Introduction

Christopher Jaynes and Robert Collins

How will we interact with our computational environments in the next 10 years? If the past is to be any indication of the future, our hands will remain restricted to a keyboard and our eyes will remain focused on a desktop display. It is true that progress in mobile computing, computer networks, and miniaturization allows users to access information through a number of less restrictive devices (e.g. video capture and playback on a cellphone). However, these devices really represent the same metaphor of interaction as the traditional desktop computer -- the keyboards and monitors are smaller, but they still sit between us the computational environments we access.

An interesting alterative to this traditional paradigm is to remove these devices from the picture altogether and build environments that directly provide access to our computational worlds. In this model, users are continuously immersed in a persistent interface and have seamless access to both the physical and non-physical (informational) aspects of their world. We refer to this paradigm as an *Interactive and Intelligent Environment* (IIE) in order to emphasize the integration of the interface with our everyday environments and the need for the interface to intelligently respond to users needs. Similar visions have been termed "Intelligent Buildings", "Augmented Environments", and "Smart Spaces".

Regardless of the name, the approach involves an environment that is able to respond to users requests, anticipate user needs, and provide interactive access to data anywhere within the augmented space. In this way the virtual world can be seamlessly integrated with our everyday physical environment without the need for user-specific devices. Instead, technology is used to endow the environment with the ability to track, recognize, and classify behavior so that is can respond to natural gestures and augment our experience with information and virtual objects. This future may not be as distant as it seems, and already one can imagine spaces covering entire buildings, augmented with projected information and interactive touch screens to provide virtual interfaces on potentially every surface.

Development of such interactive environments will require continued progress in a number of areas spanning computer vision, sensor networks, computer graphics, and human-computer interaction. Recent progress in Computer Vision and several of its subareas including vision-based human-computer interaction, gesture analysis, and tracking of human position and pose, can play an important role in this development. However, until recently, much of this research has been performed in isolation and has rarely been explored in the context of an interactive and intelligent environment.

This book represents a renewed interest by the Computer Vision community in the area of intelligent environments and the central role that camera-based monitoring and understanding of user behavior can play. In the fall of 2005, many of the top researchers in the field of Computer Vision were brought together at the Workshop on Computer Vision for Interactive and Intelligent Environments to discuss the role their research has to play and to speculate about the future of these immersive environments. Participants were challenged to envision an environment that is able to continuously monitor and respond to humans within an augmented physical space. Figure 1 depicts one such vision of an interactive environment.

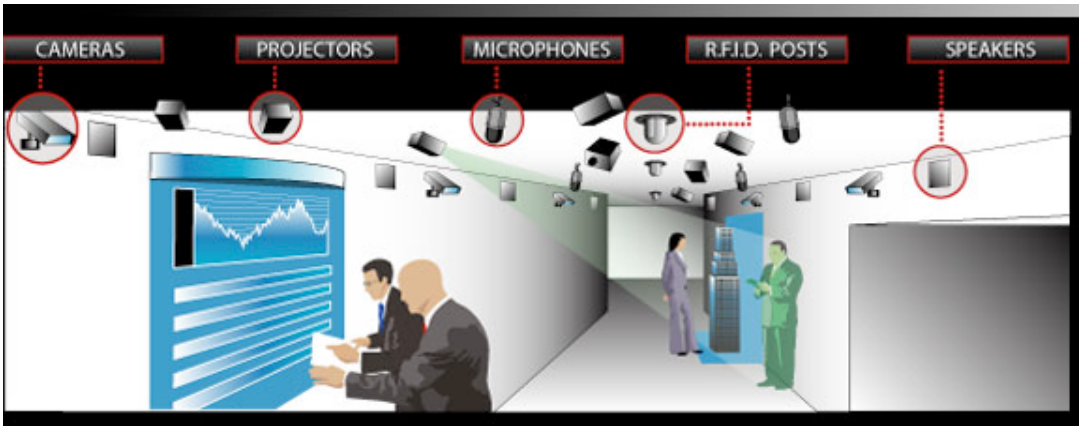

Figure 1: One vision of an interactive and intelligent environment in which the user is continuously immersed in an interface that supports collaboration, information access, and both virtual and augmented realities. Like most envisioned intelligent environments, the capabilities arise from significant infrastructure (i.e. cameras, projectors, speakers, microphones) that both senses and actuates in response to user needs.

