape and texture models in a statistical framework, providing as output a mask of face landmarks. These combined models account for all sources of variability in face images. This feature makes them suitable for face tracking and enables the tracking of both global motion and inner features. Previous work on visual tracking can be divided in two groups: deterministic tracking and stochastic tracking. Deterministic approaches [2] usually reduce to an optimization problem, i.e. minimizing an appropriate cost function, while stochastic tracking approaches often reduce to the estimation of the state for a

time series state space model. Stochastic tracking improves robustness over its deterministic counterpart, thanks to its capability to escape from local minimum since the search directions are for the most part random. Early approaches used Kalman filter (or its variants [3]) to provide solutions, while, recently, sequential Monte Carlo algorithms [4] have gained prevalence in the tracking literature, especially due to the CONDENSATION algorithm [5].

Our work is based on a direct combination of an AAM with a particle filter as first introduced by Hamlaoui [6]. In this approach the authors combine an AAM with a temporal dynamics guided by the AAM search algorithm and use a filtering scheme based on CONDENSATION. Although their stochastic approach allows to augment robustness, they rely too much on the deterministic AAM search and the resulting algorithm performs poorly in case of heavy occlusions. Our contribution consists in a customized version of the mixed-state algorithm to face the particular problem of infants face tracking.
In Figure 1, some example images show inherent difficulties when dealing with real video sequences of infants. There are two major elements, which add complexity to the tracking task: infants move continuously and in an unpredictable way, producing face self-occlusions as most undesirable effect; external objects (a hand in Figure 1) may occlude the target face either partially or totally. In this paper we propose a technique for a robust tracker of a single infant face in a video sequence with the following properties: high responsiveness to sudden movements, robustness to partial occlusions, short recovery period after distraction due to a total occlusion. Our approach, based on the CONDENSATION algorithm, integrates, in a Bayesian mixed-state framework, multiple dynamic models, allowing to cope with the limitations of previous approaches.
The rest of the paper is organised as follows: in Section 2 we briefly review the Active Appearance Model and we introduce the AAM-based CONDENSATION framework. Section 3 describes in detail our proposed mixed-state approach, while Section 4 is dedicated to the experimental results. Conclusions and future works are then reported in Section 5.

Fig. 1. Example images from input video sequences.

2 Background

2.1 Face Active Appearance Model

The AAM is a statistical method for matching a combined model of shape and texture to unseen faces. The combination of a model of shape variations with a model of texture variations generates a statistical appearance model. Principal Component Analysis (PCA) is applied to build the statistical shape and texture models:

$$\mathbf{s} = \bar{s} + \Phi_s b_s \quad \textit{and} \quad \mathbf{g} = \bar{g} + \Phi_t b_t \tag{1}$$

where $\bar{s}$ and $\bar{g}$ are the mean shape and texture, Φ_s and Φ_t are the eigenvectors of shape and texture covariance matrices. The unification of the presented shape and texture models into one complete appearance model is obtained by concatenating the vectors b_s and b_t and learning the correlations between them by means of a further PCA:

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{Q_s c} \quad \textit{and} \quad \mathbf{g} = \bar{\mathbf{g}} + \mathbf{Q_t c} \tag{2}$$

where Q_s and Q_t are the matrices describing the principal modes of the combined variations in the training set and $\mathbf{c}$ is the appearance parameters vector. By varying the appearance parameters $\mathbf{c}$, new instances of shape and texture can be generated. The matching of the appearance model to a target face can be treated as an optimization problem, minimizing the difference between the synthesized model image and the target face [1].

2.2 AAM-based CONDENSATION

The CONDENSATION is a Monte Carlo-type technique to recursively approximate the posterior state density. Approximation is done by means of the empirical distribution of a system of particles. The particles explore the state space following independent realizations from a state dynamic model, and are redistributed according to their consistency with the observations, the consistency being measured by a likelihood function (observation model). For an introduction to the subject the reader is referred to the Isard and Blake paper [5]. Using Active Appearance Models, we can represent the shape and texture of a face in terms of a vector $\mathbf{c}$. To complete the description of the face, the four pose parameters are also needed, namely $\mathbf{p} = (\alpha, \vartheta, t_x, t_y)$, representing scale, orientation and position, respectively. The state vector $\mathbf{x}$, which contains the parameters used to infer about the object (the face) is thus composed by the concatenation of the vector $\mathbf{c}$ of combined parameters and vector $\mathbf{p}$ of pose.

