Robust Multi-View Camera Calibration for Wide-Baseline Camera Networks

Jens Puwein
ETH Zurich
puweinj@student.ethz.ch

Remo Ziegler, Julia Vogel
LiberoVision
http://www.liberovision.com

Marc Pollefeys
ETH Zurich
marc.pollefeys@inf.ethz.ch

Abstract

Real-world camera networks are often characterized by very wide baselines covering a wide range of viewpoints. We describe a method not only calibrating each camera sequence added to the system automatically, but also taking advantage of multi-view correspondences to make the entire calibration framework more robust. Novel camera sequences can be seamlessly integrated into the system at any time, adding to the robustness of future computations.

One of the challenges consists in establishing correspondences between cameras. Initializing a bag of features from a calibrated frame, correspondences between cameras are established in a two-step procedure. First, affine invariant features of camera sequences are warped into a common coordinate frame and a coarse matching is obtained between the collected features and the incrementally built and updated bag of features. This allows us to warp images to a common view. Second, scale invariant features are extracted from the warped images. This leads to both more numerous and more accurate correspondences. Finally, the parameters are optimized in a bundle adjustment. Adding the feature descriptors and the optimized 3D positions to the bag of features, we obtain a feature-based scene abstraction, allowing for the calibration of novel sequences and the correction of drift in single-view calibration tracking. We demonstrate that our approach can deal with wide baselines. Novel sequences can seamlessly be integrated in the calibration framework.

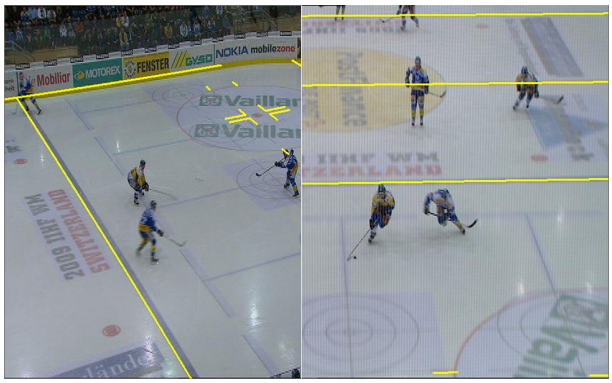


Figure 1. Starting from a calibrated frame, the proposed method calibrates a network of cameras. The images shown were captured by cameras separated by a very wide baseline. The successful calibration is visualized by projecting a wireframe model of the playing field.

1. Introduction and Related Work

Motivation Camera networks are prevalent in many areas like surveillance applications and sports broadcasts. Multiple calibrated cameras capturing the same scene can be harnessed to extract 3D data from the 2D images. This data can be measurements like the height, the 3D position or the movement of an object. It may even be used to synthesize novel views or to add virtual drawings to the scene, e.g. to enhance the experience of sports broadcasts [11, 12, 7].

The cameras capturing the scene can range from pan-tilt-zoom cameras to rail and spider cameras (i.e. moving cameras). To capture the scene from different viewpoints and to recover as much information as possible, cameras are often separated by very wide baselines. While this provides the user with extensive coverage, the feature matching between different cameras, beneficial or even necessary for camera calibration and 3D reconstruction, becomes very difficult.

Related Work Several techniques for feature extraction and description have been proposed over the past few years [8, 10, 18]. The degree of invariance towards changes in viewpoint differs and comes at the cost of distinctiveness. When captured by a camera, planar surface patches undergo perspective transformations. The synthesis of novel views and viewpoint normalization can reverse this process. Consequently, descriptors extracted from such views need not be invariant towards affine or projective transformations and can therefore be more distinctive. Normalized viewpoints have been used successfully to improve the matching of im-

978-1-4244-9497-2/10/$26.00 ©2010 IEEE

age regions [6, 19]. We use the idea of normalized patches to extract matches on dominant scene planes.

If it is known that cameras are capturing the same scene for a long periode of time, it makes sense to extract an abstract representation or to build a model of the scene being captured. The extraction of features from images to generate abstract representations of the real world is commonly used in location recognition and visual localization. Feature descriptors are not only associated with an image and 2D geometry information like scale and rotation, but they are linked to 3D geometry. Descriptors extracted from a new image are matched to a database containing feature descriptors and 3D information. The 3D information can be used for pose estimation and efficient matching [19, 2].

