Immersive Sporting Events
9807-28

Proposal for 1998-1999 Funding

W. Eric L. Grimson and Paul Viola

Abstract

In the foreseeable future, sporting events will be recorded in super high fidelity from hundreds or even thousands of cameras. Currently the nature of television broadcasting demands that only a single viewpoint be shown, at any particular time. This viewpoint is necessarily a compromise and is typically designed to displease the fewest number of viewers. We propose to create a new viewing paradigm which will take advantage of recent and emerging methods in computer vision, virtual reality and computer graphics technology, together with the computational capabilities likely to be available on next generation machines and networks. This new paradigm will allow each viewer the ability to view the field from any arbitrary viewpoint -- from the point of view of the ball headed toward the soccer goal; or from that of the goalie defending the goal; as the quarterback dropping back to pass; or as a hitter waiting for a pitch. In this way, the viewer can observe exactly those portions of the game which most interest him, and from the viewpoint that most interests him (e.g. some fans may want to have the best view of Michael Jordan as he sails toward the basket; others may want to see the world from his point of view). We propose developing a prototypical system which will allow each observer to view whatever he finds most exciting and interesting.

Scenario

In the year 2002 the World Cup will be played in Japan and Korea and perhaps 4 billion people will watch the games on television. At each game hundreds of cameras will follow every move of the players, ball and referee. But across Japan only a single viewpoint will be available to the tens of millions of viewers—the televised viewpoint. Some viewers may wish to see the ball streaking toward the goal, as Chilavert (Goal keeper for Paraguay) might see it. Others might want to follow every movement of Ronaldo (Star striker for Brazil). Still others may want to understand theintricacies of the Italian defense. And others may want to understand how the current Dutch offensive patterns compare to previous games. Instead of transmitting a single video stream, or even an HDTV stream, we propose that a new type of signal be transmitted -- one that includes all of the information necessary to compute and display any arbitrary viewpoint. We call this signal an *immersive video* stream. Received using special hardware, immersive video can be

displayed on a virtual reality headset or an ordinary television. A number of user interfaces are possible, but all will allow the user to freely move about the 3D space in the stadium in order to watch the action in a manner that best suits individual tastes.

This proposal outlines a series of technical steps required to develop a prototype immersive video sporting arena. The eventual system will have a variety of capabilities including:

• **Dynamic 3D Reconstruction of the Game:** Using the video signals from many cameras, computer vision methods will allow us to create time-varying 3D models of the field, players, referees, and ball. From these global models, we can extract the information necessary to synthesize arbitrary, time-varying viewpoints of the event.

• **Automatic Understanding and Annotation of Plays, Strategies, and Statistics:** By tracking the players and the ball, and by using variations on existing activity classification systems, game events will be automatically identified and labeled. For example one could ask the system to display an event from the past like ``Show the pass to Ronaldo that was called offsides''.

• **Detailed, High-resolution Rendering of Articulated, Dynamic Models of Players:** Using articulated models for the human body, the motions and actions of the players will be tracked. These augmented models can then be used to create high quality renderings of player motion from arbitrary viewpoint.

In addition to these key capabilities, we will develop several related technologies that will support the computation of immersive video including: automatic tracking cameras which will concentrate available resolution on critical game events; and a ``foveated'' camera which can provide higher resolution on the tracked object combined with a wide field of view.

Constructing 3D Models

The representation underlying the immersive video stream is the computation of a dynamic 3 dimensional model of the players (0f course a 3D model of the field and goals are important also important, but these can be computed before the game). From this 3D data, arbitrary viewpoints can be constructed using computer graphics rendering in real-time.

The **first stage** is the integration of multiple cameras into a single global reference frame. This will require algorithms for accurate camera calibration, pose estimation, and image mosaicing, areas in which we have considerable experience (e.g., [2], [3], [5]).

The **second stage** segments the players, officials, and ball from the stationary field. While simple image differencing may work in some cases, in many real imaging situation more complex algorithms are necessary. We will investigate techniques that build robust statistical models for the background, enabling one to reliably segment moving objects from background [4]. At this point it will be possible to record and analyze the rough 3D positions of each player, the ball and the referees.

The **third stage** takes the segmented video stream and builds a coarse, global 3D voxel model of the sporting event. In each camera's video stream, the 2D locations of background and foreground pixels strongly constrain the 3D space that a player may occupy. Casting a 3D ray from each camera's optical center through each background pixel in the camera's image plane carves out a superset of the 3D volume occupied by the player. Intersecting the volumes generated by multiple cameras produces a 3D region containing the player. The shapes will be very rough since they are computed as an intersection of silhouettes. Nevertheless if these shapes are texture mapped carefully the resulting 3D scene should yield reasonably good synthetic views.

The **fourth stage** further refines this 3D model by a process very similar to wide base-line stereo. An iterative process eliminates any voxel which when projected into two (or more) cameras yields a different intensity. The false matches which are possible in stereo are severely restricted by the presence of multiple cameras.

