MASTERCAM FVV: ROBUST REGISTRATION OF MULTIVIEW SPORTS VIDEO TO A STATIC HIGH-RESOLUTION MASTER CAMERA FOR FREE VIEWPOINT VIDEO

Florian Angehrn[2], *Oliver Wang*[1], *Yağız Aksoy*[1,2], *Markus Gross*[1,2], *and Aljoša Smolić*[1]

[1]Disney Research Zurich [2]ETH Zurich

Fig. 1: Concept of MasterCam FVV combining frameless production and multiview registration.

ABSTRACT

Free viewpoint video enables interactive viewpoint selection in real world scenes, which is attractive for many applications such as sports visualization. Multi-camera registration is one of the difficult tasks in such systems. We introduce the concept of a static high resolution master camera for improved long-term multiview alignment. All broadcast cameras are aligned to a common reference. Our approach builds on frame-to-frame alignment, extended into a recursive long-term estimation process, which is shown to be accurate, robust and stable over long sequences.

Index Terms— Free viewpoint video, camera registration, multiview video, sports visualization, alignment

1. INTRODUCTION

Free viewpoint video (FVV) is a new form of visual media, which enables interaction with the content in the sense of selection of ones own viewpoint and viewing direction of a real world scene [1]. Although FVV for the end user is still a while away from being realized, it is already widely used in production of movies and TV. Sports visualization is a particularly attractive application area [2].

Typically an FVV pipeline contains various challenging and error prone image processing tasks such as camera calibration, segmentation, depth estimation, and 3D reconstruction. Any errors in these modules influence the quality of the final output [1]. Temporal stability is particularly difficult to achieve, which is why high quality dynamic FVV, where content and camera are moving, is still very rare. The hardware system, especially the camera setup, is also of crucial importance for the design of FVV algorithms [1]. This may range from using only available broadcast cameras [2] to sophisticated additional multi-camera installations [3].

Another current trend in broadcast technology research is frameless or format-agnostic production [4]. A very high resolution, wide angle static overall view of a scene, such as a complete football stadium, is captured by panoramic imaging or 4k and beyond cameras. The actual framing of the broadcast signal, in a lower resolution like HD, happens by cropping a window of pixels out of the overall view. The cameraman may operate using a joystick to pan and zoom over the panorama.

In this paper, we present a concept for high quality dynamic FVV combined with the basic idea of frameless production as described. As illustrated in Figure 1 the center of the concept is a static high resolution master camera capturing a wide overall view of a whole stadium. The master camera is associated with a static 3D model of the stadium. All other broadcast cameras are registered independently to the master camera and with that also to the 3D model. The result is a fully registered broadcast framework with a very high quality image as reference. We believe that this will ease all error prone image processing tasks of FVV pipelines, i.e. make them more accurate, robust and reliable, and with that create a framework for high quality dynamic FVV broadcast in the future.

As a first step, we present here our approach for robust registration of multiple broadcast cameras to a common high resolution master camera. Such registration has to be accurate and reliable over a long time and avoid drift that occurs due to the accumulation of errors. Still it has to be computationally efficient and operate within a reasonable time budget. We show that this can be done using a recursive predict-and-correct homography estimation combining frame-to-frame with long-term registration. A full 3D calibration is only necessary for the first frame, while the rest of the processing is purely image-based.

2. RELATED WORK

Multi-camera synchronization and FVV bring about challenges at virtually every step of the pipeline. Although all of them have been subject to scientific investigation, the field still requires intensive efforts to overcome the existing limitations. Smolic [1] gives an overview over the whole pipeline of 3D video and FVV. It includes the steps of capturing, processing, representing, coding, transmitting, rendering and displaying.

There are several approaches to synthesize novel views from multiple cameras into FVV, focused on football scenes [5, 6, 7, 8].

However, they all use static, pre-calibrated cameras; for example, calibration using vanishing points is used in [6]. These approaches focus more on the synthesis of free viewpoint images and some do not describe the calibration step at all.

Efforts have been also undertaken in converting 2D sports content to 3D [9]. Schnyder et al. describe how to render 3D stereoscopic images from only one regular camera: based on the field ground or the camera setup a depth model is approximated and the segmented players are rendered as billboards from two nearby views of the original camera with the required disparity for 3D stereoscopic rendering.

