



# *Research White Paper*

*WHP 147*

---

*February 2007*

## **Real-Time Camera Tracking in the MATRIS project**

**J. Chandaria, G. Thomas, B. Bartczak, K. Koeser,  
R. Koch, M. Becker, G. Bleser, D. Stricker, C. Wohlleber,  
M. Felsberg, F. Gustafsson, J. Hol, T. B. Schön,  
J. Skoglund, P. J. Slycke, S. Smeitz**

***BRITISH BROADCASTING CORPORATION***



## **Real-Time Camera Tracking in the MATRIS project**

J. Chandaria, G. Thomas, B. Bartczak, K. Koeser, R. Koch, M. Becker,  
G. Bleser, D. Stricker, C. Wohlleber, M. Felsberg, F. Gustafsson, J. Hol,  
T. B. Schön, J. Skoglund, P. J. Slycke, S. Smeitz

### **Abstract**

In order to insert a virtual object into a TV image, the graphics system needs to know precisely how the camera is moving, so that the virtual object can be rendered in the correct place in every frame. Nowadays this can be achieved relatively easily in post-production, or in a studio equipped with a special tracking system. However, for live shooting on location, or in a studio that is not specially equipped, installing such a system can be difficult or uneconomic. To overcome these limitations, the MATRIS project is developing a real-time system for measuring the movement of a camera. The system uses image analysis to track naturally occurring features in the scene, and data from an inertial sensor. No additional sensors, special markers, or camera mounts are required. This paper gives an overview of the system and presents some results.

This document was originally published in *Proc. of IBC 2006*, 7-11 September 2006, Amsterdam, NL.

**Additional key words:** Virtual studio, augmented reality, TV production

White Papers are distributed freely on request.  
Authorisation of the Head of Research is required for  
publication.

© BBC 2007. All rights reserved. Except as provided below, no part of this document may be reproduced in any material form (including photocopying or storing it in any medium by electronic means) without the prior written permission of BBC Future Media & Technology except in accordance with the provisions of the (UK) Copyright, Designs and Patents Act 1988.

The BBC grants permission to individuals and organisations to make copies of the entire document (including this copyright notice) for their own internal use. No copies of this document may be published, distributed or made available to third parties whether by paper, electronic or other means without the BBC's prior written permission. Where necessary, third parties should be directed to the relevant page on BBC's website at <http://www.bbc.co.uk/rd/pubs/whp> for a copy of this document.

# REAL-TIME CAMERA TRACKING IN THE MATRIS PROJECT

J. Chandaria<sup>1</sup>, G. Thomas<sup>1</sup>, B. Bartczak<sup>2</sup>, K. Koeser<sup>2</sup>, R. Koch<sup>2</sup>, M. Becker<sup>3</sup>,  
G. Bleser<sup>3</sup>, D. Stricker<sup>3</sup>, C. Wohlleber<sup>3</sup>, M. Felsberg<sup>4</sup>, F. Gustafsson<sup>4</sup>,  
J. Hol<sup>4</sup>, T. B. Schön<sup>4</sup>, J. Skoglund<sup>4</sup>, P. J. Slycke<sup>5</sup>, S. Smeitz<sup>5</sup>

<sup>1</sup>BBC Research, UK, <sup>2</sup>Christian-Albrechts-University Kiel, Germany,  
<sup>3</sup>Fraunhofer IGD, Germany, <sup>4</sup>Linköping University, Sweden,  
<sup>5</sup>Xsens, Netherlands

## ABSTRACT

In order to insert a virtual object into a TV image, the graphics system needs to know precisely how the camera is moving, so that the virtual object can be rendered in the correct place in every frame. Nowadays this can be achieved relatively easily in post-production, or in a studio equipped with a special tracking system. However, for live shooting on location, or in a studio that is not specially equipped, installing such a system can be difficult or uneconomic. To overcome these limitations, the MATRIS project is developing a real-time system for measuring the movement of a camera. The system uses image analysis to track naturally occurring features in the scene, and data from an inertial sensor. No additional sensors, special markers, or camera mounts are required. This paper gives an overview of the system and presents some results.

