TEMPLATE TRACKING WITH OBSERVATION RELEVANCE DETERMINATION

Ioannis Patras *

Queen Mary, University of London
Department of Elec. Engineering
I.Patras@elec.qmul.ac.uk

Edwin Hancock

The University of York
Department of Computer Science
erh@cs.york.ac.uk

ABSTRACT

This paper addresses the problem of template tracking in the presence of occlusions, clutter and rapid motion. We adopt a learning approach, using a Bayesian Mixture of Experts (BME), in which observations at each frame yield direct predictions of the state (e.g. position / scale) of the tracked target. In contrast to other methods in the literature, we explicitly address the problem that the prediction accuracy can deteriorate drastically for observations that are not similar to the ones in the training set; such observations are common in case of partial occlusions or of fast motion. To do so, we couple the BME with a probabilistic kernel-based classifier which, when trained, can determine the probability that a new/unseen observation can accurately predict the state of the target (the 'relevance' of the observation in question). In addition, in the particle filtering framework, we derive a recursive scheme for maintaining an approximation of the posterior probability of the target's state in which the probabilistic predictions of multiple observations are moderated by their corresponding relevance. We apply the algorithm in the problem of 2D template tracking and demonstrate that the proposed scheme outperforms classical methods for discriminative tracking in case of motions large in magnitude and of partial occlusions.

Index Terms— Template Tracking, Occlusion Handling, Motion Estimation, Discriminative Tracking.

1. INTRODUCTION

In order to reduce the computational complexity of generative methods for visual tracking and detection (*e.g.* [3][2]) a number of methods have emerged in the so called discriminative tracking framework [1] [6] [5] [7]. In this framework, a model of the posterior $p(x|y)$ is learned from annotated or artificially generated training data, so that an observation y can deliver directly a prediction of the unknown state x. This is a major improvement over generative methods that require evaluation of the likelihood $p(x|y)$ for a large number of candidate states x.

In the recent years, discriminative methods have been used with great success for a number of problems, including 3D human pose estimation [1] and 2D template tracking [6][5]. However, none of these methods addresses explicitly the problem of assessing in advance how well an observation y can predict the state x nor do they use multiple observations in order to increase robustness. Regression-based methods are known to be sensitive to observations that do not belong to the space that is sampled by the training dataset. Therefore the accuracy of the prediction of the posterior $p(x|y)$ deteriorates sharply for observations y that come from areas that are uninformative of the state of the visual target such as occlusions and the background. Such observations are very likely to occur when the motion is large in maginitude ([5] and [6]).

In this paper, for 2D visual tracking, we extend the discriminative/regression tracking framework ([7]) in two ways:

- We explicitly address the problem of the determination of the relevance of an observation to the state estimation by learning in a supervised way the underlying conditional probability distribution.
- We explicitly devise a probabilistic framework that allows multiple observations to contribute to the prediction of the state of the target according to their corresponding relevance.

In this way, the contribution of the predictions that come from relevant observations is high, while observations that come from occluded areas or observations that can not give good predictions are largely ignored. We propose an extension of the discriminative filtering framework by introducing additional binary random variables z that are related to the observations' relevance. We use Relevance Vector Machines [8] in order to learn the conditional probability $p(z = 1|y)$ (the probability that the observation y is relevant). A Bayesian Mixture of Experts [9] is used for modelling ($p(x|y)$) (the posterior probability of the state x given an observation y).

The remainder of the paper is organised as follows. In Section 2 we provide an outline of the proposed discriminative tracking framework with data relevance determination. In Section 2.1 we briefly describe the Bayesian Mixture of Experts framework and in Section 2.2 we present our method for data relevance determination. In Section 3 we present experimental results for 2D target tracking and in Section 4 we present some conclusions.

