Machine Vision and Applications (2007) 18:207–220
DOI 10.1007/s00138-006-0063-x

SPECIAL ISSUE PAPER

Person tracking and reidentification: Introducing Panoramic Appearance Map (PAM) for feature representation

Tarak Gandhi · Mohan Manubhai Trivedi

Received: 19 November 2005 / Accepted: 26 October 2006 / Published online: 13 February 2007

Abstract This paper develops a concept of Panoramic Appearance Map (PAM) for performing person reidentification in a multi-camera setup. Each person is tracked in multiple cameras and the position on the floor plan is determined using triangulation. Using the geometry of the cameras and the person location, a panoramic map centered at the person's location is created with the horizontal axis representing the azimuth angle and vertical axis representing the height. Each pixel in the map image gets color information from the cameras which can observe it. The maps between different tracks are compared using a distance measure based on weighted SSD in order to select the best match. Temporal integration by registering multiple maps over the tracking period improves the matching performance. Experimental results of matching persons between two camera sets show the effectiveness of the approach.

Keywords Video surveillance · Camera network · Color matching · Multiple view geometry · Visual tracking

This work has been sponsored by the Technical Support Working Group (TSWG) of US Department of Defence (DoD).

T. Gandhi (✉) · M. M. Trivedi
Computer Vision and Robotics Research Laboratory,
University of California, San Diego,
La Jolla, CA 92093, USA
e-mail: tgandhi@ucsd.edu

M. M. Trivedi
e-mail: mtrivedi@ucsd.edu

1 Introduction and motivation

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

An important problem in these applications is to reidentify objects that leave one camera (or a set of overlapping cameras) and enter another in a different area, or re-enter 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 introduces the novel concept of Panoramic Appearance Maps (PAM) useful for person reidentification in multi-camera setups. The preliminary version of this research was presented in [5]. Multiple overlapping cameras perform robust detection and 3D location estimation of objects in their field of view [2,14], and cover object features from all directions. The PAM extracts and combines information from all the cameras that view the object features to form a single signature. The horizontal axis of the map represents the azimuth angle with respect to the world coordinate system, and the vertical axis represents the object height above the ground plane. The maps generated from two different events can be compared to find potential matches. Presently, we use color information for comparison, but other appearance information such as texture could also be integrated in the framework. Temporal integration can be performed to register and blend a number of PAMs generated from an object over a period of time.

2 Related research

Visual surveillance has become a very active research area in recent years. Hu et al. [7] have conducted a comprehensive survey on research in visual surveillance. In particular the problem of person identification using biometric features is articulated, and the research using gait features is described in detail. Espina and Velastin [4] have surveyed the state of art in automated visual surveillance systems, In particular, the PRISMATICA system [21] developed by an EU funded project deals with security in large distributed spaces of public transportation networks, detecting interesting events in busy conditions. Remagnino et al. [15] have introduced the concept of "intelligent agents", which are autonomous modules that merge information from multiple cameras and incrementally build the scene model.

Table 1 describes the research in object tracking relevant to the reidentification problem. Initial work on object 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 [10] 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. Kogut and Trivedi [13] use color information, along with spatial organization in the form of platoons to match the vehicles while reducing the false matches. Trivedi et al. [18] describe vehicle matching between views two miles apart on the ends of a bridge using features of color, size, and time of transit.

Person tracking and reidentification are often more complex, since persons are articulated, move arbitrarily, and often wear multi-colored dresses. Kettnaker and Zabih [12] propose to use the similarity of views of the person, as well as plausibility of transition times from one camera to another in a Bayesian framework. Javed et al. [11] use various features based on space-time (entry/exit locations, velocity, travel time) and appearance (color histogram) in a probabilistic framework to identify best matches. The field of view boundaries between overlapping cameras are automatically identified. 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 dimensionality.

Multiple cameras with overlapping fields of view increase the reliability of tracking and reidentification due to better handling of occlusions, accurate estimates of the floor position and height of the persons, observation of features from multiple perspectives, and extraction of 3D information. This paper develops on several ideas from research on multi-camera systems. For example, Mittal and Davis [14] propose a multi-camera person tracking system called "M2-Tracker". They develop a region-based stereo algorithm that finds 3D points inside an object from the knowledge of regions belonging to object in two views. The color models of horizontal sections of the person are used for segmentation and association across frames, and could also be useful for reidentification over longer time intervals. Chang and Gong [3] use Bayesian networks to fuse information from multiple cameras for tracking persons. They maintain object identities during temporary occlusions using the shape and appearance models of the people. Wu and Mastuyama [22] use multiple cameras to obtain real-time voxel-based shape reconstruction of persons by volume intersection. In order to reduce computations, reconstruction is performed separately for group of parallel planes in the voxel space and cross sections of 3D volumes in each plane is reconstructed. Utsumi and Tetsutani [20] perform head-tracking with multiple cameras by creating an appearance model of the head as a set of color patches in 3D space. For every new frame, the current model is projected back to the 2D 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 the person reidentification 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 as well as weight is proposed to integrate and compare the panoramic maps.

Our research group at Computer Vision and Robotics Research Laboratory (CVRR) at the University of

Table 1 Related research on multi-camera based tracking and reidentification

System	Objective	Techniques
Vehicle reidentification		
Huang, AI98 [10]	Matching vehicles across widely separated views on freeway	Designs a probabilistic framework to match vehicles between pair of cameras. Uses features such as color, lane position, time of transit, and speed
Kogut, ITSC01 [13]	Vehicle and vehicle group matching	Uses color features along with spatial organization in the form of platoons for matching vehicles. Platoons are modeled as labeled undirected graphs. Matching is performed using an edit distance measure
Trivedi, IS05 [18]	Vehicle matching over large distances	Uses color statistics, shape, and time of transit between two cameras to identify potential matches and eliminate false alarms
2D based person reidentification in cameras with non-overlapping FOVs		
Kettnaker and Zabih, CVPR99 [12]	Reconstruction of paths of objects moving over non-overlapping cameras	Uses similarity of person views and plausibility of transition times from one camera to another in a Bayesian framework to reidentify persons between multiple cameras
Javed, ICCV03 [11]	Person reidentification between cameras with unknown spatial configuration	Uses Space-time and appearance features in a probabilistic framework to identify best matches. The FOV boundaries between overlapping cameras are automatically identified
Bird, ITS05 [1]	Detection of loitering behavior	Matches pedestrians intermittently spotted over a long time. Horizontal slicing of detected person is used to generate feature vector based on color
3D based person reidentification and analysis in multi-camera systems with overlapping FOVs		
Mittal, IJCV03 [14] (M2-Tracker)	Tracking of people in presence of severe occlusions and clutter	Novel region-based stereo algorithm to find 3D points inside the object. Color distribution at multiple heights is used for segmentation and association across frames. Bayesian classification and occlusion analysis are used for combining evidences from cameras
Wu, ISPA03 [22]	3D shape reconstruction	Performs voxel-reconstruction based on volume intersection. Divides the object using parallel planes to reduce computations.
Chang, ICCV01 [3]	Multi-camera person tracking	Use Bayesian networks to fuse information from multiple cameras for tracking persons. Maintains object identities during temporary occlusions using shape and appearance models of people
Utsumi, FGR04 [20]	Head tracking using appearance models	Create appearance model of the head as a set of color patches in 3D space. Projects the model on each camera and matches with new image frames to localize head position.
Huang, V4HCI05 [9]	Gesture recognition using 3D voxel based shape context	Persons tracked using multiple omnidirectional cameras and a multi-layer cylindrical histogram based on voxelization is used to identify gestures
This paper	Person reidentification in multi-camera setup	Generates a PAM multiple modeling the appearance of the person from all sides using cameras. Finds matches by displacing one map (corresponding to rotation around vertical axis) and comparing with the other map

California San Diego has been working on distributed interactive video array (DIVA) systems for several years. In [19], a general framework analyzing human activities at multiple abstraction levels is presented. The infrastructure and experimental testbed are described in detail, followed by the description of modules for multiple camera based person tracking, event detection and event based servoing for selective attention, voxelization, and streaming face recognition. The subsystem dealing with person detection and tracking based on multiple rectilinear and omnidirectional cameras is described in more detail in [8]. Background subtraction is used to detect people in all cameras. Correspondence is established between the detected image locations and triangulation is performed to compute the height of the persons as well as their location on the floor plan. A novel approach of analyzing human gesture using 3D Shape Context, which is a the multilayer cylindrical histogram based on voxelization of human body, is described in [9]. The concept of PAM introduced here is complementary to the 3D Shape Context since the latter uses the volumetric information whereas the former uses the surface appearance information.

