http://dx.doi.org/10.1117/12.2027311

Feature-based automatic configuration of semi-stationary multi-camera components

Ann-Kristin Grosselfinger, David Münch, Wolfgang Hübner, and Michael Arens

Fraunhofer IOSB, Gutleuthausstraße 1, 76275 Ettlingen, Germany

ABSTRACT

Autonomously operating semi-stationary multi-camera components are the core modules of ad-hoc multi-view methods. On the one hand a situation recognition system needs overview of an entire scene, as given by a wide-angle camera, and on the other hand a close-up view from e.g. an active pan-tilt-zoom (PTZ) camera of interesting agents is required to further increase the information to e.g. identify those agents. To configure such a system we set the field of view (FOV) of the overview-camera in correspondence to the motor configuration of a PTZ camera. Images are captured from a uniformly moving PTZ camera until the entire field of view of the master camera is covered. Along the way, a lookup table (LUT) of motor coordinates of the PTZ camera and image coordinates in the master camera is generated. To match each pair of images, features (SIFT, SURF, ORB, STAR, FAST, MSER, BRISK, FREAK) are detected, selected by nearest neighbor distance ratio (NNDR), and matched. A homography is estimated to transform the PTZ image to the master image. With that information comprehensive LUTs are calculated via barycentric coordinates and stored for every pixel of the master image. In this paper the robustness, accuracy, and runtime are quantitatively evaluated for different features.

Keywords: Master-slave camera system, Video surveillance, Self calibration, Semi-stationary camera system, Weak calibration

1. INTRODUCTION

In a surveillance scenario an operator observing different places suffers from a high cognitive load. Thus, a system automatically analyzing the occurring situations assists the operator massively. Often there is the need to put the focus on an interesting region of the scene or only significant parts of the scene should be used for further treatment. Detailed parts of the scene could be used for example to recognize number plates as shown in Ref. 1 or for person identification as shown in Refs. 2 and 3. Also, a combination of different tasks can be achieved, e.g. tracking of persons with different cameras and getting detailed view of the person for identification where it is possible as shown in Ref. 4. A rough overview of an entire scene with a wide-angle camera is not enough: On the one hand a situation recognition system needs enough visual input to detect fine grained situations, and on the other hand when recognizing the occurrence of a car or the presence of a person a close-up view of that interesting agent is required to further refine the information or to identify this potentially interesting agent. To this an overview of the entire scene is needed as well as detailed information about interesting parts of the scene. Semi-stationary multi-camera components which are able to operate autonomously are the core modules of ad-hoc multi-view methods. This part can be done by an additional PTZ camera to capture high-resolution close-up views.

The Control of a multi camera system can be accomplished in different ways. According to Ref. 4 these are: Picture Domain Camera Control (PDCC), Scene Domain Camera Control (SDCC), Conceptual Level Camera Control (CLCC). In PDCC the slave camera is steered to center interesting parts of the master camera image in the slave camera view. In SDCC interesting parts are chosen via 3D world coordinates and used to choose the appropriate camera and control for detailed view.

Further author information:
Ann-Kristin Grosselfinger: E-mail: ann-kristin.grosselfinger@iosb.fraunhofer.de
David Münch: E-mail: david.muench@iosb.fraunhofer.de
Wolfgang Hübner: E-mail: wolfgang.huebner@iosb.fraunhofer.de
Michael Arens: E-mail: michael.arens@iosb.fraunhofer.de

No matter what purpose of control, in either case the camera system needs a configuration of one camera to the other to manage control in motorspace of the slave camera according to a chosen pixel in the master camera's image. This configuration can be done by *strong* or *weak* calibration of the camera system. In strong calibration all intrinsic and extrinsic parameters of the involved cameras are determined and used to get the relation between 2D coordinates of one camera and motor coordinates of the other camera via 3D world coordinates. The calibration of each camera can be done separately for each camera as described in Ref. 5 or the cameras can be calibrated as a stereo system as done in Ref. 6. In weak calibration the relation between two cameras is done directly on both images; no 3D world coordinates are determined. These methods mainly use lookup tables between pixel coordinates of one camera and corresponding motor coordinates of the other. Interpolation is used to receive all other values, e.g. trigonometric interpolation in Ref. 7, weighted interpolation function in Ref. 8 or thin plate spline interpolation in Ref. 9. Early systems like Ref. 10 use manual annotation to get the initial LUT. Other methods exploit scene characteristics such as road marking in Ref. 11 or vanishing points in Ref. 12. Feature based week calibration can be used if both cameras have a similar FOV. The used methods differ in the way they cover the master image with images from the slave camera. They either use one image from a zoomed out PTZ camera as done in Refs. 7 and 13 or build a mosaic image from moving a PTZ camera as shown in Refs. 8, 14, and 15 and use this mosaic image with stored corresponding motor coordinates to each image part. Since only feature points are needed, in Ref. 16 only features are stored instead of the mosaic image.

