Realtime Camera Tracking in the MATRIS Project

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

In order to insert a virtual object into a television 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. Today, 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. 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-visualization) is rarely attempted. To overcome these limitations, the MATRIS project is developing a realtime system for measuring the movement of a camera. The system uses image analysis to track naturally occurring features in a 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.

The MATRIS (Markerless Realtime Tracking for Augmented Reality Image Synthesis)[1] approach to camera tracking mimics the way humans orient themselves, 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 European Union's (EU's) 6th Framework program. 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, experts in developing technology for realtime virtual graphics.

TV Production

For television 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 realtime graphics such as news programs usually do this by using a specially equipped studio. The studio might use a marker-based solution such as the free-d system.[2]

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

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 realtime camera tracking systems to provide on-set visualization 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

In addition to television 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 headmounted 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 of the position of the camera or the user's head, in realtime. 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 modeled (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 modeled are chosen to be visible from a wide range of viewpoints and contain sufficient detail to enable their location to be accurately determined. Depending on the application, this modeling process may make use of existing computer-aided design (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 modeled 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 initialization, the predictive tracking module takes over, tracking the movement of the camera from frame to frame. If the predictive algorithm should fail, the initialization process can be called again, to re-establish tracking.

Figure 1 shows the main components of the MATRIS system structure. The hardware components are shown in italics and the software components are shown in bold.

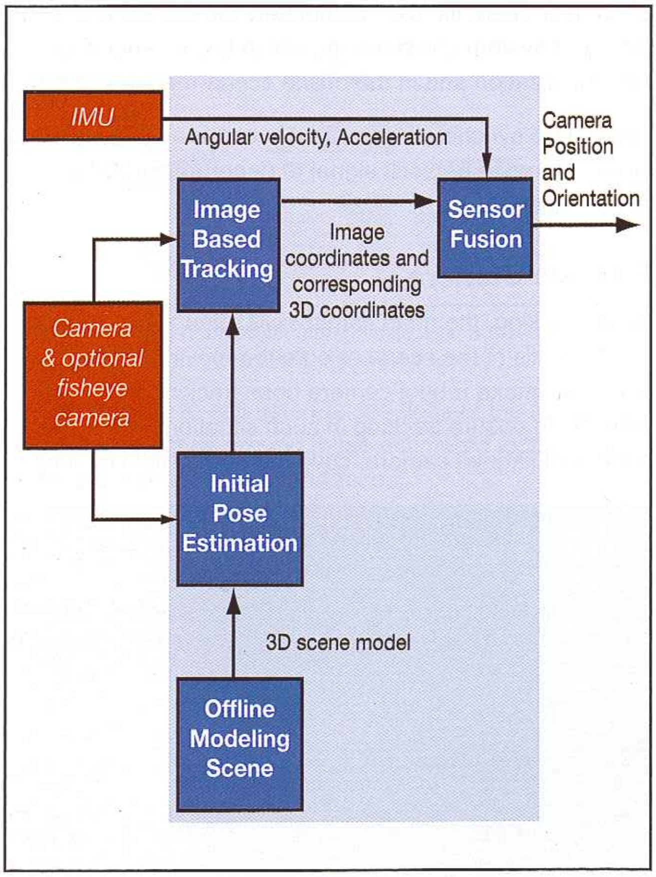
IMU
Angular velocity, Acceleration
Camera Position and Orientation
Image Based Tracking
Image coordinates and corresponding 3D coordinates
Sensor Fusion
Camera & optional fisheye camera
Initial Pose Estimation
3D scene model
Offline Modeling Scene

Figure 1. The MATRIS system structure.

The next two sections describe these components in further detail.

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 global positioning system (GPS) module, and a lens sensor. (Fig. 2). Depending on the application scenario, different setups are necessary to synchronize the IMU with the camera.

IMU

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

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

Fisheye Camera

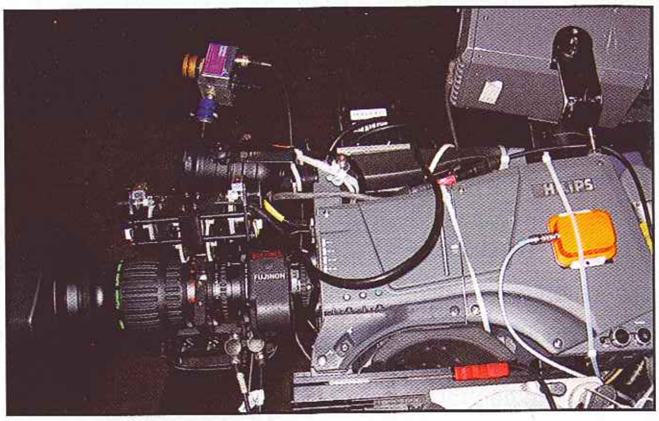

Figure 2. The TV camera with attached inertial sensor, fisheye camera, and lens sensor.