3 Person reidentification using panoramic appearance map

The block diagram of the person reidentification approach is shown in Fig. 1. The person is detected in each camera using background subtraction. Multi-camera triangulation is used to obtain the person's position on the floor. An object surface model in form of a generalized cylinder is placed at the person's location. Using this model, parametric grid of the object surface is generated along the azimuthal and height directions. The grid is then projected onto all the cameras where the object is visible, and the image patches corresponding to each of the elements in the grid are extracted. The features obtained from the image patches from all cameras corresponding to each grid element are integrated to form a PAM.

Person detection is performed in each camera using background subtraction. The background image is generated using temporal averaging 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 person's location in the floor plan is then obtained by triangulating on the image locations obtained from the cameras that can observe the person [8]. 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:

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

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 3D 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 p_0 minimizing the sum of squares of the perpendicular distance to all rays is estimated as:

$$p_0 = \arg\min_p \min_{\lambda_k} \left[p - \lambda_k R_k p_k - t_k\right] \tag{3}$$

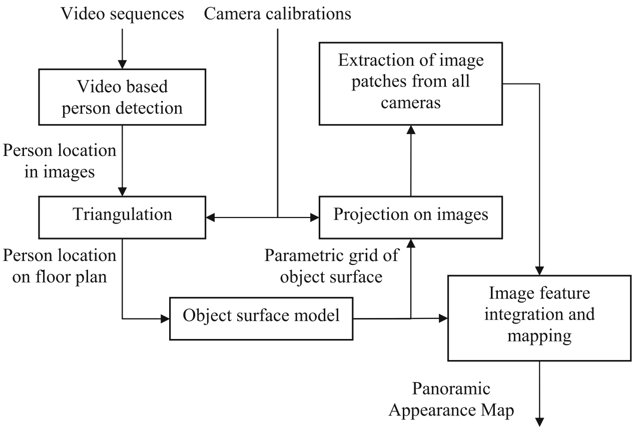

Video sequences
Camera calibrations
Video based person detection
Person location in images
Triangulation
Person location on floor plan
Object surface model
Parametric grid of object surface
Projection on images
Extraction of image patches from all cameras
Image feature integration and mapping
Panoramic Appearance Map

Fig. 1 Block diagram for generation of PAM

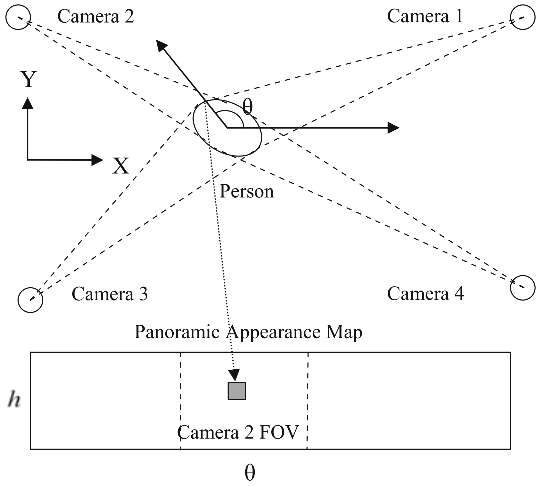


Fig. 2 Mapping from 3D space (*plan view*) to PAM

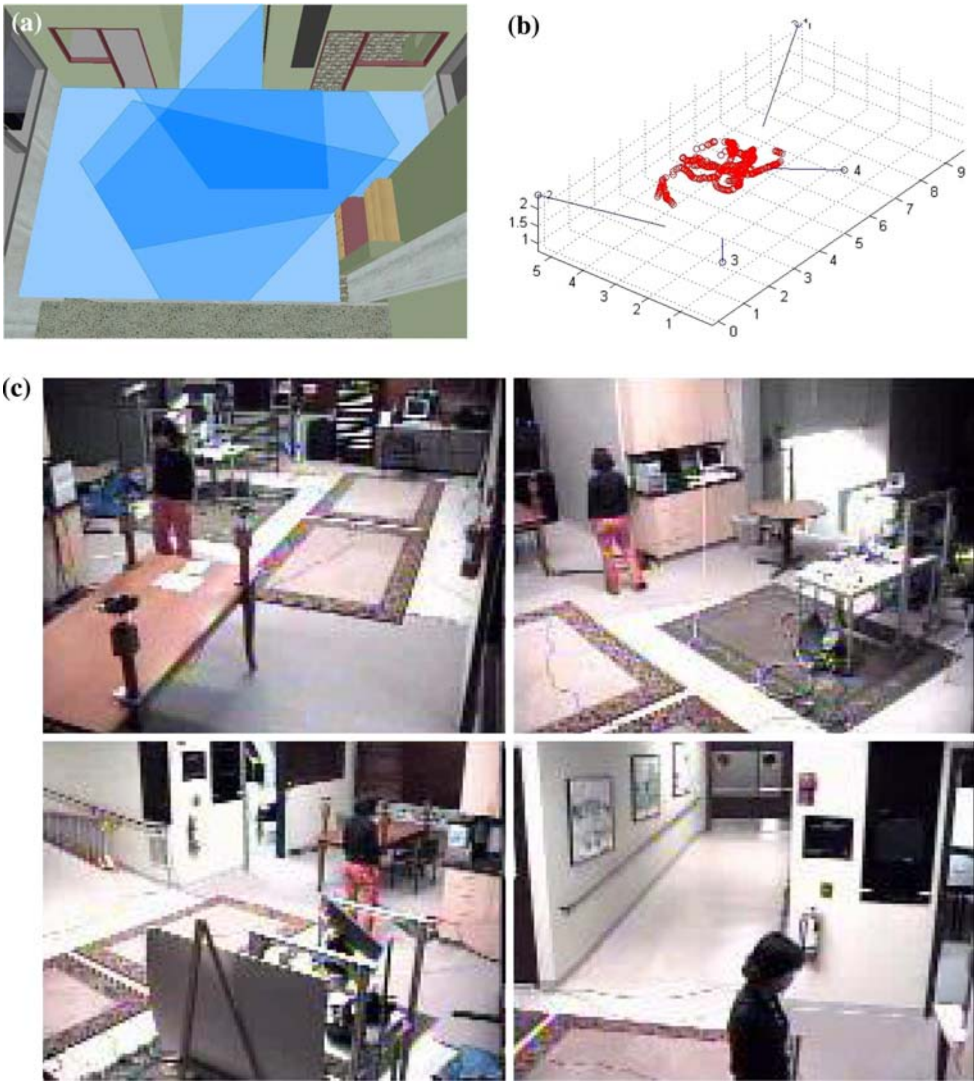


Fig. 3 **a** Camera coverage in SHIVA laboratory. Light to dark colors show coverage with 1–4 cameras. **b** Camera calibration computed using [16], showing position and orientation of the cameras in 3D. The trajectory of the laser pointer used for calibration is marked with red. **c** Images of the same person from different viewpoints acquired from the setup

The accuracy of the 3D location is very important for reliable projection and fusion of appearance information from the cameras. To ensure that the projection is accurate, only the frames where three or more cameras detect the object are used. In addition, an estimate of the triangulation error is determined by taking the maximum perpendicular distance of all of the camera rays from the estimated point. All the frames where the triangulation error was greater than a threshold of 0.15 m are rejected from further analysis, so that they do not corrupt the subsequent steps.