Observation model The Observation Model is based on the difference between the sampled pixel grey level patch at the hypothesized position in the current image and the one generated by the face model. The likelihood function $p(\mathbf{y}_k|\mathbf{x}_k)$ denotes the probability that a hypothesized state $\mathbf{x}_k = (\mathbf{c}_k, \mathbf{p}_k)$ gives rise to the

observed data. Since the observed data consist of pixel greylevel values, it is straightforward to look for a function with the following form:

$$p(\mathbf{y}_k|\mathbf{x}_k) = p(\mathbf{y}_k|\mathbf{c}_k, \mathbf{p}_k) = C \exp\left(-d[\mathbf{g}_{model}(\mathbf{c}_k), \mathbf{g}_{image}(\mathbf{c}_k, \mathbf{p}_k)]\right) \quad (3)$$

where $\mathbf{g}_{image}(\mathbf{c}_k, \mathbf{p}_k)$ is the image patch sampled at the hypothesized pose and shape, $\mathbf{g}_{model}(\mathbf{c}_k)$ is the model texture representing the hypothesized appearance of the face, and C is a normalizing constant. The texture distance $d[;]$ is an error measure, summed over all L pixels of both textures.

State transition model The state transition model characterizes the dynamics between frames. The state evolves according to

$$\mathbf{x}_k = \mathbf{f}(\mathbf{x}_{k-1}) + \mathbf{S}_k\mathbf{u}; \quad (4)$$

in this equation, $\mathbf{f}(\mathbf{x}_{k-1})$ represents a deterministic function of the previous state vector $\mathbf{x}_{k-1}$, $\mathbf{S}_k$ is the process noise covariance and $\mathbf{u}$ is a vector of normally distributed random variables.
Whereas the choice of gaussian noise is expected when accurate model uncertainty cannot be provided, choosing an appropriate function $\mathbf{f}(.)$ is not an easy task, and depends on the particular situation.

3 Proposed scheme: a mixed state CONDENSATION

In the case of this particular application the presence of heavy occlusions and the unpredictable infant movements make the choice of a single model inadequate to describe the dynamics. In this spirit we opted for a mixed-state framework with model switching.

3.1 Pose-CONDENSATION and ICONDENSATION

The first two retained models consist in a fixed constant-velocity model with fixed noise variance

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{S}_k\mathbf{u}; \quad (5)$$

and an adaptive dynamic model, guided by a deterministic AAM search:

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \Delta\mathbf{x}_{k-1,k} + \mathbf{S}_k\mathbf{u}, \quad (6)$$

where $\Delta\mathbf{x}_{k-1,k} = (\Delta\mathbf{c}_{k-1,k}, \Delta\mathbf{p}_{k-1,k})$ is the predicted shift in pose and appearance parameters, obtained by applying the AAM search to previous estimated state vector $(\mathbf{x}_{k-1})$ with respect to frame at time k. An AAM-CONDENSATION approach, using either (5) or (6) as dynamic models, has three main drawbacks: AAMs describe faces with high-dimensional vectors; the deterministic AAM search is highly sensitive to large occlusions, so robustness is achieved only by increasing the noise variance; accuracy is only accomplished at the cost

of unacceptable time performance. In order to solve the problem of dimensionality while assuring a good robustness to occlusions, Davoine [7] proposed the use of the CONDENSATION to track the four pose parameters (in the following this algorithm is referred as Pose-CONDENSATION). In this case (5) is a convenient form for the dynamic model with the state vector represented by the pose vector. Obviously, the accuracy of the algorithm, intended as the ability to generate a photo realistic synthetic replica of the face, cannot be guaranteed since the appearance parameters, $\mathbf{c}$, are not tracked. In order to track the entire state vector $\mathbf{x} = (\mathbf{c}, \mathbf{p})$ and keep the computational time acceptable, a solution could be to use an importance function, as described in the ICONDENSATION framework [8], to constrain the search in a neighbourhood of the previous state estimate. In this case a proper dynamics is the one described in (6). This approach performs well, in terms of speed and accuracy, only in presence of soft partial occlusions: the deterministic search is inherently not robust and the main feature of CONDENSATION, i.e. maintaining multiple hypothesis, is constrained in a limited regionof the state space by the importance function. The two described approaches are complementary: the first one sacrifices accuracy for robustness, while the second one allows fast and accurate tracking, but only in occlusions free situations. The integration of both dynamics into the same tracker would then allow for a wider range of motion to be supported without losing the advantages of an accurate prediction. In summary, we can say that in case of limited occlusions, the deterministic AAM search should be reasonably trusted, leaving to the importance sampling technique the task to improve the accuracy of the detection. In the second scenario, when strong occlusions occur, a less accurate tracker but more robust to occlusions should be preferred.

