Machine Vision and Applications (2003) 14: 215–217
Digital Object Identifier (DOI) 10.1007/s00138-002-0109-7

Machine Vision and Applications

Human movement capture and analysis in intelligent environments

Mohan M. Trivedi

Computer Vision and Robotics Research Laboratory, University of California at San Diego, La Jolla, CA 92093-0434, USA

Published online: 28 August 2003 –

Investigators from multiple disciplines and applications areas are interested in Human body modeling, tracking, and synthesis related topics. These topics offer fertile ground for challenging research problems as well as potential for a wide range of applications.

Our own interest and involvement in human modeling, analysis, and synthesis comes from a very specific need: that of developing "intelligent" environments or spaces [1–3]. An intelligent environment automatically derives and dynamically maintains an awareness of its composition as well as events and activities occurring within. Moreover, these spaces should be responsive to specific events and triggers. Such spaces need not be limited to rooms in buildings, but extend to outdoor environments [4] and any other spaces that humans occupy such as a performance on a stage or an automobile on a highway [5].

The spaces are monitored by multiple audio and video sensors, which can be unobtrusively embedded in the infrastructure. To avoid intrusion on the normal human activities in the space, all sensors, processors, and communication devices should remain "invisible" in the infrastructure. The system should also support natural and flexible interactions among the participants without specialized or encumbering devices.

In a conference room environment, multiple video cameras and microphones may be embedded in walls and furniture. Video and audio signals are analyzed in real time for a wide range of low-level tasks, including person identification, localization and tracking, and gesture and voice recognition [6]. Combining the analysis tasks with human face and body synthesis enables efficient interactions with remote observers, effectively merging disjoint spaces into a single intelligent environment.

We are currently embedding distributed video networks in rooms, laboratories, museums, and even outdoor public spaces in support of experimental research in this domain. This involves the development of new frameworks, architectures, and algorithms for audio and video processing as well as for the control of various functions associated with proper execution of a transaction within such intelligent spaces. These test beds are also helping to identify novel applications of such systems in distance learning, teleconferencing, entertainment, and smart homes.

There are several key elements to the development of intelligent spaces:

1 Multilevel interpretive analysis of body and movement

Intelligent environments need to support the wide array of natural interactions that their human inhabitants perform. This places great demands on the sensory system. Basically, these systems should be capable of providing multilevel descriptions of typical human activities so that semantic-level interpretation of the events can be achieved (Fig. 1). Some of the typical functionalities of such powerful sensory systems include tracking of people in 3-D, estimation of the poses and postures of the people, person recognition, event recognition, and body modeling and movement analysis [7, 8].

2 Integration of multiple video and audio streams, and multiple camera types

In order to achieve reliable and robust system performance in intelligent spaces where interactions among multiple people and other fixtures can be properly captured and analyzed, the systems require sensory information from multiple perspectives and at multiple resolutions. Researchers are beginning to consider these issues seriously. Our research is characterized by its emphasis on using large numbers of channels, both video and audio, to augment the precision and robustness of our algorithms. We believe that scene- and activity-analysis algorithms of the future will use similarly large numbers of channels as transducers become cheaper and the available I/O bandwidth of computers grows. There will be a need in the future for algorithms matched to systems offering large numbers of input channels.

Current computer techniques for sensing the environment have not yet caught up to the abilities of humans, partly because of the lack of cross-modal information sharing in computer perception algorithms. Robust audiovisual (AV) signatures of the participants, gestures, and events are required in the development of intelligent rooms. Approaches based purely