While tracking, the main camera will often 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 not only the set, but also parts of the surroundings can be viewed. 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.[4] It has been shown that the fisheye lens geometry is particularly well suited for such tracking situations.[5]

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.

Software

Offline Scene Modeling

In the offline phase, 3-D 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

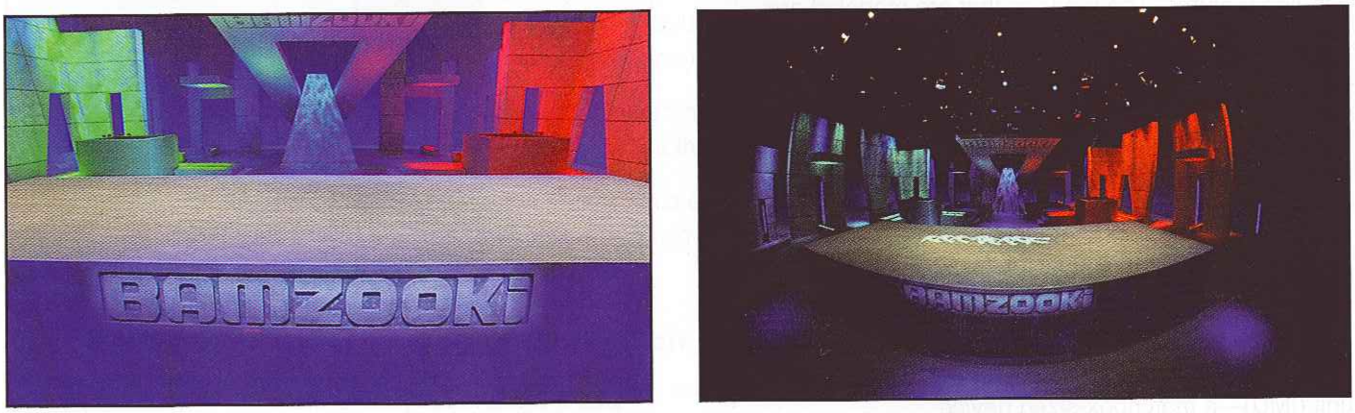

Figure 3. View from the main camera (left) and the fisheye camera (right) of the same scene.

approximate the possible views during online shooting. From all the different images, a 3-D surface model of the set is computed automatically. 2-D image features are tracked using a fast image feature tracker, and 3-D feature points are triangulated from the 2-D features and the automatically estimated camera poses. Dense depth maps are computed from all camera views with multiview stereo analysis.[6] The 3-D features are selected so that they will facilitate the online tracking.

Currently, the scene is segmented into 3-D 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 in which the 3-D features are extracted from the image sequence by means of the surface depth model. The 3-D 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 3-D features are stored in a database for later use in the online phase.

Initial Pose Estimation

When the system is started for realtime use, the initialization 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[7] between the current camera image and the closest database image. These features are extracted from the live image and compared to the 3-D features in the database associated to the reference image. The resulting "correspondences"—visible features and their corresponding 3-D locations—are employed for refining the camera pose using the well-known RANdom SAmple Consensus (RANSAC) approach,[8] which is robust against outliers.

Image-Based Tracking

This module uses computer vision techniques to locate the modeled 3-D 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.

