MODEL-FREE MARKERLESS TRACKING FOR REMOTE SUPPORT IN UNKNOWN ENVIRONMENTS

Alexander Ladikos, Selim Benhimane, Nassir Navab
Department of Computer Science, Technical University of Munich, Boltzmannstr. 3, 85748 Garching, Germany
ladikos@in.tum.de, benhiman@in.tum.de, navab@in.tum.de

Mirko Appel
Siemens Corporate Technology, Munich, Germany
mirko.appel@siemens.com

Keywords: Tracking, Real-Time Vision, Remote Support, Augmented Reality

Abstract: We propose a complete system that performs real-time markerless tracking for Augmented Reality-based remote user support in a priori unknown environments. In contrast to existing systems, which require a prior setup and/or knowledge about the scene, our system can be used without preparation. This is due to our tracking algorithm which does not need a 3D-model of the scene or a learning-phase for the initialization. This allows us to perform fast and robust markerless tracking of the objects which are to be augmented. The proposed solution does not require artificial markers or special lighting conditions. The only requirement is the presence of locally planar objects in the scene, which is true for almost every man-made structure and in particular technical installations. The augmentations are chosen by a remote expert who is connected to the user over a network and receives a live stream of the scene.

1 INTRODUCTION

Remote Support is highly interesting for the service industry, as it can help field service technicians to carry out their tasks more efficiently. Due to increasing complexity and availability demands of plants and products, quick intervention in case of failures, involving highly specialized experts, is becoming a key requirement. Remote Support enables on-site personnel, which is often rather universally trained than specialized, to obtain interactive assistance by globally distributed experts. Supported by a combination of technologies such as Remote Service, Video Conferencing and Augmented Reality, field service personnel can solve a broader range of tasks faster and more efficiently. First steps towards this direction have been taken within the ARVIKA project (Friedrich, 2004). In this case the augmentation was purely marker-based, requiring scene preparation and limiting the field of view of the on-site technician. The approach presented in this paper better fits the needs of the service industry, as it can be used in an unknown environment without any preparation and does not limit the operating range of the service technician. This is achieved by using a learning-free markerless tracking system, which is based on template-based and feature-based tracking algorithms. The tracking system can handle commonly occurring real-world tracking conditions including partial occlusions, motion blur, fast object movement and changing lighting conditions. A 3D-CAD model of the scene is not required. We use this tracking system in a remote support application, which allows a remote expert to help a local user by providing appropriate augmentations for objects in the scene. The remote support is realized by streaming a live video to the expert who can select regions in the image and choose from several possible types of augmentations including text, images and drawings.

2 RELATED WORK

Markerless tracking methods comprise different approaches. The most common are template-based tracking (Hager and Belhumeur, 1998; Baker and Matthews, 2004) and feature-point-based tracking (Lowe, 2004; Lepetit et al., 2005) since they are fast and stable under many conditions. There also exist tracking methods combining multiple visual tracking

cues such as lines, feature-points and texture (Vacchetti et al., 2004; Pressigout and Marchand, 2006; Ladikos et al., 2007).
In particular the combination of template-based tracking and feature-based tracking gives robust results. Template-based tracking works well for small interframe displacements, image blur, oblique viewing angles and linear illumination changes, while feature-based tracking works well with occlusions and large interframe displacements. Therefore, we based our algorithm on combining template-based tracking and feature-based tracking. For template-based tracking, we use the ESM algorithm of (Benhimane and Malis, 2004) due to its high convergence rate and accuracy, while the feature-based tracking is based on Harris Corner points.
In (Ladikos et al., 2007) the authors assume that they have a textured 3D-model of the object and that the camera is calibrated. In our application we do not have this information since we want to use it for interactive tracking in unknown scenes with unknown cameras. Our contributions therefore focus on making this tracking approach work under the constraints imposed by our application. To accommodate tracking in unknown scenes we parameterized the camera motion using a homography instead of the pose in the cartesian space. This is sufficient for our application since we only want to superimpose labels over the objects. This simplifies the Jacobian so that we obtain a significant speed increase for the tracking. We also included an online illumination compensation by estimating the gain and the bias for the tracked pattern. To accommodate interactivity we use a learning-free SIFT-based initialization instead of a learning-based Randomized Tree initialization.
Augmented Reality-based remote support has received much attention in the past. However, most existing work has focused on the interaction (Barakonyi et al., 2004) between the remote user and the local user and very little work (Lee and Höllerer, 2006) has been devoted to markerless tracking. Therefore, most systems make use of fiducial markers or explicit scene knowledge which precludes an ad-hoc use in unknown environments. Our contribution is to use a reference template for markerless tracking to avoid drift and combine both template-based and feature-based tracking to overcome problems with jittering, fast object motion and partial occlusions.