In the example environment depicted in Figure 1, users are able to point to the nearest wall and request the display of data to support an ad hoc collaborative discussion-. The system may already know that the data is appropriate for display in a public environment (as opposed to private email, for example) and projects the data into the space so that it appears correct for the participants. Other users may request access to a three-dimensional model that augments the user's space. Avatars of real or virtual assistants are able to "walk" through the environment with the users.

One can imagine that as a visitor steps into such an environment, he/she is observed by several cameras. The newly detected subject is tracked and an appearance model is constructed. This model is compressed and added to a library of people and objects currently monitored as part of the global state of the environment. Since no radio-frequency identification (RFID) tag is present and the acquired appearance model does not match anyone already known to the system, simple reasoning allows the environment to denote the subject as a "visitor". Perhaps activity recognition techniques in concert with expression analysis and multi-view tracking will allow the environment to recognize a "wandering" behavior and alert the environment that the visitor is lost. Given the tracked position of the visitor and the calibrated location of digital light projectors in the area, a virtual assistant is rendered into the environment so that from the subject's eye position the avatar augments the world in a real manner. The user then interacts with the avatar via virtual touch menu and voice.

This vision is a significant departure from earlier virtual reality systems that sought to revolutionize the way we interact with our environments and one another. In particular, the mechanism is less device-specific and instead offloads user-specific devices into the environment. However, the roots of interactive and intelligent environments can best be traced to early work in virtual reality and human-computer interaction.

The goals of an interactive and intelligent environment were perhaps best articulated by Ivan Sutherland's landmark paper "The Ultimate Display" in 1965 that envisioned how computers may one day generate realities whose perceptual fidelity matches that of the real world. This compelling vision helped encourage a tremendous amount of research into novel

displays, human-computer interaction, and interactive computer graphics that began to be termed "Virtual Reality" (VR) in the early 1980s and reached public consciousness by the early 1990s. By this time, virtual reality systems, typically composed of a head-mounted display, an interactive pointing device, and a head tracking system, were introduced to the public. Entertainment kiosks that supported limited interaction and simple multi-player games could be found shopping malls and convention centers.

Unfortunately, these systems were of limited success and public use of virtual reality has not materialized as promised. As a result, the virtual reality envisioned by researchers in the 80s and 90s did not have significant impact on our daily lives. This was probably partly due to issues beyond the technical aspects of the systems. However, limitations of the technology certainly played a role in the public's general dislike of early VR systems.

Researchers responded to this with a renewed focus on the shortcomings of existing VR approaches, and by searching for technical alternatives to the traditional VR system. Both of these approaches have been successful in producing tremendous advances in interactive computer graphics and virtual reality over the past several years, and today's head mounted display system is far superior in both fidelity and usability than its counterpart from the early days of VR. However, as researchers have sought out new methodologies for realizing a computer-controlled environment, the base assumptions about what a virtual environment is, how is it accessed, and how it may impact our reality has been revisited.

Virtual Reality has been expanded to include Augmented Reality, in which the virtual and real are able to simultaneously interact with a user, and Telepresence, where (multiple) users share a sense of presence in a computer-mediated world. Virtual Realities may now be better termed Virtual Environments in order to imply that users are able to move freely within both in a virtual and real space.

