Panoramic Appearance Map (PAM) for Multi-Camera Based Person Re-identification

Tarak Gandhi and Mohan M. Trivedi
Computer Vision and Robotics Research Laboratory
University of California San Diego
La Jolla, CA 92093, USA
{tgandhi,mtrivedi}@ucsd.edu

Abstract

This paper proposes a concept of panoramic appearance map to perform reidentification of a people who leave the scene and reappear after some time. The map is a compact signature of appearance information of a person extracted from multiple cameras. The person is detected and tracked in multiple cameras and triangulation is used to accurately localize the person in 3-D. A virtual cylinder is formed around the person's location and mapped onto an image with the horizontal axis representing the azimuth angle and vertical axis representing the height. Each bin in the map image gets the appearance information from all the cameras which can observe it. The maps between different tracks are matched using a weighted metric. Experimental results showing person matching and reidentification show the effectiveness of the approach.

Keywords: *Video Surveillance, Camera network, Color matching, Multiple view geometry, Visual tracking*

1. Introduction and Motivation

Recently, there has been a considerable interest in multi-camera systems for intelligent environments, surveillance, traffic analysis, and other applications. Multiple cameras with overlapping views offer superior scene coverage from all sides and enable 3-D information extraction by triangulation. On the other hand, cameras with non-overlapping views can provide coverage of wide areas without sacrificing on resolution.

An important problem in multi-camera systems is to reidentify objects that leave one camera (or a set of overlapping cameras) and enter another in a different area, or reenter in the same place after a period of time. This problem is often difficult since an object could have a number of potential matches, and it may not always be possible to disambiguate all the matches. In such cases, it may be best to identify all possible matches using coarse-level features such as color, texture, and transition time between the cameras in order to narrow down the search. Further disambiguation can then be performed manually, or by using specialized features characteristic to the objects.

This paper describes a novel approach to perform person reidentification in multi-camera setup. Multiple overlapping cameras provide robust detection as well as 3-D localization of objects in their field of view [2, 7], and covers object features from all directions. A concept of object-centered panoramic map is introduced which extracts and combines appearance information from all the cameras that view the object features to form a compact, time averaged signature of the entire body of the person.

The X axis of the map represents the azimuth angle with respect to the world coordinate system, and Y axis represents the object height above the ground plane. Temporal integration can be performed to registering and blending a number of maps generated for the person over a period of time. The maps generated from two different events can be compared to find potential matches. Presently, we use color information for comparison, but other information such as texture could also be integrated in the framework.

2. Related Research

Initial work on reidentification has taken place in traffic analysis applications where the vehicle objects are rigid, move in well defined paths and have uniform color. Huang and Russel [4] propose a probabilistic framework to match vehicles between two non-overlapping views using features such as color, size, velocity, lane position, and time of observation. Parametric models of the probability distributions of these features are are learned over time and the matching problem is solved in a maximum likelihood framework.

Proceedings of the IEEE International Conference
on Video and Signal Based Surveillance (AVSS'06)
0-7695-2688-8/06 $20.00 © 2006 IEEE

Person tracking and reidentification are often more complex, since persons are articulated, move arbitrarily, and often wear multi-colored dress. Kettnaker and Zabih [6] exploit the similarity of views of the person, as well as plausibility of transition times from one camera to another. The Bayesian formulation is translated to a linear programming problem which maximizes the posterior probability of the correspondences given the data. Javed et al. [5] use various features based on space-time (entry/exit locations, velocity, travel time) and appearance (color histogram). A probabilistic framework is developed to identify best matches. Bird et al. [1] detect loitering individuals by matching pedestrians intermittently spotted in the camera field of view over a long time. Snapshots of pedestrians are extracted and divided into thin horizontal slices. The feature vector is based on color in each slice and Linear Discriminant Analysis is used to reduce the dimension.

Multiple cameras with overlapping fields of view increase the tracking and reidentification reliability due to better handling of occlusions and accurate estimates of the floor position and height of the persons, and observation of features from multiple perspectives. Cai and Aggarwal [2] perform multi-camera tracking based on geometric and intensity features. Mittal and Davis [7] propose a multi-camera person tracking system called "M2-Tracker". They develop a region-based stereo algorithm is that finds 3D points inside an object from knowledge of regions belonging to object in 2 views. The color model is formed from of horizontal slices of the person by taking the histogram of each slice in the color space. Utsumi and Tetsutani [8] perform head-tracking with multiple cameras by creating an appearance model of the head as a set of color patches in 3-D space. For every new frame, the current model is projected back to the 2-D space of each camera, and compared with the pixel values in the camera image to locate the object in that image. The model is dynamically updated by adding information from all camera images.