3 THE TRACKING SYSTEM

Our tracking system combines template-based and feature-based tracking. This design is based on several simulations and experiments that were conducted in order to determine the properties of both tracking methods. Some of these experiments are presented in the experimental section of this paper. The results suggest that no single method can deal well with all tracking situations occurring in practice. It is rather the case that the two tracking approaches are complementary. Therefore, efficiently combining those methods yields robust and accurate tracking results. In the remainder of this section, we will first discuss the design of the tracking system and then go on to describe each component in detail.

3.1 System Design

Once the initialization module, which uses SIFT descriptors, has determined the pose of the object in the current image, our algorithm adaptively switches between template-based tracking and feature-based tracking depending on the value of the Normalized-Cross-Correlation (NCC) computed between a reference image of the object and the object's appearance in the current image. During template-based tracking, a low NCC score usually means that the object is being occluded or that the interframe displacement is outside the convergence radius of the minimization algorithm. In both cases, feature-based approaches give higher quality results. Therefore, the proposed algorithm automatically switches over to the feature-based tracking. As soon as the NCC score gets high enough, it is safe to go back to template-based tracking. If the score remains low, the feature-based tracking is invoked as long as there are enough inlier points for a stable pose estimation. If this is not the case, the algorithm invokes the global initialization.

3.2 Initialization

The initialization is performed using SIFT descriptors. The choice of SIFT descriptors over other approaches is based on the fact that they do not need a learning phase and that they have been shown to be very robust with respect to different image transformations. The runtime performance of SIFT is not real-time, but since it is only used for initialization this is not a critical issue.

3.3 Template-based Tracking

We use the ESM algorithm to perform template-based tracking because it enables us to achieve second-order convergence at the cost of a first-order method. Given the current image I and the reference image I^* of a planar target, we are looking for the homography

which transforms the reference image into the current image. The ESM algorithm iteratively updates the approximation $\widehat{\mathbf{H}}$ of the homography by finding the parameters $\mathbf{x}$ of the incremental homography $\mathbf{H}(\mathbf{x})$. In addition, we also estimate an incremental gain α to the current gain $\widehat{\alpha}$ and an incremental bias β to the current bias $\widehat{\beta}$ for the template in order to minimize the photometric error. This allows the method to be used with changing lighting conditions. The gain is parametrized through the exponential function to improve the conditioning of the Jacobian and to guarantee that it is always positive. The SSD-based cost function to be minimized is:

$$\sum_{\mathbf{p}}\left[\exp(\widehat{\alpha}+\alpha)\, I\left(\mathbf{w}\left(\widehat{\mathbf{H}}\mathbf{H}(\mathbf{x})\mathbf{p}\right)\right)+(\widehat{\beta}+\beta)-I^{*}(\mathbf{p})\right]^{2} \tag{1}$$

where $\mathbf{p}$ are the pixel coordinates and $\mathbf{w}$ performs the division by the homogeneous coordinate. In order to speed up the optimization, we use a multi-scale approach for optimizing the cost function.

3.4 Feature-based Tracking

Our feature-based tracking algorithm makes use of Harris Corner points and NCC for matching. The Harris features are extracted on the same reference template used for tracking and matched to Harris features found in the current image. However, because NCC is not scale and rotation invariant, we add an additional step to ensure that the changes in pose between the reference template and the current image are small. This is done by using the inverse of the homography $\widehat{\mathbf{H}}$ found in the previous frame to warp the current image so that the NCC can still be used as a similarity measure. The feature points are then extracted in the warped image and the pose is estimated using RANSAC.

4 THE REMOTE EXPERT APPLICATION

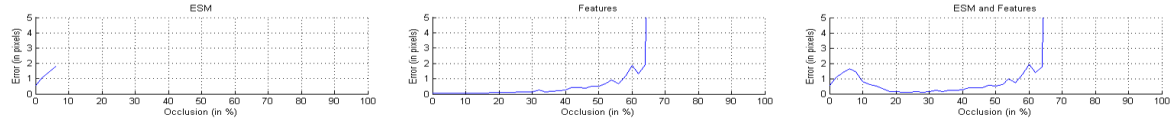
We use the tracking algorithm presented in section 3 in a remote support application. The application, called the Remote Expert, allows a user to receive instructions from a remote expert, which are augmented onto objects in the scene.
The workflow of the application is as follows. First the expert connects to the user's machine to receive a live video-stream. When the expert wants to add an annotation to an object he can select a region in the image and select either a text, a sketch or an image augmentation. When selecting the region the expert application is requesting an uncompressed frame at the full resolution from the worker, in order to select and send back the template in the actual resolution and quality of the input camera. The selected region in the image and the augmentation are sent to the user's application, which starts tracking the object using the reference template and augmenting it on screen. For every frame the user's application sends the pose of the object determined by the tracking system back to the expert application which uses this pose to augment the video locally.
To reduce bandwidth requirements the video is encoded to MPEG-2 at a resolution of 320x240 pixels and 15 fps independent of the actual size and frame rate of the user's camera. In order to keep the delay low the tracking is performed on the user's machine. This also yields superior tracking results, because the user can make use of the full resolution and speed of the camera.