In this paper we address the connection of what can be seen in an overview-camera to the motor configuration of a PTZ camera. As the proposed system is semi-stationary the hardware architecture is designed to be independent of any wired connection and additionally the methods work online and reliable.

2. PROCESSING PIPELINE

Images are captured from a uniformly moving PTZ camera until the entire FOV of the master camera is covered. Along the way a LUT is generated using corresponding motor coordinates of the PTZ camera and image coordinates (u, v) in the master camera. To match each pair of images, features are detected in both images, selected by NNDR, and matched. A 2D-2D homography is adapted to theses matches and applied to transform the center of the image from the PTZ camera to the master image. After having collected enough data in the LUT two functions - one for pan, one for tilt - as function of master image coordinates are determined using linear regression and RANSAC. With these functions two comprehensive LUTs (pan and tilt) for every pixel in the master image are calculated and stored as motormaps. Alternatively barycentric coordinates are used to determine the motormaps.

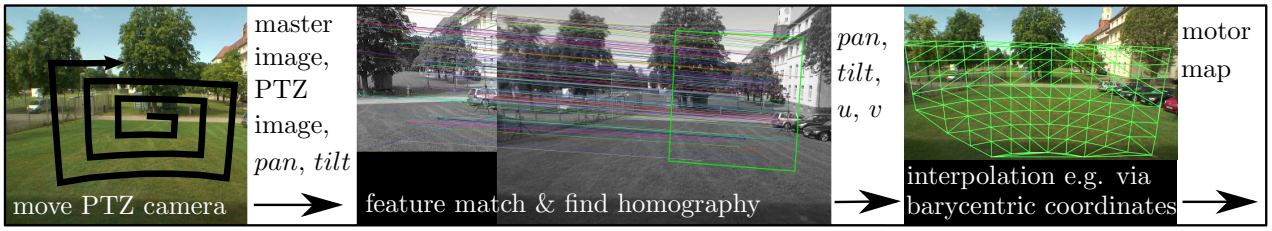

Figure 1. Processing pipeline: Images from the moving PTZ camera, related motor coordinates and the master image are captured at same time. Subsequently, the matching of the PTZ image into master image is processed: Keypoints are detected in both images, appropriate descriptor vectors are matched, and assorted with nearest neighbor distance ratios. With the obtained matches a homography is determined and applied to get the transformed center of PTZ image (u, v). All matches together build a LUT, which is used to interpolate a dense motormap, containing pan and tilt coordinates of the PTZ camera corresponding to each master image pixel.

To capture images from the PTZ camera to cover the whole FOV of the master camera we use a spiral move of the PTZ camera. Each image taken along this way with corresponding motor coordinates is used for one match with the master image. In this step features are used to determine the transformation of the PTZ image center in master image pixel coordinates (u, v). Together with the stored pan, tilt motor coordinates these (u, v) values build one entry in a lookup table (LUT). In the next step the LUT values are used to build a motormap, i.e. a dense LUT for each master image pixel. For the interpolation either barycentric coordinates or linear regression

