2005 IEEE International Conference on Systems, Man and Cybernetics
Waikoloa, Hawaii October 10-12, 2005

Tracking deformable motion

I. Patras and M. Pantic
Delft University of Technology
Electrical Engineering, Mathematics, and Computer Science
Delft, The Netherlands
{I.Patras, M.Pantic}@ewi.tudelft.nl

Abstract - This paper addresses the problem of template-based tracking of non rigid objects. We use the well-known framework of auxiliary particle filtering and propose an observation model that explicitly addresses appearance changes that are caused by local deformations of the tracked object. In addition, by adopting a colour difference that is invariant to local changes in the illumination, the proposed observation model can deal with changing lighting conditions and shadows. Experimental results with real image sequences demonstrate the efficiency of the proposed method in tracking facial features, such as mouth and eye corners.

1. Introduction

Boosted by applications in surveillance, military target tracking, multimedia analysis and, recently, by applications in the area of Human Machine Interaction, there has been a considerable interest in visual object tracking. In this framework, the goal is the estimation of the location of an object at each frame of an image sequence based on a template that models the appearance of the object in question. The template is either learned off-line from a database containing images of the object in question, or is initialised at the first frame of the sequence. The main challenge in template-based tracking is that the template cannot always find a good match at the true location of the object because of occlusions, changing lighting conditions and non rigid object motion.

In particular, the problem of changes in the object's appearance due to rigid and non rigid transformations has been in the focus of the academic community for more than a decade. A popular and robust approach, when the object or the class of the object in question (e.g. faces, or vehicles) is known in advance, is to train an appearance model from a database that contains instances the object in question. In this framework a number of approaches that lie at different parts of the spectrum between tracking and detection have been proposed [3] [2] [4] [11]. However, training databases are available only for a limited classes of objects, such as faces, and therefore the applicability of such schemes is limited. Furthermore, the construction of a database is by no means a trivial task.

On the other hand, the approaches that rely on a template that is initialized at a single (the first) frame of an image sequence attempt to deal with changes in the illumination and in the object's appearance a) by designing robust similarity functions for matching the template with image patches or b) by adopting schemes that incrementally update the information that is contained in the template. The most common choice for the definition of a similarity function is the Mean of Squared Differences defined at an appropriate colour/intensity space. However, such a similarity function is known to be sensitive to outliers (that is mismatched pixels at which the colour/intensity difference is usually very high) and therefore not robust to occlusions and changes in the appearance due to object deformations. Jepson et al [1] propose a robust appearance model that is based on a mixture of three distributions, one of which, the so called wandering component, is designed to cope with outliers. In a similar approach, Nguyen and Smeulders [10] design robust similarity functions that are based on M-estimators and reduce the effect of outliers by limiting the cost that is introduced by large mismatches. They report results for both an intensity-based appearance model and well as with the use of a normalised RGB space. Finally, Matthews *et al* [8] adopt an update strategy that aims at reducing the problem of template drifting, that is the gradual loss of tracking due to the update of the template with background information, by maintaining the template that was extracted at the first frame of the image sequence.

Although significant, these works do not explicitly address the problem of structural appearance changes caused by object deformation. Instead their similarity functions implicitly assume that the changes in the colour/intensity of a single pixel are generated by a random process and not, as is the case of deformable motion, by 'structural displacement' of colour/intensity values within the template. In what follows we present a method that explicitly addresses varia-

0-7803-9298-1/05/$20.00©2005 IEEE

tion in the appearance of the model that are caused by non rigid deformations within the template. More specifically, we propose a distance measure between a template c that is extracted at the first frame of the sequence and the image data around position α at image y. The distance measure comprises of a colour distance term and a shape deformation term and can be seen as a generalization of distance transformations used for tracking deformable binary contours. Finally, we will address appearance changes due to changes in the lighting conditions by utilizing a colour distance that is invariant to local changes to the illumination intensity. We will incorporate our observation model in the particle filtering framework and more specifically use the auxiliary particle filtering as this was introduced by Pitt and Shephard [14].