To the best of our knowledge no multi-view approaches for camera calibration have been presented in sports specific research. Large parts of the scenes are dynamic, cameras are separated by wide baselines and features are often occluded. This makes it hard to extract multi-view correspondences. Existing approaches deal with each camera separately, often relying on the known appearance and geometry of the playing field [16, 3]. There, lines are used to initialize and track the calibration or to avoid drift when using optical flow based trackers. While these approaches proved to be rather successful for single-view usage like augmented reality on the playing field, they lack the multi-view constraints, which reduce the relative error between cameras. This is important for 3D reconstruction or free-viewpoint rendering.

Contributions In this paper we present a framework to calibrate camera sequences separated by wide baselines, where additional sequences can be integrated into the system at any time, adding to its stability. Cameras are not required to capture the scene at the same time and, in particular, they do not have to be synchronized. By providing the calibration of a single frame, a method to extract and update a feature-based representation of the scene being captured is bootstrapped, taking advantage of different kinds of features.

In the feature matching stage, the large tolerance which MSER [10] features provide towards changes in viewpoint is leveraged and combined with SIFT [8] in a two-step matching procedure. The special structure of homographies mapping scene planes to images leads to a very efficient 2-point RANSAC when matching MSER features. To increase the robustness of the matching stage, features extracted from several frames are transformed and collected before matching. A feature based representation of the scene is created and updated using a very simple yet effective strategy based on the visibility and the uniqueness of features.

This paper is structured as follows. An overview of the proposed system is given in Section 2, followed by a formal description of the technical part in Section 3. The setup for our experiments and the results can be found in Section 4. Finally, we conclude the paper in Section 5.

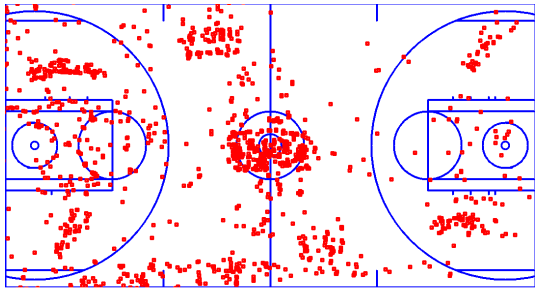


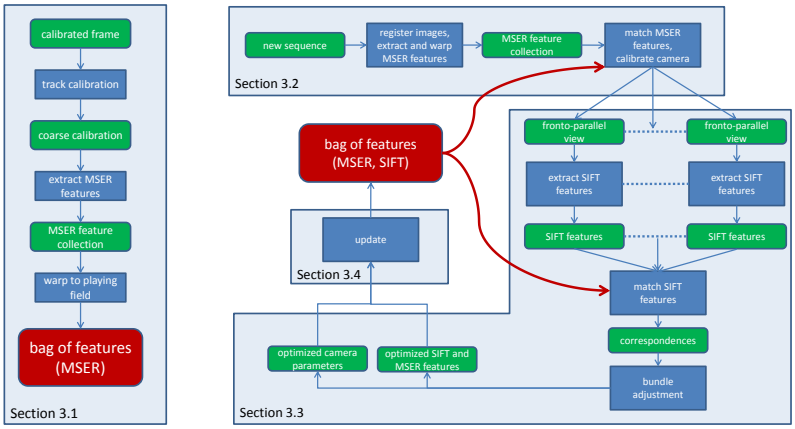
Figure 2. MSER and SIFT features are stored in a regularly updated bag of features representing the playing field. Red dots indicate positions on the playing field associated with the features.

2. System Overview

Usually the raw material of sports broadcast consists of the video streams provided by a variety of cameras, including fixed cameras, pan-tilt-zoom cameras, rail cameras, spider cameras, cameras mounted on large robot arms and cameras carried by humans. All of them except for the completely fixed cameras change zoom and orientation and some of them also change pose. In the following we work with sequences of video streams provided by such cameras. Our goal is to determine the 6 degrees of freedom of the camera extrinsics as well as the focal lengths, both for every frame. Radial distortion can be estimated in the bundle adjustment step [17].

