Automatic Unconstrained Online Configuration of a Master-Slave Camera System

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

Fraunhofer IOSB, Gutleuthausstraße 1, 76275 Ettlingen, Germany
david.muench@iosb.fraunhofer.de

Abstract. Master-slave camera systems – consisting of a wide-angle master camera and an actively controllable pan-tilt-zoom camera – provide a large field of view, allowing monitoring the full situational context, as well as a narrow field of view, to capture sufficient details. Unconstrained calibration of such a system is a non-trivial task. In this paper a fully automatic and adaptive configuration method is proposed. It learns a motor map relating image coordinates from the master view to motor commands of the slave camera. First, a rough initial configuration is estimated by registering images from the slave camera onto the master view. In order to be operational in poorly textured environments, like hallways, the motor map is online refined by utilizing correspondences originating from moving objects. The accuracy is evaluated in different environments, as well as in the visual and the infrared spectrum. The overall accuracy is significantly improved by the online refinement.

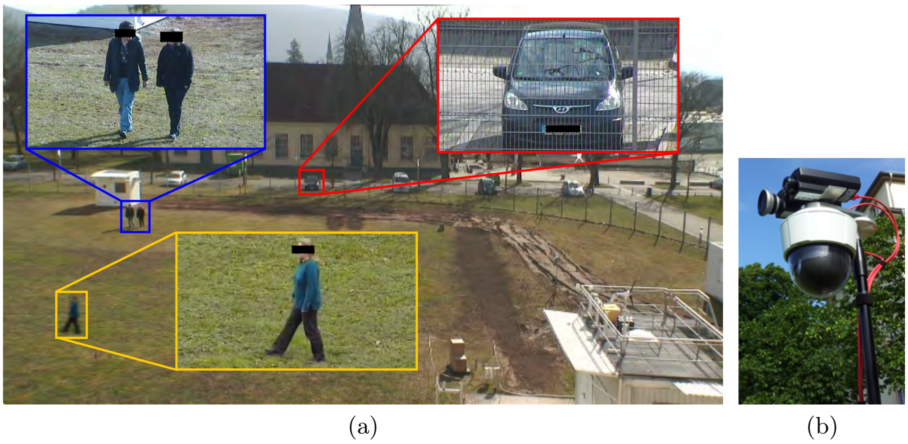
Keywords: Master-slave camera system. Self calibration. Semi-stationary camera system. Video surveillance. Weak calibration.

1 Introduction

In content-based video analysis the goal is a holistic understanding of an observed scene. We assume a surveillance scenario which means that any situation of an observed agent of interest – an object, a person, a vehicle, etc. – is automatically recognized and logged. If a situation is considered dangerous or unusual an automated warning should be raised in order to assist the human operator. Being visually supported by the surveillance system, the human operator is able to pay attention on specific predefined situations of interest.

Master-slave camera systems are a practical trade-off in order to overcome limitations of cameras with a single focus. They provide a large field of view (FOV), allowing to monitor the full situational context, as well as a narrow FOV, used to capture sufficient details of individual objects, at the same time. A typical setup consists of a wide-angle master camera with a fixed focus and an actively controllable pan-tilt-zoom (PTZ) camera.

Figure 1 depicts the camera setup and the envisioned application scenario in which detailed information about people, interactions, etc. are automatically



(a) (b)

Fig. 1. Application scenario and camera setup. The master camera captures a scene with automatically increasing the information in potentially interesting areas. (a) In this image a master camera detects several unspecified objects, while close-up views from the PTZ-camera allow, e.g. the identification of a car (red) or person (orange) or what the persons are doing (blue). (b) The master-slave camera system used in this work consists of an Axis P5534, Q1755, and Q1922.

captured by a detailed close-up view. Close-up views are required for more detailed detection and identification of objects. As mentioned in [9] there is a lower bound of the size of detecting a person in an image. Other demands for close-ups might stem from pose reconstruction [6] for action recognition of persons. Face detection for person identification can also be applied [4]. The system is intended to support a rule-based inference machine [20], which fuses all the mentioned methods above and performs semantic feedback to the active controllable cameras to even further specialize the situations of interest in a scene.

