A HYBRID HMM/PARTICLE FILTER FRAMEWORK FOR NON-RIGID HAND MOTION RECOGNITION

Huang Fei

Department of Engineering Science,
University of Oxford, Parks Road, OX1 3PJ, UK
fei@robots.ox.ac.uk

ABSTRACT

In sign-language or gesture recognition, articulated hand motion tracking is usually the first step before achieving behaviour understanding. However the non-rigidity of the hand, complex background scenes, and occlusion make tracking a challenging task. In this paper we present a novel hybrid HMM/Particle filter framework for simultaneously tracking and recognition of non-rigid hand motion. The novel contribution of the paper is that we unify the independent treatments of non-rigid motion and rigid motion into a single, robust Bayesian framework and demonstrate the efficacy of this method by performing successful tracking in the presence of significant occlusion clutter.

1. INTRODUCTION

Many multimedia signal processing applications such as surveillance, sports, and human-computer interfacing [1] require robust estimation and understanding object motion. In this paper, we address the problems of tracking and recognition of articulated hand motion in complex scenes, and scenes with significant occlusions in a single view. In these situations, simultaneous estimation and recognition of articulated hand poses can be a challenging problem but is crucial for real-world applications of gesture recognition. We explicitly decompose the articulated hand motion into rigid motion and non-rigid dynamical motion. Rigid motion is approximated as motion of a planar region and approached using a *Particle filter* [7] while non-rigid dynamical motion is analyzed by a *Hidden Markov Model (HMM) filter* [5]. Our hybrid HMM/Particle filter (Joint Bayes Filter) tracker demonstrates its strength by successfully recovering continuously evolving hand poses in complex scenes. Due to its ability to link the observations and underlying motion patterns in a generative fashion, hand articulation is correctly estimated even under significant occlusions.

The organization of this paper is as follows: in section 2, we will introduce the problem and give a brief outline of the algorithm. In section 3, the *Learning-Based Tracking* component (HMM) is examined with a geometrical interpretation. In the next section, we discuss the global motion tracking issues and details of the Joint Bayes filter (JBF) tracker. Afterwards some experimental results will be given. Finally we present a summary of this work.

2. JOINT BAYES FILTER METHOD

Visual tracking algorithms usually utilize one or several robust image cues as tracker representations. To deal with fast appearance deformation, strong occlusion clutter and complex scene, we propose the use of shape and colour as our JBF tracker representation. Global shape region provides a more reliable measure of the dynamic appearance changes than sparse boudary edges. Colour is also useful, because it can not only provide task-specific object representation (for example, skin colour can segment the hand from the shadows and form a silhouette sequence), but also provide a good measure of the moving region when we need to approximate 'rigid region' motion [4].

In our Joint Bayes Filter (JBF) method, a colour-based particle filter provides a robust estimation of non-rigid object translation and localizes the most likely hand location for the HMM filter input. In turn, the shape output from the HMM filter provides the importance weighting for the particle set before the resampling stage and the particle set updating in the prediction stage. This combination distinguishes our method from others. For illustrative purposes, we introduce the overall tracking system in Figure 2. The relationship between the two independent Bayesian filters, the HMM filter and the Particle filter, is also summarized.

3. LEARNING-BASED TRACKING

In this section, we present a detailed analysis of *Learning-Based Tracking* in our Joint Bayes Filter approach. We assume that non-rigid motion periodically causes appearance changes. The underlying motion patterns of the articulations are intractable, but the appearance changes often observe statistical constraints. We compute image moments

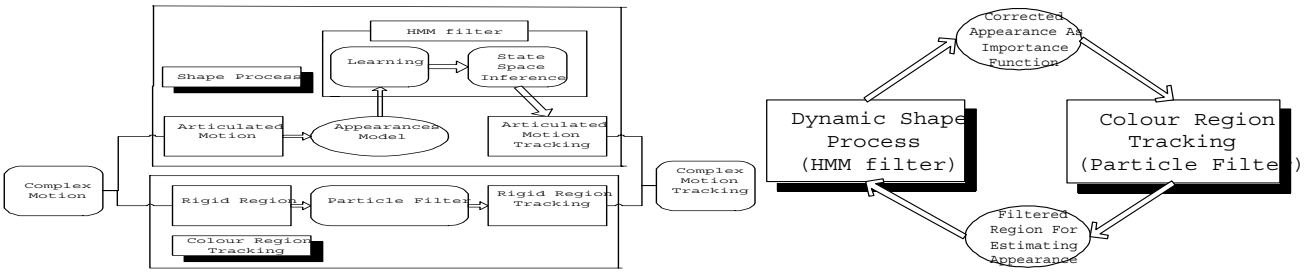

Fig. 1. (a) Flow chart of new tracking system, (b) The relationship between the two independent components.

