DISTRIBUTED VIDEO ARRAYS FOR TRACKING, HUMAN IDENTIFICATION, AND ACTIVITY ANALYSIS

Kohsia S. Huang and Mohan M. Trivedi*

Computer Vision and Robotics Research (CVRR) Laboratory
University of California, San Diego
khuang@ucsd.edu; mtrivedi@ucsd.edu
http://cvrr.ucsd.edu/

ABSTRACT

In this paper we discuss the intelligent or smart camera based systems for capturing visual contextual information at multiple levels of information abstraction. A distributed video array, with multiple omnidirectional and rectilinear cameras, is used to acquire visual information. System architecture as well as 3D tracking and human identification modules are described. Examples of the use of distributed video arrays in indoor, outdoor, and mobile environments are presented.

1. INTRODUCTION

We are interested in developing intelligent environments which automatically capture and maintain an awareness of the events and activities taking place in these spaces. Such spaces can be indoors, outdoors, or mobile. This is indeed a rather ambitious effort, especially when one considers the real-world challenges of providing real-time, reliable, and robust performance over the wide range of events and activities which can occur in these spaces. Novel multimodal sensory systems are required to realize useful intelligent spaces. Arrays of cameras and microphones distributed over the spatial (physically contiguous or otherwise) extent of these spaces will be at the front end of capturing the audio-visual signals associated with various static and dynamic features of the space and events. The intelligent environments will have to quickly transform the signal level abstraction into higher level semantic interpretation of the events and activities.

In this paper we present research efforts directed towards the development of networks of video cameras, which support a wide range of tasks of intelligent environments. Key features of these "smart" video arrays are:

1. Ability to derive semantic information at multiple levels of abstraction.
2. Ability to be "attentive" to specific events and activities.
3. Ability to actively shift the focus of attention at different "semantic" resolutions.
4. Apply different types of camera arrays to provide multiple signal-level resolutions.

In this paper we will focus on real-time tracking of single or multiple people and on coordination of multiple cameras for capturing visual information on wide areas as well as selected areas for activity analysis and person recognition. Applications of the distributed interactive video array system are also presented.

2. DISTRIBUTED VIDEO ARRAY SYSTEM OVERVIEW

Figure 1 demonstrates a general architecture of smart video array systems using both omnidirectional and rectilinear pan-tilt-zoom (PTZ) camera arrays. The omnidirectional cameras, with a full 360-degree panoramic field of view, provide a large area coverage using a relatively small number of cameras. On the other hand, the PTZ cameras provide a specific focus of attention with higher resolution. With these two types of camera arrays, the system can acquire a coarse-to-fine awareness of the events.

Camera videos are first captured and processed for low-level visual cues such as histograms, colors, edges, and object segmentations. The challenges at this level include robustness to illumination, background, and perspective variations.

On the next level of abstraction, tracking plays an important role in event analysis. It derives the current position and geometry of people as well as the histories and predictions of people's trajectories. With the semantic database which defines prior knowledge of the environment and activities, events can be detected from the tracking information, e.g., one person enters the room and sits beside a table. The activity analyzer and the semantic database could be implemented by a Bayesian net [6]. The challenges at this level include the speed, accuracy, and robustness of the tracker, as well as the scalability of the semantic database which allows incremental updating when new events are detected.

The events trigger the attention of a camera array to derive higher semantic information. Using tracking information, a suitable camera is chosen to generate a perspective that covers the event at a desired resolution, e.g., perspective on a person with an omnicam for posture and around the head area with a

* This research was sponsored by the UC Discovery (DiMI) Grants. We are pleased to acknowledge the valuable assistance of our colleagues from the CVRR Laboratory.

PTZ camera for person identification. For this purpose, necessary processing modules such as face detection and recognition should be deployed. The challenges at this level include speed, accuracy, and robustness of the view generation and recognition modules. The derived semantic information at multiple levels can also be fed back to update the semantic database.