This article is structured as follows: Section 2 provides an overview of related work on multi-camera systems and their configuration. Our proposed system and the whole processing pipeline are presented in Section 3. Section 4 gives a comprehensive evaluation and explains the online refinement. Section 5 provides a conclusion. The contribution of this article is (a) a comprehensive vision system working under low constraints (b) in different spectral ranges and (c) the automatic online refinement followed by a comprehensive evaluation.

2 Related Work

In the literature active multi-camera systems which were mainly well-defined and calibrated are addressed in [12]. Here, we mainly address the problem of setting up multi-camera systems in unconstrained environments, without extrinsic calibration and without precise knowledge about the geometry of the cameras.

Multi-camera systems. There can be differentiated between different kinds of multi-camera systems with active components. In [4] several PTZ-cameras cooperate in a scene observed by one master camera. Several master cameras and several PTZ-cameras are investigated in [23]. In an indoor environment, such as a smart control room, there are dozens of cameras among them one with a fisheye objective and two PTZ-cameras [13]. Another possibility instead of a fix master camera is the use of only PTZ-cameras with one of them operating as master camera [8].

Motivation and organization of camera control. In addition to different hardware configurations, there is a clear distinction in terms of the motivation and organization of the purpose and control of a multi-camera system. One application of a multi-camera system is to track objects over several cooperative cameras [10]. Another application is to increase the information of certain objects, such as number plate recognition [24] or person identification [27]. In [4] both methods, multi-camera object tracking and the generation of close-up views are combined in one system.

According to Bellotto et. al. [4] the organization of multi-camera control can be divided in three parts: First, there is the *Picture Domain Camera Control*, which means that the control of the cameras is only based on low-level information from 2D images. Second, there is the *Scene Domain Camera Control* making use of 3D scene models etc. And finally, there is the *Conceptual Level Camera Control* using extracted higher-level information to control the cameras intelligently.

Configuration and calibration. In a multi-camera system the mapping from a point in one camera to the corresponding point in the other camera is essential. For that the calibration of the camera system is needed. Calibration methods can be divided into weak and strong calibration methods.

Strong calibration means that both internal and external parameters of all cameras have to be determined. This allows determining the correspondences of a point in a 2D image to another point in a 2D image via 3D world coordinates. The calibration can be done for every camera [25], or as a pair-wise stereo setup [11]. In outdoor scenarios often georeferences are added [23]. Compensating the deficiencies of low-budget cameras, Jain et. al. [14] use an extended set of calibration equations, in contrast to [22] where a simplified camera model is used. Other methods need manual assistance, such as hand-drawn gridlines, or light-points [7].

Weak calibration avoids the mapping from one image to the other via 3D world-coordinates. Instead a lookup-table (LUT) with the corresponding 2D image coordinates and the PTZ motor coordinates is generated. Early approaches (e.g. [28]) make use of manually selected correspondences, such as annotated persons [18]. The interpolation of sparse LUTs can be performed e.g. geometrically [16] or with splines [1]. Other methods use specific properties of certain scenes as e.g. lane markings or vanishing points.

Methods based on local image features, e.g. [16], avoid scene dependent specialties and are independent from what is seen in an image. As the FOV of the PTZ-camera is normally only a small fraction of the master camera's FOV, several approaches generate a mosaic image of images from the PTZ-camera while storing the PTZ coordinates. Wu and Radke [26] do not save the mosaic image, instead only the features are stored.

Here we propose a method for automatically calibrating the master-slave camera system, see Figure 1. The proposed method combines several advantages of the above mentioned approaches into one configuration method. The method performs a weak calibration under the constraint that both cameras have a similar view point. In general no further assumptions about the camera geometry are made. Instead of mosaicking, a sparse initial LUT is interpolated using linear regression in order to calculate the mapping from master image coordinates to the motor space of the PTZ camera. As it is not possible to find correspondences in poorly textured regions, therefore, an online refinement using temporarily available objects during operation in these areas to incrementally improve the LUT is proposed. Online refinement turns out to be an essential processing step in most real world scenarios.

3 Methods

In the following section the whole processing pipeline of our proposed master-slave camera system is described in detail.

