PEOPLE LOCATION AND ORIENTATION TRACKING IN MULTIPLE VIEWS

Huan Jin and Gang Qian

Arts, Media and Engineering Program,
Arizona State University,
Tempe, AZ 85281

ABSTRACT

This paper presents a multi-view approach to the tracking of people location and orientation. To achieve efficient and accurate likelihood evaluation, a novel likelihood computation method is proposed. Mixtures of Gaussian (MoG) are used to represent the color models of subjects. The scaled unscented transformation is used to project the MoG color models onto the image plane to predict the color distribution for a motion sample. The efficacy of the proposed approach is demonstrated by experiment results obtained using real videos.

Index Terms— multi-view tracking, appearance modeling, unscented transformation, particle filtering

1. INTRODUCTION

Reliable location and torso orientation tracking of people remains a challenge for computer vision. Existing people tracking methods can be divided into monocular [1] and multi-view [2, 3, 4, 5, 6] approaches. An excellent review of state-of-the-art methods in both categories can be found in [2]. It is obvious that multi-view approaches are necessary when precise tracking is desired in presence of occlusions.

In visual tracking, it is important to build a color model which is accurate and friendly to efficient likelihood evaluation. Existing models do not meet this requirement. For example, in [5] only color distribution in the vertical direction of the subject is used with horizontal color distribution information missing. Such simplified color model is not able to recover torso orientation. In our proposed approach in this paper, we build a color model using mixture of Gaussian (MoG), which can be used for orientation tracking. Scaled unscented transformation (SUT) is used to find the predicted color distribution given a motion sample with point-wise projection. This method avoids projecting surface points of the 3D model onto image planes and thus is computational efficient.

To handle cross-body occlusion, most Bayesian tracking techniques make use of joint state space and utilize joint-likelihood evaluation to weight motion samples. However, joint-likelihood evaluation is not sample-efficient. Independent tracking of multiple bodies is computational efficient. However, such methods cannot handel occlusions very well. In [7], a hybrid joint-separable (HJS) tracking model is presented as an efficient and accurate multiple body tracking framework which elegantly approximate the joint dynamics using a Markov random field through message propagation and evaluate the likelihood of marginal state with respect to the observation according to an occlusion map and appearance model. In our approach, we resembles HJS in the sense that an adaptive likelihood evaluation scheme is adopted to effectively improve tracking performance. Each subject is assigned an individual particle filter. Image observation is analyzed based on subjects' motion dynamics. Joint-likelihood is needed only when subjects are close in camera views and individual likelihood can be efficiently marginalized based on joint-likelihood and particle filter's proposal distribution.

In this paper, we present a robust multi-view people tracking approach, featuring the tracking of torso orientation, and efficient likelihood evaluation. Experimental results obtained using real video show the efficacy of the proposed method.

2. APPEARANCE MODELING

Assume that people are walking on the ground plane in an upright pose. We model human body as an upright ellipsoid (prolate spheroid) with three structure parameters (r_x, r_y, r_z). A similar model is used in [8]. These structure parameters are learned off-line using known subject 3D location $\{X_t\}_{t=1}^N$ and foreground image $\{\mathbf{B}_t\}_{t=1}^N$ obtained using background subtraction, where N is the number of training frames. The optimal structure parameters maximize the objective function $\prod_{t=1}^N \frac{|E(X_t,r_x,r_y,r_z)\cap B_t|}{|E(X_t,r_x,r_y,r_z)\cup B_t|}$ where $E(X, r_x, r_y, r_z)$ denotes the projected ellipse area obtained based on the ellipse conic $\mathbf{C} = (\mathbf{P}\mathbf{Q}^{-1}\mathbf{P}^T)^{-1}$. $\mathbf{Q}$ is the 4×4 matrix representation of the ellipsoid encoding the structure and position parameters and $\mathbf{P}$ the camera projection matrix. The gradient-descent is used to solve this maximization problem.