The final result will be a rough, voxelated representation of the players, referees, ball, and playing surface in world coordinates. These shapes can be smoothed and texture mapped to produce a photorealistic appearance from a variety of virtual views at a distance. We believe that reasonable results can be obtained with fairly coarse models under the following circumstances: the models are texture mapped with the image captured from a nearby camera and the model is viewed from a reasonable distance. For closer views of individual players, the modeler would switch to a more detailed 3D modeling strategy discussed in Section Full Immersion.

We believe that a demonstration system illustrating these cabilities can be developed based on existing methods, and that a fuller system conforming to the above description is likely to commercially viable. Many of the component algorithms are relatively efficient and reliable, and can be easily extended to example domain. Nevertheless it is likely that significant research will be required to fully adapt these algorithms into an integrated system, and to cover the full range of capabilities required by a high resolution, real-time system.

In addition to leveraging our experience in computer vision methods for motion segmentation, shape recovery and scene modeling, there is great potential for the application of statistical machine learning in each of the computation stages. Algorithms must be able to correctly identify the background given changing lighting, camera position, and noise. Mistakes in these estimates will propogate throughout the 3D model

construction process, yielding ``false'' blobs and missing components. Simple smoothing is likely to improve the quality of the resulting models, but it is likely that these artifacts will be more accurately reduced by the use of statistical models for both the shapes and movements of the players. We will build on our current body of learning methods applied to vision problems to create methods of statistically modeling articulated, dynamic objects such as people [1]. Such methods will have utility in a wide range of applications beyond the virtual sports arena.

The Roaming Spectator

A 3D time-varying model of a sporting event allows the viewer to move about the sporting arena freely while watching the game in real time. At the same time, the viewer should have the option of passively viewing the game from a variety of viewpoints. We propose augmenting the model with automatically generated activity annotations describing statistics about athletes, set plays, fouls committed, etc. We have extensive experience (e.g., [4]) in developing methods that learn common patterns of activity from passively observed visual motion, which we will build on to create sports annotation capabilities. Not only viewers, but announcers with various broadcasting styles, would use these annotations to create their own versions of the game's broadcast. As well, such automatic classification and annotation capabilities support indexing into stored data. Thus, a viewer can use selected plays to find clips of similar plays from previous games, thereby creating customized archiving of replays.

Customized Sports Commentary

For people who tire of controlling their immersive experience, we envision using our system to support a range of announcing crews that espouse different styles or speak different languages, all creating their own version of the game including: which cameras are displayed, how often to switch cameras, what information to overlay, colorful commentary, etc. Every group gets the full video feed, and can select one of the "commentator/producer/director" streams from the announcement crew which contains audio, which video to display, and any overlays that should appear on the screen. This would allow more traditional fans to see a single, full-field view and hear their favorite announcer say "GOOOOOOOAAAAALLLLL!" while the younger generation may prefer a more MTV style, switching from one player's virtual head cam to another in quick succession (which might give older people a headache).

Replays on Demand

Imagine the following interface: The user is presented with a display in which an overhead "field view", showing the players, referees, and ball, covers much of the screen. A "now" button and time slider is laid out below the "field view". The time slider is color coded with events of interest (e.g. change of possession, goals, etc.). Clicking anywhere on this time slider shows the player/ball configuration at that point in time.Clicking on the "now" button

returns to the present. Dragging on that slider selects a region in time to be shown. The camera positions/orientations/fields-of-view can also be overlaid on this display allowing an editor (or casual TV watcher) to quickly select and replay the best video for a particular goal or tackle. This could also be a way to watch a 90 minute match in five minutes by dragging over the uninteresting regions or creating filters (e.g. show me the minute surrounding each time either goalie is less than 5 feet from the ball). In future work, we would add an interface to select virtual views instead of the camera views themselves.

Other capabilities that utilize the immersive video stream include examples such as:

• Color code the field to show which team has controlled more of the field.

•Measure the overall statistics of teams and individual players.

•Automatically maintain the ratio of time the ball has been on one side or the other or the ratio of possession time.

•Find similar plays in other recent matches.

•Automatically classify goals, shots on goal, corner kicks, throw-ins, touches, and penalty kicks (so people won't have to manually annotate the games). This information and manually entered information (e.g. which players were carded, why play stopped, who left the game, etc.) could be used to color code the timeline, although some people may want this turned off if they are watching a game which has already been played.

Full Immersion

While distant views of a game are ideal for observing player positioning and team strategies, close up views are essential for witnessing the intricate skills, maneuvers, and hidden fouls of individual players. Using specific information about players identities and the dynamics of the athletes' physical motion, the rough 3D models from Section Constructing 3D Models may be further refined.

Stored 3D scans of individual players and articulated, dynamic models of human motion while performing athletic feats such as kicking a ball, swinging a racket, or doing a hand spring may be created and stored off-line. Then using higher resolution footage, such cues as players' shirt numbers, faces, hair, and other distinguishing characteristics would facilitate automatic discrimination between individually tracked players and referees. Coupling these data sets and using a mixture of 3D modeling and image-based rendering techniques, detailed 3D models of players in action can be synthesized for a viewer to examine from any direction.