The rest of the paper is organized as follows. In section 2 we will briefly describe the auxiliary particle filtering and in section 2.1 we give an extended description of the proposed observation model. In section 2.2 we briefly describe the transition model that we have used. In section 3 we present experimental results for tracking deformable objects and facial features in real image sequences and compare the proposed observation model with other robust observation models. Finally, in section 4 we draw conclusions and discuss future research directions.

2. Auxiliary Particle Filtering

In the recent years, particle filtering has been the dominant paradigm [5] [6] [14] [9] [7] [13] [16] for tracking the state α of a temporal event given a set of noisy observations $Y = \{\ldots, y^{-}, y\}$ up to the current time instant; in our case we denote with Y the image sequence up to the current time instant. Its ability to maintain simultaneously multiple solutions, the so called particles, makes it particularly attractive when the noise in the observations is not Gaussian and makes it robust to missing or inaccurate data. The main idea of the particle filtering is to maintain a particle based representation of the *a posteriori* probability $p(\alpha|Y)$ of the state α given all the observations Y up to the current time instance. This means that the distribution $p(\alpha|Y)$ is represented by a set of pairs $\{(s_k, \pi_k)\}$ such that if s_k is chosen with probability equal to π_k, then it is as if s_k was drawn from $p(\alpha|Y)$. In the particle filtering framework our knowledge about the *a posteriori* probability is updated in a recursive way. Suppose that we have a particle based representation of the density $p(\alpha^{-}|Y^{-})$, that is we have a collection of K particles and their corresponding weights (i.e. $\{(s_k^{-}, \pi_k^{-})\}$). Then, the Auxiliary Particle Filtering can be summarized as follows:

1. Propagate all particles s_k^{-} via the transition probability $p(\alpha|\alpha^{-})$ in order to arrive at a collection of K particles μ_k.
2. Evaluate the likelihood associated with each particle μ_k, that is let $\lambda_k = p(y|\mu_k)$.
3. Draw K particles s_k^{-} from the probability density that is represented by the collection $\{(s_k^{-}, \lambda_k \pi_k^{-})\}$. This is the essence of the auxiliary particle filtering; in this way it favors particles with high λ_k, that is particles which, when propagated with the the transition density, end up at areas with high likelihood.
4. Propagate each particle s_k^{-} with the transition probability $p(\alpha|\alpha^{-})$ in order to arrive at a collection of K particles s_k.
5. Assign a weight π_k to each particle as follows,

$$w_k = \frac{p(y|s_k)}{\lambda_k}, \quad \pi_k = \frac{w_k}{\sum_j wj} \tag{1}$$

This results in a collection of K particles and their corresponding weights (i.e. $\{(s_k, \pi_k)\}$) which is an approximation of the density $p(\alpha|Y)$.

2.1 The observation model

We subsequently define the observation model, that is the likelihood $p(y|\alpha; c)$. This likelihood expresses how well the image content y can be explained, given that the template is at position α. This likelihood is parametrized by the template c as this was extracted at the first frame of the image sequence.

More specifically, let us denote with $y(\alpha)$ the image patch centered around the spatial position α in the image y. With $y(\alpha, i)$ we denote the colour of the pixel i of image y expressed in the coordinate system defined by α and the image axes. Similarly, let us denote with $c(i)$ the colour at the pixel i of the colour template c.

We will define the likelihood $p(y|\alpha; c)$ in terms of a novel distance metric $d(y(\alpha), c)$ between the template c and the image patch $y(\alpha)$. Classically, the colour template and the image patch $y(\alpha)$ are aligned and the distance is defined as the summation of colour distances between pixels that are at the same position. Formally,

$$d(y(\alpha), c) = \sum_i \|c(i) - y(\alpha, i)\| \tag{2}$$

where $\|.\|$ is an appropriate colour distance which is classically taken to be the L2 norm in the RGB colour space.