## INTRODUCTION

In order to insert a virtual object into a TV image, the graphics system needs to know precisely how the camera is moving, so that the virtual object can be rendered in the correct place in every frame. Nowadays this can be achieved relatively easily in post-production, or in a studio equipped with a special tracking system. However, for live shooting on location, or in a studio that is not specially equipped, installing such a system can be difficult or uneconomic. Thus the use of effects such as “on-pitch” sports graphics tend to be restricted to major events, and the generation of virtual graphics on other location shoots (either for live use or for pre-visualisation) is rarely attempted.

To overcome these limitations, the MATRIS (Markerless Real-time Tracking for Augmented Reality Image Synthesis) project (1) is developing a real-time system for measuring the movement of a camera. The system uses image analysis to track naturally occurring features in the scene, and data from an inertial sensor. No additional sensors, special markers, or camera mounts are required. The MATRIS approach to camera tracking mimics the way a human orients himself, using the vestibular organ (in the ears) – which is essentially an inertial measurement unit, and the eyes – essentially comparable to a camera.

The three-year project, which started in February 2004, is a part of the EU’s 6<sup>th</sup> Framework programme. It is led by Fraunhofer IGD, experts in industrial augmented reality. The other partners are Xsens, specialists in developing miniature inertial sensors; Linköping University, experts in sensor fusion and image processing; Christian-Albrechts University of Kiel, experts in computer vision and BBC Research, who have many years of experience in developing technology for real-time virtual graphics.

## **TV production**

For TV productions that have elements such as a virtual set or virtual objects, it is necessary to measure the precise pose (i.e. position and orientation) of each studio camera so that the virtual elements can be rendered from exactly the right viewpoint. Studio productions with real-time graphics such as news programmes usually do this by using a specially equipped studio. The studio might use a marker based solution such as the BBC developed *free-d* system (2).

Programmes that insert the virtual graphics in post production can use match-moving software but this is much slower than real time.

Location shoots such as sports events usually use special camera mounts with pan/tilt sensors, which means their use is limited to a small number of cameras at major events.

## **Film production**

Although most films rely heavily on post-production, there is still a need for real-time camera tracking systems to provide on-set visualisation of virtual elements, so that the cameraman, director and actors can relate to the virtual elements. At present, the only viable solution is the use of tracking systems that rely on markers, such as the way in which the *free-d* system was used in the making of the film *AI* (3).

## **Augmented Reality**

As well as TV and film production, augmented reality has many applications including industrial maintenance, medicine, education, entertainment and games. The central idea is to add virtual objects into a real scene, either by displaying them in a see-through head-mounted display, or by superimposing them on an image of the scene captured by a camera. Depending on the application, the added objects might be instructions for repairing a car engine or a reconstruction of an archaeological site. For the effect to be believable, the virtual objects must appear rigidly fixed to the real world, which requires the accurate measurement in real-time of the position of the camera or the user's head. Present technology cannot achieve this without resorting to systems that require the operating area to be fitted with devices such as acoustic transducers, multiple cameras or markers, severely restricting the range of possible applications.

As can be seen, all these application areas would clearly benefit from a system that does not require markers, external sensors or other special infrastructure in the environment.

## **APPROACH**

The environment in the area in which the system is to be used is first modelled (the 'offline' phase) in order to provide a set of images and features that may be used as 'beacons' by the tracking process. Many types of features could be used. So far, the project has considered texture patches (flat regions containing lots of fine detail) and line features. The features that are modelled are chosen to be visible from a wide range of viewpoints, and to contain sufficient detail to enable their location to be accurately determined. Depending on the application, this modelling process may make use of existing CAD models, or may build a model from images of the scene.

However, there may not always be enough naturally occurring features in the scene to provide sufficient references for highly-stable tracking. Therefore, the system also makes use of an Inertial Measurement Unit (IMU) – a matchbox-sized device attached to the camera, which incorporates miniature accelerometers, gyroscopes and magnetometers. An initial calibration process determines the relative position and orientation of the IMU with

respect to the camera. The lens distortion and focal length are also calibrated; in the case of a zoom lens these will be variable, so the setting of zoom and focus are measured with mechanical sensors.