Some publications present registration of dynamic scenes of a single camera and the reconstructions remain two dimensional. They typically estimate homographies to describe the frame changes. This type of registration is not sufficient to synchronize multiple cameras as it lacks the possibility to determine matching points across multiple cameras with a common 3D coordinate system. Steedly et al. [10] register videos into panoramic mosaics. In order to speed up the process they identify keyframes and following that they register and stitch the rest of the frames. Ghanem et al. [11] propose a parameter-free method to register a video for the application of field-sports analysis. They claim that the advantage of their method is the matching of image patches or entire images rather than relying on salient points on the field like points and lines. However, their algorithm does not consider any temporal stability of the estimation between frames.

Carr et al. [12] propose a method for gradient-based alignment to edge images. It requires an initial calibration with a reprojection error of only a few pixels. Although a model of the field is necessary, our tests have shown that it also works with edge response images of a sports field. It extracts the full camera calibration as well as one radial distortion parameter. According to the results of our tests, which showed that it takes several minutes to compute the calibration of one frame, it is not feasible to use this method to calibrate a whole sequence of frames. The following publications target whole video scenes [13, 14, 15, 16]. Farin et al. [14] allow the user to manually input a field model. Therefore they make their algorithm applicable for a variety of different sports. Using Hough transformations they find correspondences between the image and the model, and with gradient descent the camera parameters are refined.

Guillemaut et al. [17] build a whole FVV system for dynamic scenes. The camera calibration is based on the approach of Thomas [16]. The key idea for the further steps is the combination of the segmentation and reconstruction via graph-cut optimization. The energy function consists of color, contrast, matching and smoothness terms.

3. MASTERCAM ALIGNMENT

MasterCam alignment denotes individual registration of multiple dynamic broadcast cameras to a common reference camera view, which is high resolution, wide angle overview and static (i.e. the camera does not move or zoom, the content may move). It has to be robust and accurate over a long period, while keeping computational complexity reasonable. The reference view is registered to a 3D model of the scene (e.g. a stadium), which can be done in a pre-processing step. Having all cameras registered to a high quality static reference view with a 3D model behind, will improve error prone image processing steps to achieve high quality dynamic FVV in the future.

Our approach includes an initial calibration step, where the first frame of a certain dynamic broadcast camera is locked to the master view. Typically a broadcast view will cover only a small portion of the master view and it is unpredictable where this will be. We therefore apply an interactive process, where the user specifies corresponding points in both views that determine the initial calibration [18]. As a football field is approximately planar, this calibration is fully defined by an initial homography matrix H_0. This model assumption of planar image relations allows us to keep all following registration also in homography and image space, without the need for tracking of full 3D calibration.

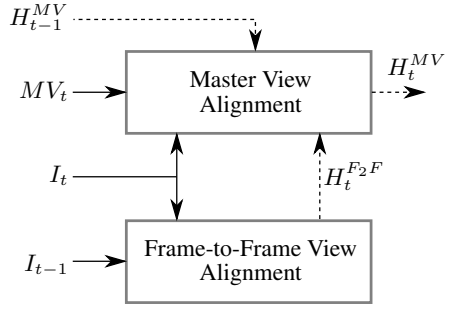

Fig. 2: Block diagram of MasterCam alignment. Current and previous frames from a camera, I_t and I_{t-1}, are used to extract the frame-to-frame homography $H_t^{F_2F}$. This homography is fed to the refinement phase together with the master frame MV_t, current frame and homography from master frame to the previous frame H_{t-1}^{MV} to estimate the current mapping H_t^{MV}.

After initial homography calibration, our method tracks homographies over time between each view and the master view, which must be stable and robust to keep each camera locked accurately to the master model. Therefore the alignment of a certain camera is divided into 2 steps as illustrated in Figure 2. First frame-to-frame alignment is performed between consecutive views of the corresponding video sequence. The result is then refined by registration to the master reference view. These algorithms are described in the following sub-sections in detail.

