SCENE-AWARE VIDEO STABILIZATION BY VISUAL FIXATION

Christian Kurz, Thorsten Thormählen, Hans-Peter Seidel

Max Planck Institute for Computer Science (MPII)
Saarbrücken, Germany
http://www.mpi-inf.mpg.de

Abstract

Visual fixation is employed by humans and some animals to keep a specific 3D location at the center of the visual gaze. Inspired by this phenomenon in nature, this paper explores the idea to transfer this mechanism to the context of video stabilization for a hand-held video camera. A novel approach is presented that stabilizes a video by fixating on automatically extracted 3D target points. This approach is different from existing automatic solutions that stabilize the video by smoothing. To determine the 3D target points, the recorded scene is analyzed with a state-of-the-art structure-from-motion algorithm, which estimates camera motion and reconstructs a 3D point cloud of the static scene objects. Special algorithms are presented that search either virtual or real 3D target points, which back-project close to the center of the image for as long a period of time as possible. The stabilization algorithm then transforms the original images of the sequence so that these 3D target points are kept exactly in the center of the image, which, in case of real 3D target points, produces a perfectly stable result at the image center. The approach is evaluated on a variety of videos taken with a hand-held camera in natural scenes.

Keywords: video stabilization, visual fixation, camera shake, camera motion estimation, structure-from-motion.

1 Introduction

When moving in an environment, the vision system of humans and several animals uses the process of ocular fixation that stabilizes the center of the visual gaze on a particular position in 3D space. Thereby, the movement of the eyes compensates the possible jitter introduced by the motion of the body [2]. Inspired by ocular fixation, in this paper we investigate, how the process of fixation can be used to stabilize the images of a video recorded with a hand-held video camera.

Current consumer cameras are usually equipped with video stabilization hardware to reduce camera shake; e.g., special lens systems or moveable image sensors in combination with gyroscopic sensors [7, 11]. However, these systems can usually compensate only small vibrations.

Software solutions offer greater flexibility and are able to remove undesired camera shakes of large amplitude. Most methods track image features [3] or estimate the optical flow [4] between successive images. This information is then used to obtain the parameters of 2D transformation between the images. The transformation parameters are then smoothed and the difference between the original and the smooth transformation is applied to compensate the undesired camera shake.

Different 2D transformations were explored, starting from a simple two-dimensional shift of the image [6, 12] to affine transformations [4]. Instead of using 2D transformations there are also approaches that employ 2.5D [9] or 3D camera models [13, 10, 1]. Various smoothing approaches exist, e.g., Kalman filters [6], particle filters [5], the Viterbi method [14], or other digital filters [12].

In this paper, we present an image stabilization approach, which simulates ocular fixation used in human and animal vision by fixating the camera orientation to a specific 3D target point in the scene. The advantage of this technique, in contrast to smoothing, is that after stabilization the target point is kept perfectly stable in the image center. The 3D target points are automatically determined by analyzing the recorded scene with a structure-from-motion algorithm. Thereby, the algorithm can either generate a virtual target point or a target point that is located on a real surface in the 3D scene. These target point extraction algorithms are simple to implement and require only a single user parameter, which controls directly how strongly the original image sequence is altered due to fixation.

The paper is organized as follows. In the next section, we describe the information that is available after employing a state-of-the-art structure-from-motion approach. Sections 3 and 4 introduce the algorithms to extract the virtual and real target points from the recorded image sequence. Section 5 explains how the target points can be used for video stabilization. These sections correspond to individual steps of the algorithm, which is illustrated in Fig. 1. In Section 6, we report results of our experiments that show the performance of the suggested algorithms. The paper ends with concluding remarks in Section 7.