Perhaps the biggest shift away from traditional virtual reality has been the realization that users would like to be immersed in environments that are not only accurate representations of some reality but are, in some sense, intelligent and responsive to the user. As Computer Vision and other sensing technologies that are able to formulate more complex models of the user emerge, virtual environments can be endowed with intelligence in support of more sophisticated interaction. These IIEs have many similarities to a traditional Virtual Environment but may also be capable of tracking user motion and behavior, recognizing user intention, and responding to user needs as they move through the environment.

This vision of a virtual reality is quite different than the traditional view in that the interaction between the user and the virtual space now plays a central role: the user is immersed in a interactive and intelligent interface whose metaphors for interaction can be virtual, augmented aspects of his/her environment, or simply the real environment itself.

This idea of IIEs has begun to take hold of a great many researchers over the past decade. Fundamental advances in head-worn see-through displays as well as digital light projectors allow researchers to explore how these devices can immerse a user in a visual space that is composed of both the real and virtual. Advances in both passive and active tracking technologies, capable of maintaining positional estimates over very wide-areas allow us to now untether the user to move through this virtual environment.

Of course, an IIE is bound to be quite complex and involve a variety of technical problems related to the system itself. It is unsurprising, then, that an active community of researchers has been focusing on IIEs over the past several years. Progress related to distributed processing, personal security issues, and collaboration within such environments has been explored by several different researchers. One goal of the Workshop was to explore the role Computer Vision can play as well. This book includes contributions from each of the Workshop participants. The chapters cover a variety of computer vision topics that are important to realizing an IIE and will provide the reader with a snapshot of the current state of the art.

In addition to describing their current efforts, workshop attendees participated in four forum discussions that focused on the future of computer vision and IIEs: 1) Displays and Interactivity; 2) Face, Gesture and Action Recognition; 3) System Challenges and Logistics; and 4) Calibration and Tracking. During these lively discussion sessions, participants offered thoughts on current technical issues, ideas for future research and tips on funding opportunities. As what was said in these sessions represent the opinions of some of the leading minds in the field, we summarize the highlights of the forums below.

Forum: Displays and Interactivity

Developing Useful Interactive Environments

Is it possible to make an intelligent, interactive environment (IIE) that isn't annoying? As anyone who has used a word processing program that attempts to "correct" while they type can attest, too much autonomy in a tool can be extremely distracting. Imagine the level of frustration caused when the building you are in starts imposing modes of interaction upon you. How productive will you be when you live inside the physical analog of an animated paperclip that keeps interrupting you with helpful reminders? We must strive to build environments that are supportive of the user's intelligence rather than trying to usurp their autonomy by thinking for them.

One way to soften the impact on the user is to make their interaction with the IIE as transparent as possible. Although the Windows desktop metaphor works well in virtual space, it may not be the most appropriate interface for interaction with the physical world. The ability to project interactive displays onto the walls of a room has led many of us to build gesture-based analogs to the point and click interface. However, the vision of standing in your living room communicating with your house through exaggerated arm motions like a navy flagman is not appealing. At a finer level, we can project menus or buttons on object surfaces for direct touch interaction, however, there are indications that the material of the surface you project on influences the likelihood of such interactions; e.g. people have a different comfort level touching paper versus milk. To twist a phrase by Rodney Brooks, the world is its own best metaphor, so perhaps we should use the world as our interface. Examples of seamless world interfaces include pointing at objects to select them and projecting illuminated footprints to guide someone through a store to the product they want.

Many prototype IIEs have implicitly assumed that the only thing you are doing at the moment is interacting with the system, and therefore can afford to do whatever the system requires for communication. This is wrong. You have more important things to do, and the system should be seamlessly supporting you in achieving these other tasks. If your hands are full

carrying something, it is inconvenient to try to use gestures to ask the room to turn on the lights or open the door for you. When walking or driving, you need your eyes to see where you are going and to avoid obstacles, and do not appreciate an interface that tries to monopolize your visual attention. Not surprisingly, the computer vision community has focused only on the visual aspects of communication, namely transmitting information through visual displays and receiving information from a camera watching the user. To go beyond such autistic IIEs, we need to explore new interface methods such as tactile and acoustic. Natural language interfaces, in particular, offer a convenient approach to interaction that is largely orthogonal to visual and gestural I/O.

