Dynamic Classifier for Non-rigid Human Motion Analysis

Huang Fei, Ian Reid
Department of Engineering Science,
University of Oxford, Oxford,UK.
fei@robots.ox.ac.uk

Abstract

Automatic analysis (parsing) of non-rigid human motion in a cluttered outdoor enviroment is a useful but challenging task. In a single view point, the lack of depth order relations causes a major ambiguity of the object identities. Coupled with the non-rigidity of articulation, 3D human motion tracking/pose estimation in one view is a formidable problem. In this paper, we present a novel solution that directly address this depth ambiguity, in which we extend a discriminative analysis (Support Vector Machine (SVM)) to non-rigid human motion classification with a temporal generative motion model (Hidden Markov Model (HMM)). This method can discriminate dynamic depth ordering as well as 3D articulated motion automatically from 2D images. Experiments with this method have demonstrated promising results.

1 Introduction

Automatic interpretation (parsing) of non-rigid human motion is essential to motion estimation, gait pattern analysis and behaviour understanding. A number of motion parsing algorithms have been put forward in related computer vision research areas: object detection [5],[6], visual tracking [7] and behaviour understanding [8]. In these fields, machine learning methods have received increasing attention in recent years. Discriminative-learning based approaches such as SVM or AdaBoost have achieved promising results in face [13] or people [6] detection; Sophisticated motion inference algorithms like Particle filtering [11], Meanshift [12], etc. have demonstrated successes in solving visual tracking difficulties; HMM-based stochastic methods [8] dominate the sign-language and gesture recognition community. All these advances seem to promise a fully-fledged human motion interpretation system in the near future.

Nevertheless, there still remain significant challenges. Take two similar frames from a video sequence (Figure 1) as an example, in the cluttered outdoor enviroment, without utilizing a kinematic model and proper initialization, few computer algorithms can easily retrieve the human figure structure from the image, let alone discern some minute but useful information. For instance, one might be interested to know: in addition to person's position, which leg is in front of the other at a particular time t? In a single view, depth information is lost when silhouettes [1] are extracted. Edge-detection might give some useful clues, but noise causes trouble when utilizing such information.

BMVC 2004 doi:10.5244/C.18.33

Figure 1: Lack of depth information requires additional discriminative ability when parsing two similar frames of human motion in a single view.

Two broad solutions to still image classification present themselves: discriminative methods such as SVM, generative methods such as factor analysis. Even so, at a single time-instant it remains an error-prone task. We propose a dynamic classifier which combines discriminative model (SVM) with a generative dynamic model (HMM), in order to ameliorate this fundamental problem.

2 Previous Work

Generative methods ususally dominate the visual motion analysis. On-line EM algorithm has been proposed for tracking [24],[23], in which case a generative model is utilized to update the appearance tracker. However they are limited to relatively rigid-bodies such as the face. To cope with the high non-rigidity of human body/hand articulation, the HMM filter is usually a better choice. Although having been widely used in sign-language analysis [8] and gesture recognition, gait dynamics classification [14], [15]. the power of the dynamic Markov model as a tracker to analyze non-rigid human motion was not addressed until Toyama and Blake's "Metric Mixture Tracker" [7]. Nevertheless in [7], two independent dynamic processes share one observation density provided by chamfer distance, thus limiting the non-rigid shape inference power considerably.

Though HMM based generative methods such as [7] and [16] have demonstrated usefulness in handling incomplete information (often caused by occlusion clutter), they have several obvious constraints:

(1). Inter-class/intra-class variability. An ideal motion classification or parsing mechanism should maximize *inter-class* variation while minimizing *intra-class* dissimilarity. In a HMM-based visual tracker [7] and [16], the statistical learning procedures usually require a large amount of training data (often not available in practice) to approximate the underlying motion dynamics. In addition, very similar shape templates (either acquired from chamfer matching or silhouette descriptors) often bear different class labels while different templates share the same class labels, a deficiency worsened by the *nearest-neighbour* like unsupervised learning algorithm. Therefore, efficient learning of the optimal motion dynamics is at best difficult and occasionally unfeasible. A tree-structure has been proposed in [17] as a plausible alternative, but only to increase the exemplar number without comparable improvement on the capacity to discern similar exemplars.

