Proceedings of the Applications and Science of Soft Computing IV,
July 29-August 3 2001, San Diego, California, USA

Networked omnivision arrays for intelligent environment

Kohsia S. Huang and Mohan M. Trivedi*
Computer Vision and Robotics Research (CVRR) Laboratory,
Dept. of Electrical and Computer Engineering, University of California, San Diego

ABSTRACT

Intelligent environments are systems that are aware of the spatial information and activities within them through sensors and interact with people in a natural and unobtrusive way. An intelligent system using networked omnivision array is proposed based on specified requirements of intelligent environments. It utilizes Omni-Directional Vision Sensor (ODVS) network as the sensory input. ODVS optical modeling is described, which allows panoramic and perspective view generation. A 3D tracker based on the ODVS network is constructed. Using the tracking information, active camera selection and dynamic perspective view generation enable real-time face tracking. Face recognition is also implemented for person identification. Current results of the modules and extensions to the system are also discussed.

Keywords: Omnidirectional camera network, 3D tracking, dynamic view generation, active camera selection.

1. INTRODUCTION

Intelligent environments are rooms or spaces which automatically derive and continuously maintain an awareness of the space, its composition, and activities taking place in them[3, 10, 11, 12]. The overall goal of intelligent environment research is to design and develop integrated sensor-based systems that allow natural and efficient mechanisms between humans and computers in order to make transactions of human interactions more productive, efficient, and enjoyable. The research challenges are to make those intelligent systems be capable of:

- ***Space Awareness***: Develop and maintain an awareness of their 3D environment,
- ***Activity Awareness***: Acquire and respond to the multimodal sensory inputs from the users in a robust manner,
- ***Televiewing***: Interact in a natural and flexible manner with both the local and remote users, and
- ***Summarization***: Summarize the events in an efficient and comprehensive way.

Among them, space awareness and activity awareness let the system understand the environment where it resides and detect events occurs within the environment. Televiewing and summarization support the system efficient interfaces to interact with the users. In this paper, we propose an intelligent system based on Omni-Directional Vision Sensor (ODVS) network. This approach is different from other systems [3, 8, 10, 11, 14] and has more functionality. Based on this infrastructure, first the optical modeling of ODVS is developed. Conversions from ODVS image to panoramic and perspective views are described. They are useful for omnidirectional televiewing. Using ODVS network, 3D person tracking can be performed. Camera selection and dynamic face tracking allows the system to actively track human face using a perspective view. The perspective face image is then identified by face recognition module.

2. THE ODVS NETWORK

The ODVS network system architecture is shown in Figure 1. The video sources are from four ODVS cameras installed in our testbed. The four ODVS are installed on the four corners of a meeting table at the center of the room. The signals of the four ODVS are fed into a video mixer called "quad," and the mixed video is shown in Figure 3. Single ODVS signal can also be selected by the camera selection unit through the multi-channel frame grabber. A tracker unit is designed to detect and track people on the quad ODVS image. Different from regular rectilinear cameras, ODVS takes 360 degrees of view in one shot. The captured ODVS image is disc-shaped and can be unwrapped into a panorama or a perspective view. If the perspective view is on a person's face, the view is then sent to the face recognition unit to identify the person.

* khuang@ucsd.edu; trivedi@ece.ucsd.edu; phone 858-822-0002; fax 858-534-1004; http://cvrr.ucsd.edu; CVRR Lab., 9500 Gilman Drive, Mail Code 0407, La Jolla, CA 92093-0407

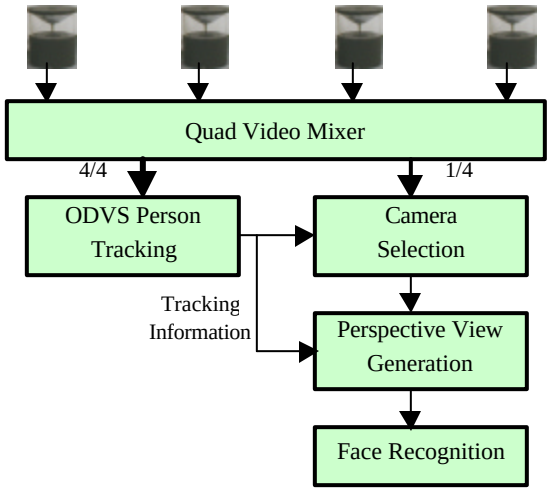

Figure 1: The ODVS network system architecture.