There has been an evolution of IIEs from systems that allow you to access static information, to systems that monitor and tell you something about the current state of the world, to systems that are capable of physically manipulating the current state of the world. Early applications of IIEs focused on information manipulation, and were therefore extensions of the typical information retrieval and display process, but with the user being untethered from the computer keyboard, mouse and monitor. More useful systems can monitor the state of the dynamic world around them and can offer timely suggestion in answer to queries. Examples are office systems that highlight papers you haven't looked at for a while, parking lots with sensing to show where the free space is, or smart home environments that can tell you where you put your car keys. However, most people believe the largest impact is likely to come when IIEs are able to directly manipulate the physical world. To do this, IIE's need to be coupled with actuators, leading directly to the metaphor of IIE as a big robot that you live inside of. Examples are intelligent buildings that control heat and light based on user profiles, and future systems that will automatically tidy up the house for you by putting items back in their appropriate place.

Are Projector Displays Dead?

Most IIE interfaces have focused on the use of projectors, and on the feedback loop between cameras and projectors. But there are signs that projector technology may be a dead end. Projectors were originally the best way to make LARGE displays possible, however flat panel screens offer better resolution, and larger sizes (30-70in) are becoming affordable. Projectors are also good for displaying information on arbitrary surfaces in the environment, for example for augmented reality applications. However, flexible displays such as E-paper have been developed that can also be put on arbitrary surfaces. Projectors still have an edge for selectively displaying patches of information at different locations over a large candidate area, so that the display can follow the user through the scene, for example. However, this could change -- is smart paint around the corner?

One up-and-coming technology involves replacing ordinary incandescent or fluorescent lights in a building with light emitting diodes (LEDs). Recent advances in creating blue LEDs means that arrays that mix red, green and blue can be made to produce white light. Such LED arrays can produce controlled lighting as lights, but could also be selectively controlled to produce color displays. A current problem with such "smart lights" is cost, and the amount of time you use them as a display versus as a light may be too low at present to be cost effective. However, the burn out rate for LEDs is very low, and the technology continues to see large increases in brightness, energy efficiency, and longevity, such that LEDs may soon become cost effective for use, even just as lighting.

Forum: Face/Gesture/Activity

Current Issues in face recognition

Face recognition is a valuable biometric for security applications since it requires little cooperation from the subject. One insight from recent work in face recognition is that, given a large number of training examples for each individual, straightforward pattern recognition methods such as LDA or correlation filters work very well. The challenge for turning this insight to practical advantage is how to collect a large amount of ground-truthed training data "in the wild." Video processing obviously has an important role to play in rapidly collecting multiple images of a person across a variety of poses and facial expressions for use in training face recognition systems. The existence of multiple training views for an individual is one example of incorporating the notion of *familiarity* into the face recognition process. In human performance, prior familiarity with the subject leads to a higher recognition rate, even when a novel pose or expression is presented for classification. One challenge for vision research is to find ways to incorporate familiarity into recognition algorithms such that prior views of a person can be generalized to recognize their identity after large changes in pose, lighting or age.

The workshop participants identified several other promising new directions for face recognition research. Multi-modal recognition strategies have not received much study. These include combining face and iris recognition, or face with gait recognition. Audio-visual person verification is another example of combining information from multiple data sources. The challenge of multi-modal recognition is the lack of availability of datasets where multi-modal recordings are available for a common set of subjects. Currently available multi-modal datasets contain only small numbers of individuals. Face recognition from low resolution imagery is an interesting area for research that could be of great practical use in long-range surveillance applications. Finally, it was noted that there has been no real work done to address recognition of uncooperative subjects. As opposed to noncooperative subjects, who merely don't know that they are being observed, uncooperative subjects are actively trying to deceive, perhaps by wearing heavy makeup or disguise. Recognizing experienced uncooperative subjects is very difficult even for human observers.