This architecture of multi-level abstraction can be further generalized to include many other applications such as object recognition, facial expression recognition, 3D human body modeling and tracking [7], and even intention estimation and prediction.

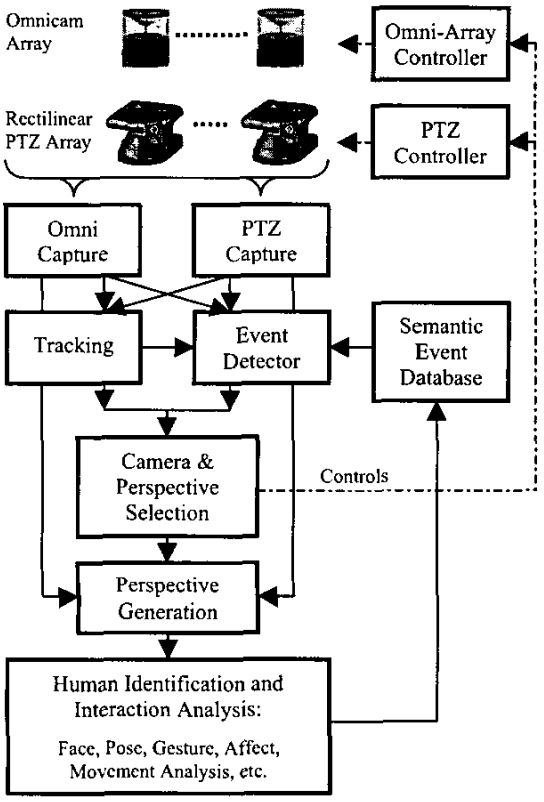

Figure 1. Distributed video arrays for tracking, human identification, and activity analysis.

3. DISTRIBUTED VIDEO ARRAYS: DEPLOYMENT EXAMPLES

In this section we present the applications of indoor, outdoor, and mobile distributed video arrays that represent the general architecture in Figure 1 to different degrees of event abstraction.

3.1 Indoor Distributed Video Array

We first present an indoor intelligent environment, the networked omni-video array (NOVA) system [4], as an example of Figure 1. It operates on four omnicameras mounted on a meeting table in the midst of a conference room for an *inside-out* panoramic coverage. As opposed to rectilinear cameras which have an *outside-in* coverage, the omnicam array has less chance of occlusion when there are more people. The omnicam array is used for 3D person tracking, as shown in Figure 2(a), as well as for head and face tracking, as shown in Figure 2(b). The 3D tracker provides location, velocity, and height information of the people for head and face tracking on the camera array. If the person is static, the most nearby omnicam is selected; otherwise the omnicam that faces the walking person is selected. Instantaneous location of the person's head relative to the selected camera is then used to compute the pan-tilt-zoom values for the perspective transformation [4] to electronically extract a perspective view from the omnidirectional video and latch upon the person's face in real-time, as shown in Figure 2(b). The face under tracking is then detected and recognized for higher semantic information.