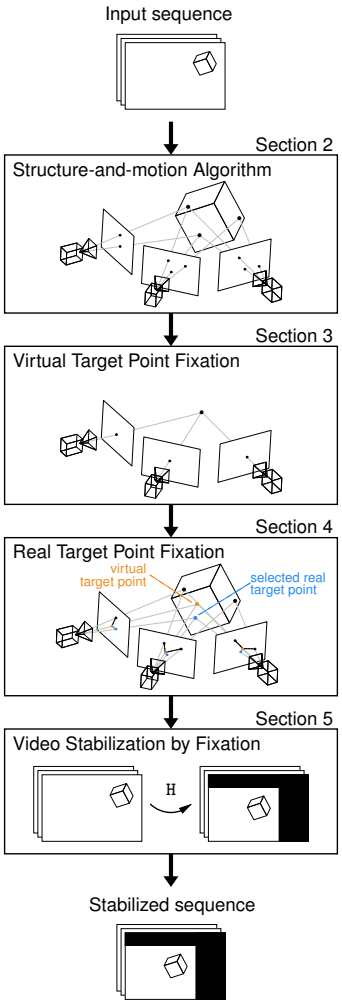

Figure 1: The processing pipeline of the stabilization by visual fixation algorithm.

2 Structure-from-Motion Algorithm

Reliable algorithms for camera motion estimation and 3D reconstruction of rigid objects from video have been developed over the last decades [8, 15, 17]. Employing such a state-of-the-art structure-from-motion algorithm is the first step in our processing pipeline.

Consider an image sequence consisting of K images I_k, with $k = 1, \ldots, K$. Let $\mathtt{A}_k$ be the 3×4 camera matrix corresponding to image I_k. First, corresponding 2D feature points $\mathbf{p}_{j,\,k}$ are determined in consecutive frames with the KLT-Tracker [16]. Using the corresponding feature points, the parameters of a camera model $\mathtt{A}_k$ are estimated for each frame. As shown in Fig. 2, for each feature track a corresponding 3D object point position is determined, resulting in set of J 3D object points $\mathbf{P}_j$, with $j = 1, \ldots, J$, where

$$\mathbf{p}_{j,\,k} \simeq \mathtt{A}_k \mathbf{P}_j \quad . \tag{1}$$

Thereby, the 2D feature points $\mathbf{p}_{j,\,k} = (p_x, p_y, 1)^\top$ and 3D object points $\mathbf{P}_j = (P_x, P_y, P_z, 1)^\top$ are given in homogeneous coordinates.

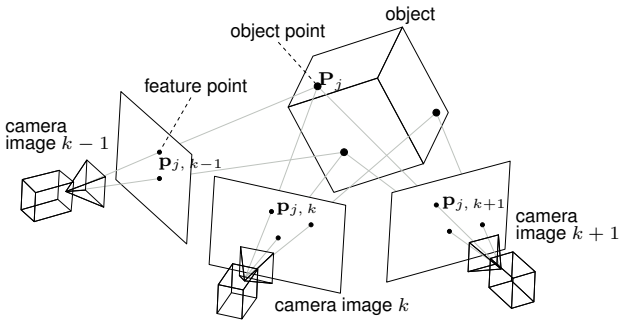

Figure 2: Result after structure-from-motion estimation. The projection of a 3D object point $\mathbf{P}_j$ in the camera image at time k gives the tracked 2D feature point $\mathbf{p}_{j,\,k}$.

The camera matrix $\mathtt{A}$ can be factorized into

$$\mathtt{A} \;=\; \mathtt{K}\,\mathtt{R}\,[\mathtt{I}\,|\,-\mathbf{C}] \quad , \tag{2}$$

where the 3×3 calibration matrix $\mathtt{K}$ contains the intrinsic camera parameters (e.g., focal length or principal point offset), $\mathtt{R}$ is the 3×3 rotation matrix representing the camera orientation in the scene, and the camera center $\mathbf{C}$ describes the position of the camera in the scene.