An overview of the presented system is depicted in Figure 3. It is initialized with a single calibrated frame (e.g. Figure 6(a)), which determines absolute scale, rotation and translation of the scene. The initial calibration is propagated to a few neighboring frames using a KLT feature tracker [9, 13]. Given the resulting coarse calibration for those neighboring frames, MSER features are extracted and warped onto the playing field, i.e. the ellipses characterizing the MSER features are backprojected onto the playing field. The warped features initialize a bag of features (Figure 2), providing an abstract representation of the playing field (Section 3.1). This concludes the initialization stage (Figure 3(a)).

Whenever a new sequence is added to the system in the main loop (Figure 3(b)), the input images are registered using homographies. MSER features are extracted in all images and collected in a reference image. This feature collection is matched with the bag of features more robustly than features extracted from a single frame only. The affine transformations extracted from MSER feature matches lead



(a) (b)

Figure 3. (a) The initialization of the bag of features (b) The main loop

to a very efficient 2-point RANSAC (Section 3.2).

From the resulting MSER matches coarse initial calibrations are obtained, which allows us to warp input images into fronto-parallel views before SIFT features are extracted and matched. In a subsequent bundle adjustment the camera parameters and the positions of the features on the playing field are refined (Section 3.3). The bag of features is updated by warping and adding the new MSER features and SIFT features along with their optimized 3D positions. In order not to grow the bag of features continously and in order to keep it discriminative, a simple updating scheme based on visibility, scale and viewpoint criteria is applied (Section 3.4).

3. Technical Part

3.1. Initializing the System

The whole system is initialized with a calibrated frame. There are several ways to obtain such a frame. While automatic line based methods exist [16], it is difficult to reliably detect field lines in sports like ice hockey and basketball. Therefore, the user is required to hand-click a few predefined points on the playing field (e.g. corners of playing field) to make sure that the initial calibration is correct. Due to the standardized geometry the absolute scale of the scene is known. The 3D coordinate frame is chosen such that the playing field coincides with the xy-plane. Starting from this calibrated frame, the calibration is propagated to a few neighboring frames using a GPU based KLT feature tracker. Since the calibration and the geometry of the stadium are known, only features lying on the playing field are tracked. Additionally, a simple chromakeying based on a Gaussian mixture model of the colors on the playing field followed by some morphological operations allows us to get rid of most of the feature points detected on players, referees and balls. Feature tracks are pruned in a RANSAC step by making sure they can be explained by a homography.

Once all frames of the initial sequence are calibrated, MSER features are extracted for several keyframes, again ignoring parts of the image not corresponding to the playing field. Every 15th frame of the sequence is chosen to be a keyframe. Applying the affine invariant MSER feature detector results in a number of interest points and ellipses [10]. Warping an image patch centered at the interest point such that the ellipse associated with it becomes a unit circle renders the image patch invariant to affine transformations, except for rotation. Using the SIFT descriptor, the ambiguity in rotation is removed by aligning the image patch according to its dominant orientation [8]. The SIFT descriptor extracted from the normalized image patch is used for matching the MSER feature with other MSER features. Combining the normalizing affine transformations from two corresponding MSER features leads to an affine transformation between them, i.e. an affine correspondence.

Since the camera calibration of the initial sequence is known, the MSER features extracted from novel query images can be matched directly with the playing field instead of the calibrated images. For that purpose the affine transformations extracted from the calibrated frames need to be updated accordingly. Each ellipse in an image corresponds to an ellipse on the playing field. In order to be able to extract affine correspondences between the playing field and a query image, the transformations mapping the ellipses in the calibrated images to oriented unit circles need to be preceeded by the transformations mapping the ellipses on the playing field to the ellipses in the images. These local mappings of image patches on the playing field to the corresponding patches in the images are given by the Jacobians

of the homographies mapping the playing field to the images.

Due to the spatial coherence within successive frames some features are extracted several times. All MSER features together with their affine transformations and 3D positions are used to initialize the bag of features. While duplicates can be detected and eliminated by matching features across frames, the task of handling multiple occurences and features not useful for matching is left to the bag of features, as explained in Section 3.4.