2.1 ODVS optical modeling and panorama generation

An ODVS is composed of a regular rectilinear camera with an optical mirror in front of the lens. In our testbed, hyperboloidal mirror is applied in the ODVS. Hyperboloid mirror has two foci. If the mirror surface is concave down, then for all light beams going through the upper focus will be reflected by the hyperboloidal mirror to go through the lower focus. A CCD camera is installed right below the hyperboloid mirror looking upward. If the lens center is positioned at the lower focus, then the image will be formed on the CCD image plane. The net effect of this image formation is that the upward narrow view of the CCD camera is enlarged to a hemisphere view looking downward from the upper focus. Thus it satisfies the criterion of single point of view[4]. Then the hemisphere view can be projected onto a cylinder to be a panoramic view, or onto a plane perpendicular to the line of sight to be a rectilinear perspective view, as shown in Figure 2.

Figure 2: An ODVS image and the unwrapped perspective view and panoramic view.

The relationship between an object point in 3D space and the corresponding image point on CCD plane can be derived[5]. With this relationship, the panoramic image can be produced in a straightforward manner. Assume a cylindrical screen exists around the upper focus of the ODVS. Each pixel on the cylindrical screen represents a point in the 3D space. By mapping the cylindrical pixels onto the CCD image plane, the panoramic pixel values can thus be obtained from the ODVS image.

2.2 ODVS perspective selection

The advantage of ODVS is that it takes an omnidirectional view in one shot. It is advantageous over rectilinear cameras because it does not need mechanical moving mount. A camera with moving mount violates single viewpoint criterion[4] and consumes time while rotating, thus could miss important shot. On the other hand, ODVS always keeps track of real-time activities around 360 degrees from a single viewing point.

The rectilinear perspective view is a planar view of the scene. The perspective view plane is always perpendicular to the line of sight from the viewing point, which is the upper focus. The direction of line of sight determines the viewing direction, whilst the distance of the perspective plane is the effective focal length of the perspective view. The direction of line of sight can be represented in terms of the horizontal pan angle and vertical tilt angle with the upper focus as the origin. Thus the pixels on the fixed-size perspective plane can be represented in Cartesian coordinate in terms of the pan angle, the tilt angle, and the effective focal length. With these 3D coordinates, the corresponding point on the CCD plane can be mapped using the relationship mentioned earlier. Thus the perspective view can be interpolated from the pixel values in the ODVS image. One example of the perspective view selection is shown in Figure 4. The user can specify the pan and tilt angles, the effective focal length, and the method of image interpolation. It is obvious that ODVS image is suitable for televiewing[2]. If transmitted to the remote site, any perspective view can be generated in the area covered by the ODVS image.

2.3 Person detection and tracking

A person tracker utilizing the 4 ODVS is shown in Figure 3. It utilizes the quad ODVS image and unwraps individual ODVS images into four panoramic views. A 1D profile of the background is formed by accumulating the pixel differences from background in each column of the panorama. Then if a person or object presents, it will be detected in terms of pan angle, as shown in the histograms of below the panoramas. Knowing the locations of the four ODVS, the planar location of the person or object can be determined by triangulation method. It is illustrated in the floor plan in Figure 3, where clips of the people or objects are also shown. This localization mechanism is called N-ocular and is extensively analyzed by Sogo, et. al.[7]. The radius of person is approximately assumed. Based on the size of intersection zone of triangulation, the final location of person is determined by incorporating error compensation measures. This tracking system has accuracy within 20cm in our 6.7m-by-3.3m room.

