A LIGHTWEIGHT MULTIVIEW TRACKED PERSON DESCRIPTOR FOR CAMERA SENSOR NETWORKS

Michael J. Quinn Thomas Kuo B.S. Manjunath

University of California - Santa Barbara
Department of Electrical and Computer Engineering
Santa Barbara, CA 93106
{mquinn,thekuo,manj}@ece.ucsb.edu

ABSTRACT

We present a simple multiple view 3D model for object tracking and identification in camera networks. Our model is composed of 8 distinct views in the interval $\left[0, \frac{7\pi}{4}\right]$. Each of the 8 parts describes the person's appearance from that particular viewpoint. The model contains both color and structure information about each view which are assembled into a single entity and is meant as a simple, lightweight object representation for use in camera sensor networks. It is versatile in that it can be gradually assembled on-line while a person is tracked. The model's ease of use and effectiveness for identification in surveillance video is demonstrated.

Index Terms— Camera Networks, Surveillance, Appearance Modeling, Tracking, Video Indexing

1. INTRODUCTION

Advances in both hardware and software are quickly making pervasive video monitoring a reality. Video sensor networks research concerns itself with the collaboration and operation of camera-equipped sensors for applications such as environmental monitoring, wildlife observation, or traditional human activity surveillance. These applications require efficient object representation for proper object tracking and activity description. This paper describes an object model designed for a multiple camera network.

Object and human representation in surveillance videos uses a variety of models and features. Models range in complexity from simple point features that represent the location of an object [1] to complex, fully articulated models of major body parts [2]. Features for these models include shape, color, and texture [3, 4, 5]. For a single view, Chien et al. [6] proposed the *Human Color Structure Descriptor* that describes the position and mean color of the torso, leg, and shoe regions of a person. Yuk et al. [7] describe objects in a single camera with dominant colors and edge direction histograms.

A camera network can take advantage of multiple views, and more reliable 3D motion and pose information. Mittal and Davis [8], for example, develop a color model for each height slice of a person, which does not account for 3D structure of a person. One can also create a detailed 3D human model; however, this is too complex for tracking and surveillance.

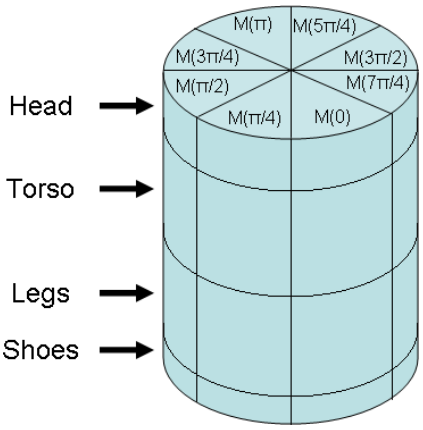


Fig. 1. Color models computed for 8 radial directions and 4 body regions.

In this paper, we present a simple 3D human model that takes advantage of the multiple camera views, yet remains compact. As shown in Fig. 1, it models a person as a cylinder and subdivides the body into four body regions and eight radial directions. Essentially, it describes each viewpoint as a 2D model. If we assume that people are walking and standing upright, we can estimate the pose and learn the models automatically and on-line.

Section 2 describes the 2D model for each viewpoint and its distance metric. Section 3 details the 3D model assembly and learning process, and section 4 shows experimental identification results obtained with our model.

2. TRACKED OBJECT MODEL

Our proposed object model uniquely captures the variation of a person's appearance about his vertical axis. It does so di-

viding a person into eight directions, and creating a 2D model for each direction. The 2D model is created in four steps: (1) detection, (2) segmentation, (3) feature extraction, and (4) assembly. An assembled 2D model is used to compare with information from a single camera.

We use background subtraction to detect an unoccluded person and to find a clear silhouette. This silhouette and tracking information helps us determine the pose and visible direction. Then, we find a bounding box for the person and segment it into head (H), torso (T), legs (L), and feet (F) regions; typical boundaries observed with people's clothing. We currently use the heuristic of $\frac{1}{6}$, $\frac{7}{20}$, $\frac{7}{20}$, and $\frac{2}{15}$ of the person's height for the H, T, L, and F regions, respectively.