3.2. Adding Additional Sequences

Once a novel sequence is added to the system, MSER features are extracted exactly like in the initial step. Since the camera parameters are unknown, the audience cannot be eliminated as before. However, assuming the camera shows the playing field to a large extent, the chromakeying, the KLT feature tracker and the RANSAC step can still be used to get homographies relating the image parts showing the playing field in neighboring frames. By accumulating the homographies, a mapping between any two images of the sequence is provided. In particular, this leads to a mapping from any image to the first image of the sequence. This means that all MSER features can be warped into the coordinate frame of the first image. Again, apart from the feature positions the affine transformations normalizing the image patches need to be adjusted as well. This is analogue to Section 3.1. Instead of using the mapping of the playing field to the image, the homography between the two images is used.

To obtain a coarse calibration for the new sequence, the collected features are matched with the bag of features. A match between MSER features f_i and f_j is a valid candidate match if and only if f_i is the most similar feature to f_j and vice versa. The resulting matches usually contain many outliers. The additional information provided by the ellipses obtained from MSER detections provides us with the means to apply an efficient 2-point RANSAC to extract the homography mapping the playing field to the image [1]. We pick up ideas from Koeser *et al*. to calculate a homography from two affine correspondences [5]. The key observation is that, locally, affine feature correspondences provide a first order Taylor approximation to homographies. While Koeser *et al*. present the case of conjugate rotations, a different method is necessary to deal with unkown poses and unknown focal lengths. The homography H describing the mapping from the playing field to the image has the following structure:

$$\begin{aligned} \mathrm{H} &= \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} \\ &= \lambda \mathrm{K} \begin{pmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mathrm{r}_1 & \mathrm{r}_2 & \mathrm{t} \end{pmatrix}. \quad (1) \end{aligned}$$

K contains known intrinsic camera parameters and f denotes the unknown focal length. r_1 and r_2 denote the first and the second column of the camera rotation matrix, respectively, and t is a translation vector. Since r_1 and r_2 are columns of a rotation matrix, the following constraint on the entries of H has to hold:

$$\begin{aligned} 0 &= (h_{11}h_{12} + h_{21}h_{22})(h_{32}^2 - h_{31}^2) \\ &\quad + h_{31}h_{32}(h_{11}^2 + h_{21}^2 - h_{12}^2 - h_{22}^2) \quad (2) \end{aligned}$$

One affine correspondence together with one point correspondence is already enough to determine the homography. Actually, it is already enough to have the x- *or* the y-coordinate of the point correspondence while the second coordinate can already be used to verify the generated hypothesis. A more detailed derivation is given in the appendix.

If the hypothesis obtained from two affine correspondences is valid, it is verified using the remaining correspondences. The threshold in the RANSAC procedure should be adapted for individual features due to the warp of the features. Each feature is warped using a homography. The determinant of the Jacobian of this mapping indicates the amount of magnification of the image area around a feature. The RANSAC threshold can be adjusted accordingly for every feature. If the number of inliers exceeds a certain threshold and the hypothesized homography does not describe a degenerate case, correspondences are assumed to be valid and they are used to initialize the camera parameters of the new sequence.

3.3. Calibration Refinement

The coarse initial calibration of all sequences added to the system so far can be refined in a bundle adjustment at any time. Since the correspondences obtained using MSER features are not very accurate and not very numerous, a different strategy is applied to extract the correspondences for the bundle adjustment.

The coarse calibrations obtained from the MSER matches provide a very good idea of which views should be overlapping. For each pair of views the amount of overlap of image regions representing the playing field is determined. Views are connected based on the overlap. All images are warped such that the image plane of the synthetic view is parallel to the playing field. For all warped views,

(a) (b)

Figure 4. (a) MSER feature matches from original images (b) SIFT feature matches from fronto-parallel views