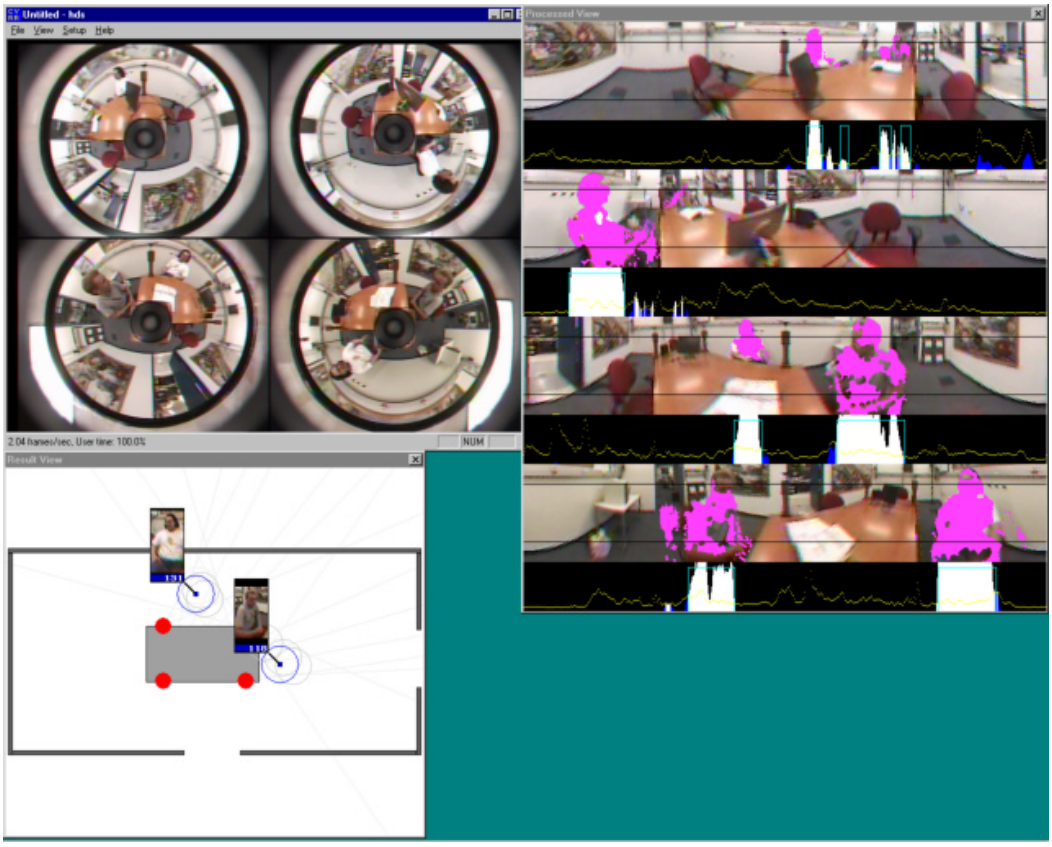

Figure 3: A human tracker based on four ODVS video streams.

This ODVS tracker can be extended from 2D to 3D tracking easily. From the panorama, the person can be segmented from the background, as shown in Figure 3, after binary morphological processing. In order to save computation, these operations are performed only in the pan angle range where a person is detected by the histogram below the panorama. Since the planar location of the person is estimated previously, the distance of the person to a specific ODVS can be computed. From the panorama, knowing the topmost point of person's blob, the height of person can be estimated by similar triangles. The final

estimate of person's height is a weighted sum of the estimates from the four ODVS. The weight is inversely proportional to the distance of the person to the corresponding ODVS. If the parameters are carefully calibrated, the accuracy of human heights can be within 6cm in our testbed. As shown in Figure 3, human height is displayed in centimeters below the clipped human image in the floor plan. With this 3D tracker, the location of head and face of a person can be estimated for face tracking.