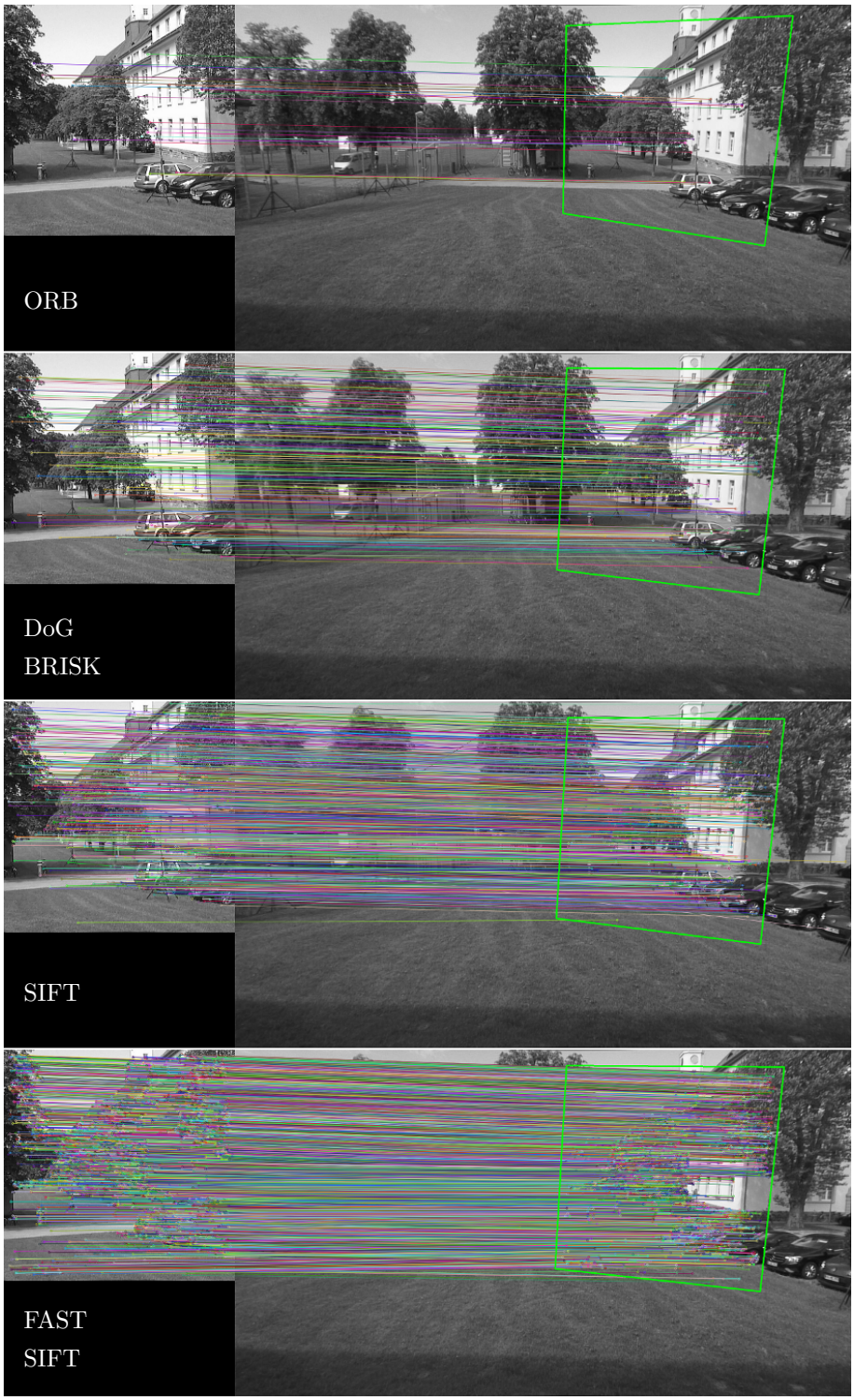

Figure 2. Qualitative matches from an experiment with zoom 100: ORB with 29 out of fixed size 500 matches fulfilling NNDR 0.7, DoG keypoint detector and BRISK descriptor extractor with 193 out of 1353 matches, SIFT with 805 out of 9815 matches and 3404 out of 15357 matches from FAST keypoint detector combined with SIFT descriptor extractor.

is used. To improve this initial motormap which lacks in areas with few structure an online refinement of this configuration could use objects passing the scene while the system is in use. These objects give more structure to the interesting parts of the image. In this paper we focus on the feature match step in the processing pipeline. Especially in terms of online refinement, where the images from PTZ camera are zoomed in and therefore harder to match with images from the zoomed out master camera. Hence we need a feature match method that performs good with various image scales. We compared different methods to find the keypoints and extract the feature descriptor. We evaluate them according to speed, robustness, and accuracy. In the reviewed step features in both images are detected, assorted with NNDR, and matched. From these matches a homography is determined if possible. For this the Random Sample Consensus (RANSAC) algorithm is used. Depending on the inlier outlier ratio and geometry of the found projection we accept or reject the result to use in LUT.