In order to capture the appearance of the person from all directions, we model the person's body as a convex generalized cylinder shown in Fig. 2. A point p on its surface is parameterized by the azimuth angle θ and height h as:

$$p = (x, y, z)^{\mathrm{T}} = (x_0 + r(\theta, h)\cos\theta, y_0 + r(\theta, h)\sin\theta, h)^{\mathrm{T}} \tag{4}$$

where r is a function representing the cross section of the cylinder at height h and $p_0 = (x_0, y_0, 0)^{\mathrm{T}}$ is the center of the cylinder projected on the ground. Hence, if r is modeled or computed for every θ and h, one has a transformation from (θ, h) to image pixels (u_k, v_k) for every camera.

In our experiments, we have modeled the cross section of the body as a circle with constant radius for the sake of simplicity. However, a voxel-based reconstruction may give more realistic model of the complex body shape at the cost of complexity.

The parameters h and θ are discretized into a grid of $M \times N$ elements. Each element indexed (m, n) corresponds to the ranges $m\Delta h \leq h < (m+1)\Delta h$ and $n\Delta\theta \leq \theta < (n+1)\Delta\theta$. The corners of each grid element in the panoramic map are transformed using Eq. (4) to the world coordinates (X, Y, Z) of the point on the surface of the cylinder. This point is then projected onto the image plane of each camera k in which the element is

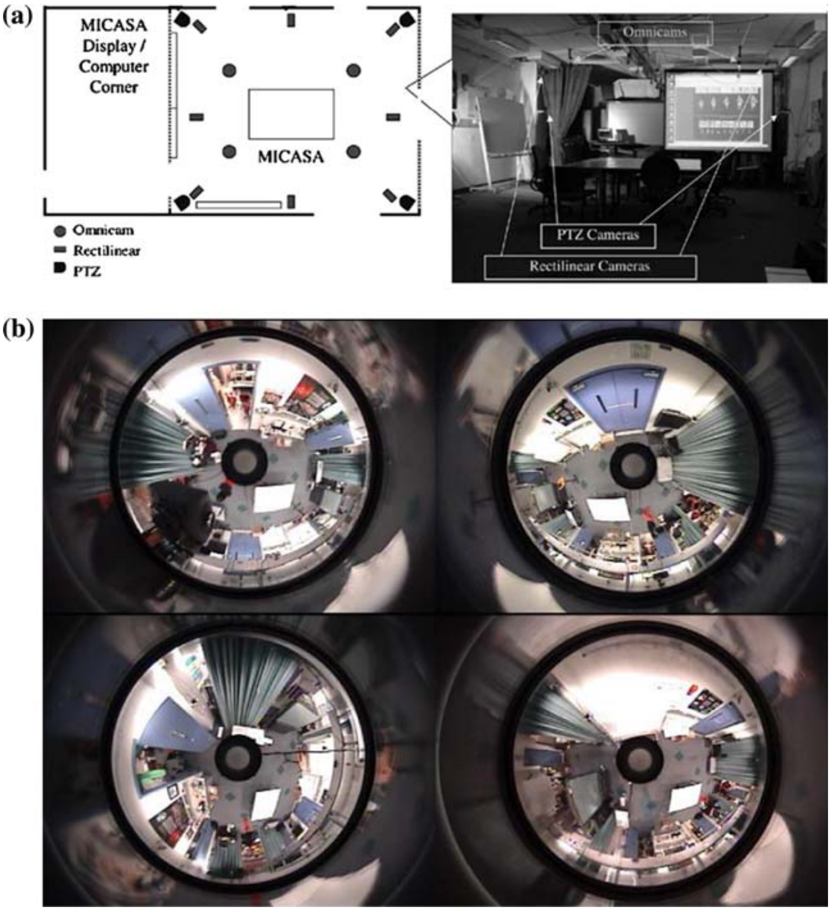


Fig. 4 **a** Floor plan and camera network in MICASA laboratory. **b** Images acquired from the omnidirectional cameras

visible, using the above world to image transformation for that camera. This defines a region in the image plane corresponding to the grid element. If $\mathbf{c}_k(m,n)$ 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 $(\mathbf{c}_k, w_k)$ is used to represent the appearance information from image k. This way, each grid element not only contains the appearance information, but the weight that specifies confidence of the grid element, which is used during integration and matching. This reduces the errors due to unreliable information from parts of the images.

The use of weights also reduces the errors due to the approximate model of the human body in form of the cylinder. For any given view, the part of the body directly facing the camera is least affected by the modeling error and contributes most to the appearance map. On the other hand, parts of the body that are on the sides are most sensitive to the error in cylindrical model but contribute least to the appearance map.

If multiple cameras can view the point p, then the individual appearance maps $(\mathbf{c}_k, w_k)$ from all cameras can be combined as:

$$w(m,n) = \sum_k w_k(m,n)$$

$$\mathbf{c} = \frac{\sum_k \mathbf{c}_k(m,n) w_k(m,n)}{\sum_k w_k(m,n)} \tag{5}$$

Two such maps from different events can be compared using a distance measure. In place of the usual sum of squares differences, we suggest a measure which takes into consideration the weights of both the maps. Also, for rotational invariance, the distance is computed for all discrete rotation angles and the minimum distance is used. The distance d between two panoramic maps $A = (\mathbf{c}^{(a)}, w^{(a)})$ and $B = (\mathbf{c}^{(b)}, w^{(b)})$, displaced by an angle $\theta(n) = n\Delta\theta$ is given by:

$$d(A,B;n) = \frac{s(\mathbf{c}^{(a)}, \mathbf{c}^{(b)}; n)}{w(\mathbf{c}^{(a)}, \mathbf{c}^{(b)}; n)} \tag{6}$$

with

$$s(\mathbf{c}^{(a)}, \mathbf{c}^{(b)}; n) = \sum_{m=0}^{M-1} \sum_{n'=0}^{N-1} \left(\mathbf{c}^{(a)}(m,n') - \mathbf{c}^{(b)}(m,n+n')\right)^2 \times \left(w^{(a)}(m,n') w^{(b)}(m,n+n')\right)$$