3.2 The third model

In the proposed solution, a third motion model from data averaging has been included and defined by:

$$\mathbf{x}_k = \bar{\mathbf{x}}_k + \mathbf{S}_k \mathbf{u} \tag{7a}$$

$$p(\mathbf{x}_k) = N(\bar{\mathbf{x}}_{k-1}, \mathbf{S}_k), \tag{7b}$$

where $\bar{\mathbf{x}}_\mathbf{k}$ is an estimate of the (fixed) *mean vector* $\bar{\mathbf{x}}$ and $\mathbf{S}_k$ is the covariance matrix of the gaussian distribution at time k. The aim of this model is to describe a kind of a priori knowledge on the face motion of the considered sequences. In each video sequence the infants are sitting, thus their movements are somehow constrained around a region in the scene. We can then assume that the face lies in a neighbourhood of such region, with probability decreasing with the distance. This third motion model (in the following referred as Reinitialization) can be used to include some probability of tracking reinitialization, particularly useful after distraction due, for example, to total occlusions.

3.3 The mixed-state algorithm

Our solution, then, consists in merging the three approaches by means of an automatic model switching procedure ([9]). The extended state is defined as

$\mathbf{X} = (\mathbf{x}, \theta),\ \theta \in 1, .., N$,where θ is a discrete variable labelling the current model, while N represents the total number of models. The process density can then be decomposed as follows:

$$p(\mathbf{X}_k|\mathbf{X}_{k-1}) = p(\mathbf{x}_k|\mathbf{x}_{k-1}, \theta_k)P(\theta_k|\theta_{k-1}, \mathbf{x}_{k-1}) \tag{8}$$

where $P(\theta_k|\theta_{k-1}, \mathbf{x}_{k-1}) : P(\theta_k = j|\theta_{k-1} = i, \mathbf{x}_{k-1}) = T_{ij}(\mathbf{x}_{k-1})$ and the T_{ij} are *state transition probabilities*. The continuous motion models for each transition are given by the *sub-process densities* $p(\mathbf{x}_k|\mathbf{x}_{k-1}, \theta_k)$. During tracking, each discrete state transition with non zero probability contributes some samples to the state distribution and the model that predicts more accurately than the others will dominate. The possible values of θ_k are: R (Reinitialization), I (ICONDENSATION), P (Pose-CONDENSATION). The particle is propagated forward in time according to the dynamics implied by the motion model, and transitions between models happen according to the transition matrix $\mathbf{T}$. We propose the following simple form for the transition matrix where the parameters α and δ control the robustness of the tracker:

$$\mathbf{T} = \begin{pmatrix} T_{RR}\ T_{RI}\ T_{RP} \\ T_{IR}\ \ T_{II}\ \ T_{IP} \\ T_{PR}\ T_{PI}\ T_{PP} \end{pmatrix} = \begin{pmatrix} \alpha\ 1-\alpha-\delta\ \delta \\ \alpha\ 1-\alpha-\delta\ \delta \\ \alpha\ 1-\alpha-\delta\ \delta \end{pmatrix} \tag{9}$$

- α is a *reinitialization* parameter, since, at each time step, a number of particles proportional to α is generated from model (7).
- The meaning of δ changes, instead, between the first and the second representation: in the first case it represents an *adaption speed* parameter, that controls how rapidly the probability flows from the model I to the model P. Thus, we can say that it trades off adaptation rate and steady state behavior. In the second case, we give it the meaning of *robustness* parameter, controlling how promptly the tracker switches into pose tracking.