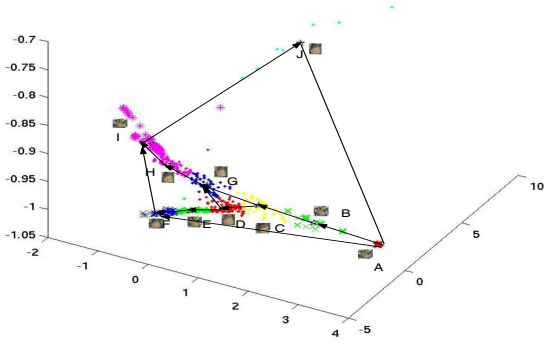

Fig. 2. Articulated hand motion embedded into a 3D metric space using LLE algorithm [2]. Trajectory of the hand motion data can be approximated by Linear Markov Chains.

($X = \{m_0, \ldots, m_n\}$ where m_i is the i^{th} moment of the shape) of the silhouettes to estimate the shape class.

Figure 2 shows the distributions of articulated hand appearances in the manifold. Here, tracing any obvious trajectory will complete a cycle of articulated hand motion. The sparseness of the clouds not only presents the evidence of possible hidden states lying under the motion sequence, but also encodes the belief of possible state transitions in the articulated motion.

3.1. Learning and Tracking

Similar to the statistical learning and inference in HMMs, we have to align the dynamic human motion to discrete states and approximate human motion dynamics from the examples. During tracking, in order to estimate what is going to be next most likely appearance correctly, the best underlying state sequence has to be decoded from current observation and a prior dynamic appearance model.

A classical VQ algorithm [3] is used in the learning stage to group the hand appearance data into clusters, here L_2 distance measure is used without strong parametric assumption. Best representative example appearances are then selected from the cluster centers. Thus we not only obtain the tracker representations, but also align articulated human motion into discrete states. The essential aspect of the *Learning-Based Tracking algorithm*: shape dynamics, is straightforward to learn. HMM provides such hand motion dynamics $P(X_{t+1}|X_t)$,

$$P(X_{t+1}|X_t) = \frac{\sum_{t=1}^{T-1} \xi_{ij}(t)}{\sum_{t=1}^{T-1} \gamma_i(t)}$$

where $\xi_{ij}(t)$ denotes the probability of being in state i at time t and at state j at time $t+1$ given the model and the observation, $\gamma_i(t)$ be defined as the probability of being in state i at time t, given the entire observation sequence and the model. This can be related to $\xi_{ij}(t)$ by summing as $\gamma_i(t) = \sum_{j=1}^{N} \xi_{ij}(t)$. N is the number of latent states, this dynamic model $P(X_{t+1}|X_t)$ is typically stronger than those from *kalman filter* or *condensation* tracker. In those trackers, the dynamics are usually built through a rather *ad-hoc* prediction (usually an AR model) of local feature measurements. In *Learning-Based Tracking*, the $P(X_{t+1}|X_t)$ is built through the statistics of latent states evolution. Estimating the expected number of transitions from state s_i to state s_j and the expected number of transitions from state s_i will determine the likelihood that one appearance evolves to another appearance. Learning human dynamics at an object-level provides a convienient and useful picture of complex human motion.

In the tracking component, the *Viterbi* algorithm [8] is adapted for decoding the best underlying state sequence as well as tracking non-rigid hand motion. In order to find the single best state sequence, $Q = (q_1, q_2, ..., q_t)$ (also known as the motion trajectory such as $A \rightleftharpoons B \rightleftharpoons C \rightleftharpoons G \rightleftharpoons H$), for the given observation $O = (o_1, o_2, ..., o_t)$ (the measurements such as $\hat{A}, \hat{B}, \hat{C}$... etc), we first define the quantity

$$\delta_t(i) = \max_{q_1, q_2, \ldots, q_{t-1}} P[q_1 q_2 ... q_{t-1}, q_t = i, o_1 o_2 ... o_t | \lambda]$$

where λ refers to a particular HMMs for hand motion analysis. $\delta_t(i)$ is the best score (highest probability) along a single path, at time t, which accounts for the first t observations

and ends in state i. Since during visual tracking, making the right predictions of the tracker states at each time instants is the major objective, we come to the *Bayesian Tracking* formula.