Most of these approaches extract the color distribution at different heights above the ground. However, the azimuth information is not considered. This approach captures appearance information at a number of height and azimuth angles around the person and produces a compact, time and camera averaged signature of the entire body of the person that is useful for reidentification. The signature contains not only the color information, but also confidence measure in form of weights. A novel metric based on color and weight is proposed to integrate and compare panoramic maps.

3. Reidentification using Object-Centered Panoramic Map

The block diagram of the reidentification approach is shown in Figure 1. Person detection is performed in each camera using background subtraction. The background image is generated using fading average of previous image frames. The foreground image is obtained by subtracting the background image from the current image. The foreground image is thresholded to generate a foreground mask. Assuming that the persons are standing vertically in the image, a projection profile is formed by taking the sum of each column. The peaks in this vertical profile give the location of the person. The pixels around the location corresponding to the person are collected to form a mask image.

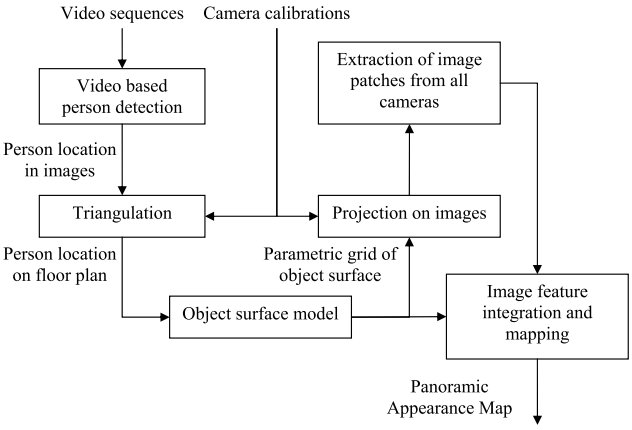

Figure 1. Block diagram for generation of Panoramic Appearance Map

The object location in the floor plan is obtained by triangulating on the image locations obtained from the cameras that can observe the person [3]. Let (R_k, t_k) be the transformation from coordinate system of camera k to the world coordinate system, so that a point having world coordinates p has the camera coordinates given by:

$$p_k = R_k^T(p - t_k) \tag{1}$$

The camera coordinates are transformed using the intrinsic parameters of the camera to obtain pixel coordinates (u_k, v_k).

For the inverse transform, each pixel (u_k, v_k) in the camera maps to a ray in the 3-D space centered at t_k along the direction corresponding to p_k. This ray is given by the parametric equation:

$$p = \lambda_k R_k p_k + t_k \tag{2}$$

If the person is detected in more than one camera, the floor position of the person can be obtained by finding the intersection of the corresponding rays. Due to localization errors, the rays may not exactly intersect at one point. Hence, the point $\hat{p}$ minimizing the sum of squares of the perpendicular distance to all rays is estimated.

Proceedings of the IEEE International Conference
on Video and Signal Based Surveillance (AVSS'06)
0-7695-2688-8/06 $20.00 © 2006 IEEE

For performing person reidentification, we model the person's body as a convex generalized cylinder shown in Figure 2. A point p on the surface of the cylinder is parameterized by the azimuth angle θ and height h:

$$p = (x, y, z)^T = (x_0 + r(\theta, h)\cos\theta, r(\theta, h)\sin\theta, h)^T \quad (3)$$

where r is a function representing the surface of the cylinder and $p_0 = (x_0, y_0, 0)^T$ is the location of the cylinder on the ground. The function r can be modeled using circular or elliptical cross section for simplicity. On the other hand, voxelization by volume intersection would be appropriate if better accuracy is desired. If (R_k, t_k) is the transformation from coordinate system of camera k to the world coordinate system, the camera coordinates of p are given by:

$$p_k = R^T(p_0 - t_k) \quad (4)$$

Intrinsic parameters of the camera can then be used to transform to pixel coordinates (u_k, v_k).