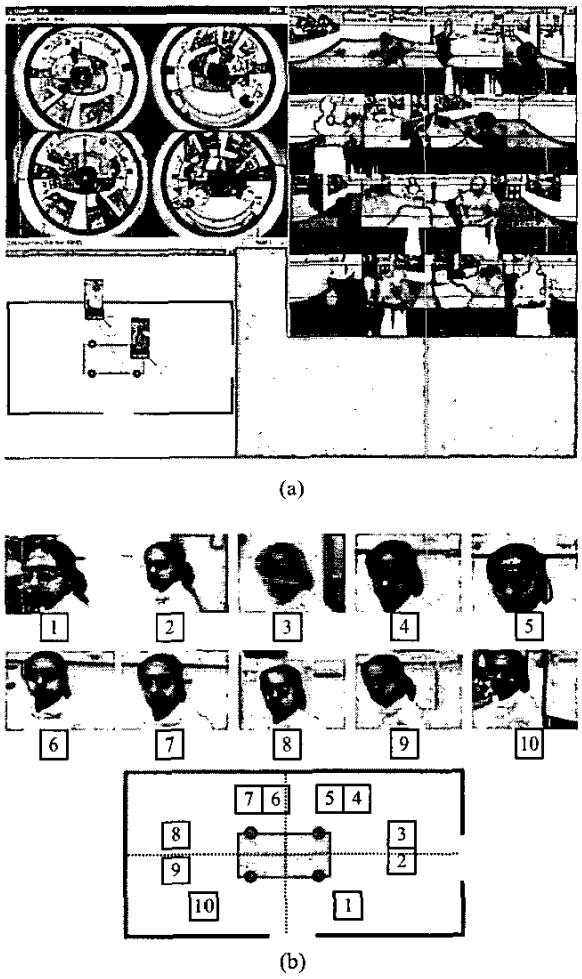

Figure 2. The real-time NOVA system with (a) 3D person tracking and (b) head and face tracking on a walking person.

The real-time 3D tracker runs on one computer and takes a quad video of the four omnicams. It plays a critical role in system accuracy, speed and robustness. As detailed in Figure 3, at the signal level, the panoramic views of the omnicams are processed to perform shadow detection and background subtraction to segment humans. From human segmentations, the location and the height of human is estimated by a sophisticated 3D triangulation process [4]. The measurements are associated to a track history and averaged to reduce measurement noise. Currently the tracking accuracy in standard deviation is within 15cm when tracking 4 people on ~20fps. If Kalman filter is used for track filtration and prediction, tracking accuracy would be improved significantly.

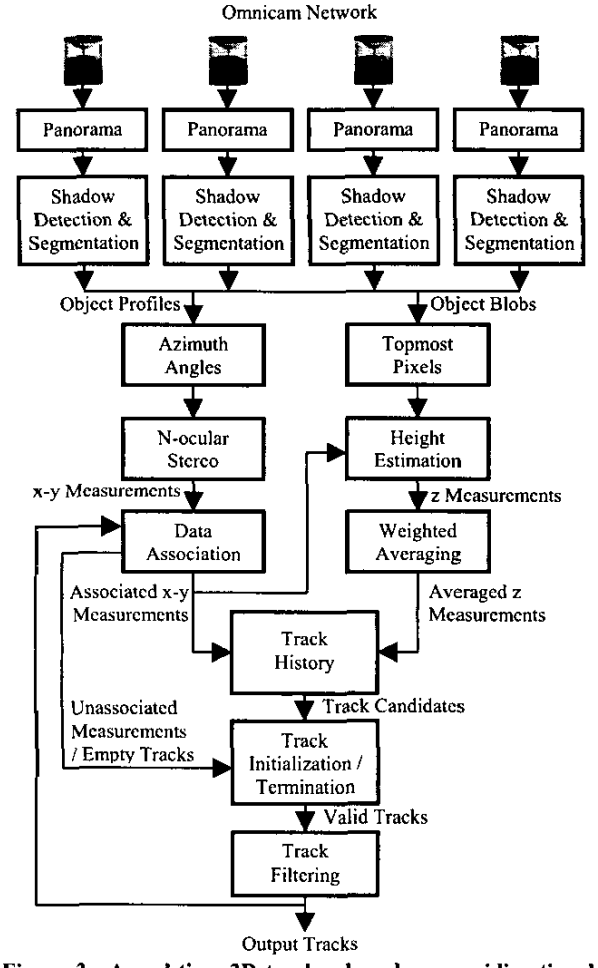

Figure 3. A real-time 3D tracker based on omnidirectional camera array.