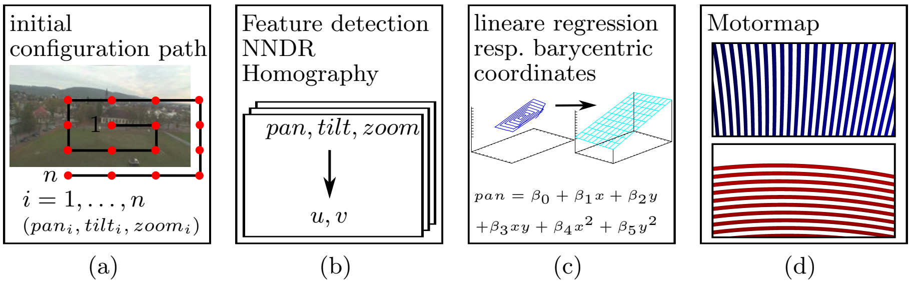

initial configuration path
1
n
$i = 1, \dots, n$
$(pan_i, tilt_i, zoom_i)$
Feature detection NNDR Homography
$pan, tilt, zoom$
u, v
lineare regression resp. barycentric coordinates
$pan = \beta_0 + \beta_1 x + \beta_2 y$
$+\beta_3 xy + \beta_4 x^2 + \beta_5 y^2$
Motormap
(a)
(b)
(c)
(d)

Fig. 2. The overall processing pipeline, used to automatically determine the motor map. (a) A rough initial set of correspondences is determined. (b) The process samples a mapping between the image coordinates (u, v) of the master camera and the motor space of the PTZ camera. (c) A linear regression resp. Barycentric coordinates are used to generate a dense LUT. (d) Visualization of the learned dense motormap (rounded to integers and alternately colored).

3.1 Determining Correspondences Between Uncalibrated Cameras

Local image features sampled around view invariant interest points have been proven to be an efficient tool in determining corresspondences without additional constraints. For a comprehensive review on local features see [15, 17]; particularly due to their efficient runtime complexity, we decided to use SURF features [2], although the proposed method is not limited to a specific feature type. In order to compensate stronger deviations in view point, our method can also be used with affine invariant features [17, 19], or features adapted to the source image [21]. Despite estimating the internal camera parameters no further preprocessing, image correction, photometric correction, or spectral adaption is required.

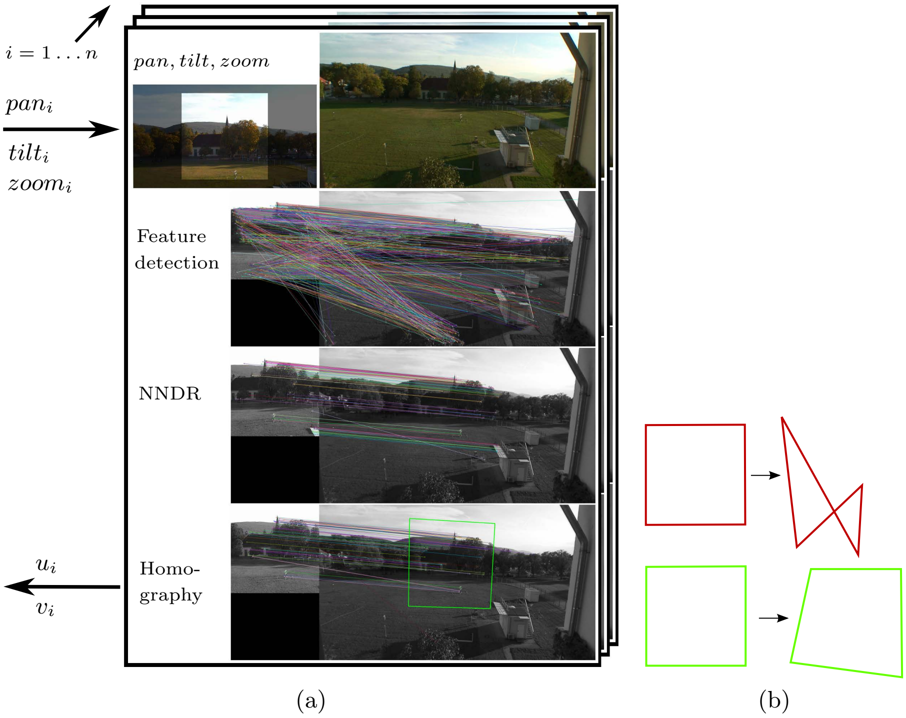


Fig. 3. (a) Visualization of detecting features in master and slave image, matching corresponding features, and estimating the homography. (b) Verifying probably unuseful homographies (upper) and only allowing probably good homographies (lower).