Event Recognition is a Developing Area

In contrast to the great deal of attention that face recognition has received, event recognition and retrieval is a wide-open area for research. Indeed, there needs to be some discussion about how to precisely define what an "event" actually is. Opinions at the workshop varied from a computable predicate, to a salient change in time varying data, to a change in steady-state situation. The consensus was that an event represents a transition between one situation and the next, and that the definition of what constitutes an event is likely to be task dependent.

The challenge for developing a principled approach to studying events is that we, as vision researchers, are good at defining primitives (e.g. feature operators), but have been relatively less successful at defining higher-level concepts. We need a high-level language or lexicon for describing events. This lexicon must be flexible enough to be applicable across a variety of domains, yet be simple enough that a wide range of users can become proficient in specifying event queries. It is likely that progress will be first made in developing context-

specific ontologies for limited domains, such as shopping, tarmac security, and sports broadcasting.

Forum: System Challenges and Logistics

Systems Issues

Development of commercial IIEs will involve spending a lot of time and energy on issues that have nothing to do with computer vision. Some lessons learned from commercializing surveillance systems are summarized below. A distinction is made between systems issues, which are technical issues beyond those involved in the vision algorithms, and logistics issues, which are mainly nontechnical in nature.

The key to success is to have a well-designed system. The design process starts with having a clear definition of what tasks the system is being built to perform. This task definition drives the rest of the design. Although the designer must consider a range of issues, including (but not limited to) choice of sensors, lighting, power consumption, user interface, communications bandwidth, and even system security (how to prevent adversaries from breaking the system), each choice is guided by considering how that selection will contribute to performance of primary tasks.

Other systems issues to consider:

- Choice of software architecture is task dependent. An API for specific tasks and vision components needs to be designed, and code should be written in a manner that allows replication and sharing across applications. Robustness of algorithms is often more important than optimality in a fielded system.
- Thought needs to be given to how the system will be installed and configured in the field. How many engineers are required to install the system? Are camera modules designed to be plug and play? How difficult or lengthy are the camera calibration procedures? These concerns do not go away after the initial installation -- it is helpful to remember that up to 50 percent of cameras typically need to be replaced each year. The speed and ease with which a "dead" camera can be swapped, calibrated, and brought online is an important factor.
- A plan is needed for integrating the system with existing infrastructure and backend systems. Existing infrastructure should be leveraged whenever possible, such as displaying results on existing security terminals. Alternatively, the system could be designed to provide remote access and display through web interfaces, cellphone, or instant messaging services.

Logistics Issues

Perhaps most frustrating to technical people are the many hazards of a nontechnical nature that must be navigated when developing a surveillance system or IIE. Systems that involve cameras run into ethical, legal and social issues. It helps to have a concrete plan for privacy protection from the outset, since even experimental systems require IRB approval for testing. Getting cameras installed involves a lot of bureaucratic red tape. Outdoor installations, in particular, need to be approved by the building architect and facilities management. Contractors need to be hired to install power outlets, place camera mounts, and run cables. As a system reaches maturity, resources are needed for training operators and educating users.

Ultimately, a key logistic issue is getting customers to buy and install your system. The people who make this decision do not understand or care about the details of your vision algorithms. They care about positive return on investment, and the go/no-go decisions are based on bottom line estimates of cost to have the system versus cost not to have.

Automated surveillance systems have met with a public resistance that we can perhaps avoid with IIEs. The main obstacle hindering acceptance of surveillance networks is the perception that they are not there to help you, but rather to watch you in case you do something wrong. The more immediate feedback of interactive interfaces will hopefully relay the impression that the user is in control, and that the system exists to do something beneficial for the user. Applications are needed that have a clear, positive social impact. We should avoid the dangers of over-hyping our systems. Whereas failures of surveillance systems are not readily noticeable by the public, underperforming IIE interfaces will be glaringly apparent to even naïve users.