Summarizing, at each time step k, the proposed algorithm chooses a particle from the previous sample set, proportionally to its weight. The particle is then propagated through one of the three dynamic models, in accordance with the current motion label. If, for example, a particle is chosen with label I, then with probability T_{II} a particle is drawn from the importance function $q(\hat{\mathbf{x}}_{k-1})$, with probability T_{IR} it is drawn from (7), and with probability T_{IP} it is propagated through (5) (where $x \equiv p$). Finally, the particle is weighted in accordance to observation, and the multiplicative factor f/q is applied if it was generated with Importance Sampling.

4 Experimental results

The implementation of the tracker is based on AAM-API, a C++ implementation of Active Appearance Model. In order to build our AAM representation of face, we have manually landmarked a set of 222 images. To test the Mixed State CONDENSATION tracker we used 50 video sequences. For all of them a visual

analysis has been done, resulting in good overall performance of the proposed tracker. Figure 2 shows consecutive frames of an example sequence characterized by a total occlusion. Given that the videos used in the experiments are under corporate proprietary rights, only a small percent of them can be published. In Figure 3, 4 and 5 we report the results applying the three methods to three different sequences. In the image grid, each column represents a tracked frame from the sequence, while each row is related to one of the 3 compared approaches. Below the image grid, a plot shows the tracking error for the entire sequences.

Fig. 2. Tracking results for consecutive frames

The results show that, although the Pose-CONDENSATION tracker is quite robust to occlusions, the tracking error is still very high, since the inner motion of the face is not tracked. The behaviour of the ICONDENSATION is exactly the opposite: it tracks well as long as the target is not occluded, but it is not able to recover the target face once distracted. The mixed-state CONDENSATION offers a good trade-off: it automatically switches model, choosing the best for each situation. It reveals high robustness and the best accuracy, from our analysis. Concerning the α and δ parameters of the transition matrix $\mathbf{T}$, their values have been empirically chosen in order to adapt the algorithm to our task. Table 1 reports the time performance of the different algorithms for 3 representative sequences. The results are obtained with a P4 1.8 GHz processor, equipped with 512MB of RAM. A further algorithm has been used for benchmarking: AAM Search. This is a simple algorithm in which the AAM Search is applied frame-by-frame: it cannot be used for tracking in practical situations, since it is neither

accurate nor robust, however it gives a kind of reference for time performance. From Table 1, it is clear that, despite the increased complexity of the tracker, the mixed-state CONDENSATION has performance comparable with those of the other algorithms.

	IC	**C-pose**	**Mixed State C.**	**AAM Search**
Sequence 1	2.62	2.58	2.53	3.18
Sequence 2	3.13	2.57	2.59	2.5
Sequence 3	3.19	2.48	2.5	2.63

Table 1. Time performance comparison (in frames per second)

5 Conclusion

In this work we presented a stochastic framework for robust face tracking using complex models of face appearance. When compared to other approaches found in literature [6], the presented tracker not only succeeds in crucial cases of occlusion, but experiments show that it outperforms in accuracy the compared methods and finds an equilibrate trade-off between robustness and use of time resources. The resulting algorithm is an adapted mixed state CONDENSATION combined with AAM. The stochastic CONDENSATION search compensates for AAM limits in handling occlusions, and due to an appropriate choice of motion models, allows an efficient reinitialization of tracking, when an exhaustive search in the image is impractical. Furthermore, the approach is general enough to be applied to other face tracking problems: an advantage of the probabilistic approach is that it is modular, in the sense that application-specific dynamic models or observation models can be seamlessly included.

References

1. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. Pattern Analysis and Machine Intelligence, IEEE Transactions on **23** (2001) 681–685
2. Comaniciu, D., Ramesh, V., Meer, P.: Real-time tracking of non-rigid objects using mean shift. CVPR **2** (2000) 142–149
3. Anderson, B.D.O., Moore, J.B.: Optimal Filtering. Dover Publications (2005)
4. Doucet, A., Freitas, N.D., Gordon, N.: Sequential Monte Carlo Methods in Practice (Statistics for Engineering and Information Science). 1st edn. Springer (2005)
5. Blake, A., Isard, M.: The CONDENSATION algorithm - conditional density propagation and applications to visual tracking. In Mozer, M., Jordan, M.I., Petsche, T., eds.: NIPS, MIT Press (1996) 361–367
6. Hamlaoui, S., Davoine, F.: Facial action tracking using an aam-based condensation approach. IEEE ICASSP ,Philadelphia, U.S.A. (2005)
7. Davoine, F., Dornaika, F.: Head and facial animation tracking using appearance-adaptive models and particle filters. Springer Verlag (2005)
8. Isard, M., Blake, A.: ICONDENSATION: Unifying low-level and high-level tracking in a stochastic framework. Lecture Notes in Computer Science **1406** (1998) 893–908
9. Isard, M., Blake, A.: A mixed-state CONDENSATION tracker with automatic model-switching. In: ICCV. (1998) 107–112