$$P(X_t|Z_t) = \max\{\delta_{t-1}(X) \cdot P(X|X_{t-1}) \cdot P(Z_t|X_t)\}$$

4. COLOUR-REGION TRACKING

Tracking non-rigid hand motion cannot be successful without a robust global motion estimator, colour-histogram based particle filter [4] provides a robust region estimation of the articulated hand. Nevertheless, particle filter tracker has some drawbacks. First it lacks a sophisticated mechanism for updating the region's scale changes. In fact, the adaptive scale corresponds to the non-rigid shape changes. In our JBF framework, we explicitly model the dynamics of the particle set as a first-order AR process, updated by output from the HMM filter. A second problem with the traditional particle filter is that *factored sampling* often generate many lower-weighted samples which have little contribution to the posterior density estimation. Accuracy and efficiency are sacrificed. However, in our approach, the shape output from the HMM filter provides an additional sensor which can reweight the particles and form an 'important' region for the particle filter.

Here we first introduce the general notations and then summarize the JBF algorithm. In the HMM filter, let X_t represent the shape tracker state (associated with examplars), and Z_t denote the image observations (image moments of the silhouette in this case) at time t. $d(X_t, Z_t)$ refers to the distance measure in feature space. The state vector of the Particle fitler is defined as $x_t = (x, y, s_x, s_y)$, where x, y, s_x, s_y refer to the rectangle location $L(x, y)$ in the image plane and scales along x,y coordinates. $R(x_t)$ is the candidate region thus defined, M is the number of particles used. $b_t(u) \in \{1, \ldots, N\}$ is the bin index associated with the colour vector $y_t(u)$ at pixel location u in frame t. Assume we have a reference colour histogram: $q^\star = q^\star(n)_{n=1,\ldots,N}$ obtained at initial frame. $q_t(x_t)$ denotes the current observation of the colour histogram [4]. $D([q^\star, q_t(x_t)])$ represents the Bhattacharyya distance.

The $g_t(X_t)$ used is similar to the one proposed in *ICondensation* [7], $g_t(X_t) \sim exp(-\lambda(C(S_t) + \Delta x_t))$ where $C(S_t)$ denotes the centroid of the shape, and Δx_t is the offset between the centroid of the shape and the colour region. In the JBF tracker, $A_H(x_t)$ denotes the most likely hand region, which is a rectangle area. $A_S(X_t)$ refers to the shape tracker output from the HMM filter.

Joint Bayes Filter Algorithm

1. **Initilization.**

   Particle Filter: Select the hand region, obtain the reference colour-histogram $q^\star$. For $i = 1, 2, \ldots, M$, select the initial particle set $x_0^{(i)}$.

   HMM Filter: Obtain $A_H(x_0)$ from the tracker initilization. Perform colour segmentation in $A_H(x_0)$ to obtain the silhouette.

2. **Prediction.**

   Particle Filter: For $i = 1, 2, \ldots, M$, draw new sample set $\tilde{x_t}^{(i)} \sim p(x_t|x_{t-1}^{(i)})$, here the dynamics process is a first order AR model. Calculate the colour-histogram distribution $q_t(\tilde{x}_t)$. Evaluate the importance weights $\tilde{\omega_t}^{(i)} = \frac{p(x_t|x_{t-1}^{(i)})}{g_t(X_t)} p(z_t|\tilde{x_t}^{(i)})$, where $p(z_t|\tilde{x_t}^{(i)}) \sim \exp(-\lambda D^2[q^\star, q_t(x_t)])$, and normalize the importance weights.

   HMM Filter: Generate the new prior $P(X_t|Z_{1:t-1})$ by propogating $P(X_{t-1}|Z_{t-1})$ through the markov chain.

3. **Update.**

   Particle Filter: Resample with replacement N particles $(x_t^{(i)}; i = 1, 2, \ldots, N)$ from the set $(\tilde{x}_t^{(i)}; i = 1, \ldots, N)$ according to the importance weights. Output the $A_H(x_t)$ from the particle filter.

   HMM filter: Obtain the $A_H(x_t)$ from the Particle filter, perform colour segmentation, get the observation density $P(Z_t|X_t) \sim \exp(-\lambda d(X_t, Z_t))$. Combine with the prior $P(X_{t-1}|Z_{t-1})$ to estimate $P(X_t|Z_{1:t})$ which is the most likely appearance at time t.

5. EXPERIMENT

We design several experiments to examine the performance of the JBF tracker.