An unavoidable logistic issue is how to pay for vision research. The workshop members discussed several funding sources and opportunities. Commercial funding is starting to become available, for example to provide services that gather demographic marketing data from video surveillance cameras. However, commercial funding is still seen as an immature funding source, as the market is concerned with what we can do today and prefers technology that can immediately begin to pay for itself. The National Science Foundation (NSF) provides funding for basic research, but in small amounts. The Department of Defense is still the best source for long-term research funding, although programs are becoming shorter and more applications oriented. The Department of Homeland Security has a natural interest in surveillance technology for border security and biometric identification. In seeking funding, we should embrace and leverage vision success stories, such as development of Mpeg2 compression, optical character recognition, and automated industrial inspection systems. Computer vision shares the same dilemma as AI -- once a problem is solved it ceases to be considered part of the field. We need to remind our sponsors (and sometimes ourselves) of the positive impact that has come from previous vision research.

Forum: Calibration and Tracking

Tracking Issues

Tracking is characterized by reassociating object identity across spatial location and time while ignoring appearance differences due to lighting and viewpoint. It is helpful to think about tracking at two levels. The data level generates trajectories of location and appearance through continuous low-level tracking of features or image patches in video across short periods in time. At the high level, longer term tracking is achieved by associating ID with data level tracks after discontinuities in space and time, based on similarity of motion and appearance. This latter *data association* view of tracking is particularly useful when considering distributed networks of sensors.

A number of current research directions in tracking were discussed:

- Figure/ground segmentation is emerging as a vital tool for tracking objects through camouflage or clutter situations. The classic approach is to do motion segmentation via background subtraction from static cameras. This is not always possible due to camera platform motion, and when looking at scene with crowds where there is no static background to model. Recent approaches explore appearance-based

segmentation of the foreground target by on-the-fly training of figure/ground classifiers using samples of pixels drawn from the foreground object region and nearby scene background.

- When tracking objects through significant changes in viewpoint, old tracking features go out of view and need to be replaced with new visible ones. Rather than 2D appearance-based modeling, approaches based on modeling 3D structure of the object are less sensitive to viewpoint changes. These approaches work by doing view-based rendering, and comparing the rendered image features to those observed in the image. Instead of tracking in image appearance space, one instead tracks the "viewpoint", resulting in a more constrained problem.
- Recent advances in object recognition have exciting implications for object tracking. Recognition is actually an umbrella term that can mean identification (who is this), re-identification (is this the same object I saw before); and categorization/classification (is this a cow). The re-identification problem in object recognition is identical to the data association problem in tracking. Local invariant feature based representations (e.g. SIFT keys) for modeling object appearance are relevant to tracking. One approach is to learn aspect graphs of feature-based appearance models on the fly during continuous tracking, for use in recognizing the object again after loss due to occlusion or switching to another camera in the network.

Camera Configurations and Calibration

Is it better to have high resolution images at low frame rate, or low resolution images at high frame rate, for tracking? If smooth trajectories are desired during the continuous tracking phase, high frame rate is better. When doing data association to reidentify an object after a lapse in coverage, high resolution is preferred. One happy medium is to use a megapixel camera that can downsample images by binning. Another approach that has proven to "sell" well is to use a hybrid system consisting of one large field of view (FOV) camera plus an active pan/tilt/zoom small FOV camera. The large FOV is essential if you require low latency detectability everywhere in the scene, whereas results from that view can be used to control pan, tilt and zoom of a second, active system to take higher-resolution close-ups.