2. REGRESSION-BASED TRACKING WITH RELEVANCE DETERMINATION

Filtering, such as Kalman filtering or particle filtering, has been the dominant framework for recursive estimation of the conditional probability of the unknown state x given a set of observed random variables $Y = \{.....y^-, y\}$ up to the current time instant. In the discriminative filtering framework (Fig. 1(a)) the filtered density can be derived as [7]:

$$p(x|Y) = \int_{x^-} p(x^-|Y^-)p(x|x^-, y). \quad (1)$$

However, this derivation ignores the fact that for certain problems different parts of the observation y can give different predictions of the state of the target. In [6], for 2D tracking where the evidence y is an image frame, the prediction of the state of the target (*e.g.* 2D location) is based on the data $y(r)$ in a single window, which (in the absence of a motion model) is centred around the estimated

*The author performed most of this work at the University of York, UK

position $r = \hat{x}^-$ of the target in the previous frame . This disregards the information that is present at other positions r. Similarly, for 3D tracking, in [1] [7] a single feature vector is extracted from the object silhouette. On the other hand, in the generative particle filtering framework for 2D tracking, it is common practice that several parts of the observation are examined. This is achieved by using multiple samples (particles) r and modelling the likelihood that is used to evaluate the importance of each particle as $p(y|r) = p(y|y(r))$. The particles r are sampled using the transition probability $p(x|x^-)$ and, in the simplest case, a number of measurements $y(r)$ around the positions of the particles in the previous frame are utilised.

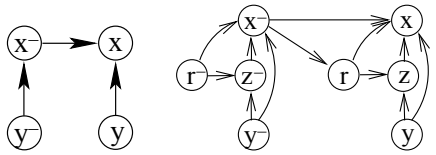


Fig. 1. Graphical models for (a) classical discriminative tracking and (b) for regression tracking with relevance determination

Here, we propose a discriminative particle filtering method that utilises the fact that several parts of the observation can yield predictions of the state of the target. We do so by introducing a random variable r that determines which parts (or in general how) the observation y will be used. In the simplest case, r will determine the positions in the current frame of image patches that will yield predictions of the position of the target. In general, r will be used for obtaining a set of candidate observations $y(r)$. In this work we condition r on x^- as we expect that the previous state can sufficiently informative on how candidate observations can be obtained. Subsequently, we introduce a binary variable z and denote with $p(z = 1|y, r)$ the probability that the observation $y(r)$ is relevant for the prediction of the unknown state x. The dependencies of the variables are depicted in Fig. 1(b) where y is observed and the rest is unknown.

In this network the filtered density can be derived as:

$$p(x|Y) = \int_{x^-} p(x^-|Y^-) \left(\int_r p(r|x^-) \int_z p(x|z, x^-, y, r) p(z|y, r) \right) \tag{2}$$

In order to deal with posteriors with multiple modes and to recover from tracking failures we maintain an approximation of the $p(x|Y)$ using a mixture of M Gaussians. In Table 1 we summarise our modelling choices. We assume that the probability distributions in Table 1 are either given or learned in the training phase (as explained in Sections 2.1 and 2.2). For example, in the testing phase, given a triple (y, r, x^-), the trained BME yields a mixture of Gaussians that is our approximation of $p(x|z = 1, x^-, y, r)$.

We will subsequently describe a computational scheme that, given an approximation of the posterior $p(x^-|Y^-)$ of the state at the previous frame, yields an approximation of the state posterior $p(x|Y)$ at the current frame. This is achieved by the following procedure:

1. Sample a state x^- from $p(x^-|Y^-)$.
2. Sample r from $p(r|x^-)p(x^-|Y^-)$ by sampling r from $p(r|x^-)$ for each of the state samples x^- obtained in step 1. Let us assume that R samples are obtained this way.
3. For each of the R samples r
   (a) Evaluate the relevance of the observation $y(r)$ as $\alpha = p(z = 1|y, r)$
   (b) Given (x^-, y, r) use the trained BME to obtain a probabilistic prediction of the state x, given that $y(r)$ is relevant (*i.e.* $p(x|z = 1, x^-, y, r)$) as a mixture of M Gaussians. Model the probabilistic prediction given that $y(r)$ is irrelevant (*i.e.* $p(x|z = 0, x^-, y, r)$) as a single Gaussian with a large variance. This leads to a mixture of $K + 1$ Gaussians, that is:

$$\int_z p(x|z, x^-, y, r) p(z|y, r) = \alpha \sum_{i=1}^{K} g_i \mathcal{N}\left(\mu_i + r, S_i\right) + (1-\alpha)\mathcal{N}\left(x^-, S_0\right) \tag{3}$$

4. Approximate the resulting mixture of L Gaussians ($L = R(K+1)$) with a mixture of M Gaussians.

$p(r\|x^-)$	Uniform around x^-.
$p(z = 1\|y, r)$	Probability that the observation $y(r)$ is relevant. It can be evaluated using Eq. 5.
$p(x\|z = 1, x^-, y, r)$	Probabilistic prediction of the target state given that the observation $y(r)$ is relevant ($z = 1$). Modelled as $\sum_{i=1}^{K} g_i \mathcal{N}(r + w_i^T y(r), S_i)$ (Eq. 4)
$p(x\|z = 0, x^-, y, r)$	Probabilistic prediction of the target state given that the observation $y(r)$ is not relevant ($z = 0$). Modelled as $\mathcal{N}(x^-, S_0)$

Table 1. Modelling choices.