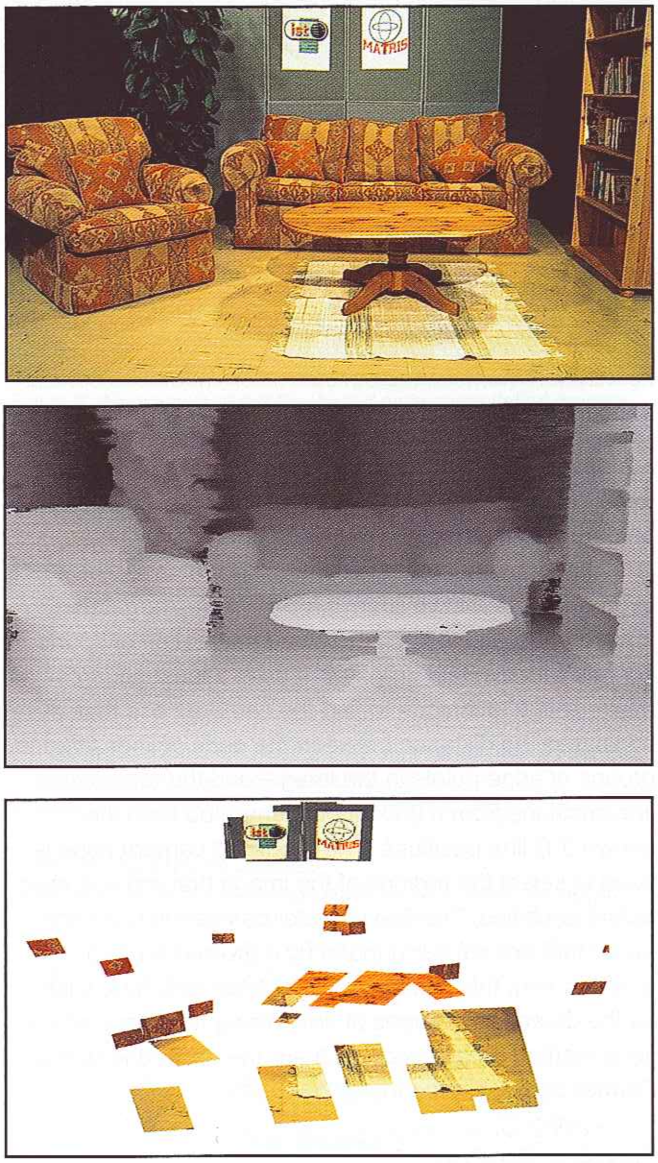

Figure 4. Image (top), depth map (middle) and 3-D planar feature patches (bottom) of the scene.

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 those 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 single instruction, multiple data (SIMD) instructions (the MMX extension on a Pentium 4)[9] to maximize performance. To increase the accuracy of the

* SIFT- Scale-Invariant Feature Transform

final result, sub-pixel interpolation is used.[10] 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.[11]

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 recognizable 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 3-D model, targeted toward applications such as the augmentation of sports scenes with overlaid graphics. It uses a least-squares minimization technique to find the camera pose that minimizes the distance between the ends of lines fitted to groups of edge points in the image, and the reprojected line positions from a 3-D model generated from the known 3-D 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 color (e.g., green grass); this helps to prevent false matches, such as the detection of edges of advertising hoardings when lines marked on a soccer pitch are the desired features. Further details can be found in Ref. 12.

Sensor Fusion

The sensor fusion process takes the correspondences from the computer vision stage and combines them with the pose estimate from the IMU.[13] By modeling the entire system, including confidence information, the best possible estimate of the camera pose can be made.

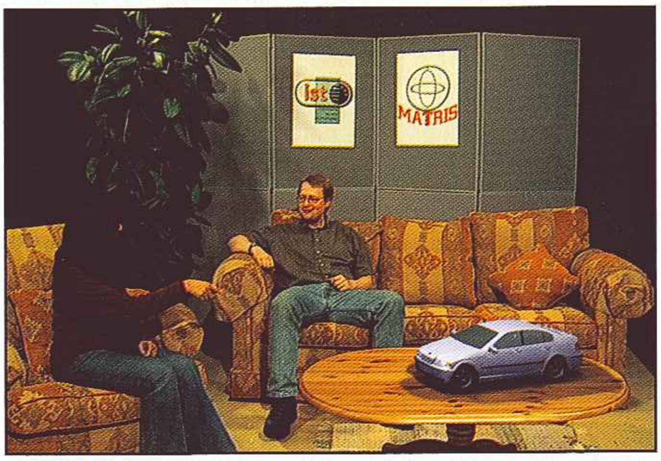

Figure 5. Augmented scene showing a virtual car on a real table.

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 (pose) estimation. Dead-reckoning using IMU data allows rapidly changing motions to be tracked, but suffers from longterm drift. For that reason, supporting sensors such as ground-based beacons and satellites in the 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 2-D and 3-D features from the computer vision stage are used to correct the integrated IMU pose estimate. A Kaman filter framework is used, just as in aircraft navigation, in which 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 Fig. 5. The system is less reliable 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.