3 Virtual Target Point Fixation

Once the camera motion parameters and 3D object points have been obtained, the 3D target points $\mathbf{T}_i$ for fixation are estimated. It is assumed that the camera operator tries to keep the respective object of interest centered in the image but introduces large jitter because of the hand-held camera. Given the principal point $\mathbf{c}_k$ of the camera view k, which is the intersection of the optical axis with the image plane, an estimate for the 3D target point $\mathbf{T}_i$ can be found by a triangulation algorithm, which minimizes

$$\underset{\mathbf{T}_i}{\arg\min} \quad \sum_{n \in \mathcal{N}_i} d(\mathbf{c}_n\,,\, \mathtt{A}_n\,\mathbf{T}_i) \quad . \tag{3}$$

where $\mathcal{N}_i$ is a subset of the whole set of images $[1 \ldots K]$ consisting of strictly consecutive images, and $d(\ldots)$ denotes the Euclidean distance.

To determine a suitable subset of images $\mathcal{N}_i$ for a target point, a multi-scale approach is employed, which evaluates the sequence at multiple time-scales.

The coarsest scale is assigned to scale index $S = 0$, while the index is incremented for the subsequent, refined scales. Given a specific scale with the corresponding scale index S, the total number N_S of consecutive images for all individual subsets $\mathcal{N}_i$ for this scale is

$$N_S = K - S \quad . \tag{4}$$

For any given scale with scale index S the maximum number M_S of possible subsets $\mathcal{N}_i$ evaluates to

$$M_S = K - N_S + 1 = S + 1 \quad . \tag{5}$$

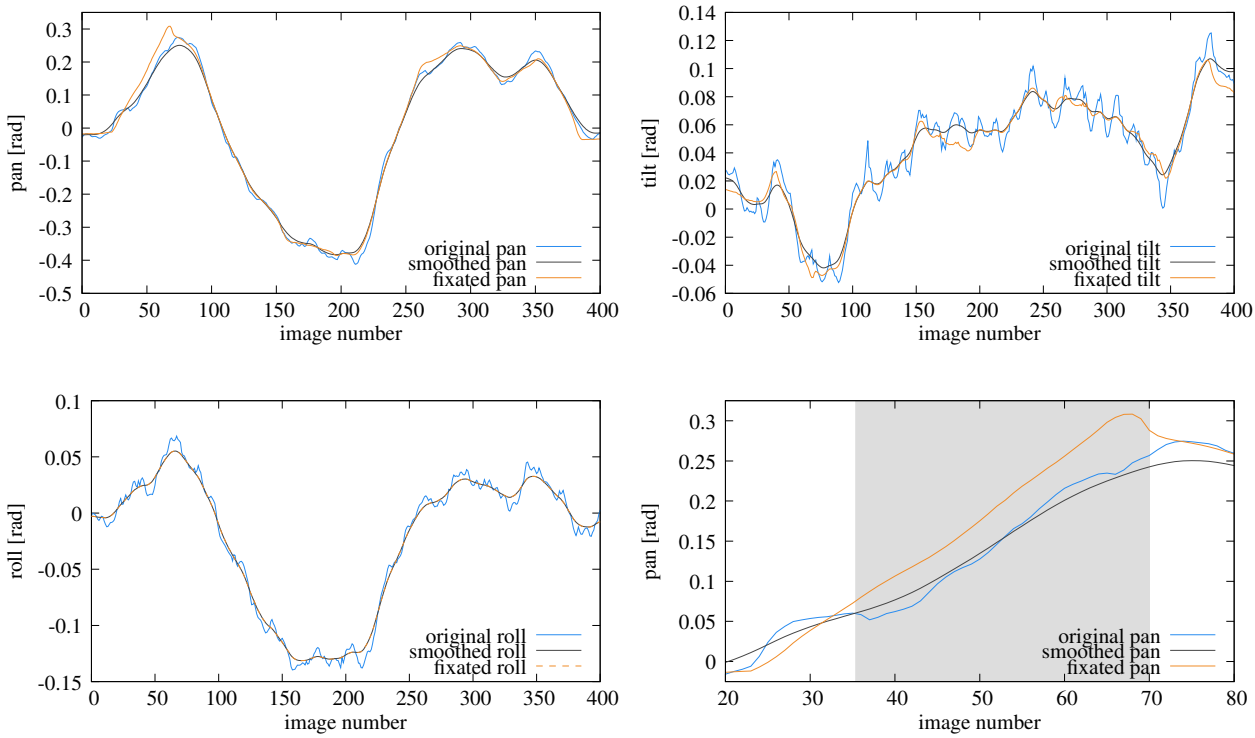

Figure 3: Example 1 - Comparison between the camera parameters estimated from the original image sequence, the smoothed parameters [4], and the fixated parameters. Results for camera parameters pan, tilt, and roll are shown. The diagram in the lower right corner shows a detail magnification for the pan parameter. The gray region indicates the fixation to a target point.