Once segmented, color descriptors are used for each section. The head and torso regions are encoded using the Color Layout Descriptor (CLD) [9], and the pants and shoes regions are represented by an HSV histogram. The CLD converts the region into the YCbCr color space and resizes it to 8 by 8 pixels. It then applies the Discrete Cosine Transform (DCT) to the image patch, of which we retain the top six DCT coefficients in each channel. This processes allows us to preserve some of the structure of the region and decouple luminance and chrominance to better handle variable lighting conditions. Since the pants and shoes of people are generally less colorful than clothing above the waist, the structure is not as important. Thus we represent these regions with an HSV histogram that has the H, S, and V channels quantized to 16, 4, and 4 bins respectively.

Finally, the CLD from the head and torso and the HSV histogram from the pants and shoes are combined to form the descriptor for a single direction.

To compare different views, the distance metric used is a weighted combination of the distances of the four separate descriptors in our model, as in [10]:

$$d = \alpha_H d_H + \alpha_T d_T + \alpha_L d_L + \alpha_F d_F \qquad (1)$$

where $\alpha_H + \alpha_T + \alpha_L + \alpha_F = 1$. The distance measure used for the head and torso regions (d_H, d_T) is defined as:

$$\begin{aligned} d(D_1, D_2) \quad = \quad & \sqrt{\sum_i w_{Yi}(Y_{i1} - Y_{i2})^2} \\ & + \sqrt{\sum_i w_{Cbi}(Cb_{i1} - Cb_{i2})^2} \\ & + \sqrt{\sum_i w_{Cri}(Cr_{i1} - Cr_{i2})^2} \end{aligned} \qquad (2)$$

where Y_{ix},Cb_{ix}, and Cr_{ix} are the DCT coefficients and the weights assign different importance to different frequency values. The distance measure for the legs and feet regions is the Bhattacharyya distance, a common measure used with histogram information. Our experiments led us to assign more weight to the torso region than to the other regions. Logically, this makes sense since it is not only the most detailed but also most likely to be unique between people.

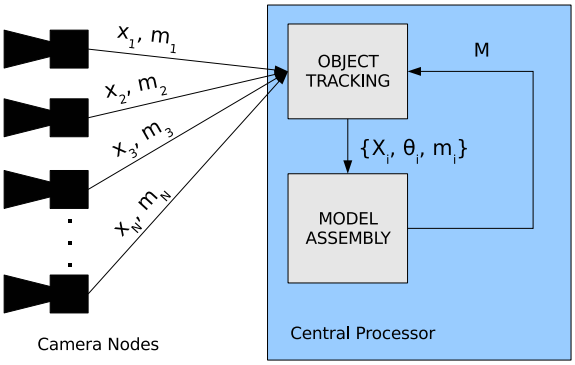


Fig. 3. Model Learning Process

3. TRACKED OBJECT MODEL LEARNING

While manual offline construction of appearance models is preferable and will yield the best results, it is impossible to do so in most circumstances. Instead, we seek an automatic, online learning procedure which will build, over time, an accurate appearance model for each tracked person.

The single-view modeling process, outlined in Section 2, is performed at each node, yielding a number of single-view appearance models of the tracked person. These nodes are in turn assembled during the learning process to produce a multiview model of the tracked person. As a person enters the system, he or she is assigned an ID and the modeling process is initiated. As the person is tracked through the system, the different sensing nodes' individual view angles are calculated. The view angle is a function of the person's position and velocity with respect to the camera's orientation. In the case that the person stops or is occluded, the learning process can be suspended until more reliable data is again available. With sufficient coverage, the system should be able to assemble a full model of the tracked person in a short time. When the model reaches a designated level of confidence, the system can switch to model-based tracking in 3D. The proposed model learning process is illustrated in Fig. 3.