It has been shown that use of a fisheye camera for tracking can improve robustness.[5]

The line-based tracking method has been tested on many long video sequences from both soccer and rugby coverage and found to perform very well. The method runs easily at full video rate, taking around 3 to 4 ms to process a standard-definition image on a single 3-GHz CPU. It was found that it could also use circles and curves for tracking, by modeling 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.[14] Figure 6 shows an example of virtual graphics being overlaid on a rugby pitch, using camera pose information derived by the line tracker.

Conclusion

The MATRIS project has developed and demonstrated a realtime 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 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 free-form surface features.

The project has also provided the opportunity to expose program-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 programs.

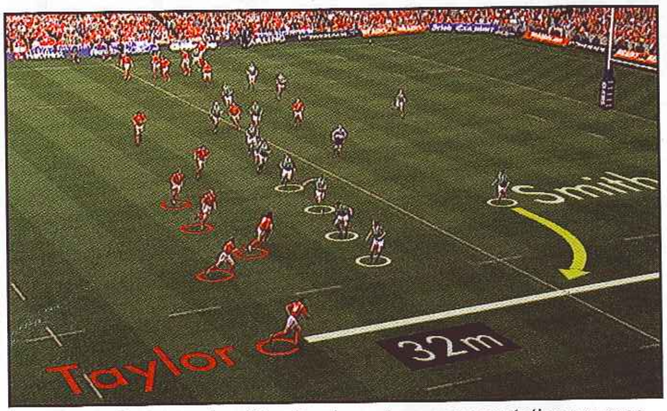

Figure 6. An example of the line-based pose computation process in use.

Acknowledgments

The MATRIS project was supported by the Information Society Technologies area of the EU's 6th Framework program. Project no. IST-002013.

References

1. www.ist-MATRIS.org/.
2. G. A. Thomas, et al., "A Versatile Camera Position Measurement System for Virtual Reality TV Production," Proc. IBC '97, pp. 284-289 1997, www.bbc.co.uk/rd/pubs/papers/paper_05/paper_05.html.
3. Warner Brothers, "Special Visual Effects, A. I. Artificial Intelligence Bonus Disc, DVD, 2002.
4. K. Koeser, B. Bartczak and R. Koch. "Drift-free Pose Estimation with Hemispherical Cameras," Conf. on Visual Media Prod. (CVMP), London, November, 29-30 2006.
5. B. Streckel and R. Koch, "Lens Model Selection for Visual Tracking." *Proc. DAGM. Symposium '05*, LNCS 3663, Springer, 2005.
6. R. Koch, M. Pollefeys, and L. van Gool, "Multi Viewpoint Stereo from Uncalibrated Video Sequences," *Proc. ECCV'98*, LNCS 1406, Springer, 1998.
7. D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Intl. J. Comp. Vision 60(2)*: 91-110, Jan. 2004.
8. M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Comm. of the ACM, 24(6)*:381-395, June 1981.
9. J. Skoglund and M. Felsberg, "Fast Image Processing using SSE2." *Proc. SSBA Symposium on Image Analysis*, March 2005.
10. J. Skoglund and M. Felsberg "Evaluation of Subpixel Tracking Algorithms," *Proc. of Intl. Symp. of Vis. Comp. (ISVC2006)* pp. 374-382 2006.
11. J. Skoglund and M. Felsberg "Covariance Estimation of SAD Matching," *Proc. SCIA 2007*, Aalborg, Denmark, 2007.
12. G. A. Thomas. "Real-Time Camera Pose Estimation for Augmenting Sports Scenes," Conf. on Visual Media Prod. (CVMP), London, Nov. 2006.
13. J. D. Hol, T. B. Schön, F. Gustafsson, and P. J. Slycke, "Sensor Fusion for Augmented Reality," The 9th International Conference on Information Fusion, Florence, Italy, 2006.
14. www.bbc.co.uk/rd/projects/virtual/piero/index.shtml.

The Authors

J. Chandaria, G. Thomas, BBC Research, U.K.; B. Bartczak, K. Koeser, R. Koch, Christian-Albrechts-University Kiel, Germany; M. Becker, G. Bleser, D. Stricker, C. Wohlleber, Fraunhofer IGD, Germany; M. Felsberg, F. Gustafsson, J. D. Hol, T. B. Schön, J. Skoglund, Linköping University, Sweden; P. J. Slycke, S. Smeitz, Xsens, Netherlands