Correspondence to: e-mail: mtrivedi@ucsd.edu

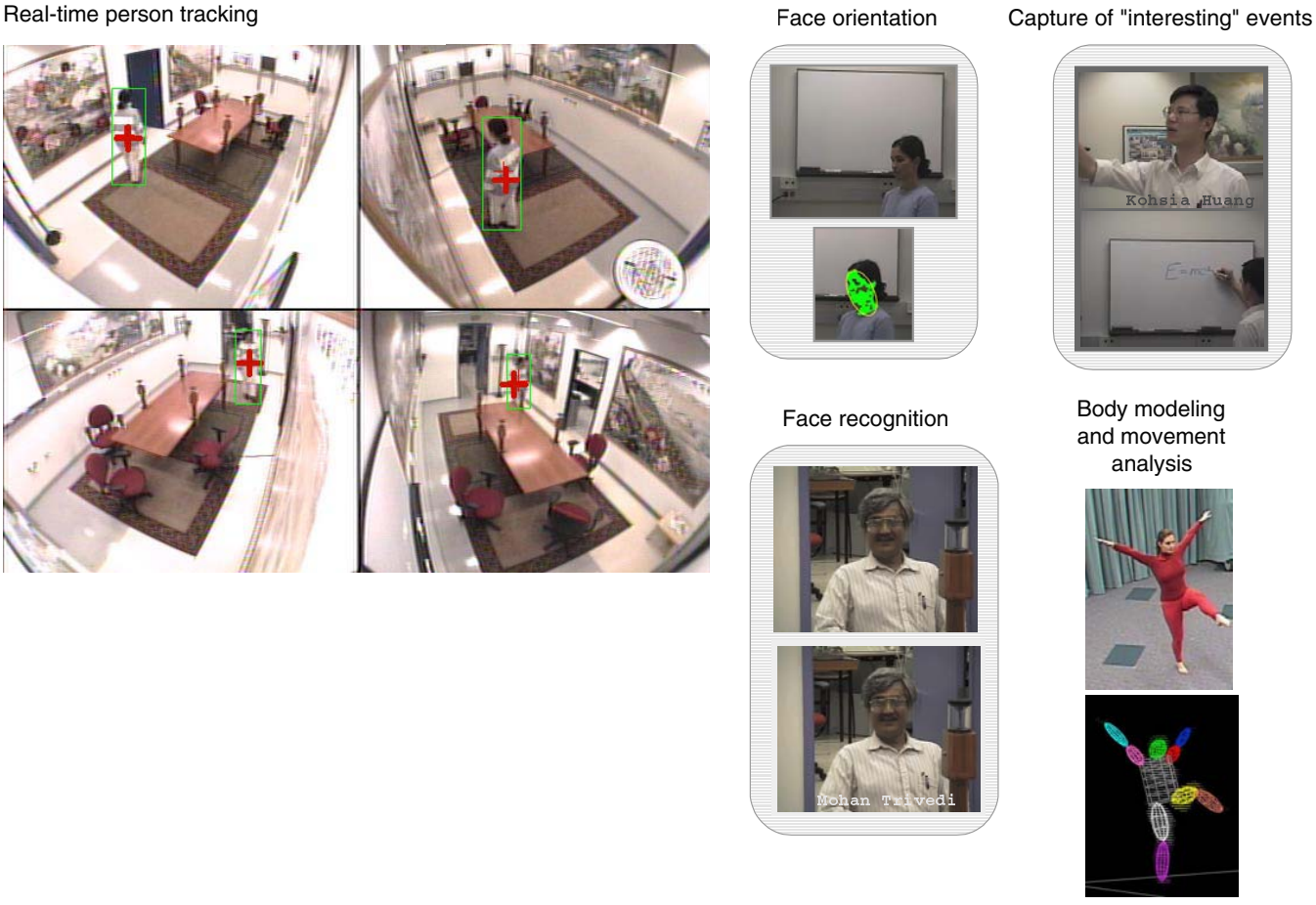

Fig. 1. Examples of a wide range of semantic level interpretations of visual signals captured in an intelligent environment

on vision or audio modalities are typically error-prone and require an extensive amount of training and computation for reliable and accurate performance. In the context of lipreading and person-identification applications, studies have shown that combining multimodal information greatly improves the system performance. We anticipate improvements of the systems for wider applications involving multiple people and dynamic events.

3 Integration of task models in sensory data processing

Scene analysis and understanding are high-level tasks, which will involve the relating of sensory information to the appropriate semantic and contextual models. Recent artificial intelligence research shows that this is still a very difficult problem which lacks a satisfactory solution. Our research is directed towards developing a hierarchical AV computational architecture, where efficient algorithms working on the raw sensory signals trigger higher level modules to evaluate potentially interesting clusters of AV sensory information fields for scene analysis. A detailed interpretation of an event would not be required from the lower level components of the system. A quick (but reliable) broad classification could be used to make decisions on the involvement of more sophisticated components. One could expect such a system to be reliable and efficient, since the time and computing resources are assigned to small areas in the environment where interesting events are occurring. Investigations are needed into the processing, representational, and control aspects governing such systems.