During tracking (the 'online' phase) images from the camera are captured live and processed in order to locate the modelled features. The data from the IMU is also processed to compensate for effects such as drift. At the start of the tracking process, the image and sensor data are used by the initial view registration process to calculate the initial camera pose; this may take several seconds as a search through the entire model space may be necessary. After initialisation, the predictive tracking module takes over, tracking the movement of the camera from frame-to-frame. If the predictive algorithm should fail, the initialisation process can be called again, to re-establish tracking.

Figure 1 below shows the main components of the MATRIS system structure. The hardware components are shown in *italics* and the software components are shown in **bold**. The next two sections describe these components in further detail.

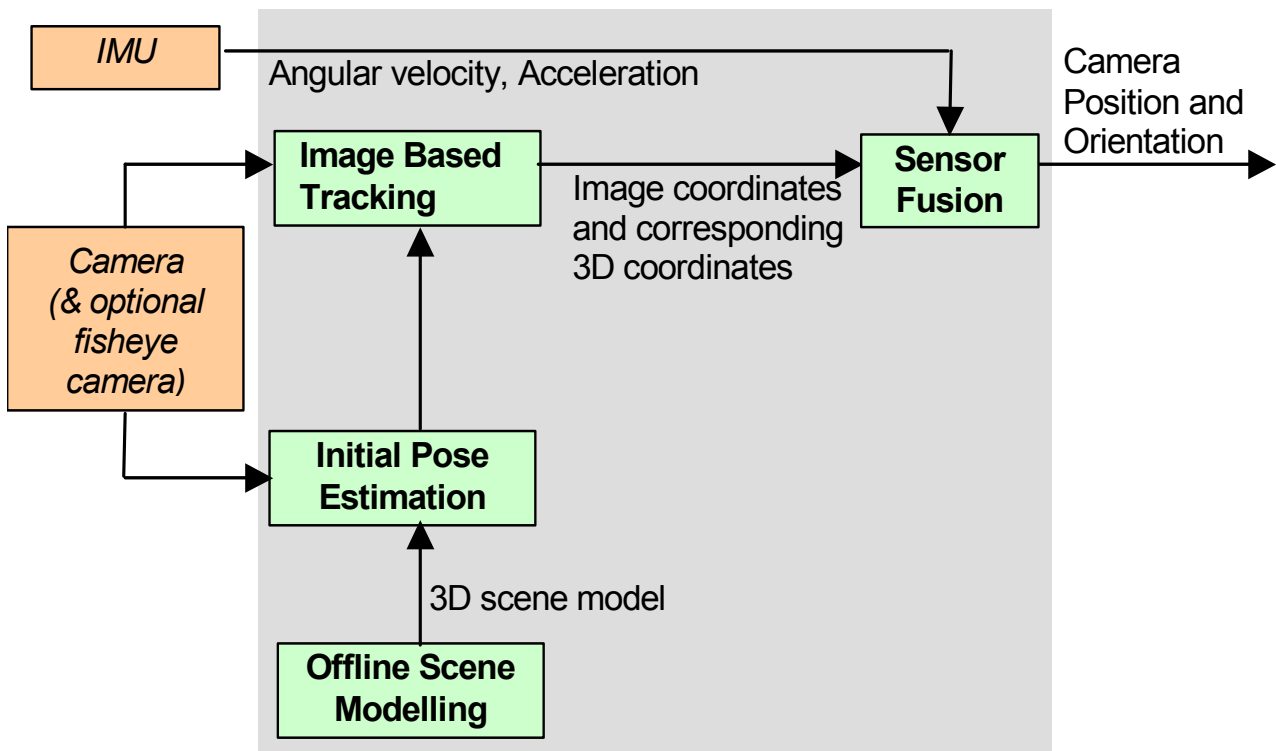


Figure 1 - The MATRIS system structure

## HARDWARE SYSTEM

The hardware components of the MATRIS system consist of a standard camera, an IMU, and a computer unit (desktop PC or Laptop). Optional components are a fisheye camera, a GPS module, and a lens sensor. (see Figure 2). Depending on the application scenario, different setups are necessary to synchronize the IMU with the camera.