2.4 Face tracking and active camera selection

The purpose of face tracking is to keep track of a human face through a perspective view generated from an ODVS image. A two-computer setup is used for human face tracking and active camera selection. One computer takes the quad ODVS image for 3D tracking and transmits the face track to the second computer through network. The second computer captures one single ODVS image and tracks human head through dynamic perspective view as in the left half of Figure 4. Single ODVS image capturing is applied because it gives higher resolution views for face recognition. The head position from the first computer is compared to the position of ODVS to estimate pan, tilt, and zoom factors for the perspective view. Meanwhile active camera selection is able to determine the ODVS that best views the human face. Thus the system is able to track human face no matter the person is walking or sitting.

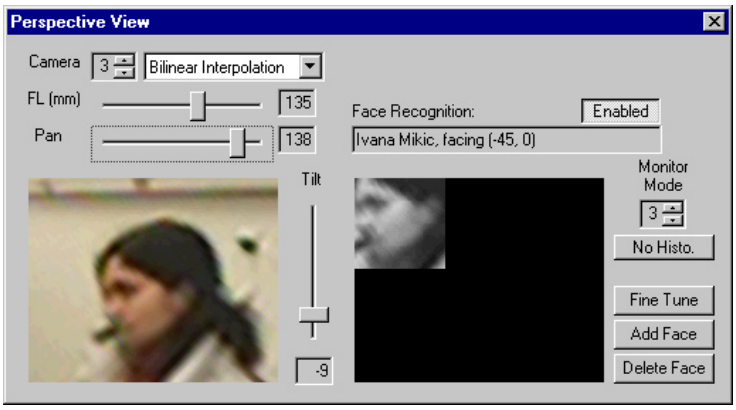

Figure 4: Face perspective view generation and face recognition.

Active camera selection tries to capture human face as clear as possible. Preliminary choice is the ODVS that is nearest to the person. If the person is walking, the nearest ODVS that the person is walking toward is chosen. This is based on the assumption that often people face the direction that they are walking. If the person is sitting beside the meeting table, then the camera that is closest to the person is first chosen. A face orientation estimation[3] is then used to estimate the facing direction of the person. If the person is facing to another direction, then the ODVS that the person is facing is chosen and face orientation is verified. This process continues until a nearby ODVS that best views the human face is found.

It should be noted that totally no mechanical motion of the camera is needed. Every view selection and camera switching are performed electronically. Thus the processing speed and system reconfigurability are increased. Also note that since the ODVS cameras are placed in the midst of the meeting participants, ODVS cameras have the advantageous angle viewing the faces around them. Therefore ODVS network is suitable for meeting room setup.

3. FACE RECOGNITION

Currently eigenface method (PCA)[13] is applied because it is simple and suitable for realtime processing. Face images are extracted from the snapshots of people by skin color segmentation[7]. Five images of each person on different facing angle are stored as the training faces. Then these faces are used to compute the eigenfaces to represent the face images. The face extraction and eigenface algorithm are programmed in a real time program to test its feasibility.

3.1 Implementation

Initially, a set of eigenfaces is computed from the training faces[13]. The projection vectors of the training faces on the eigenface set are also computed. Then face detection and recognition are performed as in Figure 5. The skin color segments are extracted[7] from the perspective view. For each skin color segment, a 64 by 64 image is cropped around the skin color segments and checked for face existence, as shown in Figure 6. False segments are rejected if the cropped image is distant

from the eigenface space[13]. Otherwise if the distance is less than a threshold, the cropped image is decided as a face image. If all skin color segments are rejected, a message indicating no person is displayed as the recognition result.