3. EVALUATION

In this paper we evaluated different features according to the robustness, accuracy, and runtime integrated in the presented master-slave camera system. The robustness and accuracy are quantitatively evaluated. A selected subset of applied features is shown in Figure 2. Features evaluated are STAR, FAST, and MSER keypoint detectors combined with BRISK and FREAK descriptor extractors, and SIFT, SURF, ORB for both. Ref. 17 gives a broad overview and evaluation of different image features.

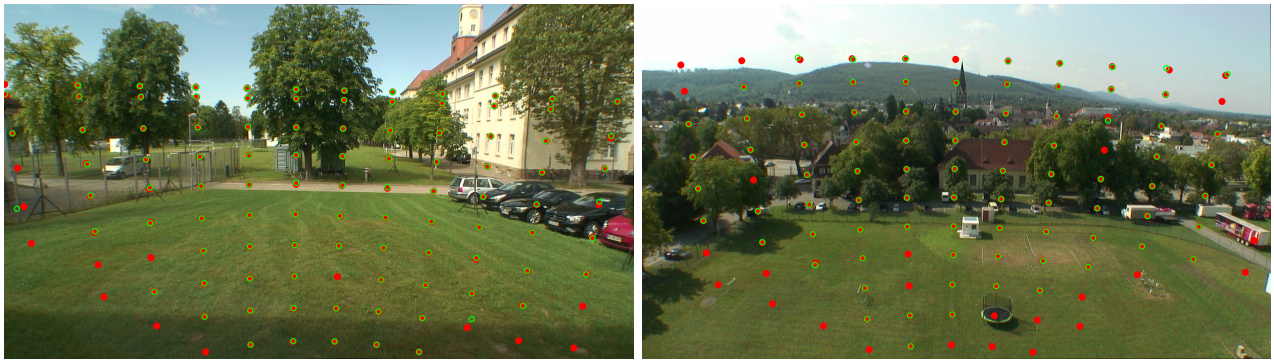

Figure 3. Ground truth (dots) from experiments with zoom 100 and zoom 500 compared with result from DoG keypoint detector combined with BRISK descriptor (circles). It can be seen that almost all matches can be determined and that the accuracy is high.

For feature detection the SIFT algorithm (Ref. 18) is a popular and powerful approach. The high accuracy of its results is paid with computational complexity. Due to that cost there is often a need for faster feature detector. A prominent derivate of SIFT is SURF (Ref. 19) which has a reduced complexity while remaining most of its accuracy. Results are shown in Figure 5. A subset of evaluated feature detectors and descriptors are shown. It can be seen that standard SURF compared to standard SIFT is several times faster but with an decreased matching performance.

The Maximally Stable Extremal Region (MSER) extractor introduced in Refs. 20 and 21 is an affine invariant region detector based on detecting boundaries around regions with extreme high or low intensity relative to all pixels around. The FAST corner detector introduced by Trajković and Headley (Ref. 22) is also based on intensity differences. To find interest points the structure of compact regions with similar brightness is compared to three types of image primitives representing 'nucleus within region', 'edge point', or 'corner point'. STAR is an OpenCV variation of CenSurE (Center Surround Extrema) feature detector introduced by K. Konolige (Ref. 23). The oriented BRIEF (ORB) is presented in Ref. 24. When comparing the results of the MSER, FAST, STAR, ORB in Figure 5 their performance compared to the computational cost is quite low.

Applying two further feature descriptors – the Binary Robust Invariant Scalable Keypoints (BRISK) from Ref. 25 and the Fast Retina Keypoint (FREAK) from Ref. 26 – increases the absolute performance and the performance vs. computation time. Except applying BRISK and freak descriptor to FAST which decreases the performance.

For ground truth information of LUT entries with PTZ motor coordinates and corresponding master image coordinates (u, v) the center of the PTZ camera image is matched into the corresponding pixel in the master camera image. Ground truth was generated in a combination of manually shrinking both images to the expected region and fitting a homography manually with the aid of results from all tested methods. Since this is a costly procedure we passed on generating ground truth where no considered method achieved to get a match result, see Figure 3: The dots correspond to the ground truth, and circles correspond to the result of DoG keypoint detector combined with the BRISK descriptor. It can be seen that both on the left image with zoom 100 and on the right image with zoom 500 the accuracy of the results is high as it is shown in Figure 4.