On another computer, 3D tracking information is received via the network to perform head tracking and face detection and recognition on the selected omnicam video for higher semantic level abstraction. A general streaming type of face detector and recognizer is shown in Figure 4. For single-frame features, edge and contour are useful to detect heads, face regions, and facial features like cheeks, eyes, mouth, and hairline. Skin tone regions could also be compared to the edges to enhance the detection and help in rejecting false positives. Template matching is then applied to detect the faces with a likelihood score. The second path in Figure 4 is a view-based approach. Distance-from-face-space (DFFS) [2] is an eigenface template matching method for face detection and recognition. Linear discriminant analysis (LDA) [1] is a good alternative to eigenface since it has better discrimination capability. The third path is based on gray-scale wavelet features. A window scanning through the test image extracts wavelet features by wavelet transform. The feature vectors are then classified by a support vector machine or a Gaussian mixture model which are trained for classifying face and non-face by bootstrap technique [2]. For spatial-temporal fusion, single-frame scores and features for each frame in the video are partitioned into segments. An accumulated likelihood of a segment is computed using HMM or Bayesian net [6]. Face detection can be reflected by the state trajectory and the face identity can be decided by maximum likelihood rules. Currently the single-frame face detection and recognition are based on the skin-tone and eigenface methods and the HMM-based streaming face recognition is able to boost the recognition rate from 76% to 99% [5].

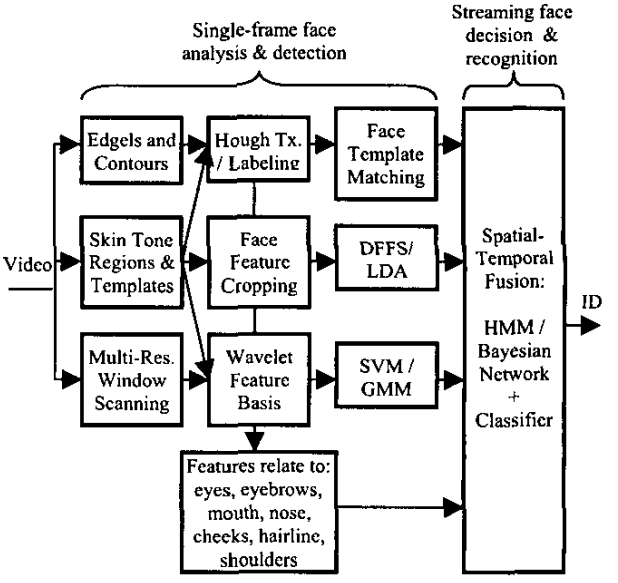

Figure 4. Video based streaming face detection and recognition.

In this example, as compared to Figure 1, only an omnicam array is used. The signal level and higher level abstraction are identical to the architecture in Figure 1. For mid level abstraction, only tracking is applied. There is no activity analysis or semantic database utilized for event recognition. The NOVA system monitors people continuously.

3.2 Outdoor Distributed Video Array

Figure 5 shows an outdoor example using a distributed PTZ camera array. This system is designed to detect stopped cars along the street. At signal-level abstraction, background statistics are updated and shadow detection is performed to

segment cars on each camera. As a car moves, the system tracks it from camera to camera. The camera hand-over is happening in an overlapping area between the two cameras. If a stopped car event is detected, the nearby camera is triggered to zoom into the car and to the driver area. Face detection and recognition is performed to detect and identify the driver by the method mentioned in the indoor example. In this example, the activity analysis module detects stopped cars from the tracking information.

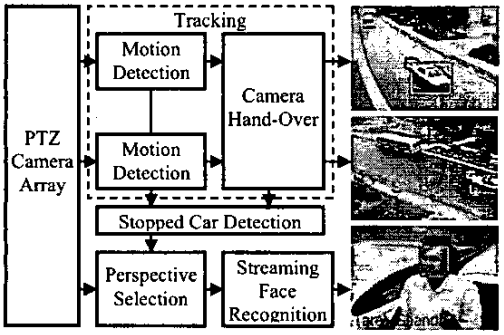

Figure 5. Example of an outdoor distributed PTZ camera array for stopped vehicle detection and driver identification.