The main problem with such colour metrics is that the correspondence between the pixels of the colour template and the pixels of the image patch $y(\alpha)$ does not hold in general due to rigid or non rigid motions that the depicted object undergoes. Such underlying motions create appearance changes that do not vary smoothly with respect to the underlying transformations and, therefore, can be bad measures for the localization of the template. As in the similar problem of the dense motion estimation, a low pass filter (i.e. smoothing) alleviates the problem. However, a low pass filtering a) leads to loss of detail and therefore can lead to bad localization and b) when applied in the colour domain introduces new colours on which colour-based invariant distances can no longer be applied. The latter is of particular importance in tracking with changing illumination conditions. Furthermore, the most common choice for a distance metric, that is the L2 norm, is sensitive to outliers, that is to large values of $c(i) - y(\alpha, i)$ that are caused when the correspondence between the pixel i in the template and the pixel i in the image patch $y(\alpha)$ does not hold.

Let us denote with $\Phi : N^2 \rightarrow N^2$ the unknown transformation that gives the correct correspondence between the pixel coordinates of the colour template c and the image patch $y(\alpha)$. We propose a novel distance metric that contains two terms. The first term, $d_c(c, y(\alpha, \Phi)))$, is a colour distance between the colour template c and the template that results after applying the non rigid shape transformation Φ on the image patch $y(\alpha)$. The second term, $d_s(\Phi)$, is a measure of the shape deformation that is introduced by the transformation Φ. The distance measure is the minimum, over all possible trasformations Φ, of the colour-based distance and the shape-based deformation cost. Formally,

$$d(y(\alpha), c) = \min_{\Phi} \left(d_c(c, y(\alpha, \Phi))(1 + \lambda d_s(\Phi)^p)\right) \tag{3}$$

where the first term of the product is used to penalize large colour-based distance, and the second term is used to penalize large shape deformations Φ. The parameters λ and the exponent p control the relative importance of the shape deformation term.

More specifically, the first term in eq. 3 is a colour-based distance between the colour template c and the colour template $y(\alpha, \Phi)$ that is formed after the transformation Φ is applied to the image patch $y(\alpha)$. Formally,

$$y(\alpha, \Phi) =< \ldots, y(\alpha, \Phi(i)), y(\alpha, \Phi(i+1)), \ldots > \tag{4}$$

where i is the pixel index.

We define the colour distance to be invariant to local changes in the intensity by normalizing each colour template with the average intensity. Finally, in order to reduce the effect of outliers, that is, the effect of large differences at certain pixels that are the result of noise or occlusions, we choose a robust error norm $\rho(.)$. Formally, the colour-based difference is defined as follows:

$$d_c(c, y(\alpha, \Phi)) = \frac{1}{\sigma_c} E_i \left\{ \rho \left(\left\| \frac{c(i)}{E\{c\}} - \frac{y(\alpha, \Phi(i))}{E\{y(\alpha)\}} \right\|_1 \right) \right\} \tag{5}$$

where $E\{c\}$ is the mean of c, that is, the average intensity of the colour template o_i, $\|.\|$ is the L_1 vector norm and as a robust function ρ we have used the absolute value. Finally, σ_c is a classical data scaling factor. It is easy to show that the proposed colour difference metric is invariant to global changes in the light intensity.

The second term of the product of eq. 3 is used to penalize large deformations Φ. Formally, $d_s(\Phi)$ it is defined as the average Euclidean distance over the pixel based displacements, that is

$$d_s(\Phi) = E_i \left\{ \sqrt{\|i - \Phi(i)\|_2} \right\} \tag{6}$$

where, with a slight abuse of notation, i denotes pixel coordinates. The essence of the shape deformation term is depicted in fig. 1 where the dashed line represents a structure in the template c and the solid line represents the same structure that at the image patch $y(\alpha)$ is slightly deformed under the (unknown) transformation Φ.

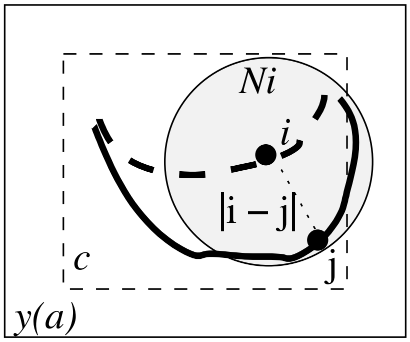


Figure 1: Shape distance for template matching.