In general the experiments depicted in Figure 5 show, that the use of the BRISK descriptor applied on SURF and DoG leads to the best results with zoom 100 and 500. Due to the similar accuracy the best feature to be used can be chosen according to its runtime. In that case the best performance with shortest runtime holds for SURF with BRISK. Nevertheless when applying the features in a whole system there are other issues which have to be considered. Due to the fact that the PTZ camera is moving mechanically the assertion that a target position is reached takes its time: in our case it takes about 2.5 seconds. Thus, when applying the features in our system all features with less than 2.5 seconds can be equally used. Still, SURF with BRISK is the best choice.

A detailed quantitative evaluation on the accuracy of the features can see in Figure 4. On top there is an experiment with FAST keypoints and SIFT descriptor and on the bottom there are results from keypoints extracted with DoG and the BRISK descriptor. On the left there is the horizontal error and on the right the corresponding vertical error. The error is the difference between the correct position and the detected position in the master camera image corresponding to the PTZ camera image center. There are only a few outliers with slightly greater errors. As already described in the section 2 the interpolation in the subsequent pipeline step is able to filter these outliers in case of using linear regression with RANSAC. The figures depict an random distribution of errors through the whole image, therefor, errors can be easily compensated.

In summary the evaluation shows that under the assumption of a similar field of view of the cameras the mean angular accuracy of the whole processing pipeline is below two degrees in an outdoor scenario. Concerning the runtime of the whole pipeline a full calibration run with e.g. over 210 markers takes about 9 minutes.

4. CONCLUSION

In this paper we have evaluated different keypoint detectors and descriptors in an master-slave camera setup. The evaluation shows, that SIFT – the most computational expensive features – is outperformed by SURF and BRISK resulting in an approximately equal accuracy with a runtime advantage by the factor eight. Using BRISK enables the matching of PTZ image into master image up to an scale difference up to the factor 7 which allows the system to refine its calibration while operating zoomed in. Future work includes the extension to master-slave camera systems with several slaves and the integration in a situation analysis framework.

REFERENCES

[1] Mohanty, K. and Gellaboina, M., "A semi–automatic relative calibration of a fixed and PTZ camera pair for master–slave control," in [*Visual Information Processing (EUVIP), 2011 3rd European Workshop on*], 229 –234 (7 2011).

[2] Hampapur, A., Pankanti, S., Senior, A., Tian, Y.-L., Brown, L., and Bolle, R., "Face cataloger: multi-scale imaging for relating identity to location," in [*Proceedings. IEEE Conference on Advanced Video and Signal Based Surveillance, 2003.*], 13 – 20 (7 2003).

[3] Yi, R. D. X., Gao, J., and Antolovich, M., "Novel methods for high–resolution facial image capture using calibrated PTZ and static cameras," in [*Multimedia and Expo, 2008 IEEE International Conference on*], 45 –48 (4 2008).

[4] Bellotto, N., Benfold, B., Harland, H., Nagel, H.-H., Pirlo, N., Reid, I., Sommerlade, E., and Zhao, C., "Cognitive visual tracking and camera control," *Computer Vision and Image Understanding* **116**(3), 457 – 471 (2012). Special issue on Semantic Understanding of Human Behaviors in Image Sequences.

[5] Tsai, R., "A versatile camera calibration technique for high–accuracy 3D machine vision metrology using off–the–shelf TV cameras and lenses," *Robotics and Automation, IEEE Journal of* **3**, 323 –344 (8 1987).