4. EXPERIMENTAL RESULTS

In order to demonstrate our proposed appearance model, we first show its ability to discriminate between people using our 2D model. We took images of ten subjects along with their silhouettes and computed their individual descriptors (for the $\theta = 0$ view) as detailed in Section 2. The people are shown in Fig. 4. The calculated distances between the different regions (H, T, L, F) are shown in Fig. 5. Fig. 5(b) shows that the torso region has the highest discrimination between people in this

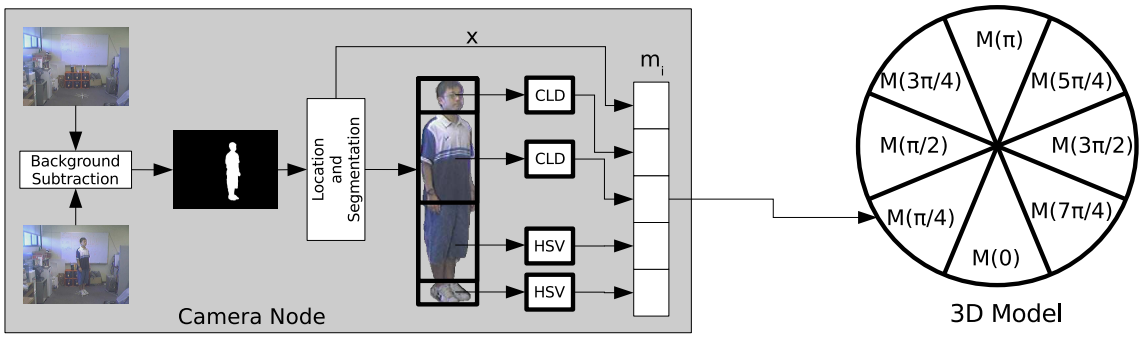
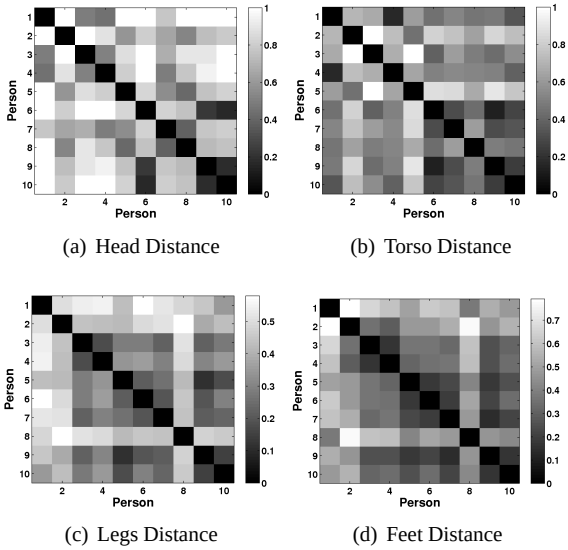


Fig. 2. Modeling a Standing Person

Fig. 4. Ten People From Discrimination Test



(a) Head Distance (b) Torso Distance

(c) Legs Distance (d) Feet Distance

Fig. 5. Ten People Individual Region Distances - (Distances between 10 people's individual regions.)

experiment. Through experimentation, we arrived at distance measure weight values of $\alpha = [0.2,\ 0.4,\ 0.2,\ 0.2]$. Fig. 6 shows the final summed distances between the ten subjects.

In order to demonstrate the ability of our multivew descriptor, we look at the subjects 9 and 10. Except for the white region on the front, these two subjects are dressed identically. The full range of the subjects' views are shown in Fig. 7.

Fig. 8 shows the distances between the corresponding views of these two people. As expected, the views near $\theta = 0$ are the most helpful for disambiguating these two people. We compare these results with that of using simply the mean color of each region (as in [6]), but also averaged across all views. This distance is also shown in Fig. 8 as the lighter line. We see that keeping each view's features separate allows better differentiation of tracked persons.

5. CONCLUSION AND FUTURE DIRECTION

We have demonstrated an extension to 2D person description techniques for use in 3D human description and tracking in camera networks. This model is composed of 2D models of the person as viewed at different angles. Each view is created by segmenting the tracked person into head, torso, leg, and feet regions. The head and torso regions are described by the MPEG-7 Color Layout Descriptor and the legs and feet regions are described by their quantized histograms in the HSV color space.

