9807-28
Variable Viewpoint Reality

Paul Viola and Eric Grimson

Progress to Date

Two years ago we proposed a project aimed at recording, in great detail, the actions of human atheletes on the playing field. At the time our goal was to construct new spectator environments which would allow for a complete immersion in the sporting event. Rather than sitting passiviely in front of a television set, the VVR user would be involved in the event - able to move about on the playing field to watch the action from their favorite viewpiont.

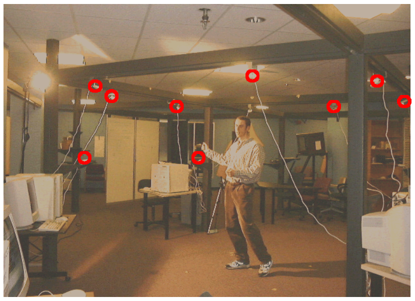
With support from NTT we have achieved a number of important milestones:

- We have constructed a parallel distributed real-time computer vision system which processes images from 16 cameras and produces accurate 3D reconstructions.

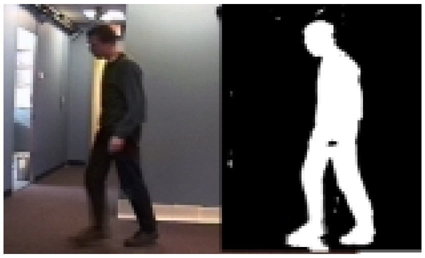
- We have developed several novel algorithms for 3D reconstruction. These algorithms work very quickly and are robust in real environments.

- We have developed automatic algorithms for camera calibration. While camera calibration is a critical component of any 3D vision system, it is especially so for a system with 16 (or more cameras). Using our technique calibration proceeds automatically while the system is in use.

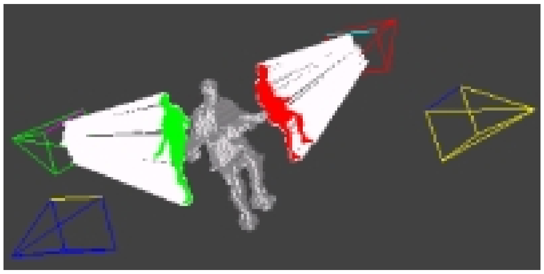
The main areas of progress are summarized in the figure below:



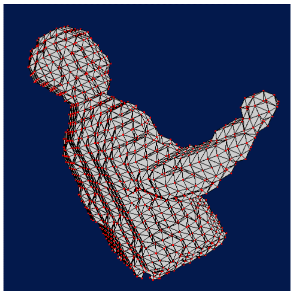
A multi-camera parallel-processing system has been constructed



Dynamic and adaptive tracking is used to find people.



Segmentation information from multiple cameras is used to perform ***real-time*** 3D reconstruction.



Reconstruction using several cameras leads to very accurate results.

In addition we have collaborated with other funded projects in the lab to develop a number of related systems:

- A real-time system for tracking people and vehicles in outdoor environments. The system has been tested for months at a time and can work in a variety of lighting and weather conditions.

- A related system which automatically clusters and classifiies people based on their actions and behaviors.

Research Proposal

Taken together we have developed a technology suite which addresses most of the central issues in contructing a VVR system. We plan to continue work on this enabling technology. Over the next year we plan to:

- Construct larger scale 3D acquisition area, including up to 32 cameras.

- Develop systems which can track the articulated movement of the human body. The current system attempts to reconstruct the shape of the human body. An attractive alternative is to estimate the positions and orientations of the joints in the body (thereby representing the human posture in 20 or 30 numbers).

- Develop systems which can track multiple people and analyze their overall behavior. This will allow us to label and analyze the plays, fouls, etc. in sporting events.

The currently funded project has as its final goal the construction of a system for viewing sporting events. We believe that this is an application area which is high profile and potentially profitable. Unfortunately, it may be beyond the capabilities of a basic research effort to develop a fully functional system. The main difficulty is one of scale. The final system must work outdoors, will include very many cameras, and will require a great deal of computer hardware. We believe that development requires several dedicated engineers with industral experience. We are very interested finding an industrial partner for this development, but since the project is understandably risky it is possible that the final system will not be built for some time.

A New Focus: The Computer Vision Macroscope

Over the next few years we hope develop a smaller scale system which nevertheless has broader scientific goals. We call this system the *computer vision macroscope*.

The development of the microscope in the 17th century initiated a fabulous stream of scientific discoveries. Bacteria, the neuron, and the charge of an electron were all found using the microscope. None of these advances was foreseen when the first microscope was assembled.

In the spirit of the microscope, we propose to construct a ***macroscope**,* a device that can record a complete three-dimensional record of every event in a large area, perhaps 20 meters square, over an extended period of time. Constructing such a device will require that we solve a number of central problems in vision, perceptual interfaces, machine learning, image synthesis and databases.

While construction of the macroscope may be interesting from a scientific and engineering standpoint, it is the *application* of the macrope which is most exciting. The macroscope will be a fundamental component of a new generation of computational interfaces, in which the macroscope can learn to recognize common gestures, actions and patterns of behavior of a person in different contexts and settings. The macroscope will allow for the study human behavior and interaction including: the control and execution of human movement, the role of information presentation in distributed collaboration, the archiving and retrieval of event information, and other issues of interactive, extended events.

Impact of the Macroscope

A macroscope will serve both as a novel tool for information acquisition and management, and as a novel tool for interaction between people and their computational partners.