Would it ever by more cost effective to have many small cameras versus a hybrid system with one large FOV and one pan/tilt/zoom camera? If they are all co-located, perhaps not. However, by spreading multiple cameras through the scene to form a sensor network, one can do cooperative multicamera tracking to avoid problems with occlusions that occur from any single viewpoint. From a calibration viewpoint, it is harder to calibrate a sensor network than a hybrid large FOV / PTZ system. Being nearly co-located with the PTZ camera, a pixel in the large FOV image maps to a viewing ray that can be likewise mapped to a pan/tilt angle, without regard to structure of the 3D world. It is much harder to get widely spaced cameras in a sensor network to look at the same object, as one has to hypothesize object distance from at least one of the cameras using known size or ground-plane constraints.

The issue of how to rapidly install and calibrate a multi-camera sensor network led to an interesting theoretical discussion. In the limit, would you rather calibrate one 10 megapixel camera, or 10 million one pixel cameras? It is far easier to calibrate one many-pixel camera, because there are fewer parameters to consider (relationships between the sensing elements

are tightly constrained by the CCD layout). However, large multi-sensor networks formed by inexpensive photocells are more like the latter. What does it even mean to calibrate a single pixel camera? Certainly its location in the world is an important piece of information, since it can then be used as a crude motion sensor via change detection. Combining information from multiple single pixel cameras could be achieved by additionally knowing their viewing rays (to compute correspondences between detections) and perhaps also their intensity point spread functions. It was agreed that a promising approach to multi-sensor calibration is to use a "firefly" approach to active calibration where some unique, distinctive, well-localizable object is waved or carried through the scene to generate object point correspondences in a simple and robust manner.

Speakers

In addition to the discussion forums, the workshop also featured short presentations by each participant on their work and how it relates to IIEs. These oral presentations were augmented with companion papers in the proceedings, revised versions of which comprise this volume. The following brief summaries of speaker presentations thus also serve as an introduction to the papers.

Rama Chellappa presented an algorithm for estimating articulated human body shape and motion from multi-camera video to achieve markerless motion capture. During his presentation, he focused on analysis of gait motion signatures, such as the helical pattern of x-t spatio-temporal slices of the legs during walking. His group has used such patterns to isolate abnormal walking behavior, and to analyze subtle variations in gait due to carrying ungainly objects, work that has applications to homeland security.

Claudio Pinhanez focused on commercial applications of IIEs. Using pan/tilt camera-projector displays, called "Everywhere Displays", he and his colleagues at IBM are exploring the paradigm of direct user interaction with merchandise by pointing and touching. He stressed the importance of estimating the gaze direction of the customer when projecting virtual displays on shelves and items.

Tanveer Syeda-Mahmood pioneered the analysis of actions as geometric shapes, specifically as generalized cylinders swept out by a varying image silhouette through time. She employs this insight to segment actions from video, and to recognize actions across variations in viewpoint and execution style by aligning these action cylinders. She pointed out during her talk that researchers studying event recognition in video are still searching for "killer apps" that will propel the field to commercial success.

Terry Boult considers the tradoff between security and privacy in video surveillance applications. His recent work combines cryptographic methods with video change detection to preserve privacy through identity obscuration. Blocks of pixels on the person are "scrambled" in the frame such that a casual observer can identify where the subject is going and what they are doing, but cannot recognize who they are. At the same time, when a crime or violation is observed, authorized personnel can recover the unaltered video data to identify the culprit.

Narendra Ahuja presented several novel sensor hardware designs. These include a spinning sensor with varying focal point to capture omnifocus imagery, a split aperture camera for high dynamic range recording, hemispherical and spherical cameras for capturing very large fields

of view, and a device for single-lens capture of depth images based on rotating a thick glass plate in front of the lens.

Tsuhan Chen has developed a sampling theory for image-based rendering in multicamera arrays, showing that nonuniform distribution of cameras can result in better rendition quality. He presented video of a recently built self-reconfigurable array of 48 IP cameras that dynamically change their locations and pan angles to optimize rendering performance with respect to current scene structure.