The definition of the deformation cost as in eq. 6 allows an efficient estimation of $d(y(\alpha), c)$, that is, the minimization of eq. 3 over all possible transformations Φ. Since both colour difference and the deformations of neighboring pixels are considered independently, the minimization of eq. 3 can be performed independently for each pixel. It is straightforward to show that the transformation Φ^*, that minimizes eq. 3 is given by

$$\Phi^*(i) = \arg\min_{j} \frac{1}{\sigma_c} \rho \left(\left\| \frac{c(i)}{E\{c\}} - \frac{y(\alpha, j)}{E\{y(\alpha)\}} \right\|_1 \right) (1 + \lambda \sqrt{\|i - j\|_2}) \tag{7}$$

In practice, given a reasonable choice for the parameter λ, large deformations, that is large $\Phi^*(i) - i$ are unlikely to minimize eq. 7. For computational efficiency, we restrict the search for $\Phi^*(i)$ in a neighborhood of i, which we denote with N_i. Fig. 1 presents the colour template c projected on the image patch $y(\alpha)$, the neighborhood N_i of a pixel i in the colour template c and a pixel j in the image patch $y(a)$ that is considered as a candidate for the minimization of eq. 7.

Notice that for $\lambda \to \infty$, the shape transformation what minimizes eq. 7 is given by $\Phi^*(i) = i$ and the proposed distance (eq. 3) reduces to a classical color template distance. In our case, this color distance is the Mean Absolute color Difference. Finally, let us note that there are also other efficient ways of combining the color-based distance and the shape-based deformation cost. In [15] we presented results with $d(y(\alpha), c) = \min_\Phi (d_c(c, y(\alpha, \Phi)) + \lambda d_s(\Phi))$. Experimentally, we have found that it is easier to tune the parameter λ for the definition of eq.3. In all of our experiments we have used $\lambda = 0.1$ and $p = 0.3$.

Finally, the observation likelihood $p(y|\alpha; c)$ is defined as:

$$p(y|\alpha; c) = \max \left\{ \epsilon, \frac{1}{Z} e^{-d(y(\alpha), c)} \right\} \tag{8}$$

where Z is a normalization term that can be ignored, since in the context of the particle filtering the likelihoods are renormalized at each iteration (eq. 1). The term ϵ is a constant that is used in order to deal with large occlusions, and in general with situations at which the dissimilarity between the template c and all candidate image patches $y(\alpha), c)$ is very large. In such situations, where the tracking is effectively lost, the presence of the term ϵ does not allow the solution to be attracted by an image patch $y(\alpha)$ that is a good match relatively to the other image patches, but bad match in absolute terms. Effectively, the term ϵ assumes a uniform likelihood $p(y|\alpha; c)$ at occlusions or when the tracking is lost. Similar terms that limit the effect of large discrepancies between the template and the data have been extensively used in the framework of robust estimation and tracking [1] [10].

2.2 The transition model

In what follows we will describe the transition model, that is the probability density $p(\alpha|\alpha^-)$ that is used to generate a new set of particles given the current one. The transition probability models our knowledge about the feature dynamics, that is our prior knowledge about their position α in the current frame given their position α^- in the previous frame. In the case that the domain is known, prior knowledge can be very effective for reliable long term tracking. In [12] multiple facial features are simultaneously tracked and their joint positions are constrained by a probability that was learned from previously tracked points. However, in the case that the domain is not known, general constraints are usually imposed in the form of a first order (constant velocity) or second order (constant acceleration) models. Here, we use a simple zero order model in which the probability density $p(\alpha|\alpha^-)$ is modelled as a Gaussian noise around the position α^- in the previous frame. Formally,