Tracking dynamic appearances using JBF. We obtain a long video sequence of cyclic hand motion. 60% of the data is used for training the dynamic appearance model $P(X_t|X_{t-1})$ and selecting the exemplar set, the rest for tracking. 200 particles are used to approximate posterior density. Near real-time performance has been achieved for the overall tracking system. The result is shown in Figure 3. Small non-rigid appearance deformations and varying changing speed between successive frames are well cap-

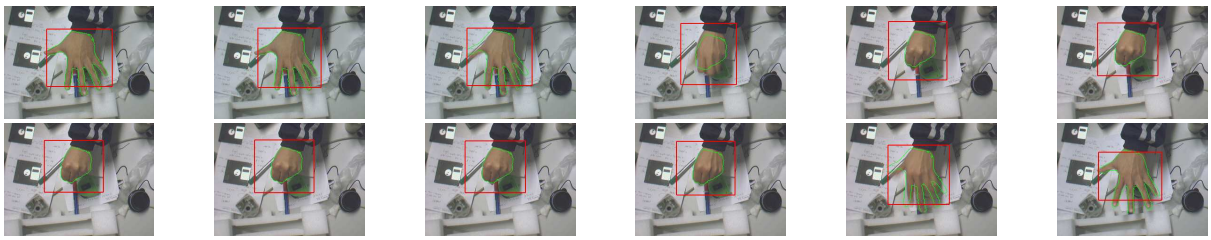

Fig. 3. Tracking results of the JBF tracker, the Particle filter determines the most likely hand region (the red rectangle), the HMM filter produce the most likely appearances (the green contours).

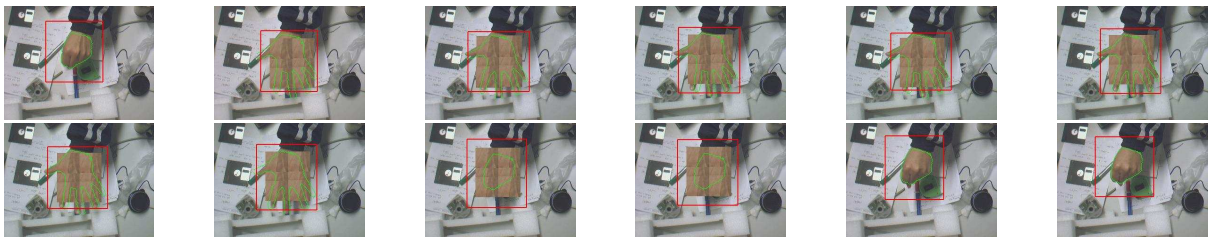

Fig. 4. The HMM filter in JBF withstands several frames of occlusion clutters.

tured by the HMM filter. In fact, the gait of the articulated hand motion is encoded in the strong appearance dynamics which is built in the learning stage. We also notice that even using the weak cue of image moments alone, tracking non-rigid hand poses in the JBF framework can achieve rather good performance.

Coping with occlusion. Experiment (Figure 4) shows that skin colour occlusions do not prevent the tracker from recovering the articulated hand poses. During occlusion, the observation density $P(Z_{t+1}|X_{t+1})$ contributes little to the shape appearance tracking. A strong dynamic appearance model $P(X_t|X_{t-1})$ obtained during the learning stage, and a correct initial estimate $P(X_0|Z_0)$ in the tracking stage, are two important factors which enable the HMM filter tracker to give an optimal estimate even under harsh conditions.

6. CONCLUSIONS

This paper presents an unifying framework of human motion tracking and understanding. The Joint Bayes filter tracker proposed extends the state of the art in visual tracking [6]. We explicitly explore non-rigid object (hand in this paper) motion analysis in the presence of scene clutter and occlusion, and demonstrate the probabilistic inference mechanism of the HMM filter. We show that state-based inference is also robust to occlusion clutter and unreliable measurements. Both components are fully Bayesian and therefore this combination (JBF filter) gives robust tracking results in real-world applications.

7. REFERENCES

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

[2] S.Roweis and L.Saul "Nonlinear dimensionality reduction by locally linear embedding" *Science*,2000

[3] A.Linde and R.Gray "An algorithm for vector quantization design" *IEEE.Trans.on Communications*,1980

[4] P. Prez, C. Hue, J. Vermaak and M. Gangnet "Color-based probabilistic tracking" *Proc.European Conference on Computer Vision*,2002

[5] R.Rabiner "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition" *Proc.IEEE*,1989

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

[7] M.Isard and A.Blake "Icondensation: Unifying low-level and high-level tracking in a stochastic framework" *Proc.5th European Conference on Computer Vision*,1998

[8] A.Viterbi. "Error bounds for convolutional codes and an asymptically optimum decoding algorithm" *IEEE. Trans.on Information Theory*,1967