If r is modeled or computed for every θ and h, one has a transformation from (θ, h) to image pixels (u_m, v_m) for every camera. The parameters h and θ are discretized into a grid of $M \times N$ bins, with each bin (m, n) corresponding to the ranges $h(m)...h(m+1)$ with $h(m) = m\Delta h$ and $\theta(n)...\theta(n+1)$ with $\theta(n) = n\Delta\theta$.

The corners of each bin are mapped onto the image plane of each camera k in which the bin is visible, using the world to image transformation for that camera. This defines a region in the image plane corresponding to the bin. If $c_k(m, n)$ is the average value of an appearance feature (such as color) in that image region and w_k is the number of pixels in the region, then the pair (c_k, w_k) can be used to represent the appearance information from image k. This way, each bin not only contains the appearance information, but the weight that specifies confidence of the bin, which is used during integration and matching. This minimizes errors due to spurious bins that do not obtain reliable information from the particular image.

Note that though currently, the weight images are based only on the count of foreground pixels, in future, it may be possible to incorporate other parameters such as variance or motion within the bin in order to further improve the performance.

If multiple cameras can view the point p, the information in (c_k, w_k) for all cameras can be combined as:

$$w = \sum_k w_k, c = \frac{\sum_k c_k}{w} \quad (5)$$

Two such maps from different events can be compared using a distance measure. We suggest the following weighted metric which gives larger emphasis on points having larger weights in both maps, while suppressing points where one or both have small weights. For rotational invariance, one

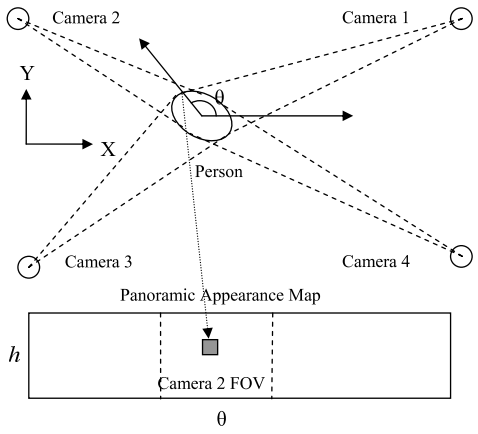

Figure 2. Mapping from 3D space (plan view) to body centered panorama

can find the metric for all discrete rotation angles and take minimum of these. The distance d between two panoramic maps $c^{(a)}$ and $c^{(b)}$, displaced by an angle $\theta(n) = n\Delta\theta$ is given by:

$$\begin{aligned} d(c^{(a)}, c^{(b)}; n) &= s(c^{(a)}, c^{(b)}; n)/w(c^{(a)}, c^{(b)}; n) \\ s(c^{(a)}, c^{(b)}; n) &= \sum_{m,n'} \left| c^{(a)}(m, n) - c^{(b)}(m, n+n') \right|^{\gamma} \\ &\quad . \ [w_1(m, n)w_2(m, n+n')]^{\gamma/2} \\ w(c^{(a)}, c^{(b)}; n) &= \sum_{m,n'} [w_1(m, n)w_2(m, n+n')]^{\gamma/2} \end{aligned}$$

where the sums $n + n'$ are taken modulo N so that they lie between 0 and $N - 1$. The best distance d is obtained by taking the minimum over all n.

The object is tracked over multiple frames and the maps for all frames are is integrated displaced averaging. The current map $(c(t), w(t))$ at time t is matched with the averaged map $(C(t-1), W(t-1))$ at time $t-1$ using distance metric. If n_0 is the index for which d gets a minimum value, the averaged map at time t is computed recursively as:

$$\begin{aligned} W(t; n) &= \alpha w(t; n + n_0) + (1 - \alpha)W(t - 1; n) \\ C(t; n) &= \frac{1}{W(t; n)}[\alpha c(t; n + n_0)w(t; n + n_0) \\ &+ (1 - \alpha)C(t - 1; n - 1)W(t - 1; n - 1)] \end{aligned}$$

with

$$n_0 = \text{argmin}_n d(C(t-1), c(t); n) \quad (6)$$

If the matching is accurate, the temporally integrated maps over the entire event corresponding to a walking person may give more reliable matching than that with a single frame.