$$p(\alpha|\alpha^-) = \mathcal{N}\left(\alpha^-, \Sigma\right) \tag{9}$$

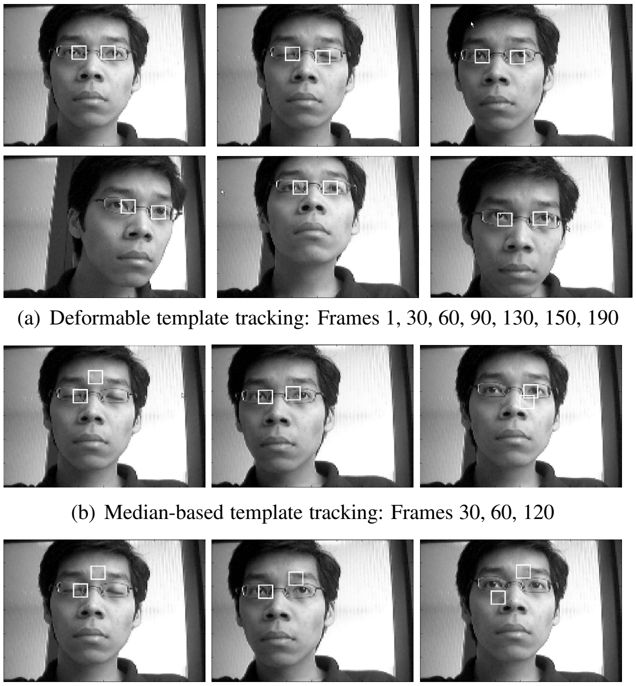
3. Experimental Results

We have applied the proposed method to a number of image sequences and here we present results for tracking facial features, that is the position of the mouth corners and the inner eye corners. We define small rectangular templates around each facial feature at the first frame and subsequently track the facial feature for the rest of the image sequence.

We have compared our method with the auxiliary particle filtering using different observation functions that are widely used in the literature, such as the mean of the absolute differences, the median and the Huber's function [10]. Here we will present results with the mean of absolute differences and with the median, which are known to perform better to the presence of outliers than the mean of squared differences.

In fig. 2 we present experimental results for tracking the corners of the eyes of a person wearing glasses in an image sequence that contains free 3D head movements. Note, that the appearance changes are due both to the difference in the appearance of the corner of the eye when viewed from different viewpoints, but also due to the relative displacement of the skeleton of the glasses that the person is wearing. While the first source of appearance variation could be explicitly modelled by a projective transformation (whose coefficients need to be estimated), the changes that are due to the relative displacement of the skeleton are more difficult to model explicitly since they depend on the relative distance of the glasses with respect to the face. Finally, non rigid appearance changes are also present and are due to unintentional blinking of the subject. In fig.2 we present some frames of the image sequence for each of the three trackers. It is clear that the proposed method performs well under all sources of appearance change and is able to track reliably the eye corners throughout the image sequence. In contrast the observation functions that are based on the median filter and on the sum of the absolute differences loose the tracking as soon as the subject blinks and although they occasionally recover, they loose regularly the track of the eyes. All three tracking schemes have been initialised at the same position and all utilise the same colour space.

Finally, in order to better illustrate the performance of the proposed method, we present results from two more image sequences, containing appearance changes that are due to scaling, 3D head rotation as well as non rigid motion of the corner of the mouth and of the eye due to blinking. In both image sequences the trackers that are based on the mean and the median absolute differences loose the track after some frames and although they occasionally recover, they are not able to track reliably the facial features for the whole image sequence. Note that the image sequence 'Emile' contains significant changes in the apparent lighting conditions due to the automatic gain control of the cam-



(a) Deformable template tracking: Frames 1, 30, 60, 90, 130, 150, 190

(b) Median-based template tracking: Frames 30, 60, 120

(c) MeanAbsoluteDifference-based template tracking: Frames 30, 60, 120

Figure 2: 'Erik' image Sequence

era and a total occlusion of the facial features for around 10 frames.