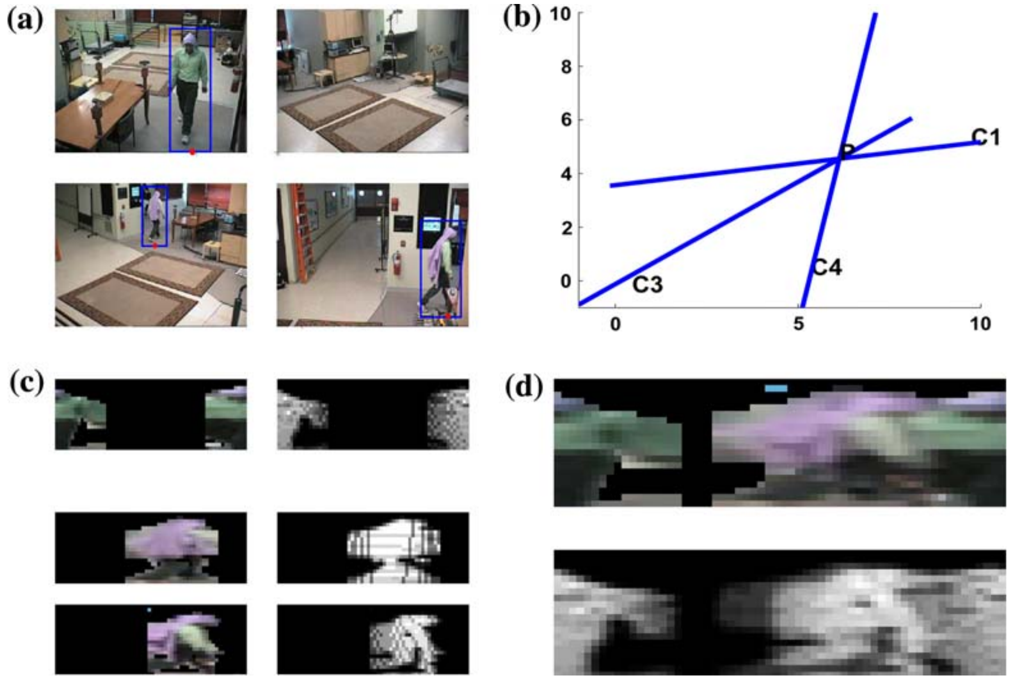


Fig. 5 **a** Detection of person from three of the four cameras capturing the same scene. **b** Estimation of person's location using triangulation from cameras in which the person is detected. **c** Appearance maps from individual cameras. The horizontal axis corresponds to the azimuth angle θ and the vertical axis corresponds to the height h. The color image is the feature component **c** and the monochrome image is the weight component (w). Each of these cover only a part of the panorama around the person's body. **d** PAM formed by combining appearance maps from all cameras

$$w(\mathbf{c}^{(a)}, \mathbf{c}^{(b)}; n) = \sum_{m=0}^{M-1} \sum_{n'=0}^{N-1} \left(w^{(a)}(m, n') w^{(b)}(m, n+n') \right) \tag{7}$$

where $n+n'$ is taken modulo N so that it lies between 0 and $N-1$. In this formula, the sum of square difference $\left[\mathbf{c}^{(a)}(m,n') - \mathbf{c}^{(b)}(m,n+n')\right]^2$ between individual grid elements of the map is multiplied by the product of the weights of both the maps $\left[w^{(a)}(m,n')w^{(b)}(m,n+n')\right]$. Due to this, a larger emphasis is given to points having larger weights in *both* maps, while suppressing points where one or both have small weights. This reduces the effect of spurious grid elements information with unreliable on the overall distance measure.

The distance $d(A,B)$ corresponding to the best match index n_0 is obtained by taking the minimum over all $n = 0, \ldots, N-1$:

$$\begin{aligned} n_0 &= \arg \min_{n=0}^{N-1} d(A,B;n) \\ d(A,B) &= d(A,B;n_0) \end{aligned} \tag{8}$$

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

$$\begin{aligned} W(m,n) &\leftarrow \alpha w(m,n+n_0) + (1-\alpha)W(m,n) \\ \mathbf{C}(m,n) &\leftarrow \frac{\alpha \mathbf{c}(m,n+n_0)w(m,n+n_0) + (1-\alpha)\mathbf{C}(m,n)W(m,n)}{\alpha w(m,n+n_0) + (1-\alpha)W(m,n)} \end{aligned} \tag{9}$$

If the matching is accurate, the temporally integrated maps over the entire event corresponding to a walking person could significantly reduce the noise and fill the blank spaces where information from per-frame maps is not reliable. In the case of slight mis-registration, the integrated map can get blurred, but still capture the basic color information from all directions around the person.

4 Experimental analysis and validation

The application of PAMs for person reidentification was tested on the distributed interactive video array (DIVA) configurations at our laboratory facility, spread over a number of rooms in the building. One of the arrays consists of four cameras mounted in a large indoor space called the Systems for Human Interaction, Visualization, and Analysis (SHIVA) at the CVRR labora-

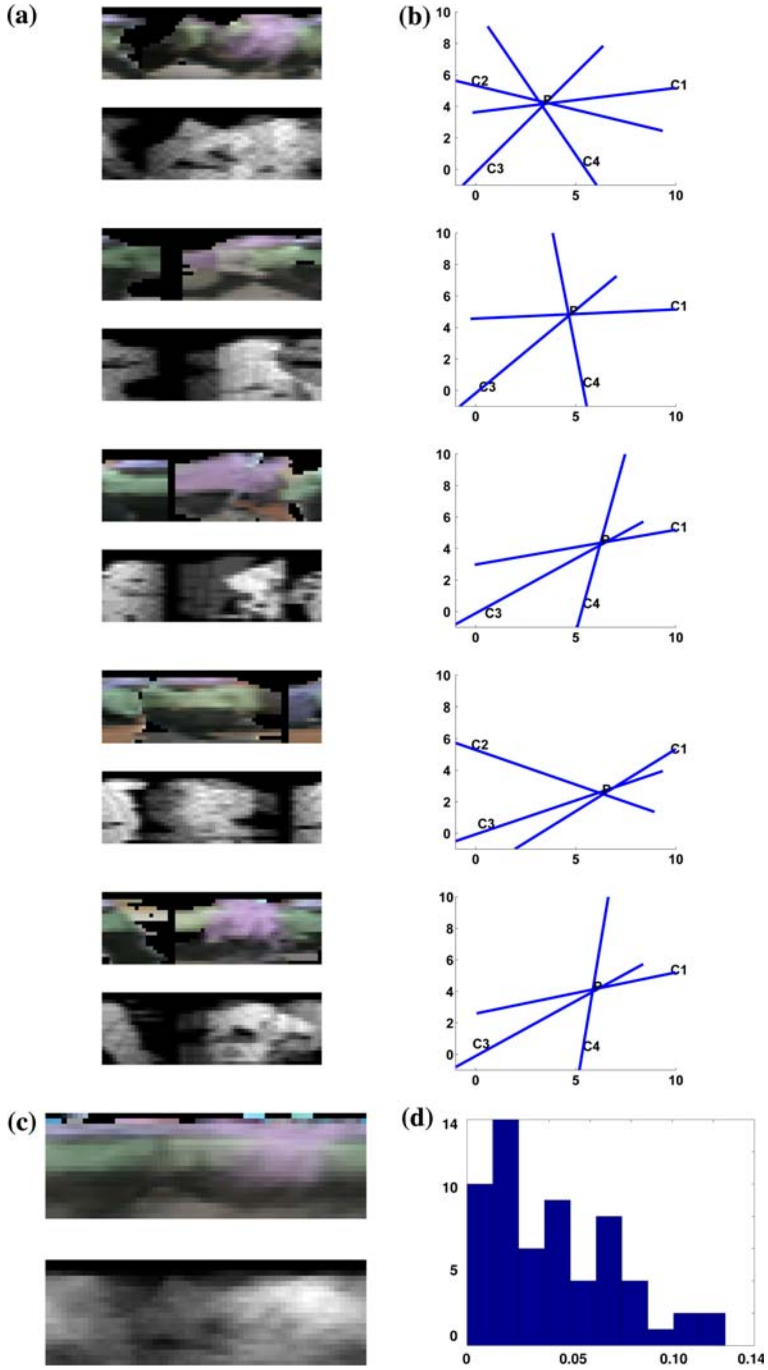


Fig. 6 **a** Multi-view PAMs from simultaneously captured images from multiple cameras. The color image is the feature component **c** and the monochrome image is the weight component w. **b** Sample results of triangulation. **c** Temporally integrated PAM that captures the features of the person's dress in more robust manner. **d** *Histogram* of triangulation error residual. Note that frames with error residual greater than 0.15 m were removed from further processing

tory. Videos from these cameras were multiplexed and collected using a single PC.