The initialization and matching of the processing pipeline are shown in Figure 2. To start the configuration a rough initial set of correspondences is determined. Thus, the PTZ-camera starts moving in a spiral-like pattern (a) in order to cover its full viewing range. Subsequently, in each correspondences step the features are extracted, matched, and a homography is estimated (b). For feature matching we avoid using fixed thresholds or simple Nearest Neighbor Search. Instead Nearest Neighbor Distance Ratio is used, as it performs best in our setups [17]. Finally, the homography between the two views is computed from the correspondences, see Figure 3 (a).

To reject ‘wrong’ homographies, see Figure 3 (b), we normalize the transformation matrix H_{sm} according to $detH'_{sm} = 1$: $H'_{sm} = H_{sm} * \frac{sgn(detH_{sm})}{\sqrt[3]{|detH_{sm}|}}$, see also [3]. $H'_{sm} \in SL_3(\mathbb{R})$ thus, we can apply thresholds to avoid outliers, in our case: $-5{,}0 < h'_{11} < 5{,}0, \quad -5{,}0 < h'_{22} < 5{,}0, \quad 0{,}55 < h'_{33} < 2{,}5$.

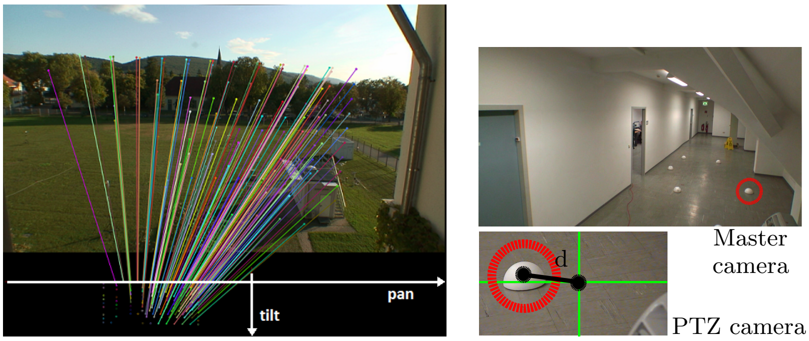
3.2 Mapping Image Coordinates to Motor Space

The adaptation process described so far generates a sparse mapping between coordinates in the master image and the motor space of the PTZ camera, see Figure 4. In order to be applicable a dense mapping has to be estimated from the sparse LUT. In general a bijective function is desired, but due to inaccuracies in the motor control of the PTZ-camera, the mapping is only injective.

Minimizing the error over different types of polynoms we get

$$\text{pan} = a_{p0} + a_{p1} * x + a_{p2} * y + a_{p3} * x * y + a_{p4} * x^2 + a_{p5} * y^2.$$

With $(p_1 \cdots p_n)^T = \boldsymbol{X a_p} + \boldsymbol{\epsilon_p}$ and $\boldsymbol{X}$ according to the chosen polynom, the least square estimate is $\boldsymbol{\hat{a}_p} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(p_1 \cdots p_n)^T = \boldsymbol{X}^+(p_1 \cdots p_n)^T$. The same procedure for tilt. To assure reliable results RANSAC is applied in this step. A prototypical visualization of the initial and dense LUT is shown in Figure 2 (c).



(a)

(b)

Fig. 4. (a) Registering the PTZ view onto the master image. This includes the sparse mapping. (zoom fixed). (b) Evaluating the error d (the distance of the ground truth PTZ motor coordinates and the actual PTZ coordinates) of the learned master-slave configuration.

3.3 Person Detection and Close-up Image Acquisition

After having learned a dense LUT in the previous section the master-slave camera system is able to communicate between the different cameras. We can differentiate the image processing into the master-camera’s part and the slave part.