In our approach, we first construct a texture map I_S on the ellipsoid surface in the spherical coordinates (θ, ϕ) (e.g. Figure 1 (a)) with $\theta \in [0, 2\pi]$ and $\phi \in [0, \pi]$. Each point (θ, ϕ) in the texture map has a corresponding 3D Euclidean location (x, y, z) where $x = r_x \cos\theta \sin\phi$, $y = r_y \sin\theta \sin\phi$, and $z = r_z \cos\phi$. I_S can be filled out from training images of the subject with known 3D location, orientation, and camera projection matrix. The top panel of Figure 1 (c) shows an example of texture map construction for one subject. Us-

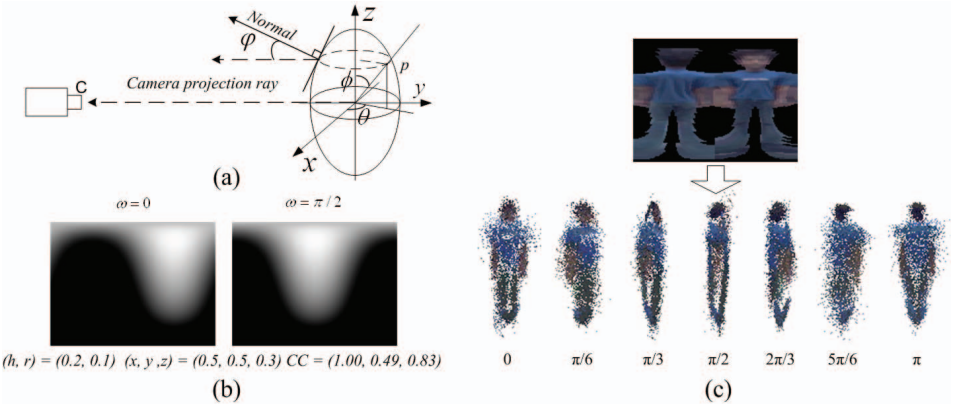

Fig. 1. (a) Illustration of visibility computation; (b) Examples of visible regions (white or gray indicates visible region); (c) Visible MoG transformed by SUT w.r.t. different torso orientations.

ing I_S, a mixture of Gaussian G is learned to represent the color distribution over (θ, ϕ) space. The HSV color space is adopted to make observations less sensitive to varying lighting conditions. In our experiment, we learned MoG with 25 components using the EM algorithm. Each subject will have her/his own MoG template, so we will have $\mathbf{G} = \{G_k\}_{k=1}^K$, where K is the number of subjects.

3. THE PROPOSED APPROACH

In this paper, our tracking task is to simultaneously find the 2D location (x, y) and torso orientation ω of these subjects. Thus, the state vector of the kth subject at time t is a 3-D vector $X_t^k = (x_t^k, y_t^k, \omega_t^k)$, and the joint state at t is $\mathbf{X}_t = \{X_t^k\}_{k=1}^K$. We cast the tracking problem into the *maximum a posteriori* (MAP) framework as follows (1)

$$\hat{X}_t^k = \underset{X_t^k}{\operatorname{argmax}}\{p(X_t^k|\mathbf{I}_t, G_k) \propto \underbrace{p(\mathbf{I}_t|X_t^k, G_k)}_{\text{Likelihood}} \underbrace{p(X_t^k)}_{\text{Dynamic}}\} \quad (1)$$

where $\mathbf{I}_t = (I_t^1, ..., I_t^C)$ are images from all C cameras at t, G_k is the appearance model of subject k. In our approach, the motion model $p(X_t)$ is given by a simple AR model $X_t = AX_{t-1} + V_{t-1}$ with A manually selected.

3.1. Adaptive Tracking

When subjects are well separated from each other, joint likelihood evaluation is not sample-efficient since more samples are needed to simulate the joint-distribution to get accurate estimates. In the spirit of the hybrid joint-separable tracking model [7], to improve tracking accuracy with less samples, we use marginalized likelihood evaluation. Separate subjects are first identified for each view based on image observation analysis and motion dynamics. If subject k is well separated from other subjects in camera c, a binary mask $B_t^{k,c}$ is generated for subject k only to obtain the foreground image for subject k, that is $I_t^{k,c} = I_t^c \odot B_t^{k,c}$, where $\odot$ represents element-wise matrix product. By evaluating $p(I_t^{k,c}|X_t^k, G_k)$, the position and orientation of subject k can be evaluated robustly with a small number of samples.

If multiple subjects are close in some camera view, occlusion may occur. In this case, we will then first compute the joint likelihood and then find marginalized likelihood of the movement state vectors of each individual subject. Let's consider a case when two subjects are close to each other. The marginalized likelihood $p(I_t^c|X_t^1)$ is given by

$$\begin{aligned} p(I_t^c|X_t^1) &= \frac{\int_{X_t^2} p(I_t^c|X_t^1, X_t^2)p(X_t^1)P(X_t^2)dX_t^2}{p(X_t^1)} \\ &= \int_{X_t^2} p(I_t^c|X_t^1, X_t^2)p(X_t^2)dX_t^2 \end{aligned} \quad (2)$$

where $p(X_t^1)$ and $p(X_t^2)$ can be simply obtained from the proposal distributions of the particle filter, respectively. This adaptive likelihood evaluation scheme is sample-efficient and able to produce better tracking results using a small set of samples than joint likelihood evaluation.