(2). Inadequate discriminative ability. As suggested previously, despite the probabilistic inference power in the temporal domain [7][16], previous applications of generative models are limited in discriminating object representations, and usually can not

explain the depth ordering information (Figure 1). Meanwhile, such discriminative ability is highly desirable; for it can not only provide a step towards automated 3D human motion capture in a single view, but also can reduce the amount of training data, and increase robustness against outliers in the motion data.

These inherent limitations of the generative model are usually well addressed by discriminative analysis methods. Though often seen in static image processing tasks such as object/non-object recognition [5],[6], in recent years, large-margin classifiers such as SVM (RVM) have been applied to motion analysis [18], [19] or human pose regression [1]. In [18], Avidan builds a vehicle/non-vehicle classifier into an optical flow based car tracker. Williams et.al. [19] use a probabilistic version (RVM) of SVM to classify linear Euclidean transformations. However these early attempts have been limited to rigid motion, how to parse non-rigid human motion using SVM is an interesting but difficult problem;this is because the theory of structural risk minimization (SRM) [3], on which standard SVM is based, is formulated for independent, identically distributed (*i.i.d.*) data while articulated human motion is highly correlated in the spatial-temporal domain. Agarwal et.al. [1] use RVM regression to estimate 3D human poses from silhouettes, however the lack depth relations and no temporal information are the major disavantages. In the following sections, we demonstrate that the integration of both a large-margin classifiers (SVM) and a dynamic models (HMM) into a *Dynamic Kernel Machine*, together with informative appearance model provide a plausible solution.

3 Dynamic Kernel Machine

3.1 HMM

In a hidden Markov Model [2], the sequence of observations $\{O_1, O_2, ..., O_T\}$ is modeled by assuming that each observation $\{O_t\}$ $(1 \leq t \leq T)$ depends on a discrete hidden state variable $\{S_t\}$. The joint probability for the sequence of states $\{S_t\}$ and observations $\{O_t\}$ given a particular HMM model, can be summarized as,

$$P(S_t, O_t) = P(S_1) \prod_{t=2}^{T} P(S_t|S_{t-1}) \prod_{t=1}^{T} P(O_t|S_t). \quad (1)$$

In a non-rigid appearance tracker, the current **maximum a posteriori (MAP)** estimate $P(S_t|O_t) \propto P(S_t, O_t)$. $P(O_t|S_t)$ is the observation density measure, which is usually weak and susceptible to noise and occlusion clutter. $P(S_t|S_{t-1})$ is the dynamic motion model, which is typically learned from training data. In contrast to traditional dynamic models (such as the AR process in a Kalman filter or Particle filter-based tracker), non-rigid motion dynamics $P(S_t|S_{t-1})$ take account of the entire appearance history, consequently it is a stonger factor than $P(O_t|S_t)$ in Eq. 1, which can enable the motion estimator $P(S_t, O_t)$ to track non-rigid human motion efficiently and handle occlusion clutter robustly [16].

3.2 SVM

The Support Vector Machine [3] is usually used as a binary classifier. The basic form of SVM which classifies an input vector $x \in R^n$ can be expressed as,

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i \phi(x_i) \cdot \phi(x) + b = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b \tag{2}$$

where ϕ is a non-linear mapping function $\phi(x)$: $R^n \rightarrow R^m$, (n $\ll m$). "$\cdot$" refers to the inner product operator, x_i, y_i and a_i are the i^{th} training sample, its class label, and its Lagrange multiplier, respectively. $K(\cdot,\cdot)$ is a symmetric positive-definite kernel function, and b is a bias term. The sign of $f(x)$ indicates class membership.