A macroscope could be used to support ***event archiving***. One could capture a complete record of actions taking place in a site over a range of time periods. A complete record includes a 3D reconstruction of participants and their actions from any viewpoint throughout the event. This would support storing, archiving, retrieving and recreating any event and would create a new kind of information record an *event-graph* (much like a phonograph or photograph).

Such event archiving could be used for many purposes. For example, one could use these event-graphs to study human interaction. Imagine capturing a complete record of the actions taken in an operating room or a disaster relief center. Researchers could go back to any point in time to study where each person was looking, to what information she had access, and most importantly what available information was overlooked because of poor information display or inadequate communications. Understanding these sorts of interactions can lead to better decision-making, better human-computer interface design, and ultimately better outcomes.

In addition to archiving events, a macroscope could be used for ***event analysis***. Given the potential inundation of data from the macroscope, we need tools to extract salient information: finding patterns of movement at all spatial and temporal scales; fitting models to extract critical parameters; and recognizing instances of similar events from stored archives. For example, a macroscope could be used to record human movement, which could then be analyzed and fit to biomechanical models, in an effort to critique and improve the performance on tasks like aircraft maintenance or emergency rescue. This could lead to better training methods and better presentation of information.

Alternatively, the macroscope could be used to make detailed diagnostic measurements of human movement such as grasping or walking. When used over time, it could be used to monitor degenerative conditions such as Parkinsons disease. A macroscope could be used to record human performance in tasks like driving a vehicle, which could lead to better understanding of information overload and efficient delivery of information during real-time tasks. With a change in time scale, a macroscope could be used to record and analyze the evolution of other physical phenomena besides people and their actions.

Thus the goals of this project are twofold. We wish to create the technological components and scientific insights needed to construct a macroscope. Once this goal is complete, we want to use the macroscope as a revolutionary tool to analyze human and other actions in a wide range of settings, leading to interfaces that intelligently respond to natural actions.

Semantic Analysis of Activity

The macroscope will generate an incredible amount of data. Even after 3D reconstruction, automatic tools will be necessary to quantify, analyze, archive, and recognize activity. We believe that this analysis is naturally decomposed into three levels of spatial resolution: far, intermediate, and near field; and three similar ranges of temporal resolution: long, medium, and short term. While the computations at each level of analysis will be similar, the goals and requirements will be different.

1.1.1 Far Field Analysis

In past work we have found that several types of important behavioral information can be deduced from afar, typically by observing over long periods of time, and we believe that additional methods for analyzing behavior can be developed. Examples of analysis include:

Learn normal patterns of activity around the site: where do people walk, at what speed do they normally walk through the site, what are the normal volumes of people in the site as a function of time of day, in what other activities do people engage, what interactions between people occur? This allows the system to create a model of expected behavior, as well as gather statistics on activities in the site. Although additional effort remains, we have already demonstrated such basic capabilities.

Check changes in normal patterns of activity: imagine a small macroscope installed within the apartment of an elderly person. By building models of normal activity, the macroscope can then detect changes in such patterns: a lack of activity for a particular time of day, a person in a place that does not normally involve people (lying beside a bed for a long period of time), even a slow degradation in the movement of a person over time. This allows us to alert human monitors, or to focus additional attention on particular events and individuals. Initial versions of such capabilities are available; however, additional work in learning and classification will improve these capabilities.

1.2 Moderate Field Analysis

At this resolution, the detailed action of an individual or small group is of primary interest. Analyses of actions over a range of times are of interest. One might simply want to categorize single gestures of a person, or to record an entire sporting event. Several key questions arise:

Recognition based on biometric and dynamic gait information: One key capability is to recognize people based on their gait or other biometric information, as well as to analyze changes in biometric parameters over time (e.g., how does this person's movement compare to movement a week ago, a month ago, a year ago). This requires the creation of multi-modal representations of individuals a profile of an individual that combines normalized face images, data about the biometrics of the person (e.g. limb lengths), data about the dynamics of the person (e.g. properties of the person's gait, and other parameters of a specific person. This will require methods for extracting components of moving objects, such as the limbs, torso, and head of a person; methods for tracking those components over time to extract trajectories relating to that person and his/her gait; methods for determining the 3D structure of the person.

Recognition of movement and action information: We also need to interpret these representations of people's motions. This requires methods for matching multi-modal representations of people has this person appeared in the site in the recent past? In addition to methods for matching images of faces, this will require methods for characterizing the variance in gait and other parameters, and methods for matching such multi-modal representations. Given that we can recover information about gait from subjects, can we recognize when we are viewing two instances of the same person?

For example, gait can be represented as a multi-dimensional signal across time. Embedded in this data is a rich set of identifying information: the inertia of limbs, the period of the walking motion, any irregularity in the gait, etc. Note that a gait is something like a temporal texture. A visual texture is semi-periodic and contains repeating structure. One intriguing idea is to attempt to recognize gait trajectories using the same algorithms we have used to recognize visual texture.

In addition to recognizing movements and actions, we want to fit them to models of such activities. We will use methods from kinesthesiology and from motor control to match biomechanical parameters to observed data, and to use such models to categorize actions into a hierarchy of semantically meaningful units.

1.2.1 Near Field Analysis

At this resolution key issues include determination of head direction, gaze direction, location of lips and perhaps lip reading, fine detail positions and movements of limbs. This is in addition to recognition tasks, such as face identification or iris recognition. These types of analyses require very high-resolution information, but also provide a great deal of information about the intent of the participants. In a command post or operating room situation it might be possible to analyze in detail the genesis of some mistaken action. For example it will be possible to determine that a surgeon could not have seen critical information on a monitor simply because he never looked in that direction. With the addition of sound localization tools, it will be possible to conclude that the anesthesiologist never reminded the surgeon to look at his screen.