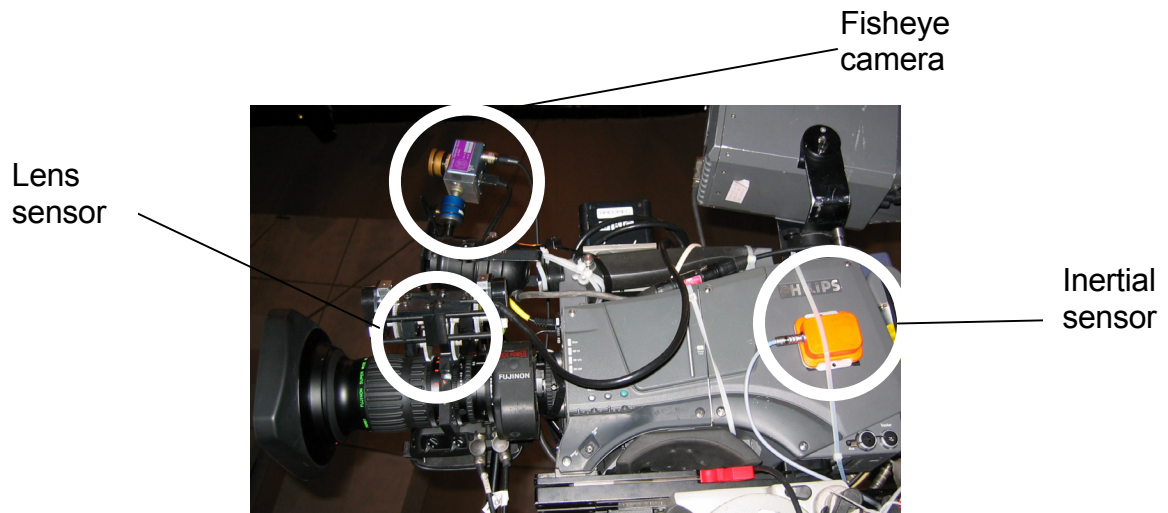


Figure 2 - The TV camera with attached inertial sensor, fisheye camera and lens sensor

### IMU

The Inertial Measurement Unit (IMU) measures the 3D angular velocity and linear acceleration using solid state rate gyroscopes and accelerometers (inertial sensors). The IMU also features an integrated 3D earth magnetic field sensor that can be used to sense heading, much like a compass. The embedded DSP processes the sensor measurements at a high rate (up to 512 Hz), compensating for sensor parameters such as offset, gain, non-orthogonality, temperature dependence, etc. The DSP also runs a local sensor fusion algorithm that enables the IMU to estimate absolute (earth referenced) 3D orientation (pan, tilt, roll) completely independent from the other system components, which is convenient for fast initialisation and in the offline scene modelling phase.

The IMU is synchronised to the TV camera by using the video reference genlock signal to generate a 100 Hz trigger signal.

### Fisheye camera

Whilst tracking, often the main camera will zoom onto small details of the scene or onto the moving actors, which will make robust camera pose tracking very difficult. To ensure tracking in such situations, an additional fisheye camera, covering a 180° field of view can be utilized as an additional tracking device. It is mounted on top of the main camera and can see not only the set but also parts of the surroundings. Because of its large field of view, some natural features in the scene and its surrounding will always be visible and can be used for tracking. It has been shown that the fisheye lens geometry is particularly well suited for such tracking situations (4).

Figure 3 compares the views from the main camera (left) and the fisheye camera (right) of the same scene and shows the much greater field of view of the fisheye camera.