3.1. Frame-to-frame alignment

Robust homography or global motion estimation between consecutive frames I_t and I_{t-1} of a video sequence is a well-studied prob-

Fig. 3: Drift in long-term estimation due to error accumulation from frame-to-frame estimation. (30th (top) and 120th frames)

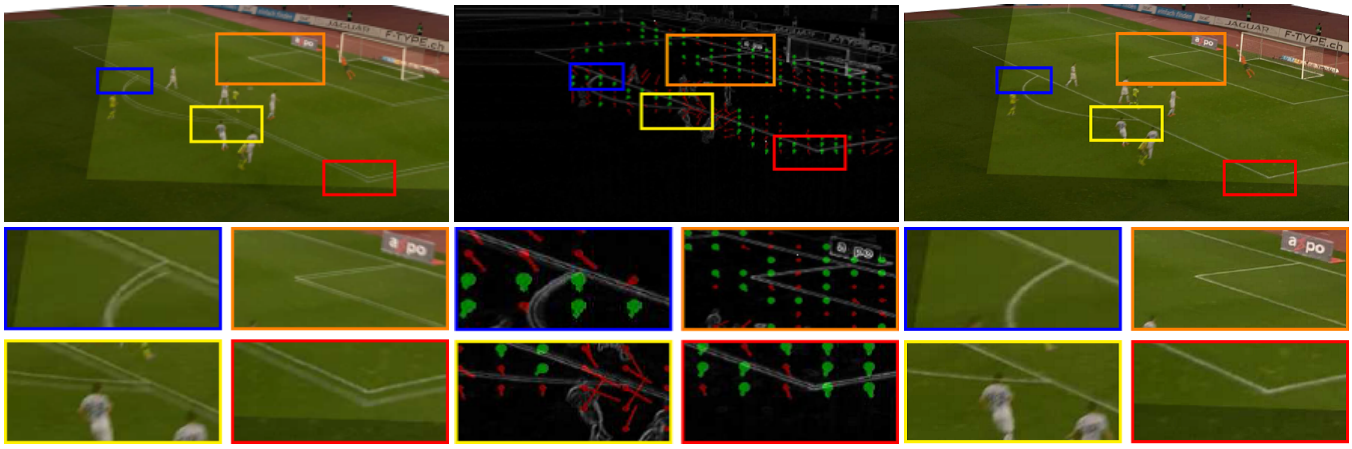

Fig. 4: Initial and final alignments of a broadcast view to the master view (left and right images, respectively) and DuctTake [19] correspondence vector fields on edge images in the middle. In the vector field, green are inliers, red are outliers.

lem, see [20] as an example. We use a state-of-the-art feature-based method combining feature matching [21], and robust estimation [22]. Typically such an approach is highly efficient and results in accurate estimates for frame-to-frame global motion $H_t^{F_2F}$, even for most fast moving sports scenes as we consider here.

A first approach for long-term alignment to the master view, expressed as homography H_t^{MV}, given an initial alignment H_0, would be to concatenate the frame-to-frame homographies $H_t^{F_2F}$ over time. However, due to even small inaccuracies, such an approach leads to error accumulation and drift. Typical results are shown in Figure 3. All images show aligned broadcast camera views overlaid over the master view. The results on the lefts are after 30 frames for 2 different examples, which are still well-aligned. However, after 120 frames as shown on the right, alignment is lost. The top example is for a single broadcast camera. The bottom example includes 3 broadcast cameras over the master view. Clearly, long-term stability cannot be guaranteed with such a frame-to-frame approach.

3.2. MasterCam refinement

Obviously, concatenation of frame-to-frame estimates cannot provide long-term stability. Interactive calibration of each single frame as done for initialization is neither an option. Our solution to this problem is a recursive prediction and correction approach that locks each view of the broadcast camera accurately to the master view, similar to long-term video mosaicking approaches [23]. According to Figure 2 our MasterCam refinement works as follows. Frame-to-frame homographies $H_t^{F_2F}$ describing the global motion between I_t and I_{t-1} are used as input as well as the current broadcast camera view It and master view MV_t. Additionally, the previous long-term homography H_{t-1}^{MV} defining the alignment between the previous broadcast view and master view is used as input. As we will see, the temporal sequence of H_{t-1}^{MV} carries by recursion the long-term information about stable alignment of the whole video sequence It-1 to the master view MV_{t-1}.