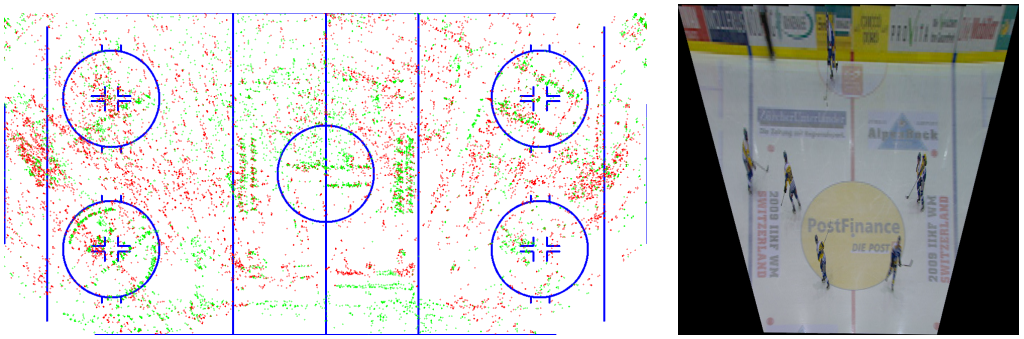
SIFT features are extracted and matched with the bag of features. Due to the already available coarse calibration only those features lying in the field of view are considered. Finally, for every pair of connected views, correspondences are extracted from the SIFT features in the warped images. Since we are dealing with fronto-parallel views of the same orientation, a very simple 1-point RANSAC or even exhaustive search can be applied. Figure 4 shows an example of MSER correspondences obtained from a pair of images and SIFT correspondences obtained from fronto-parallel views.

Once all connected views are processed, a graph having the features as nodes is constructed. Whenever two features match, an edge between them is introduced, i.e. feature matches are organized into tracks, as it was done by Snavely *et al*. [15]. In this graph, connected components with n nodes correspond to 3D points seen from n views. Together with the initial camera parameters, the DLT-algorithm provides initial values for the points on the playing field [4]. As explained before, only points on the playing field are considered. Hence, correspondences that cannot be explained by a homography are removed in a RANSAC step. The initial parameters for the cameras and the 3D points are refined in a bundle adjustment. The 3D points are constrained to lie on the playing field. Once several sequences have been added and the bundle adjustment contains a large number of images, the resulting camera parameters and 3D points are not optimized any further, but considered as fixed parameters in future bundle adjustments. After the bundle adjustment all feature positions are updated. New MSER features and SIFT features with optimized 3D positions and camera parameters are added to the bag of features. Again, MSER features not lying on the playing field are ignored. Features which are already in the bag are updated. As in the initial step (Section 3.1), the elimination of duplicates and useless features is left to the bag of features. Additionally, at some point, the number of images and MSER features becomes too large. The next subsection explains the maintenance of the bag of features.

3.4. Bag of Features

As explained before, the bag of features contains MSER features and affine transformations, as well as SIFT features extracted from fronto-parallel views. This provides an abstraction of the scene (see Figure 2), which gives a lot of flexibility for future calibration tasks. New sequences can be added at any time, enhancing the bag of features.

In order not to fill the bag of features with many similar or useless features, we propose a simple updating scheme. While features not lying inside the playing field mask were already removed in the previous steps (see Section 3.2), there might still be duplicates and many features which cannot be used for matching. For that purpose we keep track of how often an MSER feature should have been matched and how often it actually was matched. Whenever a new sequence is added and calibrated successfully, it is known which features from the bag have been visible in one of the new views and, therefore, should have been used. Since features are covered (e.g. by players) quite often, the matching frequency cannot be very high. To keep the bag as discriminative as possible, features should only be eliminated if they are not unique in terms of scale and viewpoint. Furthermore a feature should not be removed from the bag as long as it has not been matched at least 10 times although it should have been visible. Additionally, features acquired from very different viewpoints are not required to match at all, i.e. they are not considered as features that should have been used. After adding several new sequences to the bag of features a pruning step is applied. All features with a



(a) (b)

Figure 5. (a) Red dots indicate features that are removed from the bag of features during the update step. Features which are considered to be useful and thus remain in the bag are indicated as green dots. (b) The fronto-parallel view shows that logos on the playing field coincide with the green dots.

matching ratio below 0.1 are removed.

The bag of features is divided into a regular grid. Each grid cell corresponds to a patch on the playing field, 15cm by 15cm. When eliminating features, each grid cell is considered independently of all the others. Figure 5 illustrates the effect of updating the bag of features. Features which have been added to the bag only recently and therefore cannot be removed yet are not shown.