3.3 Extending SVM to probabilistic multi-class classifier

A common way [1] of extending binary classifiers to multi-class (K) classifier is by pairwise coupling (i.e. one against one). It seeks to construct independently a two-class decision boundary between every pair of the total number of classes. This can avoid introducing undesirable negative examples for motion classification as required by classifying one class against the rest. A binary classifier decides whether a point $x \equiv \hat{O}_t$ belongs to class S_i or S_j $(1 \leq i, j \leq K)$. The probability of $\hat{O}_t$ that belongs to class S_i given that $\hat{O}_t$ is in either class S_i or S_j can be written as p_{ij}=$P(\hat{O}_t \in S_i | \hat{O}_t \in S_i \cup S_j)$. With this p_{ij}, the probability estimate that $\hat{O}_t$ belongs to class i, P_i ($P_i \equiv P(\hat{O}_t \in S_i)$) is determined by using a matrix of $p_{ij}(p_{ij} > p_{ji})$. For multiple-classes and probabilistic classification, a standard 'voting' scheme is used [9]: We construct $K(K-1)/2$ binary SVMs independently to predict whether the 'winner' is class S_i or S_j. Optimal K class Bayesian maximization decision selects the class with most winning two-class decisions $\{p_{ij}\}$.

$$P_i = \frac{2}{K(K-1)} \sum_{j: i \neq j, p_{ij} > p_{ji}} p_{ij} \tag{3}$$

where the classification result is given by $\arg\max_{1 \leq i \leq K} P_i$. For each binary classifier, conversion from decision value to posterior probability output is given by J.Platt's sigmoid function fitting [22].

$$p_{ij} = P(\hat{O}_t \in S_i | O_t, \hat{O}_t \in S_i \cup S_j) = \frac{1}{1 + e^{Af(x)+B}} \tag{4}$$

Where A, B are free parameters which are needed to be estimated from the training data by cross validation, and $f(x)$ is the decision value of SVM output.

To demonstrate the effectiveness of multi-class probabilitic SVM classifiers, we perform classification test on some synthetic data. Here six component gaussian mixture data is generated with given mean values within $[0,1]$ and uniform variance $\delta = 0.03$ in 3D space. The classification accuracy of standard K-means clustering algorithm on the test data is 88.61%, six-class SVM classifier is 88.70%, while probablitic six-class SVM

[1] Tipping introduces a variation called RVM, which models the weighting of support vectors as a gaussian process but still use the same logistic sigmoid function to generate probability as in this paper. Since no significant benefits in classification rates has been reported compared to SVM, we leave the extension using RVM to future research. Nevertheless, the DKM framework in this paper is general to handle both extensions.

classifier results 88.65%. This suggests that in certain circumstances (e.g. when the data distributions are overlapping, as is the usual case for non-rigid human motion analysis), the probabilitic multi-class SVM classifiers can achieve comparable classification results as standard clustering algorithms. We deem this as an important prequisite, because when augmented with strong temporal motion prior (HMM), the multi-class probabilistic SVM classifier has the potential significant performance improvement advantages over its static counterparts.

3.4 Augmenting temporal motion prior

Here we address issues of combining temporal motion prior (HMM) with discriminant analysis (SVM). We exploit the probabilitic discriminants from the multi-class SVM as the observation process of HMM, where the observation density $P(O_t|S_t)$ is adapted from the posterior estimate P_i as in Eq.3 (strong discriminative classification), when combined with the dynamic motion prior, it gives a MAP state estimate. For clarity, we summarize the notation of the proposed DKM here: For a finite K number of states S^i $(1 \leq i \leq K)$ in DKM, we construct a $K(K-1)/2$ SVM binary classifiers, as described in the last section. We deduce probabilitic discriminant for the traditional observation likelihood $P(O_t|S_t^i)$, where

$$P(O_t|S_t^i) \equiv P(\hat{O}_t \in S_i) = \frac{2}{K(K-1)} \sum_{j:i\neq j, p_{ij}>p_{ji}} p_{ij}, \tag{5}$$

p_{ij} is defined in Eq. 4. The MAP state estimate of DKM thus be,

$$P(S_t|O_t) \propto P(S_t, O_t) = P(S_1)\prod_{t=2}^{T} P(S_t|S_{t-1})\prod_{t=1}^{T} P(O_t|S_t) \tag{6}$$

$$= \sum_{i=1}^{K} P(S_1)\prod_{t=2}^{T} P(S_t|S_{t-1})\prod_{t=1}^{T} P(O_t|S_t^i). \tag{7}$$