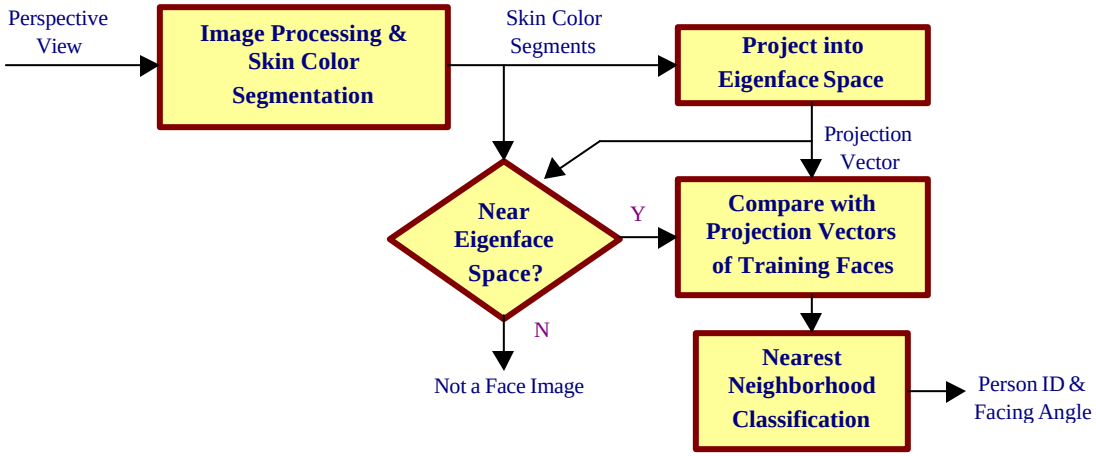

Figure 5: Face detection and recognition using eigenface method.

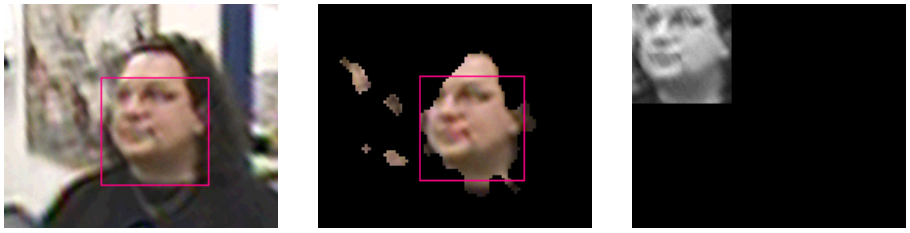

Figure 6: Face image extraction, from perspective view (left), skin color segments (middle), to extracted face image (right).

For a detected face image, its projection vector in eigenface space is compared to the projection vectors of the training faces in Euclidean distances. If the minimum distance is less than a recognition bound, then this face image is recognized as the associated training face and its identity is displayed as in the right half of Figure 4. Otherwise it is classified as an unknown person and the unknown message is displayed.

3.2 Experimental evaluation

The eigenface set was built upon only 5 people and 5 training images for each people. The accuracy was evaluated from 24 tests for each of the 5 people. Face perspective views were selected manually. Face detection was counted successful if the 64 by 64 image correctly crops the face. For face recognition results, it was counted successful if the module identified the person correctly disregarding the facing angle. A summary of the results is compared in Table 1. With 25 eigenfaces, a 74% average accuracy was achieved after calibration on skin color segmentation.

Accuracy	Face Detection	Face Recognition
Person 1	24/24 (100%)	20/24 (83%)
Person 2	24/24 (100%)	21/24 (88%)
Person 3	24/24 (100%)	14/24 (58%)
Person 4	24/24 (100%)	17/24 (71%)
Person 5	24/24 (100%)	17/24 (71%)
Average	120/120 (100%)	88/120 (74%)

Table 1: Experimental results on the accuracy of face detection and face recognition.

To improve the accuracy, it was reported that roughly an eigenface space of dimension 100 would be required in order to have a sufficient representation of faces[6]. Eigenface method is also sensitive to lighting conditions. Adaptive equalization on the face image would be useful. However, the computation may be very heavy and slows down the frame rate. Noise also fluctuates the location of face. If face image can be located very accurately, a mug-shot mask on the face can be applied to

remove background interference. To further improve the performance, ICA methods[1] may be useful and they are reported to have higher accuracy[6].

4. CONCLUSION

The proposed ODVS network architecture has 3D tracking, active camera selection, dynamic face tracking, and face recognition capabilities. Tracking allows the system to aware the space and human activities in the room. Based on the detected activities, appropriate camera selection and face tracking on ODVS perspective view keep track of the events. Face recognition further details the events. Higher level decisions can be derived from all the information. Meanwhile, it supports televiewing and summarization interfaces. Hence the ODVS network system meets the requirements of an intelligent system.