An elaborated image processing loop starts with background subtraction and blob detection to identify potentially interesting situations.

Having gathered a close-up image with the slave camera, further fine-grained image processing methods are applied, such as person identification, face detection, and object detection. In addition to it human action recognition and high-level situation recognition are applied, too. We will not further detail those methods, as we see them as generic building blocks for the work presented here.

4 Evaluation

As the proposed system operates in a closed control loop, no offline data can be used for evaluation. Therefore, we quantitatively evaluated the performance of the system in different application domains. To evaluate the accuracy more than hundreds of markers were placed all over the FOV of the master camera, see Figure 4 (b). For every marker location the PTZ is moved to a position where the marker is centered in the PTZ's view. This procedure is done manually, in order to achieve ground truth data. Next, the procedure is repeated, using the automated PTZ control. The accuracy is measured in terms of the angular deviation in the position of the PTZ camera.

Robustness with respect to the application domain. We evaluated the proposed system in different domains, see Figure 5. The comprehensive evaluation reveals the specific challenges in each domain.

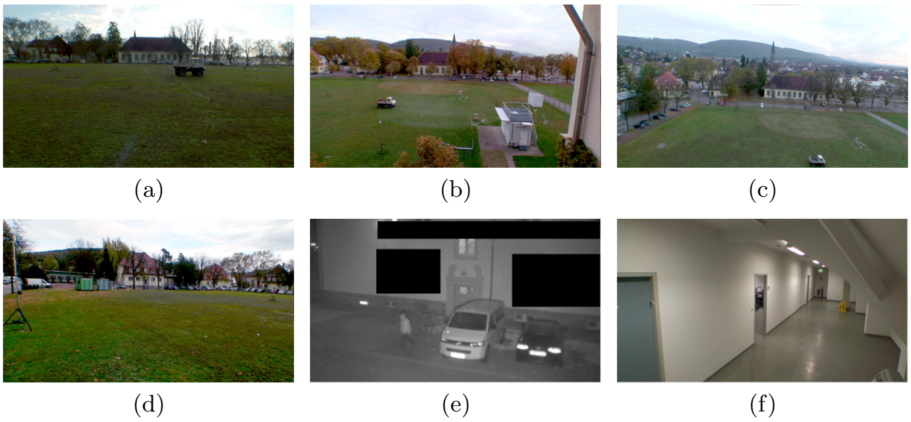

(a) (b) (c)
(d) (e) (f)

Fig. 5. The proposed system was evaluated in different domains. In a human-like surveillance scenario (a,d), in an wide-scene surveillance scenario (b,c), in a perimeter protection scenario at night (e), and indoors (f).

Figure 6 summarizes results, measured in exemplary task domains, depicted in Figure 5. Over hundred markers are placed in the FOV of the master camera.

For each marker the accuracy of the dense mapping is evaluated, see Figure 4 (b). In Figure 6 (a) resp. (b) the angular errors of pan resp. tilt of the PTZ camera are visualized (star). The x-axis are the pixels in horizontal (a) resp. vertical (b) direction. Thus, in the lower left of the FOV the error is larger than elsewhere. Comparing to Figure 4 (a), no correspondences could be found in that area of the scene.

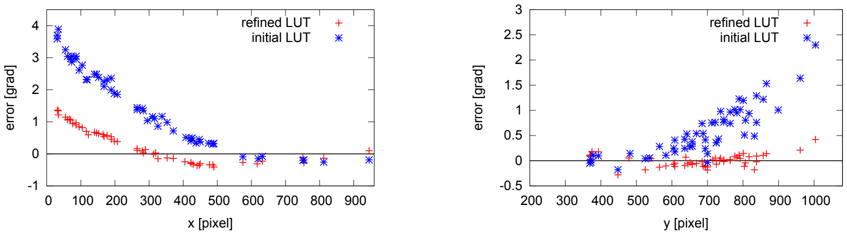


Fig. 6. Evaluation of the accuracy of pan (left) and tilt (right) in scenario Figure 5 (b) with the initial learned LUT (star) and the automatically online refined LUT (plus).