It might happen that a novel sequence cannot be matched with the bag of features. In this case the calibration of this sequence is postponed until a few other sequences have been added to the system. Since for some sequences the calibration might never succeed and every attempt to calibrate such sequences is a waste of resources, such sequences are stalled for an increasing number of new sequences. The number of intermediate sequences is increased as a power of 2, i.e. such sequences are added again after 1 new sequence, after 2 new sequences, after 4 new sequences and so on.

In order to avoid drift in single camera calibration tracking, keyframes can be matched with the bag of features.

4. Evaluation

Experiments are conducted on basketball and ice hockey footage. The initial calibration of the starting frame is done by picking 3D to 2D correspondences by hand (the geometry of the playing field is standardized and known, e.g. the freethrow line in basketball). Figure 6(a) shows the initial calibration.

In order to evaluate the method, 25 random sequences from a game of ice hockey (SD, resolution of 720x576) and 20 random sequences from a game of basketball (HD, 1280x720) are chosen. Each sequence is 60 frames long. For each sequence 4 out of the 60 frames are selected for calibration, which amounts to a total of 100 frames for the ice hockey footage and 80 frames for the basketball footage. The robustness of the system is inspected by counting the number of successfully calibrated frames. A calibration is considered successful if the average reprojection error is below 5 pixels. Additionally the results of the calibration are visualized by rendering the lines of the playing field model onto the original input images (Figures 6(b) and 6(c)).

For the ice hockey game, 52% of the sequences are calibrated successfully. While sequences showing large parts of the field are always calibrated successfully (see Figure 6(b)), cameras undergoing large zooms and fast movements are almost impossible to calibrate. The same holds for cameras filming at the height of the playing field, i.e. level with the players (see Figure 6(d)). Since the sequences are chosen completely at random without any subsequent selection and the footage at disposal often shows closeups of the goal area, the rate of successful calibration is not as high as one would expect. Nevertheless, the system manages to calibrate difficult sequences captured from new viewpoints separated by very wide baselines. The success rate of the camera shown in Figure 6(c) is at 67%, despite of fast camera movements.

The cameras capturing the basketball game are calibrated at a success rate of 51%. Apart from the typical overview camera the scene was captured by three more cameras. Again many sequences show closeups of the players. The field of view is often crowded with players occluding most of the features all the time. In such cases collecting features from different frames before matching does not help.

The implementation is not completely optimized for speed. Much time is spent on the extraction of features and the calculation of descriptors. Once extracted, descriptors are matched quite easily and efficiently on the GPU. While there exists a publicly available GPU implementation for SIFT feature extraction and matching [14], MSER features are handled by the CPU.