As we know the alignment of the previous frame to the master view H_{t-1}^{MV} and the alignment of the current frame to the previous frame $H_t^{F_2F}$, we can compute an initial estimate $\hat{H}_t^{MV}$ of the alignment of the current frame to the master view by concatenation of the homographies as:

$$\hat{H}_t^{MV} = H_{t-1}^{MV} \times H_t^{F_2F}$$

This initial estimate $\hat{H}_t^{MV}$ is in general already a good approximation of the final estimate H_t^{MV}, however, it also carries inaccuracies of the frame-to-frame alignment, which could accumulate over time and cause drift. Left image in Figure 4 shows results of such an initial alignment of a broadcast view to the master view. Although the input views are very far off, which could hardly be estimated automatically with any calibration algorithm, our recursive propagation algorithm already provides a good match. However, inaccuracies as visible in the detail views would cause drift.

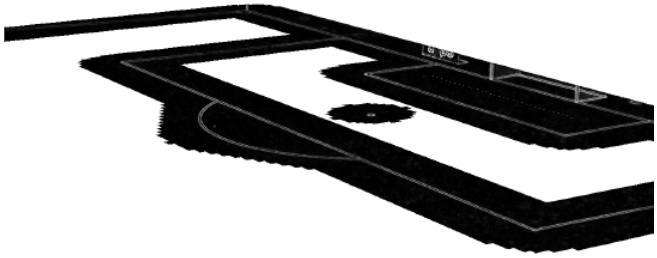

Fig. 5: Edge image for master view with predefined search areas containing distinctive line features.

We therefore add a refinement step that estimates a homography between the aligned (warped by $\hat{H}_t^{MV}$) broadcast image $\hat{I}_t$ and the master view MV_t. The first idea would be to use the same feature-based algorithm as for the frame-to-frame alignment described in 3.1. However, such an approach tends to be unstable due to lack of sufficient image features between image $\hat{I}_t$ and the master view MV_t (unlike I_t and I_{t-1}). We therefore use an approach based on edge alignment. Figure 5 shows an edge image computed using a simple Sobel 3×3 operator for the master view. A football field contains distinctive line features, and we a priori know the approximate locations from the predefined 3D model and initial calibration. We can restrict processing to those areas. A corresponding edge image is also computed for $\hat{I}_t$.

To compute correspondences between the edge images we use a recently presented patch-based method called Hierarchical Compass Search [19], which operates on a regular grid instead of features and was shown to be highly accurate and robust. Middle image in Figure 4 shows a result of such a correspondence vector field. Next, we compute a refinement homography H_t^R from such a vector field using again robust statistics, which describe alignment of the predicted image $\hat{I}_t$ and the master view MV_t. Figure 4 also shows which correspondence vectors were classified as inliers or outliers

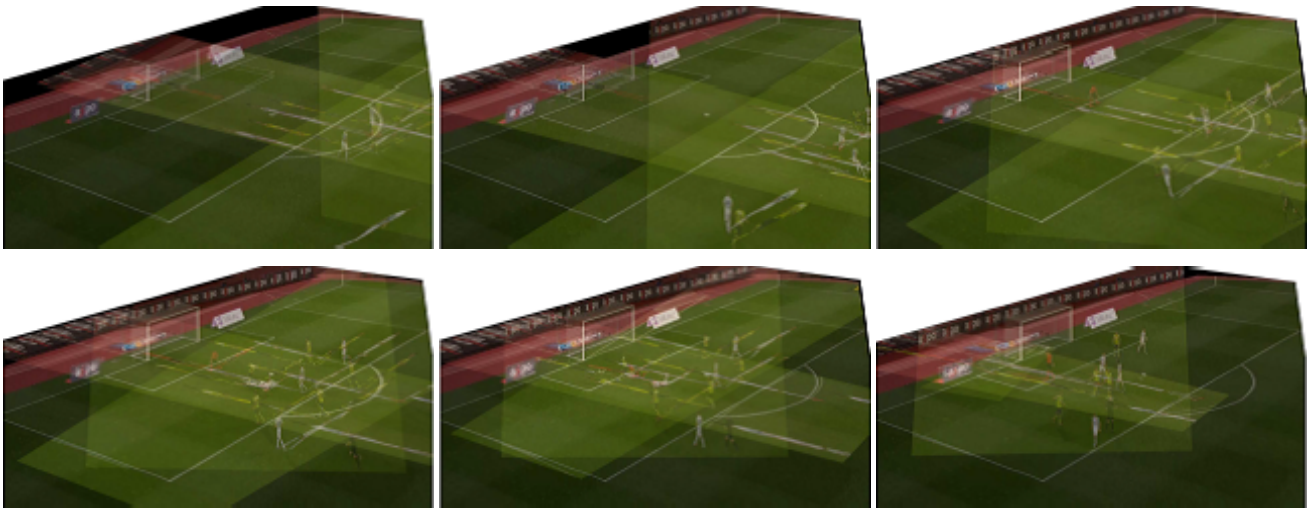

Fig. 6: Alignment of 3 cameras to master view for frames 0-271.

using RANSAC [24].