2.1. Bayesian mixture of experts for regression

In what follows we will describe a method that, given an observation $y(r)$ and the target state at the previous frame x^-, yields a probabilistic prediction of the state x at the current frame. For notational simplicity, let us ommit x^- in the subsequent derivations. Following the work of Sminchisescu *et al*. [7] we use for regression a Bayesian Mixtures of Experts. The rationale behind our choice, over other regression methods (*e.g.* RVMs [8]) is that the BME can model well predictive distributions that are multimodal. Such distributions arise often in the case of 3D tracking due to for example front/back and left/right ambiguity [7][1] and are also expected to arise in the case of 2D tracking due to the aperture problem. The (Hierarchical) Mixtures of Experts [4] is a method for regression and classification that relies on soft probabilistic partitioning of the input space (that is y). This is determined by gating coefficients $g_i(y; \xi_i)$ one for each expert i that are input dependent and parametrised by an unknown vector ξ_i. The coefficients have a probabilistic interpretation, that is the coefficients sum up to one. The prediction of each expert i is then moderated by the corresponding gating coefficient. After training [9], for a new observation y the predictive distribution is a mixture of Gaussians $\mathcal{N}()$ given by:

$$\hat{p}(x|y) = \sum_i g_i(y; \xi_i) \mathcal{N}\left(w_i^T y, S_i\right), \tag{4}$$

where ξ_i, w_i and S_i are the unknowns that are learned.

2.2. Relevance determination

For the determination of the relevance $p(z|y, r)$ of an observation $y(r)$ we use a classification scheme with the Relevance Vector Machines (RVM). The goal is to obtain an *a priori* assessment of whether

the probabilistic prediction $\hat{p}(x|y(r))$ (Eq.4) of the state of the target is expected to be good. To this end, we train an RVM classifier in a supervised way with a set of positive examples that yield good predictions and with a set of negative examples which yield bad predictions. Let us denote with $sigm$ the sigmoid function, with $\{\tilde{y}_i\}$ the training set of the classifier and with $\phi(y_i, y_j)$ a kernel function.

After training and when presented with a new observation $y(r)$, the RMV yields a prediction of its relevance as

$$p(z = 1|y(r)) = sigm\left(\sum_i w_i^{rvm}\phi(y(r), \tilde{y}_i)\right), \qquad (5)$$

where w^{rvm} is a sparse weight vector that is learned during training.

The training set $\{y(r)\}$ is constructed as follows. A candidate observation $y(r)$ is generated by artificially transforming (*e.g.* translating) the visual target with a transformation which we denote here with r. Then, for each of the candidate observations, a probabilistic prediction is made using Eq. 4. We put in the set of positive examples candidate observations for which, an appropriate norm of the difference between the true transformation r and the mean of the prediction $\hat{p}(x|y(r))$ is less than a threshold. That is,

$$\|r - E_x\left[\hat{p}(x|y(r))\right]\| < \theta_r \qquad (6)$$

As $\hat{p}(x|y(r))$ is a Gaussian mixture, the mean in Eq. 6 is obtained in closed form. Alternative schemes for constructing the positive training set, such as thresholding the distance between the true transformation r and the mode of $\hat{p}(x|y(r))$, or by thresholding the probability of the ground truth transformation r (*i.e.* $\hat{p}(r|y(r)) > \theta_r$) are also possible. The set of the negative examples comprises the observations for which Eq. 6 is not satisfied. Other examples, such as observations from background regions could also be added in the negative training set. Clearly, the transformations r that generate the candidate training set need to explore larger parts of the state space than the ones that are used to construct the training set of the BME.