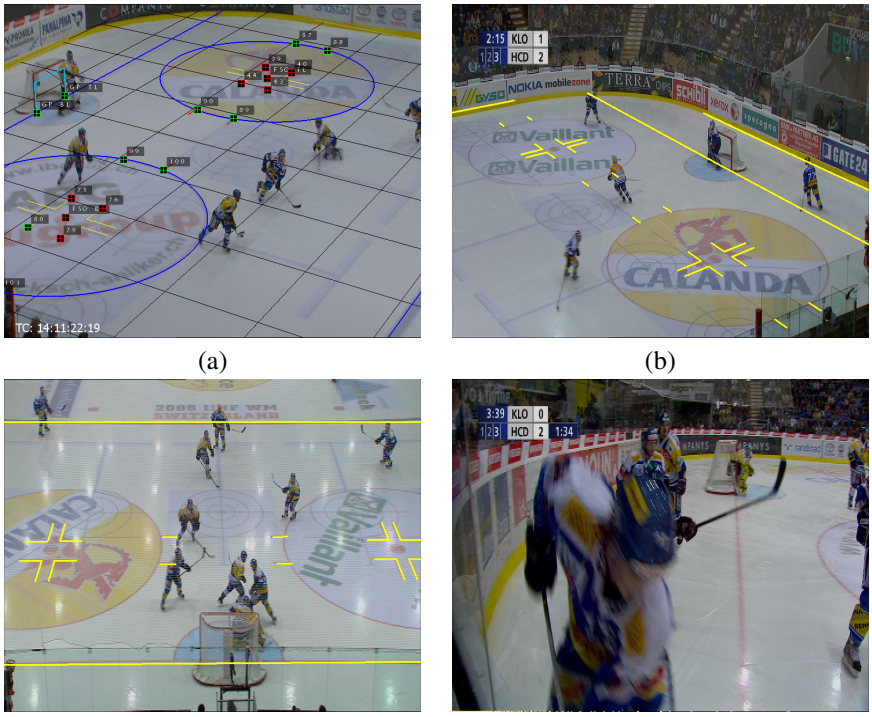


Figure 6. (a) Initial calibration from hand-clicked points, indicated as green squares (b) Typical overview camera (c) New view, separated by a wide baseline (d) Typical failure case

5. Conclusion and Future Work

In this paper we have presented an approach capable of dealing with the calibration of wide-baseline camera networks. We have shown its use for robust calibration and we have introduced a framework to seamlessly integrate novel sequences whenever necessary. The correspondences between images are obtained in several steps, coarse to fine, by collecting features and matching features in a two-step method involving MSER features and SIFT features extracted from fronto-parallel views. Each sequence added is used to update a bag of features representing the playing field, increasing the robustness of the system.

To further improve the quality of our method we plan to include field lines in the calibration procedure. This should increase the robustness in cases where field lines are visible. Additionally some stadiums offer diverse banners, which could be added to the system to further increase its robustness. With some restrictions the same holds for the audience in the stadium. Since both the banners and the audience approximately represent large scene planes, efficient matching should be possible.

Most of the cameras filming team sports such as basketball and ice hockey are mounted on some kind of device. Examples are pan-tilt-zoom cameras, cameras mounted on robot arms and cameras mounted on rails or wires. Modeling the resulting reduced space of possible calibrations could increase both robustness and accuracy.

6. Acknowledgments

The data is courtesy of Teleclub and LiberoVision. This project is supported by a grant of CTI Switzerland and the 4DVideo ERC Starting Grant Nr. 210806.

References

[1] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.

[2] F. Fraundorfer, C. Wu, J.-M. Frahm, and M. Pollefeys. Visual word based location recognition in 3d models using distance augmented weighting. In *International Symposium on 3D Data Processing, Visualization and Transmission*, 2008.

[3] S. Gedikli, J. Bandouch, N. von Hoyningen-Huene, B. Kirchlechner, and M. Beetz. An adaptive vision system for tracking soccer players from variable camera settings. In *Proceedings of the 5th International Conference on Computer Vision Systems (ICVS)*, 2007.

[4] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[5] K. Koeser, C. Beder, and R. Koch. Conjugate rotation: Parameterization and estimation from an affine feature corre-

spondence. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[6] K. Koeser and R. Koch. Perspectively invariant normal features. In *Proceedings IEEE International Conference on Computer Vision*, 2007.

[7] Liberovision. http://www.liberovision.com/.

[8] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[9] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: Proceedings of the 7th International Joint Conference on Artificial Intelligence*, pages 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc.

[10] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004.

[11] N. Owens, C. Harris, and C. Stennett. Hawk-eye tennis system. In *VIE: International Conference on Visual Information Engineering*, pages 182–185, 2003.

[12] Piero. http://www.redbeemedia.com/piero/.

[13] J. Shi and C. Tomasi. Good features to track. In *"Proceedings IEEE Conference on Computer Vision and Pattern Recognition"*, 1994.

[14] S. N. Sinha, J.-M. Frahm, M. Pollefeys, and Y. Genc. Gpu-based video feature tracking and matching. Technical report, In Workshop on Edge Computing Using New Commodity Architectures, 2006.

[15] N. Snavely, S. Seitz, and R. Szeliski. Photo tourism: Exploring image collections in 3d. In *Proceedings SIGGRAPH*, 2006.

[16] G. Thomas. Real-time camera tracking using sports pitch markings. *Journal of Real-Time Image Processing*, 2(2-3):117–132, 2007.

[17] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment: A modern synthesis. In *ICCV '99: Proceedings of the International Workshop on Vision Algorithms*, 2000.

[18] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Found. Trends. Comput. Graph. Vis.*, 3(3):177–280, 2008.