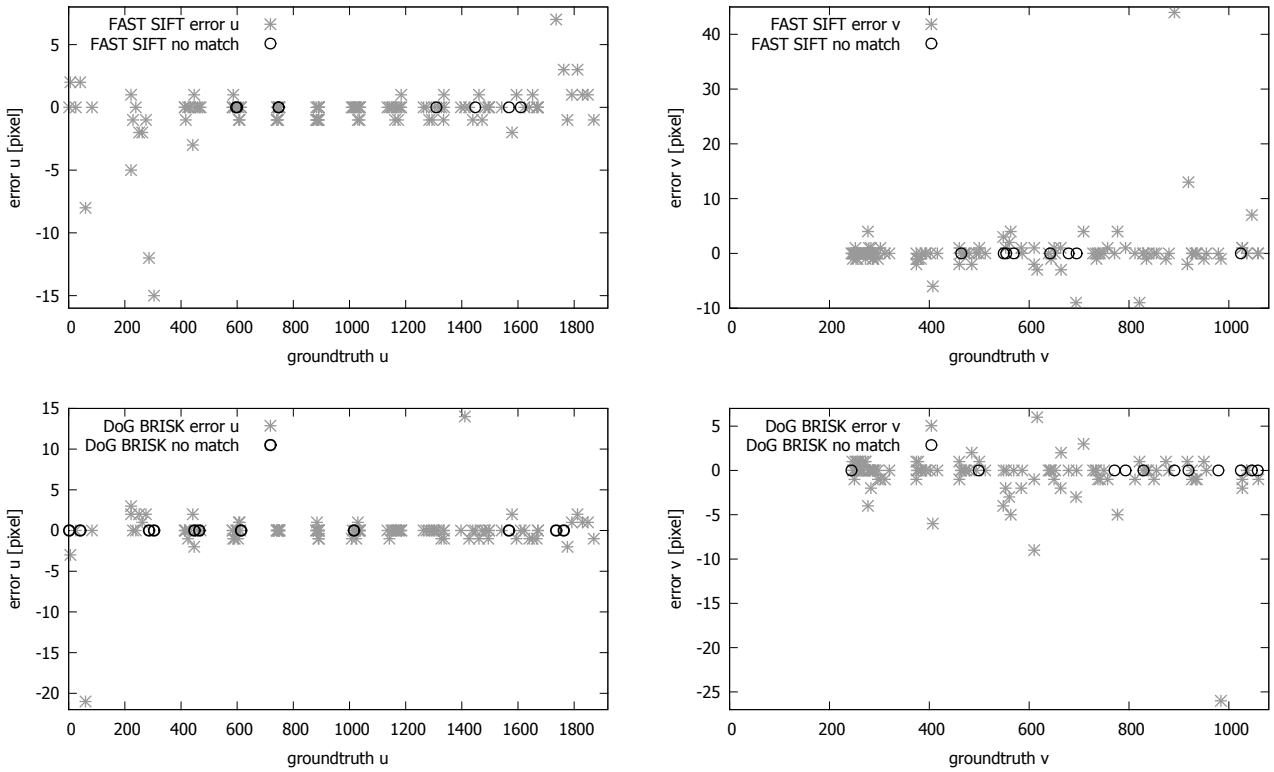

Figure 4. Quantitative error from a match using FAST SIFT and DoG BRISK in an experiment with zoom 100.

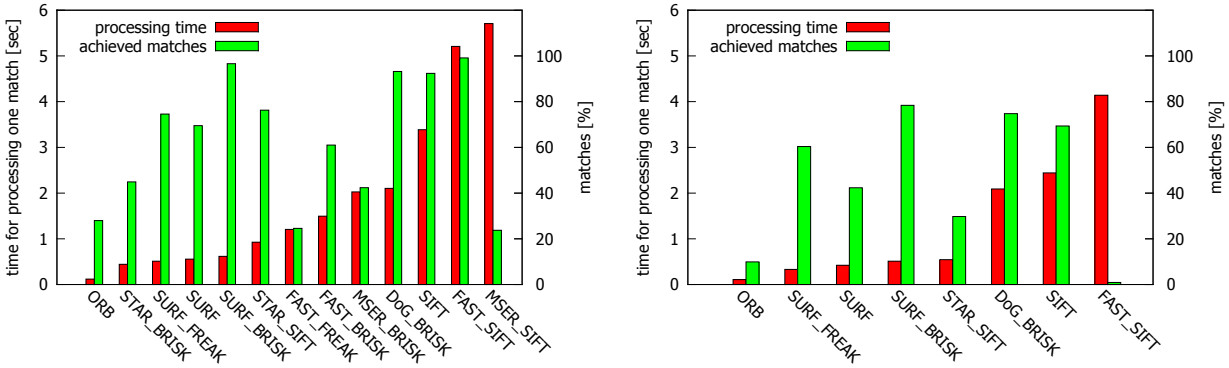

Figure 5. The left bars in the diagrams show for each method the average processing time for one image match. Only successful matches are used for this calculation. The right bars show the percentage of matches with accepted results compared to the number of all used input sets with images and motor coordinates. Experiment with zoom 100 on the left figure, zoom 500 on the right.

[6] Horaud, R., Knossow, D., and Michaelis, M., "Camera cooperation for achieving visual attention," *Machine Vision and Applications* **16**(6), 1–2 (2006).

[7] Liao, H.-C., Pan, M.-H., Hwang, H.-W., Chang, M.-C., and Po-Cheng, C., "An automatic calibration method based on feature point matching for the cooperation of wide–angle and pan–tilt–zoom cameras," *Information Technology And Control* **40**(1), 41–47 (2011).