This is due to the fact that the subsets $\mathcal{N}_i$ are required to consist only of strictly consecutive frames. As an example, consider a sequence containing a total of $K = 90$ images for a scale with scale index $S = 30$, there are at most $M_{30} = 31$ different subsets $\mathcal{N}_i$ with a length of $N_{30} = 60$ images each.

Starting at the coarsest scale, the algorithm evaluates all possible subsets of consecutive images, by checking if the residual error of Eq. (3) is below a certain user defined threshold τ. If this condition is satisfied, a target point candidate is created and stored in a candidate list, which is sorted ascendingly according to the residual error.

After processing all subsets, the target point candidate with the lowest residual error is selected and moved to the list of accepted target points. The corresponding image set is assigned to the accepted target point and is excluded from further processing. All target point candidates, which share images with the accepted target point are removed from the candidate list. The process is repeated for the next target point candidate in the candidate list until the list is empty.

At the next finer time-scale all remaining possible subsets $\mathcal{N}_i$ containing N_S consecutive images are considered. Once all subsets of a given scale have been processed, the scale index S is increased and the corresponding subsets of the next finer time-scale are considered, where it is made sure that only subsets not containing images assigned to subsets on coarser time-scales are selected. This reduces the number of possible subsets for all finer scales.

The algorithm terminates after all images have been assigned to an accepted target point or further refinement is no longer possible.

4 Real Target Point Fixation

Only a 3D target point on a real surface permits a perfectly stable projection of the surface at the image center. Therefore, it is often desirable that the selected target point corresponds to a real 3D object point of the scene. When the user activates this real target point fixation, a suitable 3D object point is selected form the set of all J 3D object points $\mathbf{P}_j$ for each virtual target point. Thereby, it is evaluated whether the back-projection of the 3D object points in the subset of images, which is assigned to the current virtual target point, is close to the principal point $\mathbf{c}_n$:

$$\epsilon_j = \sum_{n \in \mathcal{N}_i} d(\mathbf{c}_n \, , \, \mathtt{A}_n \, \mathbf{P}_j) \quad . \tag{6}$$

The 3D object point $\mathbf{P}_j$ with the smallest error ϵ_j is selected.

Undesired results might be obtained for image sequences where 3D object points in the vicinity of the virtual target points were not generated during the structure-from-motion step due to a lack of interest points in the respective image regions. This problem can be solved by enforcing an additional threshold on the residual error ϵ_j and by reverting to the virtual target point if necessary.