[19] C. Wu, B. Clipp, X. Li, J. Frahm, and M. Pollefeys. 3d model matching with viewpoint-invariant patches (vip). In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

A. Homography Computation and Verification

A homography of the form shown in Equation 1 does not have the usual 8 degrees of freedom [4]. It is constrained by the special structure of r_1 and r_2, which are the first two columns of the camera rotation matrix:

$$\mathrm{r}_1{}^T\mathrm{r}_1 = 1,\ \mathrm{r}_2{}^T\mathrm{r}_2 = 1 \text{ and } \mathrm{r}_1{}^T\mathrm{r}_2 = 0. \tag{3}$$

Since K is known, it can be eliminated by transforming image coordinates accordingly. What remains is a scaling of the first two entries of r_1 and r_2 by f, which leads to

$$\mathrm{H} = \lambda \begin{pmatrix} fr_{11} & fr_{12} & ft_1 \\ fr_{21} & fr_{22} & ft_2 \\ r_{31} & r_{32} & t_3 \end{pmatrix}. \tag{4}$$

From $\mathrm{r}_1{}^T\mathrm{r}_2 = 0$ follows $h_{11}h_{12} + h_{21}h_{22} + f^2h_{31}h_{32} = 0$. Solving for f^2 leads to

$$f^2 = -\frac{h_{11}h_{12} + h_{21}h_{22}}{h_{31}h_{32}}. \tag{5}$$

Similarly, $\mathrm{r}_1{}^T\mathrm{r}_1 = \mathrm{r}_2{}^T\mathrm{r}_2$ means that $h_{11}^2 + h_{21}^2 + f^2h_{31}^2 = h_{12}^2 + h_{22}^2 + f^2h_{32}^2$. Inserting the value of f^2 from Equation 5 finally leads to the desired constraint given in Equation 2.

Treating the homography as a mapping f_H from $\mathbb{R}^2$ to $\mathbb{R}^2$, image coordinates $\mathrm{x} = (x, y)^T$ are transformed as

$$f_\mathrm{H}(\mathrm{x}) = \frac{1}{h_{31}x + h_{32}y + h_{33}} \begin{pmatrix} h_{11}x + h_{12}y + h_{13} \\ h_{21}x + h_{22}y + h_{23} \end{pmatrix}. \tag{6}$$

Since homographies are defined up to scale, h_{33} can be set to 1. Given an affine correspondence and the associated affine transformation f_A mapping x_1 to x_2, the homography H should fulfil $f_\mathrm{H}(\mathrm{x}_1) = \mathrm{x}_2$. By translating the source and the target coordinate systems such that the homography maps $(0, 0)^T$ to itself, h_{13} and h_{23} become 0. After setting h_{13} and h_{23} to 0, the Jacobian of f_H evaluated at $\mathrm{x}_0 = (0, 0)^T$ is

$$\left.\frac{\partial f_\mathrm{H}}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{0}} = \begin{pmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{pmatrix}, \tag{7}$$

which leads to the following first order Taylor approximation:

$$f_\mathrm{H}(\mathrm{x}_0 + \delta\mathrm{x}) \approx f_\mathrm{H}(\mathrm{x}_0) + \left.\frac{\partial f_\mathrm{H}}{\partial \mathrm{x}}\right|_{\mathrm{x}=\mathbf{0}} \delta\mathrm{x} \tag{8}$$

Setting this equal to the affine transformation $f_\mathrm{A}(\mathrm{x}) = \mathrm{A}\mathrm{x} + \mathrm{t}$ obtained from the affine correspondence leads to

$$\begin{pmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \tag{9}$$

since the coordinate systems are chosen such that the translation vector $\mathbf{t}$ is $\mathbf{0}$. Therefore the only unknowns left to determine are h_{31} and h_{32}. In addition to Equation 2 one more constraint is needed. This constraint is provided by a second (point) correspondence, i.e. $f_\mathrm{H}(\mathrm{y}_1) = \mathrm{y}_2$. Using one of the two resulting equations leads to three possible solutions for H, while the second equation can already be used to verify the correctness of H. What is especially nice about this verification is that the remaining equation which needs to be fulfilled is given in pixel units.