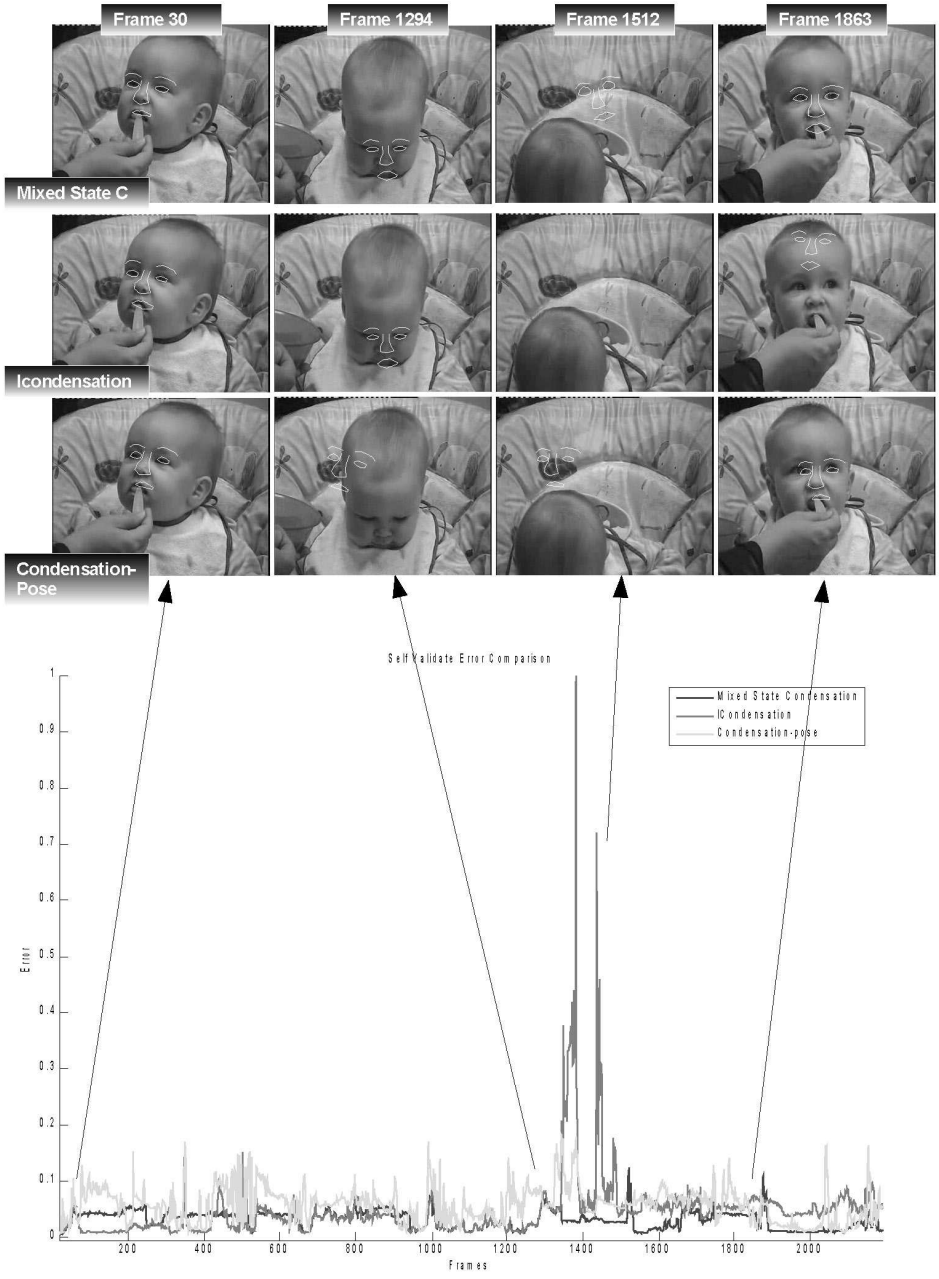

Fig. 3. Tracking results and tracking errors (Lorentzian norm) for Sequence 1