[8] You, L., Song, L., and Wang, J., "Automatic Weak Calibration of Master–Slave Surveillance System Based on Mosaic Image," in [*Pattern Recognition (ICPR), 2010 20th International Conference on*], 1824 –1827 (8 2010).

[9] Badri, J., Tilmant, C., Lavest, J.-M., Pham, Q.-C., and Sayd, P., "Camera–to–Camera Mapping for Hybrid Pan–Tilt–Zoom Sensors Calibration," in [*Image Analysis*], Ersbøll, B. and Pedersen, K., eds., *Lecture Notes in Computer Science* **4522**, 132–141 (2007).

[10] Zhou, X., Collins, R. T., Kanade, T., and Metes, P., "A master–slave system to acquire biometric imagery of humans at distance," in [*First ACM SIGMM international workshop on Video surveillance*], *IWVS '03*, 113–120, ACM, New York, NY, USA (2003).

[11] Song, K.-T. and Tai, J.-C., "Dynamic Calibration of Pan–Tilt–Zoom Cameras for Traffic Monitoring," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* **36**, 1091 –1103 (10 2006).

[12] Khan, S. and Shah, M., "Consistent labeling of tracked objects in multiple cameras with overlapping fields of view," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **25**, 1355 – 1360 (10 2003).

[13] Low, Y.-Q., Lee, S.-W., Goi, B.-M., and Ng, M.-S., "A New SIFT–Based Camera Calibration Method for Hybrid Dual-Camera," in [*Informatics Engineering and Information Science*], Abd Manaf, A., Zeki, A., Zamani, M., Chuprat, S., and El-Qawasmeh, E., eds., *Communications in Computer and Information Science* **252**, 96–103 (2011).

[14] Del Bimbo, A., Dini, F., Lisanti, G., and Pernici, F., "Exploiting distinctive visual landmark maps in pan-tilt-zoom camera networks," *Computer Vision and Image Understanding* **114**(6), 611 – 623 (2010). Special Issue on Multi-Camera and Multi-Modal Sensor Fusion.

[15] Sinha, S. N. and Pollefeys, M., "Pantiltzoom camera calibration and high–resolution mosaic generation," *Computer Vision and Image Understanding* **103**(3), 170 – 183 (2006). Special issue on Omnidirectional Vision and Camera Networks.

[16] Wu, Z. and Radke, R., "Using scene features to improve wide–area video surveillance," in [*Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*], 50 –57 (6 2012).

[17] Mikolajczyk, K. and Schmid, C., "A performance evaluation of local descriptors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **27**, 1615 –1630 (10 2005).

[18] Lowe, D. G., "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision* **60**, 91–110 (Nov. 2004).

[19] Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L., "Speeded–Up Robust Features (SURF)," *Computer Vision and Image Understanding* **110**(3), 346 – 359 (2008). Similarity Matching in Computer Vision and Multimedia.

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

[21] Nistr, D. and Stewnius, H., "Linear Time Maximally Stable Extremal Regions," in [*Computer Vision ECCV 2008*], Forsyth, D., Torr, P., and Zisserman, A., eds., *Lecture Notes in Computer Science* **5303**, 183–196 (2008).

[22] Trajković, M. and Hedley, M., "Fast corner detection ," *Image and Vision Computing* **16**(2), 75 – 87 (1998).

[23] Agrawal, M., Konolige, K., and Blas, M., "CenSurE: Center Surround Extremas for Realtime Feature Detection and Matching," in [*Computer Vision ECCV 2008*], Forsyth, D., Torr, P., and Zisserman, A., eds., *Lecture Notes in Computer Science* **5305**, 102–115 (2008).

[24] Rublee, E., Rabaud, V., Konolige, K., and Bradski, G., "ORB: An efficient alternative to SIFT or SURF," in [*Computer Vision (ICCV), 2011 IEEE International Conference on*], 2564–2571 (2011).

[25] Leutenegger, S., Chli, M., and Siegwart, R., "BRISK: Binary Robust invariant scalable keypoints," in [*Computer Vision (ICCV), 2011 IEEE International Conference on*], 2548–2555 (2011).

[26] Alahi, A., Ortiz, R., and Vandergheynst, P., "FREAK: Fast Retina Keypoint," in [*Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*], 510–517 (2012).