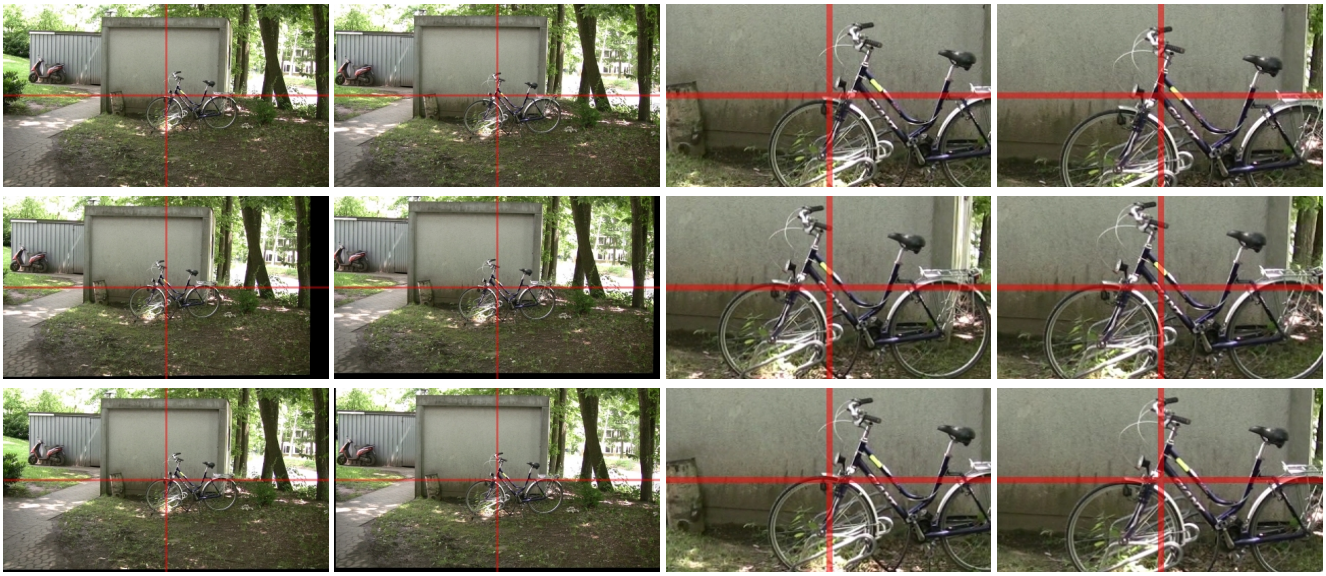

Figure 4: Example 1 - Original image sequence (top), result of stabilization by fixation (middle), result of smoothing with an affine model [4] (bottom). The images on the right are magnifications. With the stabilization by fixation approach the center of the image is kept perfectly stable. The red marker lines were added to facilitate visual verification.

5 Video Stabilization by Fixation

To stabilize the image sequence, a 2D transformation, given by the 3×3 matrix $\mathtt{H}_k$, is applied to all images I_k of the sequence. If $(x', y')^\top$ and $(x, y)^\top$ are the pixel positions in the stabilized and unstabilized images, respectively, this operation can be written as

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} \simeq \mathtt{H}_k \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \quad , \tag{7}$$

with

$$\mathtt{H}_k = (\mathtt{K}_k \, \mathtt{R}_k^{(s)})(\mathtt{K}_k \, \mathtt{R}_k)^{-1} \quad . \tag{8}$$

The calibration matrices $\mathtt{K}_k$ and the rotation matrices $\mathtt{R}_k$ are known from the structure-from-motion algorithm. The rotation matrices $\mathtt{R}_k^{(s)}$ are the smoothed versions.

A camera rotation matrix can be represented by three Euler angles, pan φ, tilt ϑ, and roll ρ with

$$\mathtt{R} = \mathtt{R}_y(\varphi) \, \mathtt{R}_x(\vartheta) \, \mathtt{R}_z(\rho) \quad , \tag{9}$$

where $\mathtt{R}_y$, $\mathtt{R}_x$, and $\mathtt{R}_z$ are rotations around the y, x, and z axis, respectively. Note that in Eq. 9 the index k is omitted for the sake of readability.

To find the smoothed rotation matrices $\mathtt{R}_k^{(s)}$, a regularization framework, as presented in [4], is employed. The regularization framework smoothes each of the three Euler angles independently and smoothed rotation matrices are generated from the smoothed Euler angles, as outlined in Eq. 9. Using this approach yields a smooth stabilization similar to the results presented in [4].

In our case, however, the fixation on a target point constraints the pan and tilt angle, and only the roll angle can still be chosen arbitrarily. Therefore, the pan and tilt angle are not smoothed but are are directly obtained from the fixation on the target point. As the fixation does not gives us any information about the roll angle, in absence of other knowledge, the smoothed roll angle as given by the regularization framework is employed.

Since our approach perfectly stabilizes the given target point in the center of the corresponding images, it is clear that the transitions between adjacent target points can be very abrupt. In most cases this effect is not desired and a smooth transition between adjacent targets is preferred. This can be achieved by applying the regularization framework mentioned above on a short image sequence covering the transition. With the same technique smoothed parameters can be calculated for longer parts of the image sequence that where not assigned to any target point.