3.2. Likelihood Evaluation

Background subtraction is used to produce binary foreground masks $\mathbf{B}_t = (B_t^1, ..., B_t^C)$ from $\mathbf{I}_t$. Given binary masks $\mathbf{B}_t$ and input images $\mathbf{I}_t$, we can get the blob $\mathbf{F}_t$ which has 2D coordinates and color values for foreground pixels. We decompose the likelihood into two parts: color matching and spatial matching, as shown by (3).

$$p(\mathbf{I}_t|X_t^k, G_k) = \underbrace{p(\mathbf{B}_t|X_t^k)}_{\text{Spatial matching}} \underbrace{p(\mathbf{F}_t|X_t^k, G_k)}_{\text{Color matching}} \quad (3)$$

Spatial matching mainly serves 3D localization, while color matching is useful for 3D localization, torso orientation retrieval and identity maintenance.

3.2.1. *Spatial matching*

Spatial matching indicates how well the 2D ellipse projections computed from X_t^k can match subject foreground $\mathbf{F}_t$. Spatial matching score is computed by (4).

$$p(\mathbf{B}_t|X_t^k) \propto \prod_{c=1}^{C} M^c(X_t^k) = \prod_{c=1}^{C} \frac{|B_t^c \cap E^c(X_t^k)|}{|B_t^c \cup E^c(X_t^k)|} \quad (4)$$

where $E^c(X_t^k)$ is 2D ellipse projection of subject k in camera view c. When subjects are close in camera view c, the joint spatial matching score of the joint state $\mathbf{X}_t$ is

$$p(\mathbf{B}_t|\mathbf{X}_t) \propto M^c(\mathbf{X}_t) = \frac{|B_t^c \cap (\cup_{k=1}^N E^c(X_t^k))|}{|B_t^c \cup (\cup_{k=1}^N E^c(X_t^k))|} \quad (5)$$

3.2.2. Color matching using SUT

To compute the color matching part in the likelihood (3), we need to compare the similarity of the observed image and the predicted color distribution. The predicted color distribution given a motion sample is obtained using the learned color template and the scaled unscented transformation (SUT). By using SUT, while preserving spatial information in matching we avoid point-wise projection of the texture map to the image plane and thus reduce computational cost and maintain good tracking accuracy.

Due to self-occlusion of the subject, only part of color template I_S is visible to one camera at one time instant. To find the predicted color distribution in this camera given the motion sample, we need to identify the visible portion of I_S first. In this approach, this is done by finding the visible components of the corresponding MoG. A point s on the ellipsoid is visible to a camera only if the angle between surface normal direction at s and camera projection ray through s is less than 90°. In this way, we can obtain G_v, the visible MoG by keeping the components with visible centers. Figure 1 (a) and (b) shows examples of visible region in texture map.

The mapping from the (θ, ϕ) space to image plane is a nonlinear transformation due to the perspective projection. In our approach, we use SUT to map G_v from the (θ, ϕ) space to the 2D image space. SUT [9] is a method for calculating the statistics of a random variable that undergoes a nonlinear transformation that can overcome the dimensional scaling effects. The MoG transformation algorithm is given below.
Input: Mixture component's mean $\mu_{\mathbf{x}} = (\theta, \phi, h, s, v)^T$ and covariance matrix $\Sigma_{\mathbf{x}}$; 3D location $X = (x, y, \frac{r_z}{2})$, where r_z is taken from the body structure parameters and used as an estimate of the height of the subject and torso orientation ω; camera projection matrix $\mathbf{P}$.
Output: Transformed mixture component's mean $\mu_{\mathbf{y}} = (\tilde{x}, \tilde{y}, \tilde{h}, \tilde{s}, \tilde{v})^T$ and covariance matrix $\Sigma_{\mathbf{y}}$.
1. Dimension of vector $d = 5$; compute sigma points with weights for mean and covariance $\{\mathbf{x}_i, w_i^{(m)}, w_i^{(c)}\}, i = 0, ..., 2d$ according to the unscented transformation [9].
2. Transform each sigma point: $\mathbf{y}_i = \mathbf{h}(\mathbf{x}_i)$, where $\mathbf{h}(\mathbf{x})$ is given by $(\tilde{x}, \tilde{y}) = f(\theta, \phi, X, \omega, \mathbf{P})$; $\tilde{h} = h$; $\tilde{s} = s$; $\tilde{v} = v$, and $f(\cdot)$ is a function encoding camera projection.
3. Compute the transformed mean $\mu_{\mathbf{y}}$ and covariance $\Sigma_{\mathbf{y}}$ using the weighted sigma points [9].