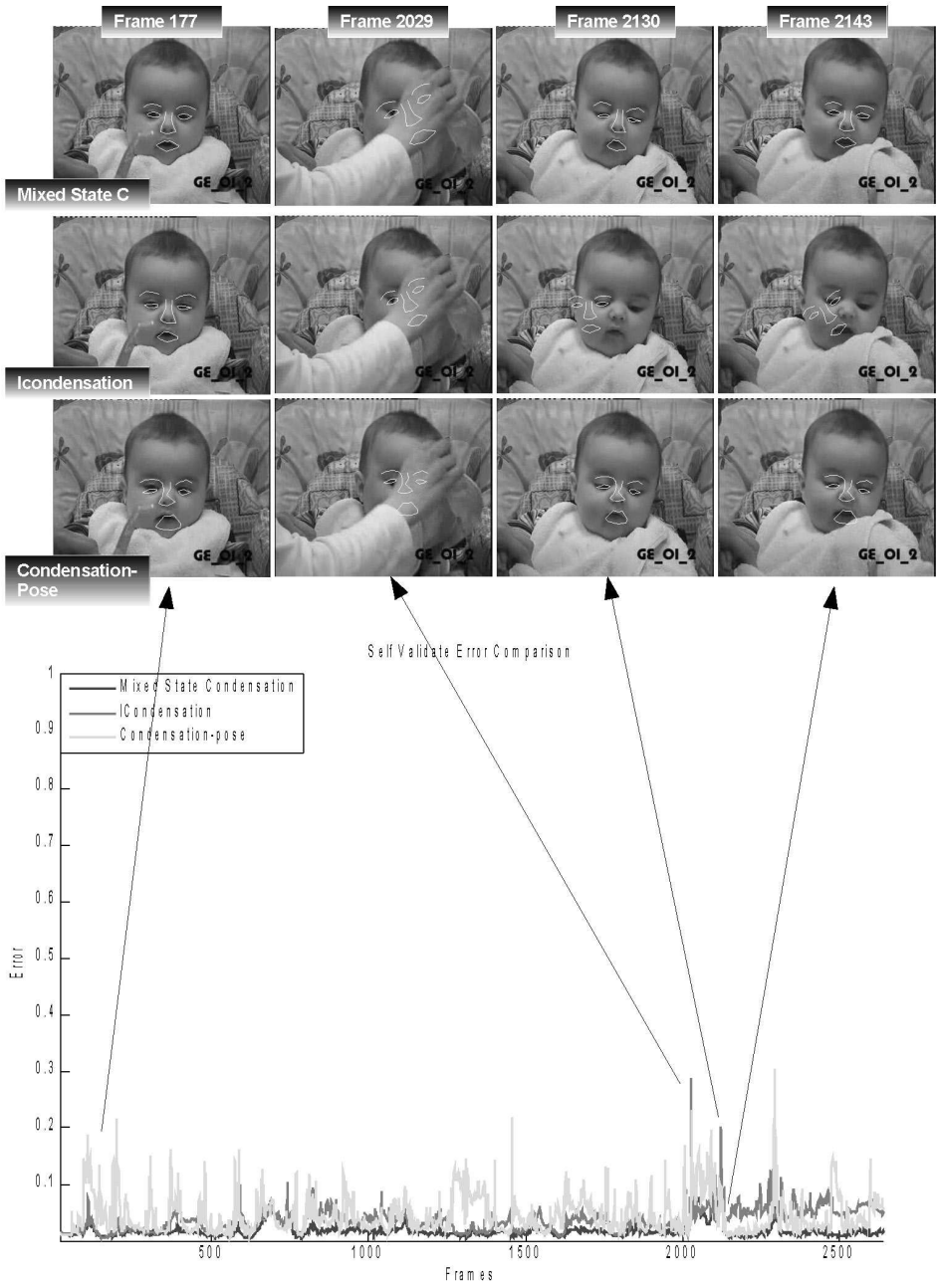

Fig. 4. Tracking results and tracking errors (Lorentzian norm) for Sequence 2