In order to obtain accurate 3D locations of the tracked people, the cameras are calibrated with respect to each other. Two approaches were explored for calibrating the cameras. Svoboda et al. [16] have designed a fully automatic approach for calibrating multiple cameras using a freely moving bright spot generated from a laser pointer as a calibration object. The calibration software is publicly available at [17]. On the other hand, [6] use the image location of moving person in all cameras over multiple frames to obtain an approximate calibration of the cameras. Figure 3 shows the computed positions and orientations of the cameras in a virtual environment.

The second array consists of four omnidirectional cameras in the MICASA room [8] of the CVRR laboratory as shown in Fig. 4. Since the omni cameras cover a 360° field of view, the object size in the image was comparatively smaller. Calibration of this setup was

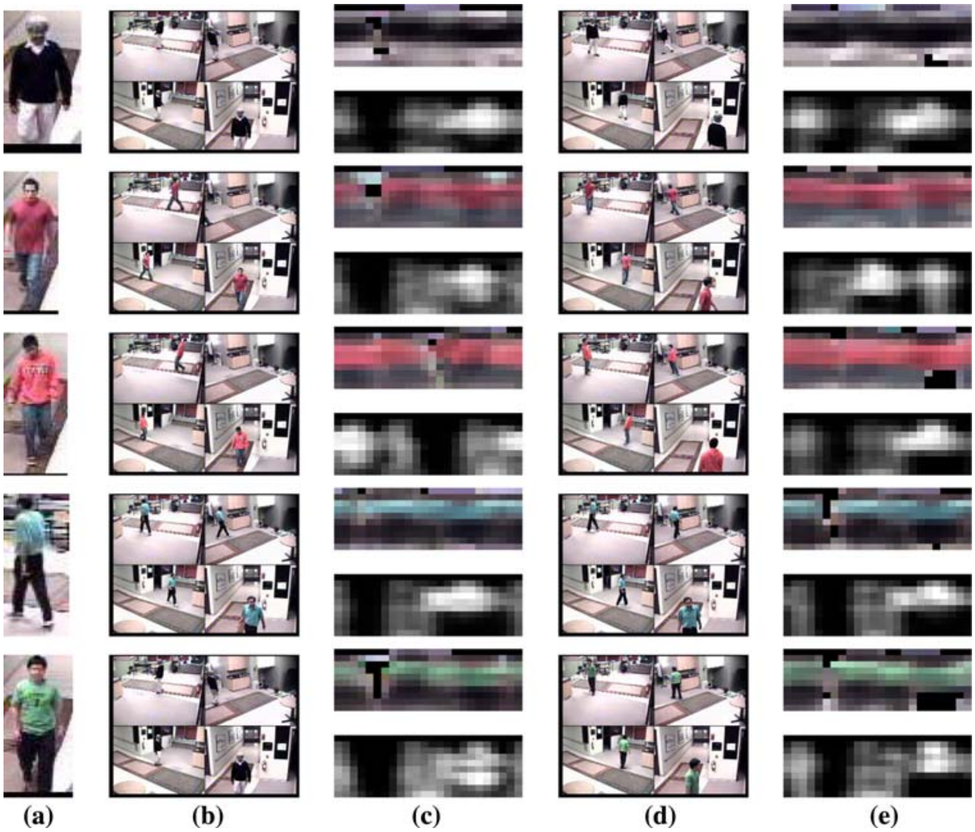
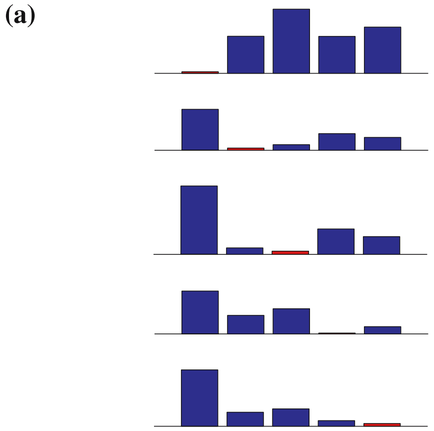


Fig. 7 **a** Test subjects **b** multi-camera image frames from training video sequences **c** Panoramic color and weight maps from temporal integration of training sequences. **d** Image frames from testing video sequences, **e** Panoramic color and weight maps from temporal integration of testing sequences



(b)

	1	2	3	4	5
1	**0.0113**	0.2752	0.4769	0.2740	0.3435
2	0.3038	**0.0160**	0.0414	0.1234	0.0955
3	0.5071	0.0483	**0.0241**	0.1884	0.1315
4	0.3177	0.1368	0.1862	**0.0060**	0.0531
5	0.4183	0.1040	0.1290	0.0414	**0.0198**

Fig. 8 Distance between the training and testing samples shown as **a** bar chart and **b** table. *Each row* corresponds to a testing video and *each column* to a training video. The *bar* showing the distance for the same training and testing videos is marked *red* and is shorter than the other bars

performed manually by physically measuring the camera locations, and inferring the orientations from the omni images.

Accordingly, we study two reidentification scenarios:

1. Person leaving a camera array comes back in the same array.
2. Person leaving one camera array goes into another array.

Figure 5 illustrates the generation of PAMs Fig. 5a shows sample images simultaneously captured from the cameras in the rectilinear array looking at a person from different directions. The person is detected in three of the four cameras using background subtraction. The location of the person on the floor plane is obtained using triangulation as shown in Fig. 5b. The appearance maps generated by the individual cameras are shown in Fig. 5c, each of these capturing the person's appearance as viewed from one direction. These individual maps are combined in Fig. 5d to form the PAM, which captures the appearance of the person from all sides.

A number of PAMs obtained from individual frames (all cameras) taken over a time interval were integrated using registration and temporal averaging described in the previous section. Figure 6a shows some of the per-frame PAMs with triangulation shown in Fig. 6b. The temporally integrated PAM is shown in Fig. 6c. It is seen that the per-frame PAMs are sharper but noisier and have more blank spaces where no information is avail-