The model was first demonstrated for tracked person identification and differentiation in laboratory surveillance video. The proposed individual view descriptor was shown to work well. The model was then shown to effectively differentiate two targets with a high percentage of similarity. The power of

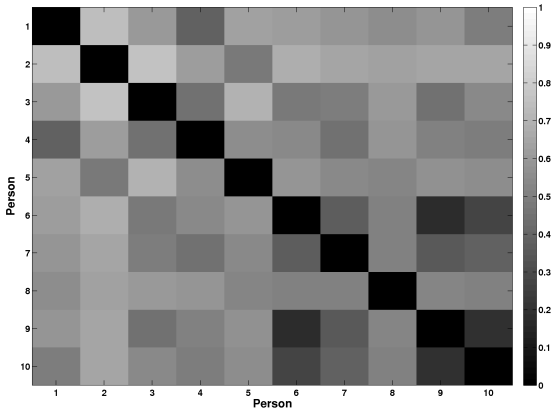
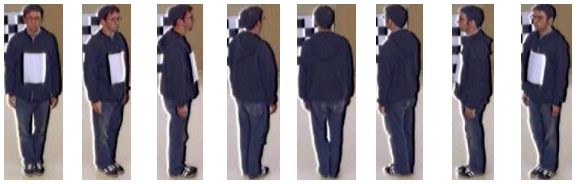


Fig. 6. Ten People Total Distances - (Total distance between appearance models of 10 example people)

(a) Person 9 Views



(b) Person 10 Views

Fig. 7. People 9 and 10 Views

the model stems from the view-dependent nature of the features contained within it.

Our next step will be to complete the real-time implementation of our on-line learning method in our experimental ten-node camera network. In order to effectively incorporate data from all cameras within the network, we will be forced to address the issue of inter camera color calibration as well. Our end-goal for the network is simultaneous appearance-based multiple people tracking.

6. REFERENCES

[1] Tao Zhao and R. Nevatia, "Tracking multiple humans in complex situations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1208–1221, Sept. 2004.

[2] D. Ramanan, D. A. Forsyth, and A. Zisserman, "Tracking people by learning their appearance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 65–81, Jan. 2007.

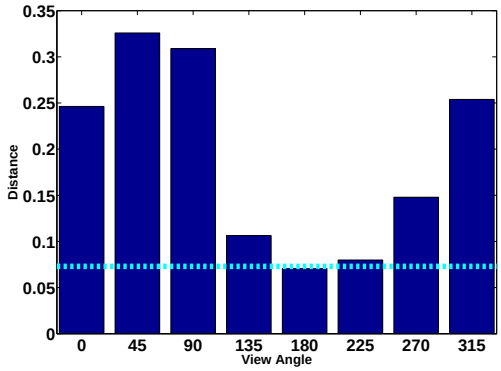


Fig. 8. People 9 and 10 View Distances - (Dark bars indicate distance at each view in multiview model. Light line indicates distance using cumulative model.)

[3] Andrew Senior, "Tracking people with probabilistic appearance models," in *PETS 2002*, 2002.

[4] Jianpeng Zhou and Jack Hoang, "Real time robust human detection and tracking system," in *Proceedings of CVPR 2005*, 2005.

[5] Rafik Bourezak and Guillaume-Alexandre Bilodeau, "Object detection and tracking using iterative division and correlograms," in *Proceedings of CRV 2006*, 2006.

[6] Shao-Yi Chien, Wei-Kai Chan, Der-Chun Cherng, and Jing-Ying Chang, "Human object tracking algorithm with human color structure descriptor for video surveillance systems," in *ICME 2006*, Toronto, Ont., July 2006, pp. 2097–2100.

[7] J. S. C. Yuk, K.-Y. K. Wong, R. H. Y. Chung, K. P. Chow, F. Y. L. Chin, and K. S. H. Tsang, "Object-based surveillance video retrieval system with real-time indexing methodology," in *ICIAR 2007*, Montreal, Canada, August 2007, pp. 626–637.

[8] A. Mittal and L. Davis, "M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene," *IJCV*, vol. 51, no. 3, pp. 189–203, 2003.

[9] P. Salembier B. S. Manjunath and T. Sikora (editors), Eds., *Introduction to MPEG-7, Multimedia Content Description Interface*, John Wiley and Sons, Ltd., Jun 2002.

[10] Till Quack, Ullrich Mönich, Lars Thiele, and B. S. Manjunath, "Cortina: a system for large-scale, content-based web image retrieval," in *ACM MULTIMEDIA 2004*, New York, NY, USA, 2004, pp. 508–511, ACM.