Finally, we concatenate the prediction $\hat{H}_t^{MV}$ and the refinement H_t^R to the final homography H_t^{MV} describing alignment of the current view to the master view:

$$H_t^{MV} = H_t^R \times \hat{H}_t^{MV}$$

A result of such a high quality alignment is shown also in Figure 4. The final homography H_t^{MV} is then used in the next step of the recursive process.

4. EXPERIMENTAL EVALUATION

We tested our algorithms with professional multiview footage (HD 720p) kindly provided by Teleclub AG and Swiss Radio and Television (SRF). We had access to several minutes of synchronized broadcast video showing a professional Swiss league football game in a stadium. In our experiments we mainly used the center camera at the midfield line and the offside cameras at the penalty area, all showing wide field views. In some experiments we also used cameras behind goals. As during that production there was no master camera available, we used single static high resolution pictures showing half of the field, captured using a DSLR as static master views, and registered the broadcast views to these images.

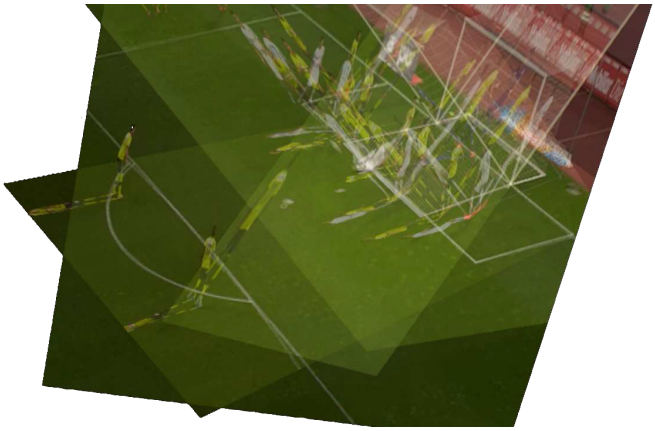

Fig. 7: Alignment of 4 cameras to master view.

We selected in total 7 different scenes of 201 to 406 frames length at 25 fps. Figure 6 shows an example of alignment of 3 broadcast cameras (center, offside and goal) to the master view, which is very accurate and stable over the long period with rapid camera operation (pan, zoom) including motion blur and limited overlap of views. In all our experiments we were able to align the center camera accurately to the master view over the whole sequence. Offside cameras sometimes showed small inaccuracies. Cameras behind goals were more difficult to register. Here the success rate was limited.

Figure 7 shows an example of 4 cameras (center, 2 offside, goal) aligned to the master view. Despite totally different views, zoom levels, etc. alignment is accurate as can be seen from match of field lines and player feet positions. This also illustrates the power of the master camera concept as we now have all cameras registered accurately to a common reference which enables a lot of consecutive processing as discussed before. Complete results as well as video examples can be found online [25].

5. CONCLUSIONS AND FUTURE WORK

We introduced the concept of a master camera for FVV, which is inspired by frameless production approaches. The idea is that a common, very high resolution static camera defines the image-based reference for a whole FVV broadcast framework. Error prone video processing tasks can largely benefit from such a setup. The first task for realization of such a framework covered in technical detail in this paper was registration of multiple broadcast cameras to a common master view. We presented a powerful approach based on a recursive prediction and correction loop to estimate homographies, which define the image-based alignment of each broadcast view to the master view. Tested with professional broadcast footage, we achieved high accuracy and long term stability in our experiments despite challenging properties of sports content (motion, zoom, blur, etc.). Our future work will include development of the whole MasterCam FVV system. We also plan for a realistic test under production conditions in a stadium, integrating a 4k camera as MasterCam. Based on that, also further improvements of the registration algorithms will be in focus, e.g. developing a Kalman filter approach for recursive long-term estimation.

Acknowledgements: We'd like to thank Teleclub AG and Swiss Radio and Television (SRF) for providing the professional multiview footage used in our experiments and this paper.