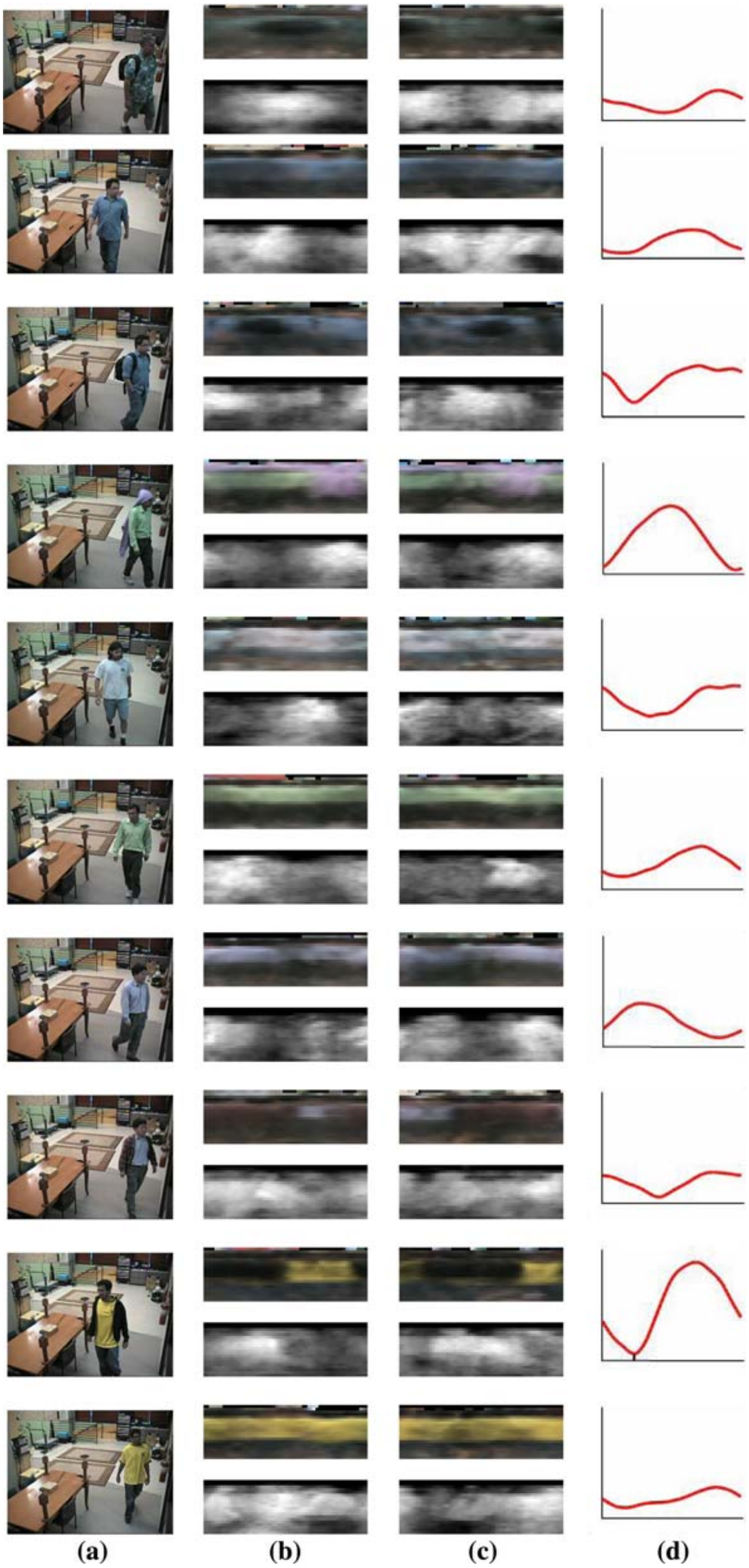
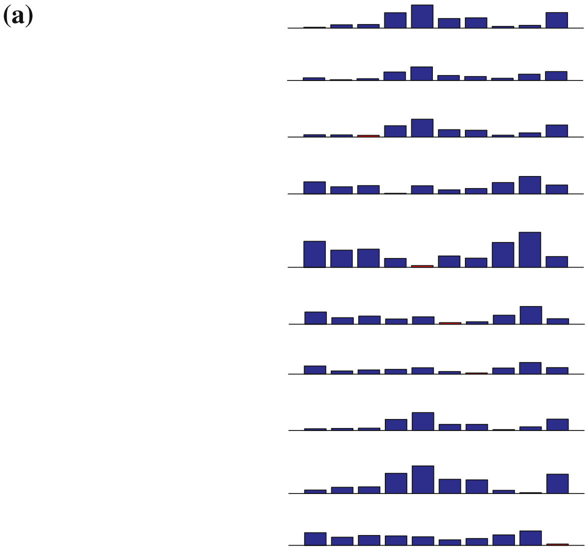


Fig. 9 Experiment showing formation of PAMs for persons wearing multi-colored dresses and backpacks. **a** Sample snapshots of test subjects. **b** Temporally averaged panoramic color and weight maps from training sequence. **c** Temporally averaged panoramic color and weight maps from testing sequence. **d** Distance metric between **b** and **c** plotted against relative rotation between them (0 to 360 degrees). It is seen that the valley corresponding to best alignment is prominent when there is sharp color variation in longitudinal direction as in tests 4 and 9

able. On the other hand, the temporally averaged PAM is some what blurred but has less noise and blank spaces. In this case, the integrated PAM captures the hood on the person's head as well as the robe on the side. On the other hand, the per-frame PAMs are noisier and do not always capture this information. Note that the frames with triangulation error residual greater than a threshold of 0.15 m were rejected to minimize corruption of the

Fig. 10 Distance matrix between the training and testing samples in **a** graphical and **b** tabular form. *Each row* corresponds to a testing video and *each column* to a training video. The *bar* for the training sample from same person is marked *red* and has minimum height in all cases



(b)

	1	2	3	4	5	6	7	8	9	10
1	**0.0107**	0.0404	0.0436	0.1895	0.2855	0.1175	0.1269	0.0210	0.0334	0.1911
2	0.0316	**0.0075**	0.0203	0.1036	0.1670	0.0598	0.0484	0.0268	0.0771	0.1084
3	0.0278	0.0258	**0.0200**	0.1408	0.2226	0.0916	0.0834	0.0223	0.0517	0.1498
4	0.1480	0.0875	0.1012	**0.0059**	0.1007	0.0490	0.0658	0.1399	0.2162	0.1086
5	0.3216	0.2118	0.2252	0.1092	**0.0210**	0.1389	0.1124	0.3064	0.4340	0.1317
6	0.1510	0.0805	0.1014	0.0637	0.0915	**0.0182**	0.0296	0.1119	0.2204	0.0672
7	0.0997	0.0385	0.0505	0.0594	0.0787	0.0318	**0.0131**	0.0755	0.1436	0.0794
8	0.0206	0.0240	0.0285	0.1360	0.2207	0.0754	0.0739	**0.0092**	0.0433	0.1395
9	0.0428	0.0772	0.0822	0.2502	0.3438	0.1756	0.1698	0.0395	**0.0099**	0.2398
10	0.1550	0.0995	0.1223	0.1163	0.1061	0.0676	0.0841	0.1280	0.1773	**0.0150**

temporally integrated PAM with unreliable 3D location information. The histogram of triangulation errors in the reliable frames is shown in Fig. 6d.

The concept of PAMs was applied for performing reidentification in both types of scenarios. In the first configuration (Fig. 3), four cameras spanned the indoor space with overlapping FOVs giving coverage from all directions. Figure 7a shows the test subjects and Fig. 7b shows typical frames from the video sequence corresponding to the particular subject. The person was modeled using a circular cross section of fixed radius. Figure 7c shows the temporally integrated PAM based on color feature. Figure 7d,e 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 between all test samples and training samples are shown graphically and numerically as a matrix in Fig. 8. It is seen that the true matches give the best distance measure in most cases. However, there is some ambiguity due to different persons wearing similar colored dresses. For example, subjects 2 and 3 as well as 4 and 5 have small distance measure between training and testing samples. Other features such as person height and texture may be used in such cases to disambiguate such matches.

The experiment shown in Fig. 9 describes the generation of PAM in cases where many of the persons are wearing multi-colored dresses or having backpacks that result in multiple colors around the periphery of the person. Figure 9a shows a typical snapshot of the person acquired from one of the cameras. Figure 9b,c show the PAM generated in the training and testing cases. It is seen that the PAM captures the multiple colors and their relative positions along the θ direction. The PAMs were compared using the distance metric by performing circular shift of one of the maps. Figure 9d shows the plot of the distance metric against the rotation angle. Valley of this graph corresponds to the best alignment between the PAMs. It is seen that the valley is more prominent when there is sharp color variation in the longitudinal direction as seen for Examples 4 and 9. Figure 10 shows the distance matrix obtained by matching the test image with all of the training images, with red showing the case of same person. It is seen that the reidentification is correct in most of the cases.