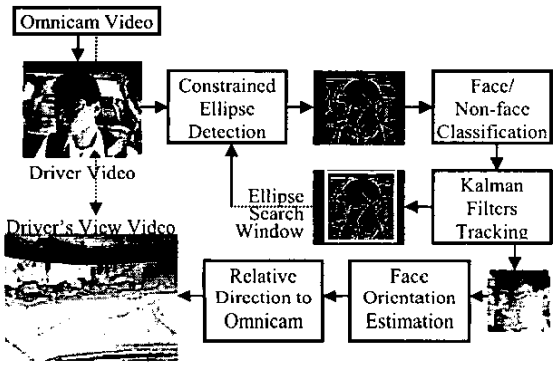

Figure 6. Mobile intelligent environment for driver's head detection and tracking and instantaneous driver's view generation using single omnicam.

3.3 Mobile Distributed Video Array

Figure 6 shows a smart car application of the intelligent environment [3]. It uses one omnicam mounted near the wind-shield pole on the driver's side to perform driver's head tracking and driver's view generation. A perspective view is generated on the driver's seat to detect driver's head since the driver cannot move too much while driving. Low level information of ellipse detection is first done on the edge map of the driver image. For mid-level abstraction, the ellipse is tracked across the frames by Kalman filters and a video of driver's face is extracted by the center and size of the ellipse. These ellipse parameters also define an ellipse search window for the next frame to enhance the head detection rate from 50% to 87% and speed it up in the meantime. On higher level of abstraction, the face orientation is estimated by a template matching in view-based feature subspace, followed by another Kalman filter. The face orientation templates have a 15° step size. The averaged estimation error is about 4° with an 8° standard deviation. The estimated face orientation is then used to derive the relative direction of the driver's face from the omnicam and generate a perspective video along this direction as the instantaneous driver's view.

In this example, the activity analyzer is the face/non-face classification and the semantic database is the templates for face classification. In the future other higher level semantic processing could be counting the number of cars in the driver's field of view and estimating the driver's cognitive load. This enables the smart car system to detects dangers and have reactions.

4. CONCLUDING REMARKS

We presented a general architecture of intelligent environment systems. The system derives semantic information at the signal level and the tracking and activity awareness level, and attentively zooms into a higher semantic resolution. Finally, several indoor, outdoor, and mobile applications of the intelligent environment system are compared to the architecture

REFERENCES

[1] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. Patt. Analysis and Mach. Intelli.*, Vol. 19, No. 7, pp. 711-720, Jul. 1997.

[2] E. Hjelmas and B. K. Low, "Face Detection: A Survey," *Comp. Vis. and Image Understanding* **83**, pp. 236-274, 2001.

[3] K. S. Huang and M. M. Trivedi, "Driver Head Pose and View Estimation with Single Omnidirectional Video Stream," To appear in *the 1st. Int'l. Workshop on In-Vehicle Cognitive Computer Vision Systems*, Graz, Austria, April 3, 2003.

[4] K. S. Huang and M. M. Trivedi, "Video Arrays for Real-Time Tracking of Persons, Head and Face in an Intelligent Room," To appear in *Machine Vision and Applications, Special Issue*, Jun. 2003.

[5] K. S. Huang and M. M. Trivedi, "Streaming Face Recognition using Multicamera Video Arrays," *Proc. Int'l Conf. on Patt. Recog.*, vol. 4, pp. 213-216, Aug. 2002.

[6] F. V. Jensen, *Bayesian Networks and Decision Graphs*, Springer-Verlag, 2001.

[7] I. Mikic, M. M. Trivedi, E. Hunter, and P. Cosman, "Articulated Body Posture Estimation from Multi-Camera Voxel Data", *IEEE Int'l. Conf. on Comp. Vis. and Patt. Recog.*, vol. 1, pp. 455-460, Dec. 2001.

[8] M. M. Trivedi, K. S. Huang, and I. Mikic, "Intelligent Environments and Active Camera Networks, " *Proc. IEEE Int'l. Conf. Sys. Man Cyber.*, pp. 804-809, Oct. 2000.