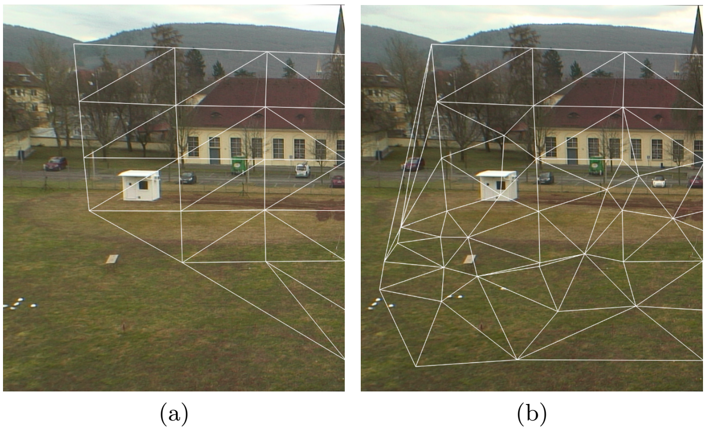
Incremental online refinement. Having evaluated the proposed system in the above domains reveals a weak point: In poorly texture regions it is not possible to establish sufficient corresspondeces. Therefore, the dense mapping has to be interpolated over large regions, resulting in an increasingly inaccurate PTZ control. That is a main drawback in the proposed system, as in that case, e.g., it is not applicable in the domains in Figure 5.

In order to overcome this limitation, we use the sparsely sampled FOV as an initial estimate and refine it using correspondences originating from moving objects. In typical scenes these are persons walking around, bikes and cars moving around. The idea is to use the temporal occurrence of persons or objects in low-textured areas to gather further correspondences for the sparse LUT and to refine the dense LUT incrementally and online.

In Figure 7 (a) a triangulation of the initial LUT is shown. In (b) it is extended by additional values gathered during a short time of operation. In Figure 6 the automatically refined LUT is evaluated again (plus). As a result, the error could be decreased by a factor of 5.

5 Conclusion

In this paper we have shown a semi-stationary master-slave camera system which is capable of self-configuration under non-cooperative low textured conditions. The effectiveness of the proposed system has been shown in different scenarios over time. Increasing the amount of cameras should include a georeferencing resulting in a global cover map. Further work include the extension of modalities e.g. [5] and the integration into a situation understanding framework [20], including high-level information inference as semantic feedback for lower level processes, c.f. first work on high-level semantic feedback is presented in [4].



(a)

(b)

Fig. 7. Triangulation of the initial learned LUT (a); refined by additonally online found correspondences (b). The triangulation is used alternatively for interpolation with Barycentric coordinates instead of linear regresssion.

References

1. Badri, J., Tilmant, C., Lavest, J.M., Pham, Q.C., Sayd, P.: Camera-to-Camera Mapping for Hybrid Pan-Tilt-Zoom Sensors Calibration. In: Ersbøll, B., Pedersen, K. (eds.) Image Analysis, LNCS, vol. 4522, pp. 132–141 (2007)
2. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-Up Robust Features (SURF). Computer Vision and Image Understanding 110(3), 346 – 359 (2008)
3. Begelfor, E., Werman, M.: How to Put Probabilities on Homographies. Pattern Analysis and Machine Intelligence, Transactions on 27(10), 1666 –1670 (2005)
4. Bellotto, N., Benfold, B., Harland, H., Nagel, H.H., Pirlo, N., Reid, I., Sommerlade, E., Zhao, C.: Cognitive Visual Tracking and Camera Control. Computer Vision Image Understanding 116(3), 457 – 471 (2012)
5. Bodensteiner, C., Hebel, M., Arens, M.: Accurate Single Image Multi-Modal Camera Pose Estimation. In: ECCV Workshop on Reconstruction and Modeling of Large-Scale 3D Virtual Environments (2010)
6. Brauer, J., Hübner, W., Arens, M.: Generative 2D and 3D Human Pose Estimation with Vote Distributions. In: Proc. of 8th International Symposium on Visual Computing. Rethymnon, Crete, Greece (2012)
7. Davis, J., Chen, X.: Calibrating Pan-Tilt Cameras in Wide-Area Surveillance Networks. In: Proc. 9th IEEE Int. Conf. on Computer Vision. pp. 144 –149 (2003)
8. Del Bimbo, A., Dini, F., Lisanti, G., Pernici, F.: Exploiting Distinctive Visual Landmark Maps in Pan-Tilt-Zoom Camera Networks. Computer Vision Image Understanding 114(6), 611 – 623 (2010)
9. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian Detection: An Evaluation of the State of the Art. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34(4), 743 –761 (2012)
10. Everts, I., Sebe, N., Jones, G.: Cooperative Object Tracking with Multiple PTZ Cameras. In: Image Analysis and Processing. 14th Int. Conf. on. pp. 323 –330 (2007)