Proceedings of the IEEE International Conference
on Video and Signal Based Surveillance (AVSS'06)
0-7695-2688-8/06 $20.00 © 2006 IEEE

4. Experimental Results

The above approach was applied for person reidentification in a laboratory setup with 4 cameras giving coverage from all directions. Figure 3(a) shows typical frames from video sequence corresponding to the particular person. The person was modeled using a circular cross section of fixed radius. The temporally integrated panoramic map based on color feature is shown in Figure 3(b). Figure 3(c) and (d) show the same person reappearing in the scene after some time. It is seen that the color-maps can be compared to find potential matches. The distance metric is applied to the panoramic maps and the distances for each test sample to all training samples are shown against the test sample in Figure 3(e). It is seen that the true matches give the best similarity measure in most cases. However, there is some ambiguity due to different persons wearing similar colored dress. In future, other features such as person height and texture may be incorporated to disambiguate the matches.

Another experiment was performed in which a number of persons moved from one multi-camera setup to another. The first setup was same as the previous experiment, but the second one consisted of four omnidirectional cameras that obtained 360 degree views of the room. Figure 4 shows the results of this experiment. For clarity, only one omni image is shown for every person. It is seen that even though the color responses for the cameras are different, the matching does give promising results. Further improvement can be expected by accounting for color differences between the camera sets.

5. Conclusion

This paper described a novel framework using weighted object-centered panoramic maps for performing person reidentification in multi-camera setups. Metrics for registering and comparing these maps accounting for the weighting were proposed. Experimental results using color features show promise of this approach.

Future work consists of performing a detailed theoretical analysis and experimental evaluation of the approach. An important issue is that the photometric responses of different cameras can be different due to the use of auto gain control and auto white balance. For such cases, one would have to perform photometric calibration or compensation before integrating the information from the cameras. However, even with the current experiments using two different sets of cameras, the matching does seem to be quite satisfactory to show the promise of using panoramic maps for reidentification. Voxel-based estimation of body shape could be used to improve the matching accuracy. Other features such as height and width of the person, texture of the clothes, as well as spatio-temporal constraints based on camera location and object velocities could be used to augment the matching and improve the reliability of the system.

The current system has been shown to handle a single person moving in the scene with no occlusion. However, due to the use of multiple cameras, the approach can be extended to work on scenes with multiple persons occluding each other. Multi-camera multi-person tracking can be easily implemented as described in [3]. In fact, if there is occlusion in one of the cameras, the other camera(s) would usually be able to disambiguate the people. In such case, one should be able to suppress the information obtained from the occluding camera by reducing the weights and use only the information from non-occluding camera. Since the information is accumulated over time, some loss of information should be tolerated.

Acknowledgment

We are grateful for the support of TSWG. We are most appreciative of the contributions and assistance of their colleagues from the CVRR laboratory, which made the experimental phase of the research possible. In particular, we thank Dr. Kohsia Huang for his valuable contributions in design of the experimental testbed and Shara Ebrahimi for her help in multi-camera calibration.

References

[1] N. D. Bird, O. Masoud, N. P. Papanikolopoulos, and A. Isaacs. Detection of loitering individuals in public transportation areas. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):167–177, June 2005.

[2] Q. Cai and J. K. Aggarwal. Tracking human motion in structured environments using a distributed-camera system. *IEEE Trans. Pattern Recogntion and Machine Intelligence*, 21(11):1241–1247, November 1999.

[3] K. C. Huang and M. M. Trivedi. Video arrays for real-time tracking of persons, head and face in an intelligent room. *Machine Vision and Applications*, 14(2):103–111, June 2003.

[4] T. Huang and S. Russell. Object identification: A bayesian analysis with application to traffic surveillance. *Artificial Intelligence*, 103(1-2):1–17, August 1998.

[5] O. Javed, Z. Rasheed, K. Shafique, and M. Shah. Tracking across multiple cameraswith disjoint views. In *Proc. IEEE International Conference on Computer Vision*, pages 1–6, June 2003.

[6] V. Kettnaker and R. Zabih. Bayesian multi-camera surveillance. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages II: 253–259, June 1999.

[7] A. Mittal and L. Davis. M2 tracker: A multi-view approach to segmenting and tracking people in a cluttered scene. *International Journal of Computer Vision*, 51(3):189–203, 2003.

[8] A. Utsumi and N. Tetsutani. Human tracking using multiple-camera-based head appearance modeling. In *Proc. IEEE Conf. Automatic Face and Gesture Recognition*, pages 657–662, May 2004.