When subjects are far apart from each other, the likelihood of the state vector of individual subjects is computed separately. Given the foreground image of the kth subject $\mathbf{F}_t^k = \{F_t^{k,c}\}_{c=1}^C$, the motion sample X_t^k, and G_k the MoG template for subject k, the color matching part of the likelihood is given by

$$p(\mathbf{F}_t^k | X_t^k, G_k) = \prod_{c=1}^{C} \prod_{i=1}^{N_c} G_{X_t^k}(\xi_{c,i}) \quad (6)$$

where $\xi_{c,i}$ is the ith foreground pixel feature including color and location in the cth camera view, and N_c is number of foreground pixels and $G_{X_t^k}(\cdot)$ is the probability density function (pdf) of the transformed visible MoG after SUT using X_t^k. When subjects are close to each other, cross-occlusion needs to to handled. To this end, we first estimate the occlusion relationship for each pair of subjects given the joint motion samples $\mathbf{X}$. The ray/quadrics intersection (RQI) method [10] is used to infer the cross-occlusion relationship, based on which the joint likelihood $p(\mathbf{F}_t | \mathbf{X}_t, \mathbf{G})$ can be computed. Due to space limitation, details on how we handel cross-occlusion have to be omitted in this paper.

4. EXPERIMENTAL RESULTS

Three calibrated color cameras with 320×240 resolution were used in our experiments. In our experiment, $1\ unit \cong 427cm$. Each subject is tracked by a particle filter with 150 particles. Our system was tested on two video sequences (1031 and 1100 frames). Since 3D ground truth is not available, we take advantage of 2D ground truth, i.e. 2D foreground segmentations F_{GT}^c instead. We generate the estimated 2D ellipse regions E^c for the camera c by projecting subjects' ellipsoids w.r.t. the tracking results (locations and orientations). Two performance measures, the precision ($P = F_{GT}^c \cap E^c / E^c$) and the recall ($R = F_{GT}^c \cap E^c / F_{GT}^c$), can be used to evaluate the tracking performance. Tracking results using two other approaches in [8] (no color model) and [5] (vertical layered color model) were also obtained for comparison. The average (P, R) pairs of the two testing videos are $(0.84, 0.59), (0.83, 0.63)$ for the method without color model [8], $(0.82, 0.87), (0.84, 0.89)$ with vertical layered color model [5], and $(0.85, 0.90), (0.84, 0.92)$ for the proposed method. It can be seen that three approaches have similar average P values. But the average recalls show that tracking degrades without a good color model. Figure 2 shows tracking results using one test video. The tracking without color model wrongly labels the subjects after they passed by each other closely in frames 274 and 354. The tracking with vertical layered color model is also incompetent in torso orientation recovery in most of frames. However, our approach is able to correctly track subjects' locations and torso orientations. Subjects' identities are well maintained as well under partial/full occlusions.

5. CONCLUSION

We show in this paper that in multi-view tracking, MoG can be used to represent the color appearance model and the scaled unscented transformation can be used to obtain the predicted color distribution in likelihood evaluation. Encouraging results have been obtained using the proposed method.

6. ACKNOWLEDGEMENT

The work in this paper is supported by U.S. National Science Foundation on CISE-RI no. 0403428 and IGERT no. 0504647. [1]

[1] Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. National Science Foundation (NSF).