4 Parsing Dynamic Human Motion

Traditional human motion analysis has often been considered as a generative process in the spatial-temporal domain. Inference at a particular time t is driven by the previous states $P(S_{1:t-1})$ and the current observation $P(O_t|S_t)$. Like [7] [16], we adopt the appearance exemplar representations as a simplification of the probabilistic mixture component model, together with the DKM algorithm, it resolves the depth ambiguity problem as highlighted in Figure 1.

4.1 Learning structural exemplars

Although learning probabilistic dependency of body parts [20], [21] is a popular choice, in a typical pedestrian walking scenario (Figure 1), such a dependency is relatively strong that we can treat it as if *"coincidental"* (for example, the right arm tends to be in the front when left leg is in the front). Hence the latent state S of non-rigid human motion is represented by the quadruple $\{S_{LF}, S_{RF}, S_{LT}, S_{RT}\}$, where $S_{LF/RF} \in \{0,1\}$ are indicators denoting if left/right foot is off/on the ground. $S_{LT/RT} \in [-\Theta_{max}, +\Theta_{max}]$, $\Theta_{max} \in (0', 90')$, which refer to the angle between the left or right thigh with the vertical axis passing

through the centroid, depth ordering information is determined by the sign of Θ. Thus we obtain a sufficient set of six exemplars encoding distinctive non-rigid motion states (Figure 2).

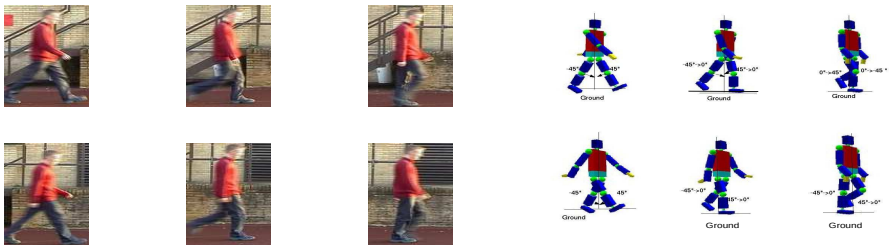

Figure 2: A minimum set of six exemplars which can sufficiently characterize human walking and depth ordering. Six 3D human models are rendered to represent the six exemplars, manifesting the depth relations and structure information.

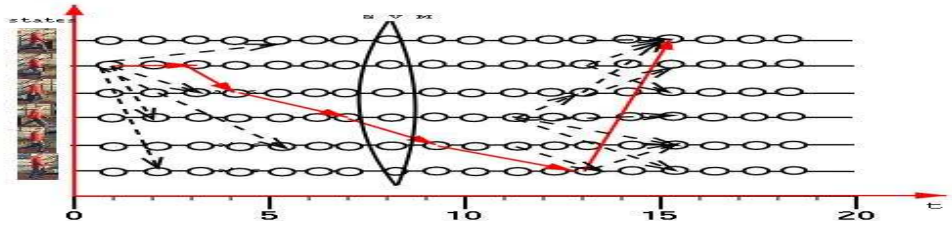

Figure 3: State sequence estimation as non-rigid motion parsing using DKM.

4.2 Learning dynamics and classifying non-rigid motion

Having supervised the exemplar classes, we propose to use multiclass probabilistic SVM classifier to discriminate the decision boundary between predefined appearance classes. At each video frame, appearance (depth order) classification results can be interpretated as a linear combination of the similiarity measure (e.g. inner product) between the data and support vectors (the exemplars lying on the decision boundaries).