3. EXPERIMENTAL RESULTS

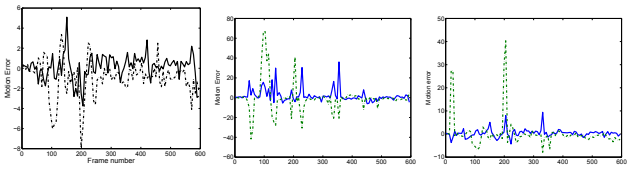
We have performed a number of experiments in order to illustrate the performance of the proposed method under different conditions, including occlusions, fast motion and moderate deformations. Here, we present results for sequences that are manually annotated. In addition, we compare our algorithm to discriminative tracking when a single observation is used (*e.g.* [7][6]) and to the degenerate version of the proposed algorithm in which the data relevance determination mechanism is not used. We do not use any dynamic model, or temporal filtering in order to judge the performance when large deviations from the motion model are present. That is we drop the dependency of $p(x|z, x^-, y, r)$ on x^-.

For training the BME we used pairs $(y(x), x)$ in which the observations $y(x)$ are produced by artificially transforming (*e.g.* translating) the visual target with the transformation x. We used translational transformations of up to 11 pixels, that is, transformations that generate observations $y(x)$ that had some overlap with the target. The examples that were used for training the RVM were generated using transformations with range 2-3 times the range that was used for training the BME. In order to deal with changes in the illumination intensity we normalise the data by the average intensity in the window in question. Unless stated otherwise, we track 5 Gaussians (*i.e.* $M = 5$) and use 25 samples r (*i.e.* $R = 25$).

We first present results for tracking a facial feature (*i.e.* an eye corner) under changes in the illumination, large head motion and deformations due to facial expressions and head rotations. We have used 600 frames which are annotated every 6 frames. We have experimented by down-sampling the image-sequence spatially (by a factor $DSS = 1, 2$) and temporally ($DST = 1, 2, 3$) in order to create sequences with different motion magnitudes. In all cases we track an 11×11 window. In the leftmost image of Fig. 2 we depict with a white rectangle the window that we used to track the eye corner at $DSS = 1$ (obviously when $DSS = 2$ the 11×11 window will cover a larger area).

Fig. 2. Tracking results for frames 75, 119 and 357 of the 'Head' sequence ($DSS = 1$ and $DST = 2$).

In Fig. 2 we present tracking results for the 'Head' sequence for some characteristic frames. The tracking is consistently good throughout the image sequence even at the presence of large motion (equal to the window size), occlusion and some deformations. In Fig. 3(a) we depict the horizontal and vertical components of the error in pixels. In order to illustrate the benefits of using multiple observations and the benefit of data relevance determination we present here comparatively results with two degenerate cases of our algorithm. The first (ALG1) is similar to classical regression-based tracking methods [6][1] that use a single observation. The second (ALG2) is a degenerate version of our algorithm in which the relevance determination is not used, that is, the probabilistic predictions of all candidate observation $y(r)$ are used. For both algorithms, we reinitialise the tracking at the ground truth position when the error is larger than 20 pixels (almost twice the window size).



(a) (b) (c)

Fig. 3. Motion error in pixels for (a) the proposed method (RMS = 1.7), (b) for ALG1 (RMS = 12) and (c) for ALG2 (RMS = 4.8) for the 'Head' sequence. Note the difference in the axes scale.

In Fig. 3(b) and Fig. 3(b) we depict the estimation error for each of the motion components for the ALG1 and ALG2. It can be observed that both ALG1 and ALG2 fail at large motion magnitudes as (some) observations come from areas far from the visual target. Both are also sensitive to occlusions (around frame 350) but the multiple observation algorithm automatically recovers as the target is occluded only for a couple of frames. In Table 2 we summarise the RMS error for a number of different spatial and temporal subsamplings of the original image sequence. Note the fact that after frame 230 there is practically no motion (the sequence contains some facial expressions and closing/opening the eyes) which makes less prominent the differences in the performance.

Similar differences in performance have been observed for a number of image sequences. Here we present results for sequences that we used to test the performance under large and persistent occlusions. First, we created an image sequence depicting a moving rigid

Algorithm vs Parameters	Proposed	ALG1	ALG2
DSS = 1, DST = 2	1.7	12	4.8
DSS = 1, DST = 3	4.8	12.1	5.8
DSS = 2, DST = 6	2.12	6.5	7.1

Table 2. RMS errors for the 'Head' sequence

object. In this sequence we manually annotated the position of the target every 6 frames and subsequently created image sequences in which up to half the visual target was occluded. In Fig. 4 we present some frames of the sequences that depict the occluded target and the estimated position of the target.