In the second configuration, a number of persons moved from the rectilinear camera array (Fig. 3) to the omnidirectional camera array (Fig. 4). The results of this experiment are shown in Fig. 11. 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. The omni videos were then replaced by other videos from the rectilinear array. Since the testing and training set are now

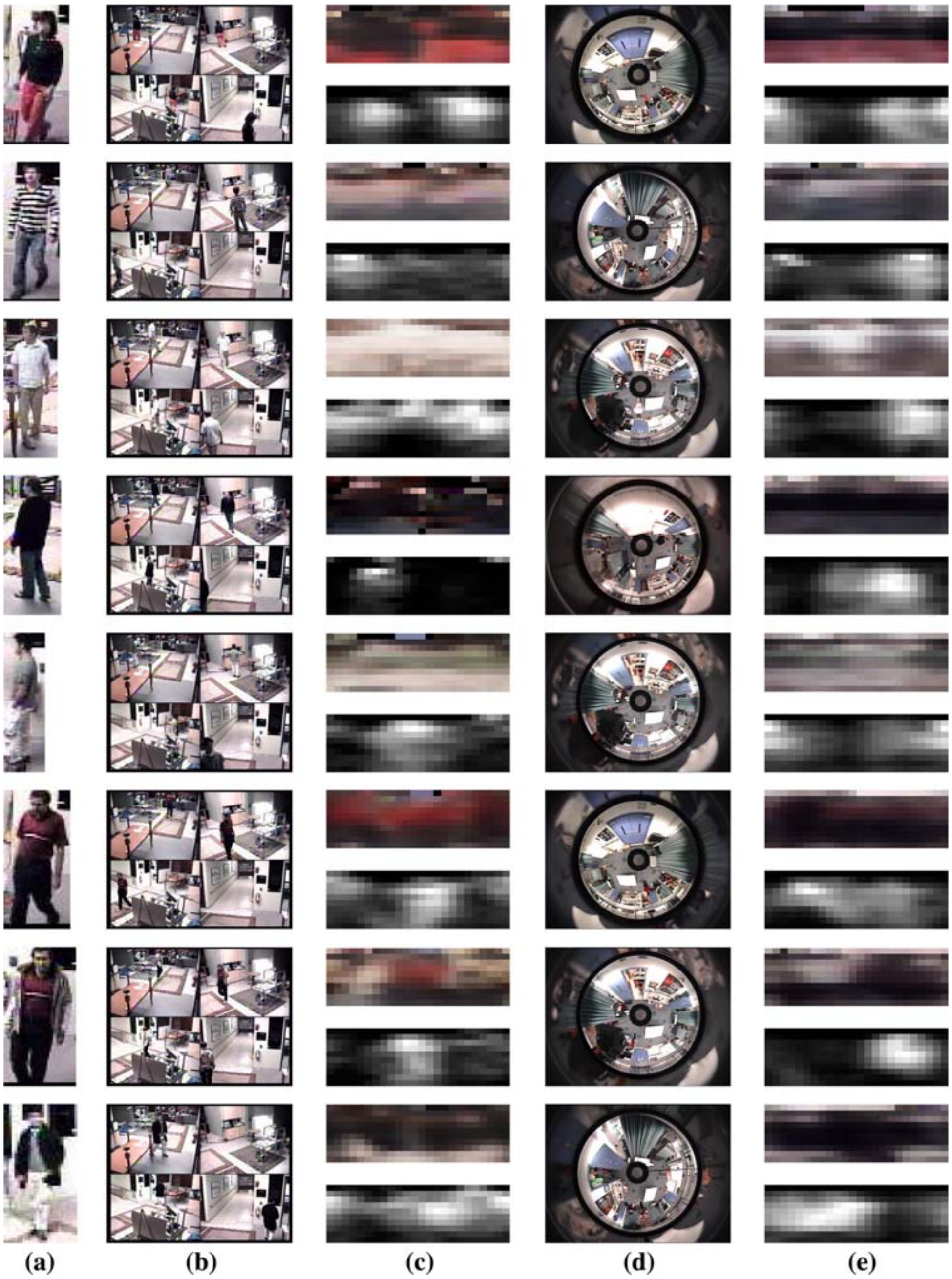


Fig. 11 **a** Test subjects **b** Multi-camera image frames from training video sequences in SHIVA lab. **c** Panoramic color and weight maps from training video. **d** Image frames from testing video sequences from omni cameras in MICASA lab. **e** Panoramic color and weight maps from testing video

from the same camera, the results are obviously better as seen from the distance matrices in Fig. 12 a,b.

5 Conclusion

This paper described a novel framework introducing PAMs for performing person reidentification in multi-camera setups. These maps can capture appearance information from all around a person's body and represent it as a compact signature. A novel distance measure based on introducing weights to the sum of squared differences was proposed for registering and comparing these maps. The use of weighting enables discarding information from parts of the map where it is unreliable. Experiments were performed with persons moving from one scene spanned by multiple overlapping cameras and then appearing in the same scene or another scene. Color was used as appearance feature in these experiments where people wore a variety of multi-colored dresses. It was observed that the panoramic appearance map captures the colors as well as

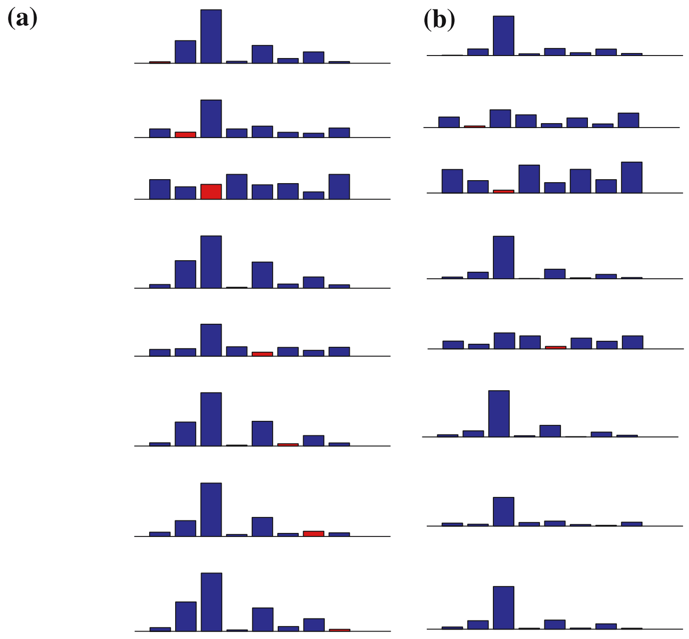


Fig. 12 Confusion matrix of distances between test samples (*rows*) and training samples (*columns*) for **a** Matching between omni camera for test samples and conventional cameras for training samples. **b** Matching with omni test samples replaced by those from conventional camera. *Bars* from the same training and test samples are marked *red*. As expected, these bars are shorter than the other bars. Also, performance in **b** is better than **a** due to similar camera and lighting conditions. **c**,**d** Corresponding tables showing distance numerically

(c)

	1	2	3	4	5	6	7	8
1	**0.0297**	0.4389	1.4312	0.0381	0.3451	0.0899	0.2188	0.0329
2	0.1670	**0.1016**	0.7258	0.1649	0.2186	0.0987	0.0819	0.1835
3	0.3821	0.2414	**0.2887**	0.4827	0.2810	0.3025	0.1429	0.4831
4	0.0707	0.5348	1.5763	**0.0168**	0.5089	0.0781	0.2168	0.0644
5	0.1292	0.1453	0.6220	0.1798	**0.0774**	0.1697	0.1126	0.1719
6	0.0625	0.4639	1.4514	0.0164	0.4801	**0.0433**	0.2005	0.0579
7	0.0832	0.3084	1.1359	0.0375	0.3690	0.0571	**0.1003**	0.0700
8	0.0701	0.5701	1.5532	0.0244	0.4531	0.0915	0.2437	**0.0388**

(d)

	1	2	3	4	5	6	7	8
1	**0.0052**	0.1239	0.7508	0.0324	0.1339	0.0531	0.1225	0.0389
2	0.1972	**0.0290**	0.3353	0.2392	0.0726	0.1796	0.0674	0.2717
3	0.4463	0.2355	**0.0527**	0.5313	0.1957	0.4504	0.2544	0.5887
4	0.0311	0.1242	0.8085	**0.0045**	0.1822	0.0161	0.0830	0.0235
5	0.1496	0.0881	0.3033	0.2480	**0.0457**	0.2057	0.1443	0.2481
6	0.0407	0.1136	0.8785	0.0229	0.2182	**0.0039**	0.0919	0.0306
7	0.0555	0.0355	0.5470	0.0671	0.0937	0.0274	**0.0122**	0.0720
8	0.0409	0.1597	0.8124	0.0167	0.1735	0.0196	0.0998	**0.0151**

their spatial arrangement. The persons were reidentified successfully by comparing the appearance maps.