To expand the capability of the ODVS network, gesture recognition allows the system to accept gesture commands. As in other intelligent systems[3, 10, 12, 14], microphone arrays can also be included. Sound localization, microphone array beamforming, speech detection and enhancement, and speech and speaker recognition capabilities can cooperate with the ODVS network modules and make the system more robust and efficient.

ACKNOWLEDGMENTS

Our research is supported in part by the California Digital Media Innovation Program (DiMI). The authors would like to thank Brett Hall, Dr. Hiroshi Ishiguro, Sadahiro Iwamoto, Ivana Mikic, Steve Roche, Takushi Sogo, and Jonathon Vance for their assistance and collaborations.

REFERENCES

1. M. Bartlett, H. Lades, and T. Sejnowski, “Independent Component Representations for Face Recognition,” *Proc. of the SPIE* **3299**, pp. 528-539, Jan. 1998.
2. B. Hall, K. Huang, and M. Trivedi, "A Televiewing System for Multiple Simultaneous Customized Perspectives and Resolutions," *Submitted to IEEE Int’l Conf. on Intelligent Transportation Systems*, Oakland, California, Aug. 2001.
3. I. Mikic, K. Huang, and M. Trivedi, “Activity Monitoring and Summarization for an Intelligent Meeting Room,” *Proc. IEEE Workshop on Human Motion*, pp. 107-112, Dec. 2000.
4. S. Nayar, “Catadioptric Omnidirectional Camera,” *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 482-488, Jun. 1997.
5. Y. Onoe, N. Yokoya, K. Yamazawa, and H. Takemura, “Visual Surveillance and Monitoring System Using an Omnidirectional Video Camera,” *Proc. IEEE Int’l. Conf. on Pattern Recognition*, pp. 588-592, Aug. 1998.
6. P. Phillips, H. Wechsler, J. Huang, and P. Rauss, “The FERET Database and Evaluation Procedure for Face-Recognition Algorithms,” *Image and Vision Computing* **16**, pp. 295-306, Apr. 1998.
7. T. Sogo, H. Ishiguro, and M. Trivedi, “Real-Time Target Localization and Tracking by N-Ocular Stereo,” *Proc. IEEE Workshop on Omnidirectional Vision*, pp. 153-160, Jun. 2000.
8. R. Stiefelhagen, J. Yang, and A. Waibel, "Simultaneous Tracking of Head Poses in a Panoramic View, " *Proc. IEEE Int'l. Conf. on Pattern Recognition*, pp. 722-725, Sep. 2000.
9. M. Storring, H. Andersen, and E. Granum, “Skin Color Detection Under Changing Lighting Conditions,” *Proc. 7th Symp. on Intelligent Robotics Systems*, pp. 1-9, Jul. 1999.
10. M. Trivedi, K. Huang, and I. Mikic, “Intelligent Environments and Active Camera Networks, “ *Proc. IEEE Int’l. Conf. on Systems, Man, and Cybernetics*, pp. 804-809, Oct. 2000.
11. M. Trivedi, I. Mikic, and S. Bhonsle, “Active Camera Networks and Semantic Event Database for Intelligent Environments,” *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Jun. 2000.
12. M. Trivedi, B. Rao, and K. Ng, “Camera Networks and Microphone Arrays for Video Conferencing,” *Proc. Multimedia Systems and Applications Conference*, pp. 384-390, Sep. 1999.
13. M. Turk and A. Pentland, "Face Recognition Using Eigenfaces," *Proc. IEEE Int’l. Conf. on Computer Vision and Pattern Recognition*, pp. 586-591, Jun. 1991.
14. D. Zotkin, R. Duraiswami, L. Davis, and I. Haritaoglu, “An Audio-Video Front-End for Multimedia Applications,” *Proc. IEEE Int’l. Conf. on Systems, Man, and Cybernetics*, pp. 786-791, Oct. 2000.