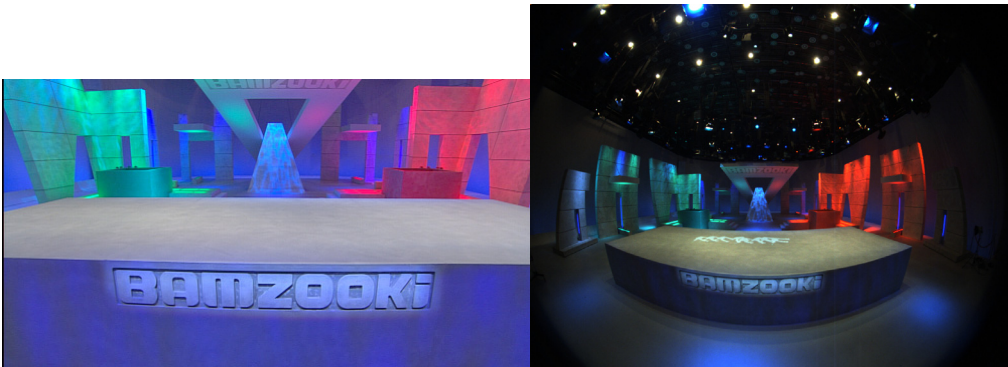


Figure 3 - View from the main camera (left) and the fisheye camera (right) of the same scene

## SOFTWARE

### Offline Scene Modelling

In the offline phase, 3D features that are suitable for efficient and reliable online tracking will be generated automatically from the scene. This allows the features to be chosen for a specific scene. Images of the scene without actors are captured along with data from the IMU and lens sensor and the poses of the camera system are estimated using robust structure from motion approaches. The camera system is moved in such a way as to approximate the possible views during online shooting.

From all the different images, a 3D surface model of the set is computed automatically. 2D image features are tracked using a fast image feature tracker, and 3D feature points are triangulated from the 2D features and the automatically estimated camera poses. Dense depth maps are computed from all camera views with multi-view stereo analysis (5). The 3D features are selected such that they will facilitate the online tracking.

Currently, the scene is segmented into 3D planar patches with highly textured surfaces, which can be used as robust and reliable tracking anchors. This approach will be generalized to free-form surface patches in the final system.

Figure 4 shows an example of a scene model, where the 3D features are extracted from the image sequence by means of the surface depth model. The 3D coordinates of the scene model are aligned to the IMU sensor data with respect to gravity and magnetic north. This alignment will allow an easy initialization of the view direction for the online tracking phase. The 3D features are stored in a database for later use in the online phase.

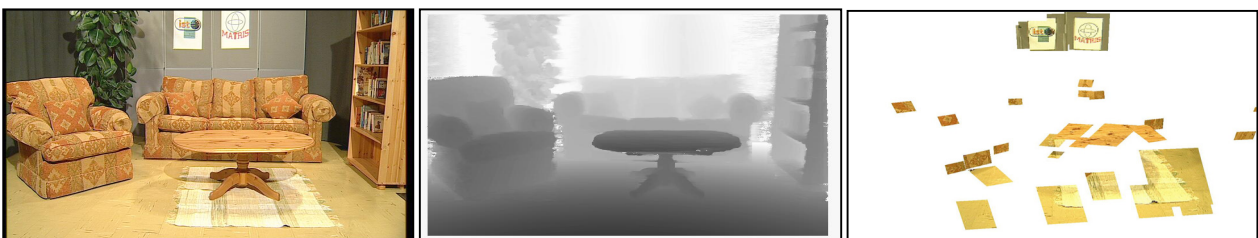


Figure 4 - Image (left), depth map (middle) and 3D planar feature patches (right) of the scene.

### Initial Pose Estimation

When the system is started for real-time use, the initialisation process generates an initial hypothesis of the camera pose. It compares the current camera image with the reference images taken from known positions and orientations in the database, and can use other available information such as orientation from the IMU and likely position (from *a priori* knowledge such as the last known camera position, or a probability distribution of positions

that the camera has been known to occupy previously).

The initial hypothesis for the camera pose is then refined by matching SIFT features (6) between the current camera image and the closest database image. These features are extracted from the live image and compared to the 3D features in the database associated to the reference image. The resulting 'correspondences' - visible features and their corresponding 3D locations - are employed for refining the camera pose using the well-known RANSAC approach (7) which is robust against outliers.