For future work, we would like to explore the voxel-based estimation of body shape for improving the model accuracy. Other features such as height and width of the person, texture and edges from 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.

Acknowledgments We thank the Technical Support Working Group (TSWG) of the US Department of Defence (DOD) for the sponsorship of the research. We acknowledge the support of the members at the CVRR laboratory.

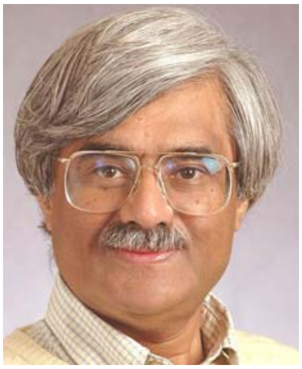
References

1. Bird, N.D., Masoud, O., Papanikolopoulos, N.P., Isaacs, A.: Detection of loitering individuals in public transportation areas. IEEE Tran. Intell. Transportation Sys. **6**(2), 167–177 (2005)
2. Cai, Q., Aggarwal, J.K.: Tracking human motion in structured environments using a distributed-camera system. IEEE Trans. Pattern Recogn. Mach. Intell. **21**(11), 1241–1247 (1999)
3. Chang, T., Gong, S.: Tracking multiple people with a multi-camera system. In: Proceedings of IEEE ICCV Workshop on Multi-Object Tracking. Vancouver (2001)
4. Espina, M.V., Velastin, S.A.: Intelligent distributed surveillance systems: A review. IEE Proceedings — Vision, Image and Signal Processing **152**(2), 192–204 (2005)

5. Gandhi, T., Trivedi, M.M.: Panoramic Appearance Map (PAM) for multi-camera based person re-identification. In: Proceedings of IEEE International Conference on Advanced Video and Signal-Based Surveillance. Sydney (2006)
6. Gandhi, T., Trivedi, M.M.: Reconfigurable omnidirectional camera array calibration with a linear moving object. Image Vis. Comput. **24**(9), 935–948 (2006)
7. Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. IEEE Trans. Syst. Man Cybern. Part C **34**(3), 334–352 (2005)
8. Huang, K.C., Trivedi, M.M.: Video arrays for real-time tracking of persons, head and face in an intelligent room. Mach. Vis. Appl. **14**(2), 103–111 (2003)
9. Huang, K.S., Trivedi, M.M.: 3D shape context based gesture analysis integrated with tracking using omni video array. In: IEEE Workshop on Vision for Human-Computer Interaction (V4HCI). San Diego, CA (2005)
10. Huang, T., Russell, S.: Object identification: A Bayesian analysis with application to traffic surveillance. Artif. Intell. **103** (1–2), 1–17 (1998)
11. Javed, O., Rasheed, Z., Shafique, K., Shah, M.: Tracking across multiple cameras with disjoint views. In: Proceedings of IEEE International Conference on Computer Vision, pp. 1–6 (2003)
12. Kettnaker, V., Zabih, R.: Bayesian multi-camera surveillance. In: IEEE Conference on: Computer Vision and Pattern Recognition **II**, 253–259 (1999)
13. Kogut, G., Trivedi, M.M.: Maintaining the identity of multiple vehicles as they travel through a video network. In: Proceedings of IEEE International Conference on Intelligent Transportation Systems, pp. 29–34. Oakland, California (2001)
14. Mittal, A., Davis, L.: M2 tracker: a multi-view approach to segmenting and tracking people in a cluttered scene. Int. J. Comput. Vis. **51**(3), 189–203 (2003)
15. Remagnino, P., Shihab, A., Jones, G.: Distributed intelligence for multi-camera visual surveillance. Pattern Recogn. **37**(4), 675–689 (2004)
16. Svoboda, T., Martinec, D., Pajdla, T.: A convenient multi-camera self-calibration for virtual environments. Presence Teleop. Virtual Environ. **14**(4), 407–422 (2005)
17. Svoboda, T., Martinec, D., Pajdla, T., Bouguet, J.Y., Werner, T., Chum, O.: Multi-Camera Self-Calibration. Czech Technical University, Prague, Czech Republic. http://cmp.felk.cvut.cz/ svoboda/SelfCal/
18. Trivedi, M.M., Gandhi, T., Huang, K.S.: Distributed interactive video arrays for event capture and enhanced situational awareness. IEEE Intell. Sys. Spec. Issue AI Homeland Security **20**(5), 58–66 (2005)
19. Trivedi, M.M., Huang, K.S., Mikic, I.: Dynamic context capture and distributed video arrays for intelligent spaces. IEEE Trans. Syst. Man Cybern. Part A **35**(1), 145–163 (2005)
20. Utsumi, A., Tetsutani, N.: Human tracking using multiple-camera-based head appearance modeling. In: Proceedings of IEEE Conference on Automatic Face and Gesture Recognition, pp.657–662 (2004)
21. Velastin, S.A., Boghossian, B.A., Lo, B., Sun, J., Vicencio-Silva, M.A.: Prismatica: Toward ambient intelligence in public transport environments. IEEE Trans. Syst. Man Cybern. Part A **35**(1), 164–182 (2005)
22. Wu, T., Matsuyama, T.: Real-time active 3D shape reconstruction for 3D video. In: Proceedings of 3rd International Symposium on Image and Signal Processing and Analysis, vol. 1, pp. 186–191 (2003)

Author Biographies

Tarak Gandhi received his Bachelor of Technology (B. Tech.) degree in Computer Science and Engineering at the Indian Institute of Technology, Bombay. He earned his M.S. and Ph.D. from the Pennsylvania State University in Computer Science and Engineering, specializing in Computer Vision. He worked at Adept Technology, Inc. on designing algorithms for robotic systems. Currently, he is an Assistant Project Scientist at California Institute for Telecommunications and Information Technology (CalIT2) in University of California at San Diego. His interests include computer vision, motion analysis, image processing, robotics, target detection, and pattern recognition. He is working on projects involving video surveillance, intelligent driver assistance, motion-based event detection, traffic flow analysis, and structural health monitoring of bridges.



Mohan Manubhai Trivedi is a Professor of Electrical and Computer Engineering and the founding Director of the Computer Vision and Robotics Research Laboratory at the University of California in San Diego. He has been a Charter Member of the Executive Committee of the University of California System's UC Discovery: Digital Media Innovation Program since 1998. He was elected Vice Chair of the Executive Committee in 2006. Trivedi has a broad range of research interests in the intelligent systems, computer vision, intelligent ("smart") environments, intelligent vehicles and human–machine interfaces areas. In close collaboration with regional transportation and first responder agencies, Trivedi regularly participates in projects dealing with infrastructure protection and physical security. He is also active in research in privacy preserving technologies as well as in dialogs related to balancing privacy with security using video technology at various multidisciplinary forums. He served as the Program Co-Chair of the IEEE Intelligent Vehicles Symposium in 2006, the Editor-in-Chief of the *Machine Vision and Applications* Journal (1996-2004) and currently serves as an editor of the *IEEE Transactions on Intelligent Transportation Systems*. Trivedi has received the Distinguished Alumnus Award from the Utah State University, Pioneer (Technical Activities) and Meritorious Service Awards from the IEEE Computer Society. Trivedi regularly serves as a consultant to various Industrial and Government Agencies in the US and abroad.