Although a strong discriminative analysis is useful for appearance classification, it only address the limitation of observation process of a HMM tracker. To parse non-rigid human motion, temporal motion dynamics are also complementary. Learning the non-rigid motion dynamics using HMM has already been established in a visual motion analysis context [16], [7], which we refer the reader to for details. In sum, taking advantages of both prediction (forward) and smoothing (backward) procedures can avoid being trapped at local maximum, therefore learning non-rigid motion (dynamic appearances) is globally optimal. The benefits of temporal alignment are twofold: firstly, we obtain piecewise stationary motion data (corresponding to different classes), which effectively relax the i.i.d. assumption underlying SVM. Secondly, estimating the expected number of transitions from state S_i to state S_j, and expected number of transitions out from state S_i

will determine the likelihood that one appearance evolves into another apperance, thus approximate non-rigid human motion. The inference (parsing) algorithm of DKM is given in Eq. 6. The parsing process of DKM (similar to HMM belief propogation) is illustrated in Figure 3.

4.3 Experimental Results

In this section, we provide experimental results to examine the proposed framework in non-rigid human motion (depth order relations) classification.

(1). **Parsing Non-Rigid Human Motion.** We obtain three short video sequences A,B,C of the same person walking, each about 50 frames length, and manually select six classes of appearance patches according to the position of the left/right foot, and the relative angles of left/right thigh (Figure 2).

Here we use one sequence for testing while two others are for learning, the over-complete Haar wavelet features (which highlight the edge properties in multiple scale, achieving better results than intensity and gradient cues) from the sub-images are utilized to train the SVM classifiers and dynamics model. Figure 5 illustrates the depth inference ability of DKM parser. Figure 4 shows the confusion matricies for SVM, DKM and dynamical model given sequence C is the testing data, it shows that similar appearance classes (only different in depth order relations) are mostly resolved to a reasonable extent.(see also Figure 7). Detailed comparation of the overall motion classification of DKM and static SVM classifier is summarized in table 1. We notice that the lighting condition in video sequence A varies from B,C, this results in lower performance in column 1. This comparation suggests acquiring more varied training data, will further increase the robustness of the algorithm.

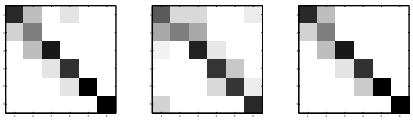

Figure 4: From left to right are Confusion Matrix for SVM (one v.s. one), Markov Dynamics, Confusion Matrix for DKM(SVM/HMM). The horizontal axis of Confusion Matrix denotes the ground truth; the vertical axis denotes the prediction.

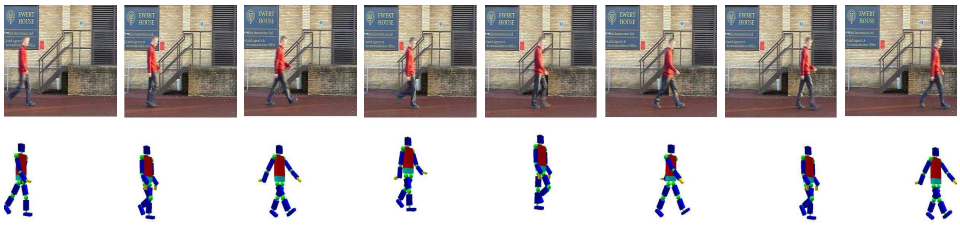

Figure 5: Parsing 3D human motion using the DKM algorithm.

(2). **More Difficult Situation.** In Figure 6, we follows the same training procedure as above, and show that DKM algorithm can successfully parse 3D human poses when the target walks towards the camera, despite the scale of the target changes over time. Here we normalize the scaling of region of interests. Three training video sequences, and one testing video sequence are used for the experiment. It also seems possible in the

Table 1: **The parsing (classification) results.**

Result	Percentage Correct		
	Test A Train B,C	Test B Train A,C	Test C Train A, B
SVM	61.90%	86.95%	91.11%
DKM	67.90%	93.47%	93.33%

future to integrate a scale discriminant into current depth inference system. In Figure 7, we provides an evaluation of the temporal belief propogation and classification results using DKM. Since a strong marginal classifier (SVM) strengthens the observation process, DKM algorithm can not only perform automatic motion classification without prior initilization of the poses, but also can recover from the error (see around frame 20). which are required by most other motion tracking/pose estimation algorithms. Besides this, it can provide temporal filtering on the static multi-class probabilistic SVM classification result (see frame 30-40). These are significant benefits over traditional tracking/motion classification algorithms.