11. Horaud, R., Knossow, D., Michaelis, M.: Camera Cooperation for Achieving Visual Attention. Machine Vision and Applications 16(6), 1–2 (2006)
12. Hu, W., Tan, T., Wang, L., Maybank, S.: A Survey on Visual Surveillance of Object Motion and Behaviors. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 34(3), 334–352 (2004)
13. IJsselmuiden, J., Stiefelhagen, R.: Towards High-Level Human Activity Recognition through Computer Vision and Temporal Logic. In: Proceedings of the 33rd Annual German Conference on Advances in Artificial Intelligence (2010)
14. Jain, A., Kopell, D., Kakligian, K., Wang, Y.F.: Using Stationary-Dynamic Camera Assemblies for Wide-area Video Surveillance and Selective Attention. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. vol. 1, pp. 537 – 544 (2006)
15. Li, J., Allinson, N.M.: A Comprehensive Review of Current Local Features for Computer Vision. Neurocomputing 71, 1771 – 1787 (2008)
16. Liao, H.C., Pan, M.H., Hwang, H.W., Chang, M.C., 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)
17. Mikolajczyk, K., Schmid, C.: A Performance Evaluation of Local Descriptors. Pattern Analysis and Machine Intelligence, Transactions on 27(10), 1615–1630 (2005)
18. Mohanty, K., Gellaboina, M.: A Semi-Automatic Relative Calibration of a Fixed and PTZ Camera Pair for Master-Slave Control. In: Visual Information Processing, 3rd European Workshop on. pp. 229 –234 (2011)
19. Morel, J.M., Yu, G.: ASIFT: A New Framework For Fully Affine Invariant Image Comparison. SIAM Journal on Imaging Sciences 2(2), 438–469 (2009)
20. Münch, D., Michaelsen, E., Arens, M.: Supporting Fuzzy Metric Temporal Logic Based Situation Recognition by Mean Shift Clustering. In: Glimm, B., Krueger, A. (eds.) KI 2012: Advances in Artificial Intelligence. vol. 7526, pp. 233–236 (2012)
21. Ozuysal, M., Calonder, M., Lepetit, V., Fua, P.: Fast Keypoint Recognition Using Random Ferns. Pattern Analysis and Machine Intelligence, Transactions on 32(3), 448–461 (2010)
22. Sinha, S.N., Pollefeys, M.: Pan-Tilt-Zoom 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
23. Szwoch, G., Dalka, P., Ciarkowski, A., Szczuko, P., Czyzewski, A.: Visual Object Tracking System Employing Fixed and PTZ Cameras. Intelligent Decision Technologies 5(2), 177–188 (2011)
24. Tian, Y.l., Brown, L., Hampapur, A., Lu, M., Senior, A., Shu, C.f.: IBM Smart Surveillance System (S3): Event Based Video Surveillance System With an Open and Extensible Framework. Machine Vision and App. 19(5-6), 315–327 (2008)
25. 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(4), 323 –344 (1987)
26. Wu, Z., Radke, R.: Keeping a Pan-Tilt-Zoom Camera Calibrated. Pattern Analysis and Machine Intelligence, Transactions on 99, 1 (2012)
27. Yi, R.D.X., Gao, J., 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. pp. 45 –48 (2008)
28. Zhou, X., Collins, R.T., Kanade, T., Metes, P.: A Master-Slave System to Acquire Biometric Imagery of Humans at Distance. In: First ACM SIGMM International Workshop on Video Surveillance. pp. 113–120. IWVS '03 (2003)