6. REFERENCES

[1] Aljoscha Smolic, "3d video and free viewpoint videofrom capture to display," *Pattern Recognition*, vol. 44, no. 9, pp. 1958 – 1968, 2011.

[2] "Viz libero," http://www.vizrt.com/products/viz_libero/, [Accessed on 2014.02.12].

[3] "Carnegie mellon goes to the super bowl," http://www.ri.cmu.edu/events/sb35/tksuperbowl.html, [Accessed on 2014.02.12].

[4] O. Schreer, I. Feldmann, C. Weissig, P. Kauff, and R. Schafer, "Ultrahigh-resolution panoramic imaging for format-agnostic video production," *Proceedings of the IEEE*, vol. 101, no. 1, pp. 99–114, Jan 2013.

[5] Y. Kameda, T. Koyama, Y. Mukaigawa, F. Yoshikawa, and Y. Ohta, "Free viewpoint browsing of live soccer games," in *IEEE International Conference on Multimedia and Expo (ICME)*, June 2004.

[6] K. Hayashi and H. Saito, "Synthesizing free-viewpoint images from multiple view videos in soccer stadium," in *International Conference on Computer Graphics, Imaging and Visualisation (CGIV)*, July 2006.

[7] Cornelius Malerczyk, Konrad Klein, and Torsten Wiebesiek, "3d reconstruction of sports events for digital tv," in *International Conference in Central Europe on Computer Graphics, Visualization, and Computer Vision (WSCG)*, 2003.

[8] C. Malerczyk, "3d-reconstruction of soccer scenes," in *3DTV Conference*, May 2007.

[9] L. Schnyder, O. Wang, and A. Smolic, "2d to 3d conversion of sports content using panoramas," in *IEEE International Conference on Image Processing (ICIP)*, Sept 2011.

[10] D. Steedly, Chris Pal, and R. Szeliski, "Efficiently registering video into panoramic mosaics," in *IEEE International Conference on Computer Vision (ICCV)*, Oct 2005.

[11] B. Ghanem, Z. Tianzhu, and N. Ahuja, "Robust video registration applied to field-sports video analysis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012.

[12] Peter Carr, Yaser Sheikh, and Iain Matthews, "Point-less calibration: Camera parameters from gradient-based alignment to edge images," *IEEE Workshop on Applications of Computer Vision (WACV)*, 2012.

[13] Flvio Szenberg, PauloCezarPinto Carvalho, and Marcelo Gattass, "Automatic camera calibration for image sequences of a football match," in *International Conference on Advances in Pattern Recognition (ICAPR)*, 2001.

[14] Dirk Farin, Susanne Krabbe, Wolfgang Effelsberg, Peter H.N. de With, and Peter H. N. De, "Robust camera calibration for sport videos using court models," in *SPIE Proceedings on Storage and Retrieval Methods and Applications for Multimedia*, 2004.

[15] N. Krahnstoever and P.R.S. Mendonca, "Bayesian autocalibration for surveillance," in *IEEE International Conference on Computer Vision (ICCV)*, Oct 2005.

[16] Graham Thomas, "Real-time camera tracking using sports pitch markings," *Journal of Real-Time Image Processing*, vol. 2, no. 2-3, pp. 117–132, 2007.

[17] J. Y Guillemaut, J. Kilner, and A. Hilton, "Robust graph-cut scene segmentation and reconstruction for free-viewpoint video of complex dynamic scenes," in *IEEE International Conference on Computer Vision (ICCV)*, Sept 2009.

[18] G. Bradski, "The opencv library," *Dr. Dobb's Journal of Software Tools*, 2000.

[19] Jan Regg, Oliver Wang, Aljoscha Smolic, and Markus Gross, "Ducttake: Spatiotemporal video compositing," *Eurographics*, 2013.

[20] A. Smolic and J.-R. Ohm, "Robust global motion estimation using a simplified M-estimator approach," in *International Conference on Image Processing (ICIP)*, 2000.

[21] J. Shi and C. Tomasi, "Good features to track," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 1994.

[22] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 1981.

[23] A. Smolic, T. Sikora, and J.-R. Ohm, "Long-term global motion estimation and its application for sprite coding, content description, and segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 8, pp. 1227–1242, Dec 1999.

[24] Martin A. Fischler and Robert C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[25] "Mastercam," http://people.inf.ethz.ch/smolica/MasterCam.htm, [Accessed on 2014.02.12].