Figure 6: Parsing human motion of walking towards the camera.

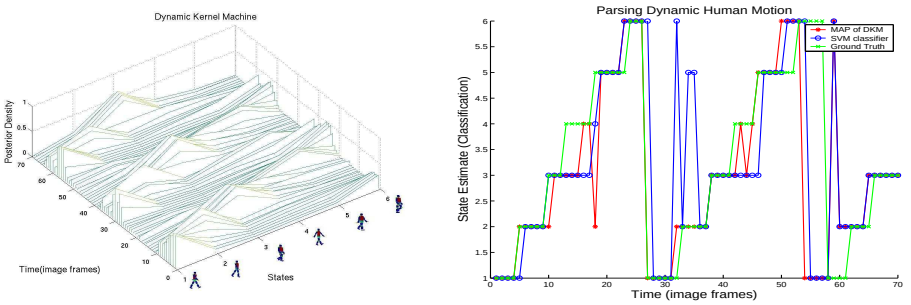

Figure 7: Left: Belief propogation of DKM over time; Right:Temporal motion(appearance) classification.

4.4 Discussion

The discerning reader may wonder whether a generative classifier (mixture of factor analyzers (MOF) for example) can perform dynamic depth order estimation. The answer is plausible but not as effective as discriminative SVM classifier. We implement a generative (motion) depth classifier as follows: Given the same appearance data, an additional generative model is used where depth order relations are kept as latent variables.

In this case, finding contributing factors of human structures from high dimensional data requires simultaneous dimensionality reduction. We use a single mixture of factor analyzers (effectively reduced dimension mixture of gaussian) to capture the variation within each appearance class including noise. Significant appearance changes (e.g. articulation, self-occlusions) are naturally modeled as different mixtures.

We combine such mixture of factor analyzers with HMM in a similar way as we propose DKM. During training, for each appearance class, cross validation ('leave one out') is used to determine free parameter: mixture component number m, factor dimension n ($[m,n] = [3,150]$). 50 iterations of EM algorithms with stopping critera $\varepsilon = 0.001$ are used to estimate the parameter set of each mixture of factor anlyzer. Six mixture of factor analyzers are used to model the six appearance classes. During parsing, the testing data is evaluated against $K = 6$ different mixtures, the overall observation probability density is a normalized log likelihood given one mixture by sum from K mixtures.

We find out that the MOFs as a classifier is significantly inferior to SVM (MOFs achieve avg. 67% classification accuracy rates), although the results can be further improved by the temporal motion prior (HMM) to 85%, it is difficult to generalize on novel testing data due to the gaussian model assumption. In contrast, since the SVM explictly model the decision boundaries between each class, our DKM not only generalize well with relatively little training data, but also can avoid the parametrical overfitting hazard. In brief, discriminative model gain advantages over generative model on these practical issues.

5 Conclusion

Traditional human motion analysis usually utilizes a 2D/3D kinematic model, acquiring depth information from multiple-views. Appearance models such as exemplar methods [7] suffer from inadequate depth information and miss-labeling errors, motion estimation and interpretation is therefore not optimal.

In this paper, we present a novel method: Dynamic Kernel Machine, which extends discriminative classifier (SVM) with the geneartive temporal motion prior. It provides an effective way to model and discriminate depth ambiguity automatically, without using a 2D/3D kinematic model and multiple-views. In the future, we plan to investigate discriminative features in addition to the appearance model we adopt here, in order to achieve person (appearance) independent non-rigid human motion classification.