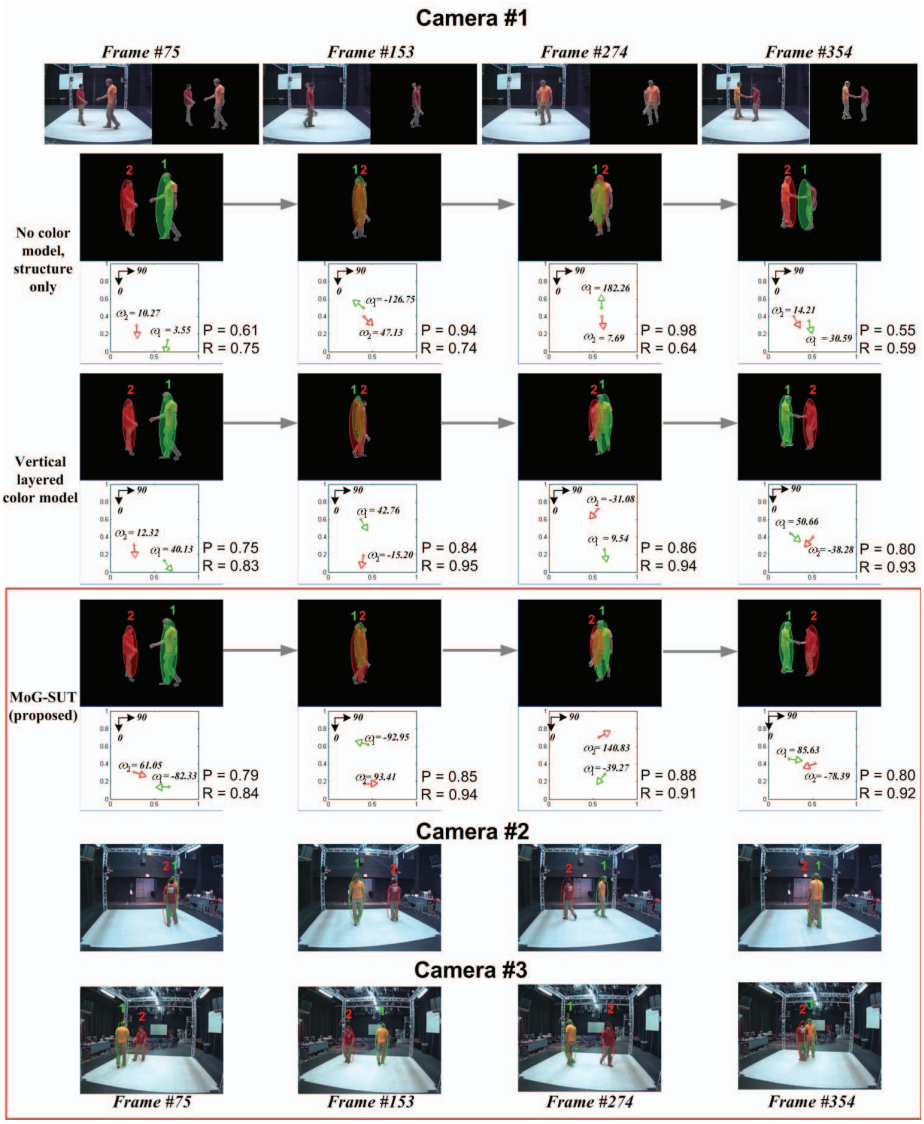

Fig. 2. Comparison of three tracking approaches. For each approach, we superimpose the estimated ellipses on the ground truth images and illustrate subjects' torso orientations in the top-down views. The last two rows show the tracking results in cameras 2 and 3 with estimated ellipses superimposed on raw images.

7. REFERENCES

[1] I. Haritaoglu, D. Harwood, and D. Davis, "Who, when, where, what: A real time system for detecting and tracking people," in *Automated Face and Gesture Recognition*, pp. 222–227, 1998.

[2] J. Berclaz, F. Fleuret, and P. Fua, "Robust people tracking with global trajectory optimization," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2006.

[3] J. Black, T. Ellis, and P. Rosin, "Multi view image surveillance and tracking," in *IEEE Workshop on Motion and Video Computing*, 2002.

[4] O. Javed, Z. Rasheed, O. Alatas, and M. Shah, "Knight: A real time surveillance system for multiple overlapping and non-overlapping cameras," in *Int'l Conf. on Multimedia and Expo*, 2003.

[5] A. Mittal and L. Davis, "M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo," *Int'l J. of Computer Vision*, 2003.

[6] K. Nummiaro, E. Koller-Meier, T. Svoboda, D. Roth, and L. Van Gool, "Color-based object tracking in multi-camera environments," in *Pattern Recognition Symposium*, 2003.

[7] O. Lanz, "Approximate bayesian multibody tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1436–1449, September 2006.

[8] T. Osawa, X. Wu, K. Wakabayashi, and T. Yasuno, "Human tracking by particle filtering using full 3D model of both target and environment," in *Int'l Conf. on Pattern Recognition*, 2006.

[9] S. J. Julier, "The scaled unscented transformation," in *American Control Conference*, pp. 4555–4559, 2002.

[10] A. Wood, B. McCane, and S. King, "Ray tracing arbitrary objects on the GPU," in *Image and Vision Computing New Zealand*, 2004.