### **Image Based Tracking**

This module uses computer vision techniques to locate the modelled 3D features in the current camera image. It uses a prediction of the camera pose from the Sensor Fusion process to estimate the positions of the features, and then searches around these estimated locations to generate a list of correspondences.

In order to match the planar patch features with those in the scene model, it is first necessary to transform each stored texture patch, so that its scale and orientation matches that expected from the predicted camera position. The transformation process includes correction for lens distortion. For each transformed texture patch, an exhaustive search is carried out around the predicted position, using the sum of absolute difference criterion. This is implemented using SIMD instructions (the MMX extension on a Pentium 4) (8) to maximise performance. To increase the accuracy of the final result, sub-pixel interpolation is used. An estimate of the confidence of the match is also needed for the subsequent sensor fusion processing; a fast covariance estimation method that uses the absolute differences computed during the search process was developed for this purpose.

A RANSAC-like outlier rejection is performed on the initial set of correspondences and the valid correspondences are then passed to the Sensor Fusion process together with their uncertainties. The correspondences are used as measurements in the Sensor Fusion module to correct the current pose estimate.

For several kinds of scene, line features are expected to provide the most stable and easily-recognisable kind of feature. A method has therefore been developed to estimate the pose of the camera by matching line segments with known lines in a 3D model, targeted towards applications such as the augmentation of sports scenes with overlaid graphics. The approach is similar to that described in (9). It uses a least-squares minimisation technique to find the camera pose which minimises the distance between the ends of lines fitted to groups of edge points in the image, and the re-projected line positions from a 3D model generated from the known 3D line positions. The predicted camera pose is used to select the regions of the image that are searched to find each line. The line detector can be set to ignore pixels that are not surrounded by a given colour (e.g. green grass); this helps to prevent false matches, such as the detection of edges of advertising hoardings when lines marked on a football pitch are the desired features.

### **Sensor Fusion**

The sensor fusion process takes the correspondences from the computer vision stage and combines them with the pose estimate from the IMU (10). By modelling the whole system including confidence information, the best possible estimate of the camera pose can be made.

The state-of-the-art for online pose estimation is based on computer vision algorithms, where other sensors are used as support. The MATRIS project has taken a reverse approach, inspired by high-performance aircraft navigation algorithms. In such applications, the IMU is the primary navigation sensor for attitude and position estimation (that is, pose). Dead-

reckoning using IMU data allows rapidly-changing motions to be tracked, but suffers from long term drift. For that reason, supporting sensors such as ground based beacons and satellites in the global positioning system (GPS) are used whenever possible. This external information should be compared to the offline scene model used in the system described here. The important points are that fast movements can be handled without difficulty and not all degrees of freedom for pose estimation have to be available in each image frame. Rather, the corresponding 2D and 3D features from the computer vision stage are used to correct the integrated IMU pose estimate. A Kalman filter framework is used, just as in aircraft navigation, where sensor offset parameters in the IMU (accelerometer bias and gyro drift) can also be estimated to further increase performance.

## EVALUATION

The MATRIS system has been shown to work well in highly textured scenes with many planar surfaces and constant lighting. The camera pose calculated can be used to augment the scene with a virtual object locked in position in the scene as shown in Figure 5. The system works less reliably in scenes lacking in texture. In order to increase the range of possible operating environments and robustness, future work will look at the use of free form surfaces as the main feature type.



Figure 5 - Augmented scene showing a virtual car on a real table

It has been shown that the use of fisheye camera for tracking can improve robustness (4).

The line-based tracking method has been tested on many long video sequences from both football and rugby coverage and found to perform very well. The method runs easily at full video rate, taking around 3-4ms to process a standard-definition image on a single 3GHz CPU. We have found that it can also use circles and curves for tracking, by modelling them as a series of line segments.