References

[1] A.Agarwal and B.Triggs. "3D Human Pose from Silhouettes by Relevance Vector Regression" *IEEE International Conference on Computer Vision and Pattern Recognition*,2004

[2] L.Rabiner and B.Juang. "An Introduction to Hidden Markov Models" *IEEE Acoustics, Speech and Signal Processing Magazine*, 1986

[3] V.Vapnik. "Statistical Learning Theory" *Wiley*, 1998

[4] Z.Ghahramani and G.Hinton. "The EM Algorithm for Mixtures of Factor Analyzers" *Technical Report,University of Toronto,Department of Computer Science*, 1997

[5] C.Papageorgiou, M.Oren and T.Poggio. "A General Framework for Object Detection" *Proc. Int. Conf. on Computer Vision*, 1998

[6] P.Viola, M.Jones, and D.Snow. "Detecting Pedestrians Using Patterns of Motion and Appearance" *Proc. Int. Conf. on Computer Vision*,2003

[7] K.Toyama, A.Blake. "Probabilistic Tracking with Exemplars in a Metric Space" *Proc. Int. Conf. on Computer Vision*, 2001

[8] T.Starner and A.Pentland. "Visual Recognition of American Sign Language Using Hidden Markov Model" *Proc.Int. Workshop on Automatic Face and Gesture Recognition*,1995

[9] J.Friedman. "Another Approach to Polychotomous Classification" *Tech Report, Dept.Statist., Stanford University*, 1996

[10] A.Baumberg and D.Hogg. "An Efficient Method for Contour Tracking Using Active Shape Models" *Research Report, School of Computer Studies, University of Leeds*,1994

[11] M.Isard and A.Blake. "Contour tracking by stochastic propagation of conditional density" *Proc. European Conf. on Computer Vision*, 1996

[12] D.Comaniciu, V.Ramesh and P.Meer. "Real-time Tracking of Non-Rigid Objects Using Mean-Shift" *Proc. Computer Vision and Pattern Recognition*, 2000

[13] E.Osuna, R.Freund and F.Girosi. "Training Support Vector Machines: an Application to Face Detection" *Proc. Computer Vision and Pattern Recognition*, 1997

[14] C.Bregler. "Learning and Recognition Human Dynamics in Video Sequences" *Proc.Computer Vision and Pattern Recognition*, 1997

[15] M.Brand. "Shadow Puppetry" *Proc.Int.Conf.on Computer Vision*, 1999

[16] Huang Fei, Ian Reid. "Joint Bayes Filter: A Hybrid Tracker for Non-Rigid Hand Motion Analysis" *Proc. Euro. Conf. on Computer Vision*, 2004

[17] B.Stenger, A.Thayananthan, P.Torr and R.Cipolla. "Filtering Using a Tree-Based Estimator" *Proc.Int.Conf.on Computer Vision*, 2003

[18] S.Avidan. "Support Vector Tracking", *Proc.Conf.Computer Vision and Pattern Recognition*,2001

[19] O.Williams, A.Blake, and R.Cipolla. "A Sparse Probabilistic Learning Algorithm for Real-Time Tracking" *Proc.Int.Conf.on Computer Vision*, 2003

[20] G.Mori and J.Malik. "Estimating Human Body Configurations Using Shape Context Matching" *Proc.Euro.Conf.on Computer Vision*, 2002

[21] D.Ramanan and D.Forsyth. "Finding and Tracking People from the Bottom Up" *Proc.Conf.Computer Vision and Pattern Recognition*,2003

[22] J.Platt. "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods" *Advances in Large Margin Classifiers, A. Smola, P. Bartlett, B. Schlkopf, D. Schuurmans, eds., pp. 61-74, MIT Press*, 1999

[23] B.Frey and N.Jojic. "Estimating Mixture Models of Images and Inferring Spatial Transformation Using The EM Algorithm" *CVPR 1999*

[24] A.Jepson, D.Fleet and T.El-Maraghi. "Robust Online Appearance Models for Visual Tracking" *CVPR 2001*