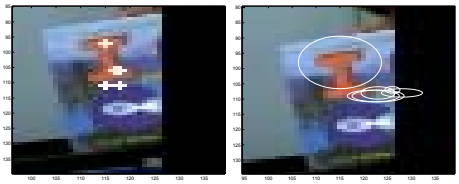
Fig. 4. Tracking results for frames 1, 55 and 302 of the 'CD cover' sequence ($DSS = 4$ and $DST = 6$). A quarter (QRT) of the target is artificially occluded.

In Table 3 we summarise the results by reporting the RMS error for a number of subsamplings and for different occlusions of the target. In the first row, the target is completely visible, in the second a quarter of the target is occluded and in the last row half of the target is occluded. The target is occluded at the frames for which there is available annotation, that is every 6 frames. This means that in the experiment in the last row of Table 3 the target is completely visible in half of the frames. A larger number of candidate observations are used here ($R = 50$). It is clear that the method is capable of tracking under partial occlussions and that it clearly outperforms the method that uses a single observation. For the latter, we used a more realistic re-initialisation scheme that is initiated when the true error is larger than 10 pixels (that is, almost equal to the template size). The last column indicates how much the often a validation scheme as the one proposed in [6] would fail. Note, that when a large part of the target is occluded a validation scheme is more likely to fail even when the prediction is accurate. In this case a full scale detection scheme [6] is also likely to fail.

Algorithm vs Parameters	Proposed	ALG1	ALG1 fails
DSS = 4, DST = 6	3.2	11.2	5 %
DSS = 4, DST = 6 (QRT)	4.2	17.8	38 %
DSS = 4, DST = 6 (HLF)	11.1	17.5	78 %
DSS = 4, DST = 3 (HLF)	2.9	19.2	28 %

Table 3. RMS errors for the 'CD Cover' sequence

Finally, in Fig. 5, we illustrate the ability of the algorithm to overcome large occlusions. In this case, observations that are located at areas neighbouring to the true target position are used to deliver reliable predictions of the target state. In Fig. 5(a) we depict the relevant observations and in Fig. 5(b) the corresponding probabilistic predictions (each ellipse represents a Gaussian). Note that our relevance determination scheme suppressed observations that were on the true target location, a result that indicates that a validation scheme using the trained RVM classifier would also fail.



(a) (b)

Fig. 5. Tracking results for the last frame of the 'CD cover' sequence ($DSS = 4$ and $DST = 6$). Half of the target is artificially occluded. (a) Relevant observations (b) Probabilistic predictions.

4. CONCLUSIONS

In this paper we have presented a method for efficient and robust visual tracking. We propose a discriminative framework in which multiple observations provide predictions of the state of the target. Each prediction is moderated by the relevance of the corresponding observation, as this is determined by a probabilistic classification scheme. To the best of our knowledge this is the first work that utilises multiple observations for discriminative tracking or uses a classification scheme to access in advance the relevance of an observation (as opposed to *a posteriori* validation of the prediction). We have illustrated the efficiency of our approach in a number of image sequences for the problem of 2D tracking. For future work we intend to extend the proposed scheme for tracking 3D human pose under occlusions and background clutter.

5. REFERENCES

[1] A. Agarwal and B. Triggs. Recovering 3d human pose from monocular images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(1), Jan. 2006.

[2] J. Deutscher, A. Davison, and I. Reid. Automatic partitioning of high dimensional search spaces associated with articulated body motion capture. In *Int' Conf. Computer Vision and Pattern Recognition*, Dec. 2001. Kauai, Hawaii.

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

[4] M. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 1994.

[5] F. Jurie and M. Dhome. Hyperplane approximation for template matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002.

[6] O.Williams, A.Blake, and R.Cipolla. Sparse bayesian regression for efficient visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (8):1292–1304, Aug 2005.

[7] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative density propagation for 3d human motion estimation. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[8] M. Tipping. The relevance vector machine. In *Advances in Neural Information Processing Systems, San Mateo, CA*. Morgan Kaufmann, 2000.

[9] S. Waterhouse, D. MacKay, and T. Robinson. Bayesian methods for mixtures of experts. In *Advances in Neural Information Processing Systems*, volume 8, pages 351–357, 1996.