Zhengyou Zhang summarized his comprehensive body of work dealing with collaborative spaces containing a whiteboard, projector and camera. Tools range from digitizing high resolution snapshots of white board content, recording and browsing dynamic whiteboard content, hypothesizing keyframes for indexing the evolution of a whiteboard presentation, and displaying real-time annotations made by a remote user. An interesting issue regarding the latter is visual echo cancellation -- how to tell which strokes are real markings on the whiteboard versus virtual strokes projected from the remote user.

Joe Mundy seeks to make object recognition more of an engineering discipline through careful testing and analysis of different representations and learning algorithms on controlled datasets. His current work demonstrates these principles on vehicle object class recognition. He posited a potential future funding opportunity along the lines of a “genome project” for training object recognition systems

Marc Pollefeys considers multiview calibration and motion capture. He has developed algorithms for calibration and time synchronization of multiple cameras, calculation of lens distortion parameters, and location of occlusion regions in video by watching people move through the scene. Body pose recovery is performed using a factorization algorithm designed for articulated objects. He discussed several virtual reality applications, including medicine, telepresence, forensics, and preservation of cultural heritage.

Rainer Lienhart also considered the problem of multiview calibration. He has designed auto-calibration algorithms for self-organizing camera networks, and tools to support rapid calibration of large multimedia systems. Calibrating camera geometry can be as simple as waving a laptop display pattern in front of each camera node. Wide-baseline sensor arrays are time synched to an accuracy of 10-20 microseconds by broadcasting a wireless sync signal. This is currently good enough for synching up audio recordings, and methods with higher accuracy are being developed for video frame synchronization.

Greg Welch presented an efficient and general model-based approach to track a 3D object from either single or multi-view video sequences in real-time. Views of the object model rendered under current and perturbed pose estimates are used to guide the search for matching features and to empirically determine the Jacobian of feature measurements with respect to pose parameters. This information is passed to an extended Kalman filter to maintain an accurate estimate of object pose over time.

Rahul Sukthankar gave an overview of his work on camera-projector systems. The technology includes methods for rectifying projector output via homographies, eliminating the shadow of the speaker on the screen, and avoiding shining bright lights in the speaker’s face. He has demonstrated that fine tracking and calibration can achieve stable projection of content on small, hand-held display surfaces carried by a user. In conjunction with Carnegie Mellon University, he is developing algorithms for detecting objects and events, as well as

pursuing the futuristic concept of small "robots" that self-assemble into larger shapes (smart matter).

Amit Roy-Chowdhury showed that the joint effects of lighting and motion on object appearance in video can be approximated in a bilinear space. This allows him to estimate 3D pose and illumination over time, leading to a 3D model-based tracker that is invariant to large changes in illumination. He is currently working to apply this approach to pose and illumination invariant recognition of objects and activities.

Jeremy Cooperstock develops technologies for human-to-human interaction using the computer as an intermediary. Many traditional vision and graphics problems must be solved to achieve this goal, including object segmentation, tracking, layer compositing, perspective correction and low-latency rendering. A key problem is how to generate views that look correct from the perspective of multiple, widely spaced users, since that ultimately must be achieved to support realistic interactions among large numbers of people. A highlight of the talk was video of a demo system for virtually bringing musicians together for coast-to-coast jamming.

Jim Crowley gave a thought-provoking talk on social computing. He stressed the need for designing digital devices with an awareness of social context in order to avoid the "ubiquitous disruption" that surely looms as a consequence of the increasing density of digital devices per cubic meter. Working by analogy to concepts from theater, he has developed a framework and situational theory for social interaction, and a software architecture that uses perception of social situations and roles to provide non-disruptive user services.

Ramesh Raskar motivates the use of RFID tags in libraries, warehouses, and other spaces to "solve" the problem of object recognition, but notes that the technology merely indicates the presence of an object while not offering clues to its location or orientation. He proposes a novel RFIG (radio frequency identity and geometry) approach that uses a combination of photodetectors and projected structured light to compute pose of the tags. One item of interest after his presentation was a small, hand-held pocket projector from Mitsubishi that he brought to show the group.