IEEE Workshop on Learning, Representation, and Context for Human Sensing in Video, June 2006

Panoramic Appearance Maps for Human Body Representation and Analysis

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

Multiple overlapping cameras help to provide robust detection as well as 3-D localization of people in their field of view [2], cover their features from all directions, and handles the problems of occlusion. We propose a novel representation called Panoramic Appearance Map (PAM) for combining appearance information from multiple cameras tracking the same person. The map is a discretized 2-D array with horizontal axis representing the azimuth angle around the body center and the vertical axis representing the height above the ground. Each bin in the map contains appearance information such as color or texture from the cameras which can observe it, as well as a weighting factor that denotes the reliability of the information.

The human body is modeled as a convex generalized cylinder shown in Figure 1. A point p on its surface is parameterized by the azimuth angle θ and height h as:

$$p = (x, y, z) = (x_0 + r(\theta, h)\cos\theta, y_0 + r(\theta, h)\sin\theta, h) \quad (1)$$

where r is a function representing the cross section of the cylinder at height h and $p_0 = (x_0, y_0, 0)$ is the location of the cylinder. Currently, we are modeling r by assuming human cross section as a cirle or ellipse with parameters according to the normal dimensions of the body. However, for better accuracy, we are working on using the 3-D shape of the body obtained from voxelization [3]. Using the value of p for every θ and h, one has a transformation from (θ, h) to image pixels for every camera k, based on its calibration. The panoramic map is formed by discretizing this transformation into a grid of $M \times N$ bins, The bin corners are mapped onto the image plane defining a region in the camera image for every bin.

If $c_k(m, n)$ is the average value of a feature in that image region and w_k gives the weight equal to the number of pixels in the region, then the pair $C_k = (c_k, w_k)$ is used to represent the information from image k. The use of weights minimizes errors during comparison due to spurious bins that do not obtain reliable information from the particular image. If multiple cameras can view the point p, the information in C_k for all cameras can be combined into one pair $C = (w, c)$ using weighted average.

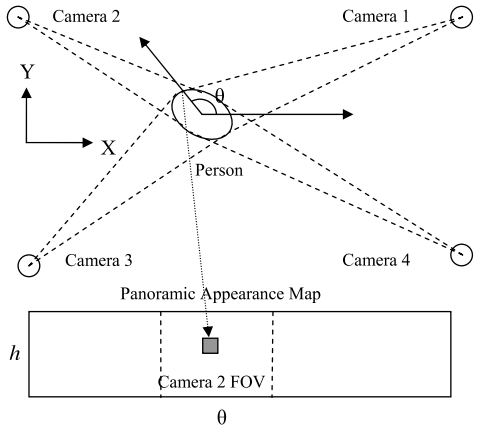

Figure 1. Mapping from 3D space (plan view) to Panoramic Appearance Map (PAM)

Matching is performed using a weighted metric which is a modification of the standard Sum of Squares Difference (SSD) or Sum of Absolute Difference (SAD) where each pair of bins is weighted by the geometric mean of the weights of both bins. This metric gives larger emphasis on bins having larger weights in both maps, while suppressing bins with small weights. For rotational invariance, the metric is computed by translating one of the maps by all discrete rotation angles and then taking the minimum. If the person is tracked over multiple frames, one can integrate the maps by displaced averaging to give more reliable matching than with single frame.

The PAM representation was applied for performing person reidentification between two multi-camera setups using color information. The first setup contained 4 rectilinear cameras in a large indoor space, whereas the second one consisted of four omnidirectional cameras that obtained 360 degree views of a room. Each set of cameras spanned the space with overlapping FOVs giving coverage from all directions. Moving persons were detected using background subtraction and their 3-D positions estimated by triangulation between multiple cameras. The PAMs were formed