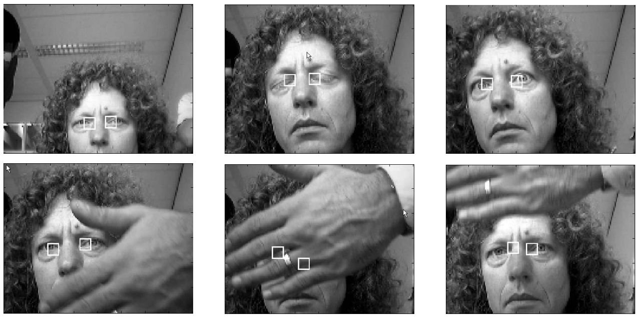


Figure 3: 'Emile' image sequence

4. Conclusions

We have presented a novel observation model for tracking non rigid objects in changing lighting conditions. The proposed observation model is robust to moderate changes in the illumination and to moderate deformations of the tracked object. We have integrated the observation model in the particle filtering framework and we were able to robustly track facial features and other objects under moderate appearance changes that are due to facial expressions and 3D motion. Although our approach did not explicitly model geometrical transformations due to rigid 3D motion, our explicit modelling of local deformations proved robust in dealing with in plane and out of plane object rotations and scaling. We have experimentally demonstrated the robustness of the proposed method in a number of image sequences and have provided comparative results with two other popular appearance models. Finally, we have devised a colour-based scheme that is invariant to changes in the intensity of the illumination and experimentally proved to be robust to the camera gain control in indoor environments. For future work we intend to use the proposed observation model in an online scheme for learning object appearances for object recognition.

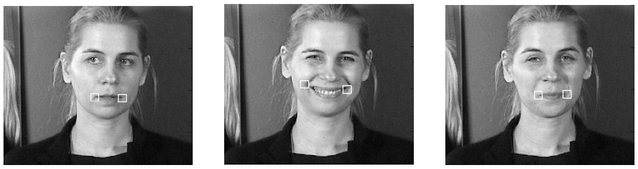


Figure 4: 'Maja' image sequence

Acknowledgements

The work of I. Patras is supported by the Netherlands BSIK-MultimediaN-N2 Interaction project. The work of M. Pantic is supported by the Netherlands Organization for Scientific Research Grant EW-639.021.202.

References

[1] Thomas F. El-Maraghi Allan D. Jepson, David J. Fleet. Robust online appearance models for visual tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(10):1296–1311, Oct. 2003.

[2] S. Avidan. Support vector tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Dec. 2001. Hawaii.

[3] M.J. Black and A.D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. In *European Conference on Computer Vision*, pages 329–342, 1996.

[4] D. W. Hansen, M. Nielsen, J. P. Hansen, A. S. Johansen, and M. B. Stegmann. Tracking eyes using shape and appearance. In *IAPR Workshop on Machine Vision Applications - MVA*, pages 201–204, dec 2002.

[5] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *Int'l Journal of Computer Vision*, 29(1):5–28, 1998.

[6] M. Isard and A. Blake. Icondensation: Unifying low-level and high-level tracking in a stochastic framework. In *European Conf. on Computer Vision*, pages 893–908, Jun. 1998. Freiburg, Germany.

[7] J. MacCormick and A. Blake. Probabilistic exclusion and partitioned sampling for multiple object tracking. *Intl J. Computer Vision*, 39(1):57–71, 2000.

[8] Iain Matthews, Takahiro Ishikawa, and Simon Baker. The template update problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):810–815, Jun. 2004.

[9] K. Murphy and S. Russell. *Rao-Blackwellised Particle Filtering for Dynamic Bayesian Networks*. Springer-Verlag, 2001.

[10] Hieu Tat Nguyen and Arnold W. M. Smeulders. Fast occluded object tracking by a robust appearance filter. *IEEE trans. PAMI*, 26(8):1099–1104, 2004.

[11] Constantine Papageorgiou and Tomaso Poggio. A trainable system for object detection. *International Journal of Computer Vision*, 38(1):15–33, 2000.

[12] I. Patras and M. Pantic. Particle filtering with factorized likelihoods for tracking facial features. In *Int'l Conf. Face and Gesture Recognition*, May 2004. Seoul, South Korea.

[13] P. Perez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *European Conf. on Computer Vision*, volume 1, pages 661–675, May 2002. Copenhagen, Denmark.

[14] M.K. Pitt and N. Shephard. Filtering via simulation: auxiliary particle filtering. *J. American Statistical Association*, 94:590–, 1999.

[15] M.F. Valstar, I. Patras, and M. Pantic. Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data. In *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR'05), Workshop on Vision for Human-Computer Interaction*, June 2005. San Diego, USA.

[16] Y. Wu, G. Hua, and T. Yu. Tracking articulated body by dynamic markov network. In *Int'l Conf. on Computer Vision*, Oct. 2003. Nice, France.