This tracking method has been licensed to Red Bee Media Ltd and incorporated into Piero, a sports graphics system based on earlier work at BBC Research (11). It has been in regular use for BBC sports programmes since September 2005 and is now used by other broadcasters as well. Figure 6 shows an example of virtual graphics being overlaid on a rugby pitch, using camera pose information derived by the line tracker.

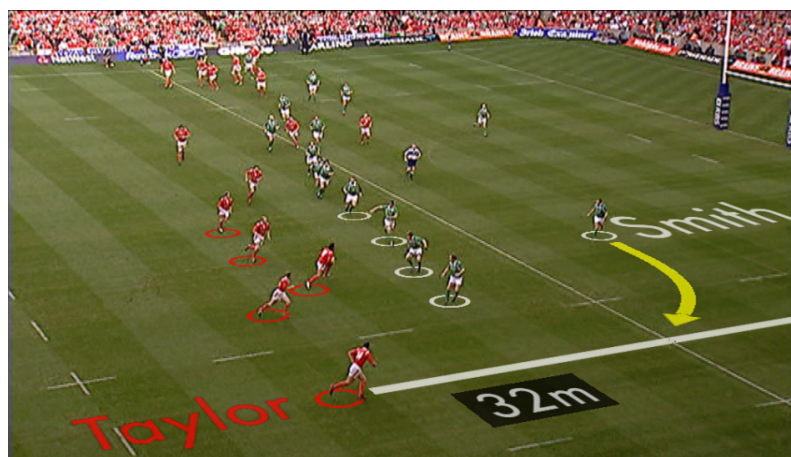


Figure 6 – An example of the line-based pose computation process in use

## CONCLUSION

The MATRIS project has developed and demonstrated a real-time camera tracking system that does not require markers or special studio infrastructure. It has shown that by combining image analysis with an IMU, better results can be achieved than with image analysis alone. Furthermore, it has shown that the use of a fisheye camera can be used to improve robustness.

Future work will aim to increase the range of environments in which the system can be used by investigating the use of freeform surface features.

The project has also provided the opportunity to expose programme-makers to some of the possibilities that technology will offer in the near future. Some of the project results are already in use in sports programmes.

## ACKNOWLEDGEMENTS

The authors would like to thank all their colleagues in the MATRIS project for their contributions to this paper. The project was supported by the Information Society Technologies area of the EU's 6<sup>th</sup> Framework programme - Project no. IST-002013.

## REFERENCES

1. <http://www.ist-matris.org/>
2. Thomas, G.A. et al. 1997. A versatile camera position measurement system for virtual reality TV production, IBC'97 Conference Publication, pp. 284-289 [http://www.bbc.co.uk/rd/pubs/papers/paper\\_05/paper\\_05.html](http://www.bbc.co.uk/rd/pubs/papers/paper_05/paper_05.html)
3. Warner Brothers, 2002. Special Visual Effects, A. I. Artificial Intelligence Bonus Disc, DVD
4. Streckel, B. and Koch, R., 2005. Lens Model Selection for Visual Tracking. Proc. DAGM Symposium '05, Springer LNCS 3663,
5. Koch, R., Pollefeys, M., and van Gool, L., 1998. Multi Viewpoint Stereo from Uncalibrated Video Sequences, Proc. ECCV'98, LNCS 1406, Springer
6. Lowe, D.G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision , 60, 2 (2004), pp. 91-110.
7. Fischler, M.A., Bolles, R.C., 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Comm. of the ACM, Vol 24, pp 381-395
8. J. Skoglund and M. Felsberg., 2005. Fast image processing using SSE2. Proceedings of the SSBA Symposium on Image Analysis.
9. Vacchetti, L., Lepetit, V., Fua, P., 2004. Combining Edge and Texture Information for Real-Time Accurate 3D Camera Tracking". ISMAR 2004, pp. 48-56.
10. Sensor fusion for augmented reality, 2006. J.D. Hol, T.B. Schön, F. Gustafsson, P.J. Slycke, accepted paper, The 9th International Conference on Information Fusion, Florence
11. <http://www.bbc.co.uk/rd/projects/virtual/piero/index.shtml>