Personal Cameras

A key question that must be addressed is ``how many cameras and what resolutions are needed to support a range of immersive video experiences''. In practice the system must support the high resolution needed to provide realistic closeup detail of a player, while still covering the full action of the entire field. A potential solution is to construct a camera which can physically track the players. Calibration of such a system is often very difficult. One approach is to rigidly attach a high resolution, narrow field of view camera to a wide angle camera. The narrow field camera would provide high resolution information while the wide field camera could be used to calibrate the assembly to world coordinates.

Ideally one could construct a ``foveated'' camera which could be used for both tasks. Such a camera would have a non-uniform spatial sampling which gives a smooth transition from a high resolution fovea to wide angle viewing. As a byproduct, the output from such a camera would include all the information required to view the area of interest at varying amounts of camera zoom over a continuous scale.

This basic camera technology would also be very useful for a variety of other computer vision tasks. For example, in 3D reconstruction one needs high resolution for accurate shape recovery, but also wide angle for accurate camera motion estimates. There are some imaging situations where a fovea is particularly useful such as a forward facing camera in a car. The image flow at the focus of expansion (FOE) is much smaller than at the edges of the image, so typically no depth information can be gathered around the FOE {\em which is the most important region in the image}. If on the other hand, the FOE had much higher resolution one could get depth information very near to the FOE as well.

We have initial ideas on a several novel technologies for creating foveated cameras based on conventional technology, which we will develop under this project.

A Laboratory Testbed

Given the present lack of infrastructure for obtaining hundreds of simultaneous video feeds from a single sporting event, we recommend setting up a testbed at the MIT AI Laboratory in which proposed solutions may be developed, tested, and demonstrated. The environment would feature continuous video capture from a least 20 cameras of a mock arena in which lab members would play games such as table soccer, table tennis, or air hockey. In this setting we would test and demonstrate simultaneous motion tracking from multiple sources and develop active cameras that search for moving objects to fixate on. From these video streams, we would build coarse 3D dynamic game models, and then test the synthesis of arbitrary virtual view by positioning a mobile camera to give the ground truth view from a desired virtual camera position. The continuous capture would also provide accumulated data for detecting recurrent activities and gathering game statistics.

The initial demonstration of the technologies described in sections 2 and 5 can be tested on a static scene. We will set up a table with a scale model soccer playing field and toy players in various poses. Given multiple images of this scene we can construct rough 3D models. To test the results, additional calibrated images can be captured from various locations on and around the field. These views provide ground truth to which the synthesized images can be compared.

Collaboration with NTT

The research will be conducted by researchers and students from the Vision Group, under the direction of Prof. Paul Viola and Prof. Eric Grimson. However, we expect that we will collaborate closely with other researchers here at MIT (especially Prof. Olivier Faugeras and Prof. Leonard McMillan) and with researchers at the Human Interface Laboratories of NTT. Dr. Tohkura has suggested the Human Interface Laboratory as one potential source of collaboration. We have encountered work by several NTT researchers which have bearing on this proposal.

The Director of the NTT Information Science Research Laboratory, Dr. Kenichiro Ishii, has also published a number of papers, which are directly related to this research. One of the areas of overlap is his work on the analysis of video taken from tennis matches. Dr. Shoji Kurakake and Mr. J. Yamato have also worked on this problem. Dr.\ Hiroshi Murase, of NTT-BRL, is also working on related problems in statistical computer vision.

Immersive Video

To summarize, we propose to leverage extensive experience in computer vision, virtual reality and computer graphics, to demonstrate an *immersive video* capability. Such a system will require state-of-the-art methods in motion analysis, camera coordination, shape recovery, activity recognition and image indexing. While the methods developed under this proposal will be tuned to support a demonstration of immersive video in the context of the World Cup, the methods will clearly be of considerable utility in a variety of other domains, including surveillance and monitoring, computer-assisted medicine, site reconstruction and process control and evaluation.

BIBLIOGRAPHY

1. De Bonet, J. S. and Viola, P. ``A non-parametric multi-scale statistical model for natural images,'' In Michael~Jordan, M.~M. and Perrone, M., editors, *Advances in Neural Information Processing*, volume~10. 1197

2. I. Zoghlami, O. Faugeras and R. Deriche, ``Using geometric corners to build a 2D mosaic from a set of images,'' CVPR 97, June 1997, San Juan, Puerto Rico

3. Sylvian Bougnoux and Luc Robert, ``Total Calib: A fast and reliable system for off line calibration of image sequence,'' CVPR 97 Demo Session, June 1997, San Juan, Puerto Rico

4. W.E.L. Grimson, C. Stauffer, R. Romano, L. Lee, ``Using adaptive tracking to classify and monitor activities in a site,'' CVPR, 1998, Santa Barbara, CA.

5. Gideon P. Stein, "Geometric and Photometric Constraints: Motion and Structure from Three Views" Feb 1998. MIT PHD thesis.