6 Results

In this section, we present four real-world examples of video stabilization by fixation. Except example 2, all examples are recorded with off-the-shelf consumer HDV cameras at a resolution of 1440×1080 pixels and a frame rate of 25 Hz. In example 2 a SD camera with a resolution of 720×576 pixels was employed. The examples are also shown in the video provided with this submission.

Example 1 has a total length of 700 frames. With a threshold of $\tau = 5.0$ pixels eleven real target points were found. In Fig. 3 a comparison between the camera parameters estimated from the original image sequence, the smoothed parameters generated using the approach described in [4], and the fixated parameters is shown. The deviation of the fixated parameters from the smoothed parameters is visible, especially in the shown detail magnification. Because the roll parameter is also smoothed

Figure 5: Example 1 - Original image sequence (top), result of stabilization by fixation (middle), result of smoothing with an affine model [4] (bottom).

during fixation the smoothed and fixated roll parameter curve are on top of each other.

For comparison, sample images of the stabilization by fixation approach are shown in Figures 4 and 5, along with the corresponding images obtained through stabilization with an affine model. To facilitate verification of the visual fixation, a red cross-hair at the center of the images is superimposed. It can be observed that the fixation approach, in contrast to the affine stabilization, keeps the same 3D location perfectly in the image center.

Example 2 presents a sequence of 250 images with an approximate orbit motion around a dredger. A threshold of $\tau = 0.5$ pixels generated three real target points for stabilization by fixation. Sample images from the original and stabilized video are shown in Fig. 6.

Figure 6: Example 2 - Original image sequence (top), result of stabilization by fixation (bottom).

In example 3 and 4 very strong camera shakes are compensated by our video stabilization approach. Therefore, a large threshold of $\tau = 50.0$ pixels was chosen. In example 3, shown in Fig. 7, two target points were established over a sequence of 212 images. In example 4, shown in Fig. 8, three target points where established over a sequence of 150 images.

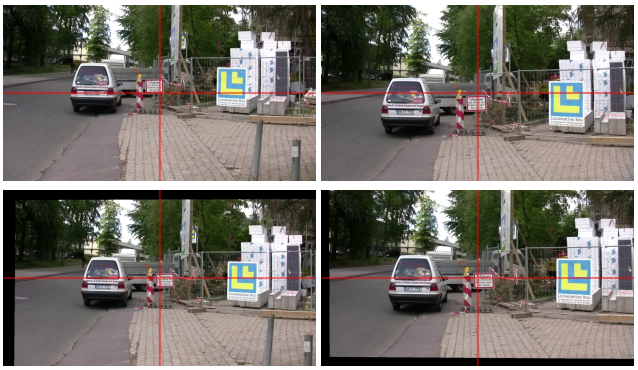

Figure 7: Example 3 - Original image sequence (top), result of stabilization by fixation (bottom).

7 Conclusion

In this paper we presented a video stabilization approach that fixates the center of the image to a specific 3D target point. After analyzing the scene with a structure-from-motion algorithm, these target points are automatically detected within the scene. The user can control how much the original sequence is altered by adjusting a single parameter τ. This user-supplied parameter specifies the maximum offset value of the projected target point to the image center in the original image sequence.

In contrast to existing automatic approaches, our approach can achieve an absolutely stable result in the center of the

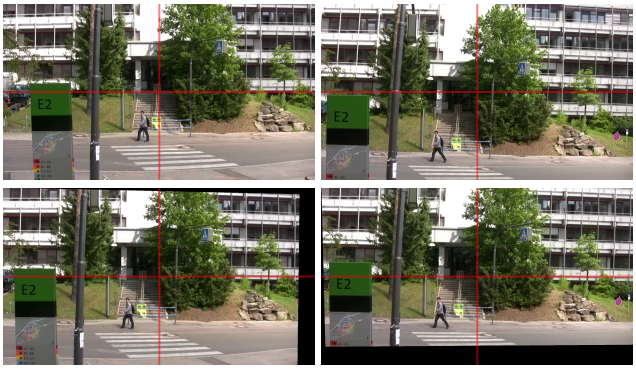

Figure 8: Example 4 - Original image sequence (top), result of stabilization by fixation (bottom).