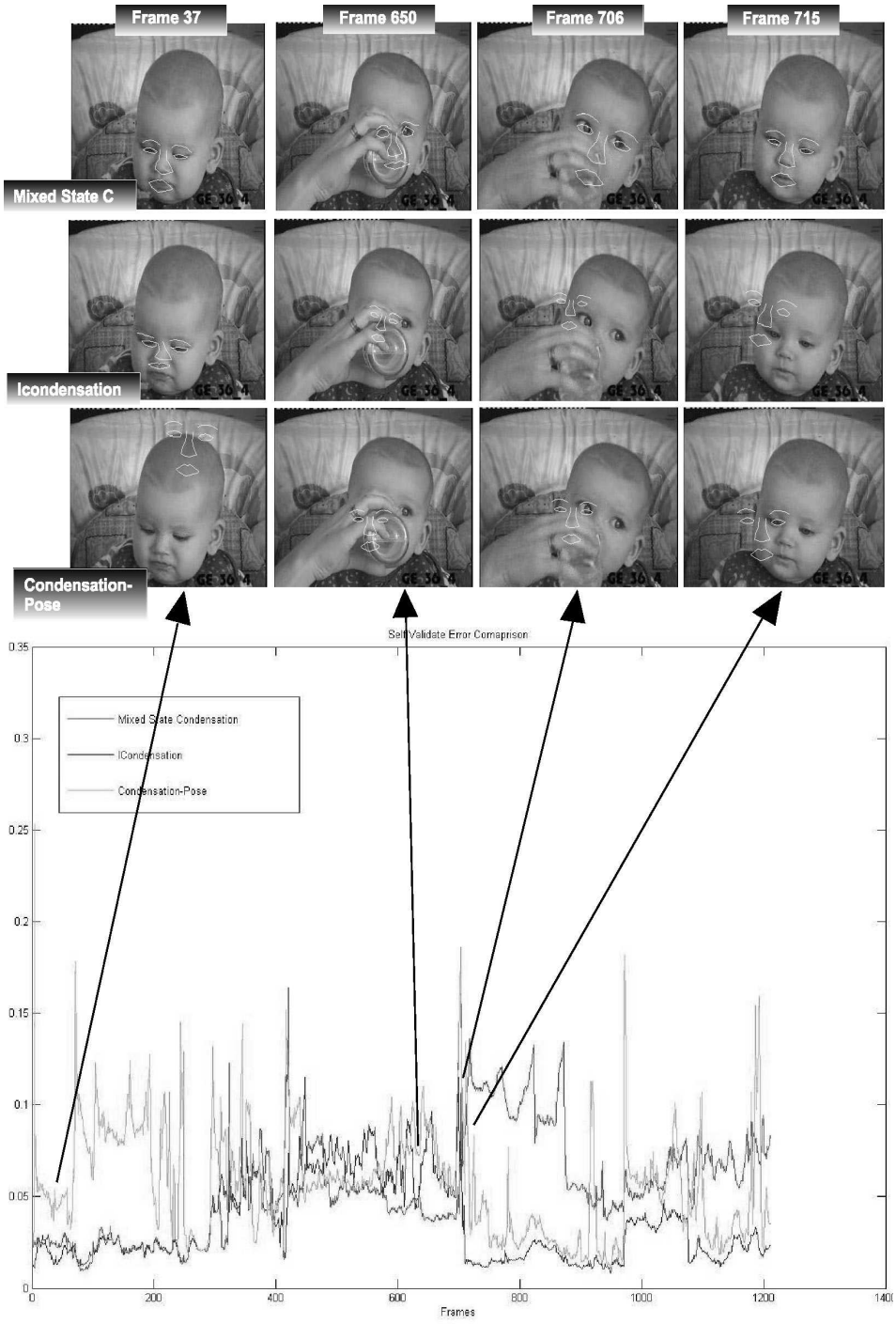

Fig. 5. Tracking results and tracking errors (Lorentzian norm) for Sequence 3