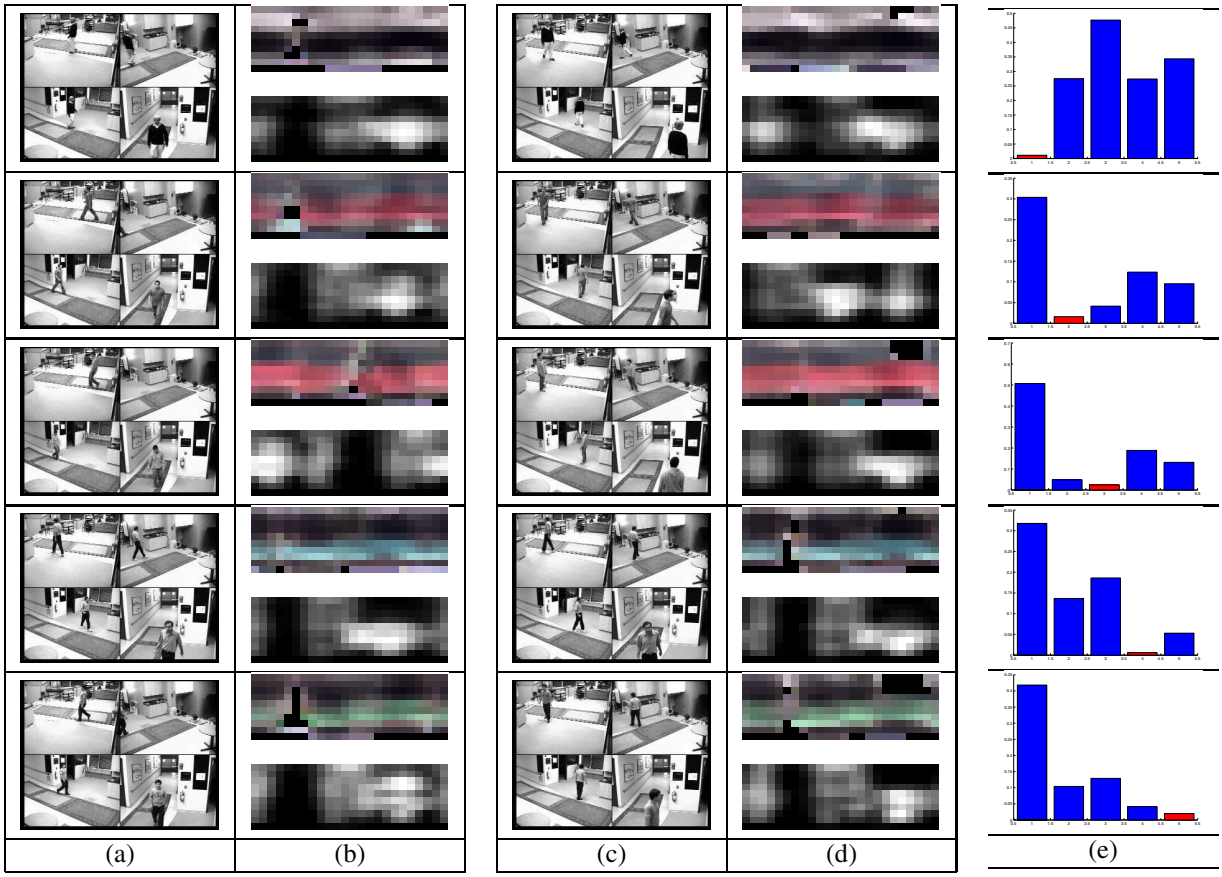

Figure 3. (a) Multi-camera image frames from training video sequences (b) Temporally integrated panoramic color and weight maps. The horizontal axis corresponds to the azimuth angle θ and the vertical axis corresponds to the height h. (c) Image frames from testing video sequences (d) Panoramic color and weight maps. (e) Distance metric from the test sample (c,d) to every training sample in (a,b). The bar for the training sample from same person is marked red.