4 Integration of sensory data processing and databases

The robustness and control flow of the system can be systematically enabled by use of a distributed multiresolution database. A database is a persistent store of relevant past states of the overall system. This state contains various spatial entities, such as landmarks, signatures, historical patterns, tracking, localization results, and other useful parameters, and raw images. Because the state is spatially extensive, there are many issues related to the storage of such a state. The database solves such issues in a systematic and robust way. The databases define the architecture and the language for describing and recognizing semantic events and activities of interest. The database framework provides high-level abstractions for the storage and retrieval of semantic entities (e.g., an intruder or specific participant in a teleconference) that are flexibly defined by the users of this architecture. This enables flexibility and a comprehensive and robust means of interpretation of numerous results produced by different components of the system [9].

In the future, we will view the spaces we occupy not as passive but as intelligent entities. These rooms will be viewed as disembodied intelligent robots that will support an impres-

sive range of capabilities to help us to be more productive and efficient. There are many exciting and fundamental research challenges in realizing such intelligent environments, and they also have significant commercial potential.

References

1. Trivedi MM, Rao B, Ng KC (1999) Camera networks and microphone arrays for video conferencing. Proc. Multimedia Systems and Applications Conference (SPIE Vol. 3845), Boston, Mass., September, pp 384–390
2. Trivedi MM, Huang KS, Mikic I (2000) Intellivent and active camera networks. Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, October, pp 804–809
3. Huang KS, Trivedi MM (2003) Video arrays for real-time tracking of person, head, and face in an intelligent room. Machine Vision and Applications 14(2): 103–111
4. Trivedi MM, Mikic I, Kogut G (2000) Distributed video networks for incident detection and management. Proceedings of IEEE Intelligent Transportation Systems Conference, October, pp 155–160
5. Huang KS, Trivedi MM, Gandhi T (2003) Driver's view and vehicle surround estimation using omnidirectional video stream. Proceedings of IEEE Intelligent Vehicles Symposium, June, pp 444–449
6. Mikic I, Huang K, Trivedi MM (2000) Activity monitoring and summarization for an intelligent meeting room. Proceedings of the IEEE Workshop on Human Motion, December, pp 107–112
7. Ng KC, Ishiguro H, Trivedi MM, Sogo T (1999) Monitoring dynamically changing environments by a ubiquitous vision system. IEEE Workshop on Visual Surveillance, June, pp 67–73
8. Mikic I, Trivedi MM, Hunter E, Cosman P (2003) Human body model acquisition and tracking using voxel data. International Journal of Computer Vision 53(3): 199–223
9. Trivedi MM, Mikic I, Bhonsle S (2000) Active camera networks and semantic event databases for intelligent environments. IEEE Workshop on Human Modeling, Analysis and Synthesis, June

Mohan Manubhai Trivedi was born in Wardha, India, on October 4, 1953. He received a BE (Honors) in electronics from the Birla Institute of Technology and Science in Pilani, India, in 1974, and MS and PhD degrees in electrical engineering from the Utah State University in 1976 and 1979, respectively. Trivedi is currently a professor in the Electrical and Computer Engineering Department of the University of California, San Diego (UCSD), where he serves as the director of the Computer Vision and Robotics Research Laboratory (http://cvrr.ucsd.edu). He and his team are engaged in a broad range of research studies in active perception and novel machine vision systems, intelligent ("smart") environments, distributed video networks, and intelligent systems including intelligent highways and intelligent vehicles. At UCSD Trivedi also serves on the Executive Committee of the California Institute for Telecommunication and Information Technologies, Cal-(IT)^2 (http://www.calit2.net/), leading the team involved in intelligent transportation and telematics projects. He also serves as a charter member of the Executive Committee of the University of California system-wide digital media innovation (http://www.dimi.ucsb.edu/) (DiMI) Program. Trivedi serves as the Editor-in-Chief of the journal *Machine Vision and Applications*. He is a recipient of the Pioneer Award (Technical Activities) and the Meritorious Service Award of the IEEE Computer Society and the Distinguished Alumnus Award from the Utah State University. He is a fellow of the International Society for Optical Engineering (SPIE). He is listed in the Who's Who in the Frontiers of Science and Technology, Who's Who of in American Education, American Men and Women of Science, Who's Who in the World, and other similar publications. He has published extensively and has edited over a dozen volumes including books, special issues, video presentations, and conference proceedings. He servers regularly as a consultant to various national and international industry and government agencies.