5 RESULTS

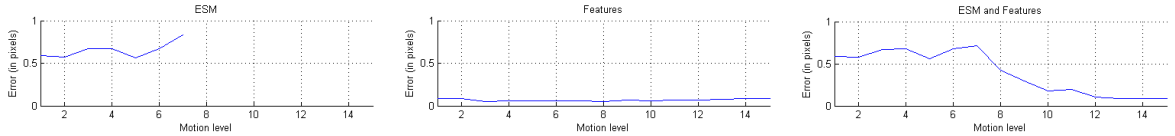
In order to evaluate both the tracking algorithm and the remote support application, we performed two sets of experiments. The first set was designed to test the accuracy and the robustness of the tracking algorithm while the second set was meant to evaluate the Remote Expert application.
To test the robustness of the tracker we performed several kinds of simulations, whose results are shown in Figure 1. In all experiments the performance of the ESM algorithm and the feature-based tracking by themselves were compared to the combined approach. The error was measured by computing the displacement of the corner points of the template from their true position after optimization. Apart from synthetic

Figure 2: Tracking results on a real-world sequence.

experiments, we also performed real-world tests. We found that our algorithm can successfully deal with partial occlusions, blur, oblique viewing angles and low lighting conditions (see figure 2).
The Remote Support application was used in a remote



(a) Tracking accuracy under occlusions. Our method switches to feature-based tracking under significant occlusions.



(b) Tracking accuracy under fast motion. Our method adaptively switches between template- and feature-based tracking.

Figure 1: Tracking results on simulated sequences.

support scenario. Figure 3 shows some screenshots of the Remote Expert application during the remote support session. The expert and the user are located at two different locations and are connected over the Internet. The user has a measuring device and a piece of hardware and the remote expert assists him in performing the measurement correctly. The user follows the instructions given by the remote expert and finishes the experiment successfully.

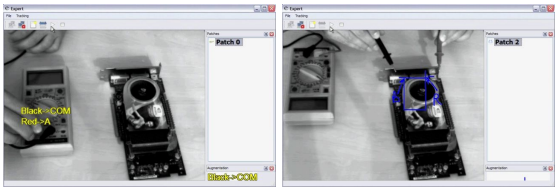


Figure 3: Screenshots from the Remote Expert application.

6 CONCLUSION

We presented a complete system based on a real-time markerless tracking algorithm which enables us to perform Augmented Reality in a priori unknown environments containing approximately planar patches. The presented system combines template-based and feature-based tracking algorithms in a homography-based model-free tracking framework and includes illumination compensation as well as a learning-free initialization procedure. This yields fast and robust tracking results which overcomes the problems most tracking algorithms have under real-world conditions, including partial occlusions, motion blur, fast object motion and changing lighting conditions. The remote support application allows a remote expert to interact with a local user by providing suitable augmentations for objects in the scene without requiring markers or knowledge about the scene.

In order to validate the proposed method, we performed multiple tracking experiments on both synthetic and real data. In addition, we evaluated the remote expert application with a support scenario.

REFERENCES

Baker, S. and Matthews, I. (2004). Lucas-kanade 20 years on: A unifying framework. *IJCV*, 56(3):221 – 255.

Barakonyi, I., Fahmy, T., and Schmalstieg, D. (2004). Remote collaboration using augmented reality videoconferencing. In *Proceedings of the Conference on Graphics Interface*, pages 89–96.

Benhimane, S. and Malis, E. (2004). Real-time image-based tracking of planes using efficient second-order minimization. In *IROS*, pages 943–948.

Friedrich, W. (2004). *ARVIKA - Augmented Reality für Entwicklung, Produktion und Service*. Publicis, first edition.

Hager, G. and Belhumeur, P. (1998). Efficient region tracking with parametric models of geometry and illumination. *PAMI*, 20(10):1025–1039.

Ladikos, A., Benhimane, S., and Navab, N. (2007). A real-time tracking system combining template-based and feature-based approaches. In *VISAPP*.

Lee, T. and Höllerer, T. (2006). Viewpoint stabilization for live collaborative video augmentations. In *ISMAR*, pages 241–242.

Lepetit, V., Lagger, P., and Fua, P. (2005). Randomized trees for real-time keypoint recognition. In *CVPR*, pages 775–781.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110.

Pressigout, M. and Marchand, E. (2006). Hybrid tracking algorithms for planar and non-planar structures subject to illumination changes. In *ISMAR*, pages 52–55.

Vacchetti, L., Lepetit, V., and Fua, P. (2004). Combining edge and texture information for real-time accurate 3d camera tracking. In *ISMAR*, pages 48–57.