Proceedings of the IEEE International Conference
on Video and Signal Based Surveillance (AVSS'06)
0-7695-2688-8/06 $20.00 © 2006 IEEE

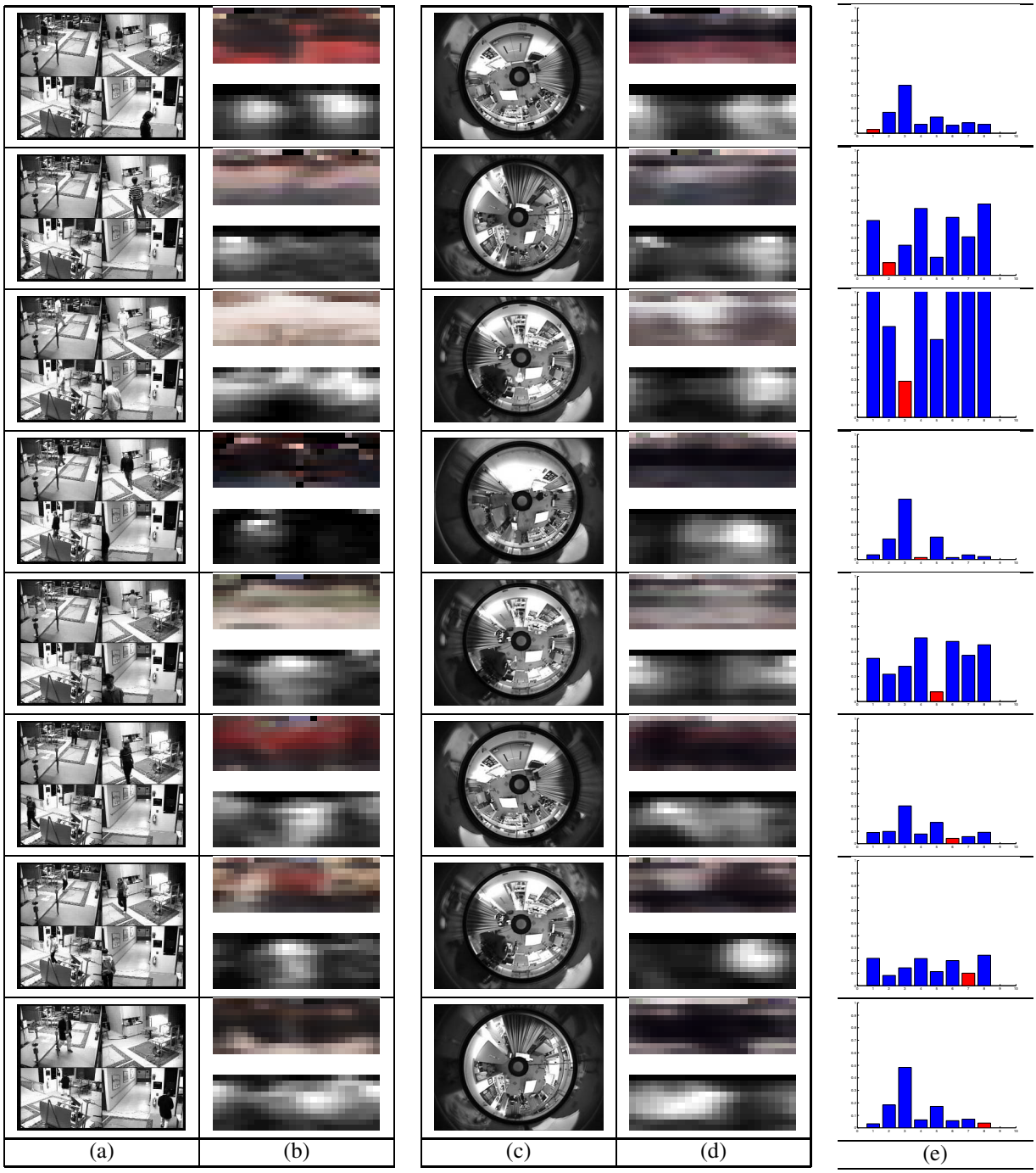

Figure 4. (a) Multi-camera image frames from training video sequences (b) Temporally integrated panoramic color and weight maps. The horizontal axis corresponds to the azimuth angle θ and the vertical axis corresponds to the height h. (c) Image frames from testing video sequences (d) Panoramic color and weight maps. (e) Distance metric from the test sample (c,d) to every training sample in (a,b). The bar for the training sample from same person is marked red.

Proceedings of the IEEE International Conference
on Video and Signal Based Surveillance (AVSS'06)
0-7695-2688-8/06 $20.00 © 2006 IEEE