images. One limitation of the approach is its dependency on the structure-from-motion algorithm. If this processing step provides wrong parameters, unpredictable results may occur. However, other automatic stabilization approaches are also dependent on reliable feature tracking. For scenes where the tracking of features is possible, state-of-the-art structure-from-motion also seldomly fails. If the camera performs a pure rotational motion, target points can not be found with the presented technique. However, similar techniques could be developed for this special case in future.

A general problem, which occurs with all image stabilization techniques that apply a 2D transformation to the image, is that the translational motion of the camera and the resulting motion parallax can not be compensated. This can be perceived as residual jitter artifacts in some of the presented videos. These artifacts could only be removed if a high quality depth map with occlusion information would be available for every pixel of all images. This is left for future research.

Acknowledgements

This work has been partially funded by the Max Planck Center for Visual Computing and Communication (BMBF-FKZ01IMC01).

References

[1] C.J. Buehler, M. Bosse, and L. McMillan. Non-metric image-based rendering for video stabilization. In *CVPR*, pages II:609–614, 2001.

[2] Roger H. S. Carpenter. *Movements of the Eyes*. Pion, London, 2nd edition, 1988.

[3] A. Censi, A. Fusiello, and V. Roberto. Image stabilization by features tracking. In *International Conference on Image Analysis and Processing*, pages 665–667, 1999.

[4] H.C. Chang, S.H. Lai, and K.R. Lu. A robust real-time video stabilization algorithm. *Journal on Visual Communications and Image Representation*, 17(3):659–673, June 2006.

[5] C.R. del Blanco, F. Jaureguizar, L. Salgado, and N. Garcia. Automatic feature-based stabilization of video with intentional motion through a particle filter. In *Proc. International Conference on Advanced Concepts for Intelligent Vision Systems (ACIVS 2008)*, pages 356–367, 2008.

[6] S. Erturk. Real-time digital image stabilization using kalman filters. *Journal of Real Time Imaging*, 8(4):317–328, August 2002.

[7] A. Fournier. Image stabilizing apparatus for a portable video camera. US Patent 5012347, 1991.

[8] Simon Gibson, Jon Cook, Toby Howard, Roger Hubbold, and Dan Oram. Accurate camera calibration for off-line, video-based augmented reality. In *IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR 2002)*, Darmstadt, Germany, 2002.

[9] J.S. Jin, Z.G. Zhu, and G.Y. Xu. Digital video sequence stabilization based on 2.5d motion estimation and inertial motion filtering. *Journal of Real Time Imaging*, 7(4):357–365, August 2001.

[10] W. Krüger. Robust real-time ground plane motion compensation from a moving vehicle. *Mach. Vision Appl.*, 11(4):203–212, 1999.

[11] T. Kudo, H. Kawahara, and J. Murakami. Vibration correction apparatus. US Patent 6734901, 2004.

[12] C.T. Lin, C.T. Hong, and C.T. Yang. Real-time digital image stabilization system using modified proportional integrated controller. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(3):427–431, 2009.

[13] C. Morimoto and R. Chellappa. Fast 3d stabilization and mosaic construction. In *CVPR*, pages 660–665, 1997.

[14] M. Pilu. Video stabilization as a variational problem and numerical solution with the viterbi method. In *CVPR*, pages 625–630, 2004.

[15] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch. Visual modeling with a hand-held camera. *International Journal of Computer Vision*, 59(3):207–232, 2004.

[16] Jianbo Shi and Carlo Tomasi. Good features to track. In *CVPR*, pages 593–600, 1994.

[17] Thorsten Thormählen, Nils Hasler, Michael Wand, and Hans-Peter Seidel. Merging of feature tracks for camera motion estimation from video. In *5th European Conference on Visual Media Production (CVMP 2008)*, London, UK, 2008.