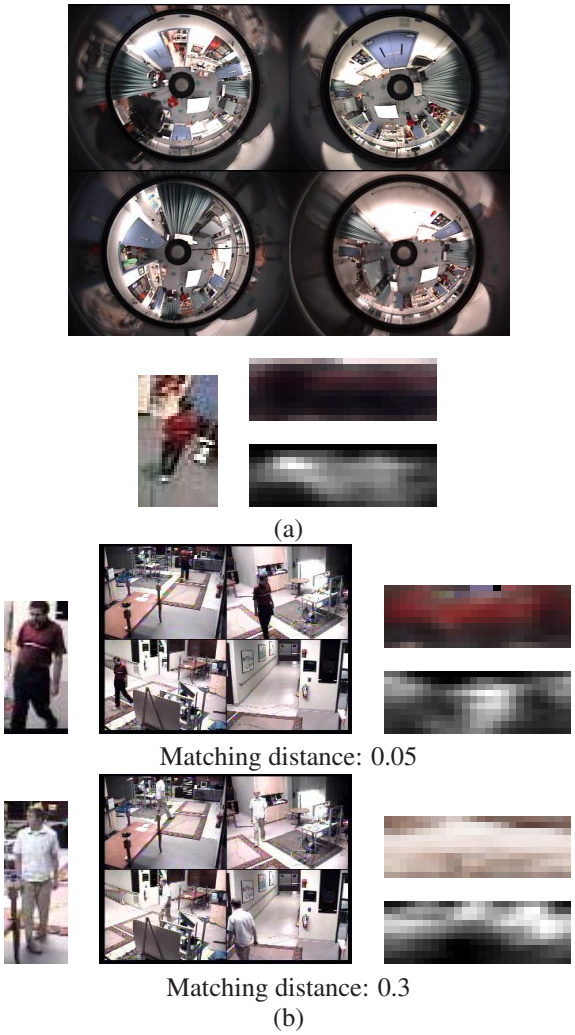

Figure 2. Sample of test results from application of PAM for multi-camera based person reidentification. (a) Images from setup containing 4 omni cameras with person's snapshot and the PAM showing the feature and weight arrays. (b) Comparison with PAMs obtained from rectilinear camera setups. The matching distance between these and the omni PAMs show that the first case is a better match than the second.

using color information from all images that had the person in the FOV. Figure 2 shows example results of this experiment. The PAM obtained from omni camera scene is compared with two PAMs from rectilinear camera scene. It is seen that even though the color responses for the cameras are different, the matching give promising results. Further improvement can be expected by accounting for color differences between the camera sets.

Recent efforts are directed towards systematic and synergestic integration of PAM with our research on voxel based human shape analysis [1]. Figure 3 shows the application of 3-D shape context representation for human body generated from voxelization with multiple cameras. A

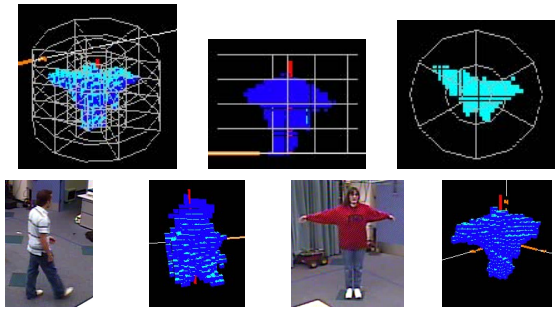

Figure 3. The 3D shape context capture of human body voxels by multilayered cylindrical histogram. [1].

cylinder is fitted around the body center and discretized in the height, azimuth, as well as radial dimensions. The number of voxels in each bin is counted and stored. This gives an efficient representation of the body shape that is useful for applications such as gesture analysis which uses volumetric information. On the other hand, the PAM represents surface information by suppressesing the radial dimension. We are exploring the combination of the appearance and the shape context to improve the accuracy and robustness for person reidentification.

In addition, we are also working towards handling issues such as occlusion during multi-person cases, color invariance between cameras, as well as the incorporation of other scene and image-based features. For example, when there is occlusion in one of the cameras, the other camera(s) would 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.

References

[1] K. S. Huang and M. M. Trivedi. 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, June 2005. 2

[2] 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. 1

[3] T. Wu and T. Matsuyama. Real-time active 3d shape reconstruction for 3d video. In *Proc. 3rd International Symposium on Image and Signal Processing and Analysis*, volume 1, pages 186–191, September 2003. 1