Automated Multi-Camera Surveillance for the Prevention and Investigation of Bank Robberies in Austria: A Case Study

S. Zambanini, P. Blauensteiner, M. Kampel

Vienna University of Technology - Pattern Recognition and Image Processing Group
Favoritenstr. 9/1832, 1040 Vienna, Austria
{zamba,blau,kampel}@prip.tuwien.ac.at

Keywords: Visual surveillance, camera mapping, case study.

Abstract

In this paper we present the recently granted *tripleB-ID* project that aims to detect suspicious actions and criminal offences in bank foyers. Based on the situation in Austria and on the actual needs of our partner bank, we deduce scenarios of interest and analyse the technical challenges. As a preliminary work, we furthermore present a fully automated method for the mapping between a static and a dynamic camera and show how this can be used to easily and fast set up a master-slave camera system.

1 Introduction

CCTV cameras in bank foyers are a common view. Up to now most of the surveillance is accomplished by recording and monitoring systems based on analogous camera technology. In case of an incident, the security staff has to browse through the recorded material manually in order to get video material of the crime scene.

It is well known that robbers are investigating the bank they intend to rob in advance. Therefore it would make sense to sift through the recorded videos in order to look for persons behaving suspiciously. However, the huge amount of data renders such a search practically impossible. State-of-the-art digital video recorders (DVRs) support this cumbersome task by providing elementary functionality like basic motion detection and camera tamper detection (i.e. the system detects if a camera is vandalised). However, if the scene is explored by the criminals during the regular business hours, these features are of little help.

This paper presents the recently granted *tripleB-ID* project which aims to improve this situation by a smart system able to recognise abnormal behaviour and set the appropriate actions (e.g. informing the responsible members of staff but also automatically operating a pan-tilt-zoom (PTZ) camera in order to get higher resolution images of a suspicious activity). Furthermore, the event information will be archived, so that in case of an incident the screening of the video data can be supported.

The use of a dynamic PTZ camera can support the task of surveilling a bank foyer since the optical zoom makes it possible to yield higher quality images (e.g. face images) or to get a better view of suspicious activity. For such a scenario, it is necessary to know the relation between the static and the dynamic camera. As a starting point, we review existing methods and present a fully automatic mapping algorithm capable to construct this relationship for collocated cameras.

The remainder of this paper is organised as follows: Section 2 analyses the current situation in Austria and gives an overview of the project and its main objectives. In Section 3 we present an approach for a fully automatic mapping between a static and a dynamic camera as preliminary work. Section 4 concludes this paper.

2 Project overview

In order to get an overview of the concrete requirements for crime prevention and detection, the current situation was analysed with the project partners, namely the Erste Bank group, the Department of Crime Prevention of the Austrian Federal Criminal Police Office (*Bundeskriminalamt*) and the Institute for Advanced Studies (*IHS*).

2.1 The situation in Austria

Austria, especially the Austrian capital Vienna with its 1.6 million inhabitants, has become one of the hot spots of bank robbery in Europe. In the last decade, the number of robberies has increased by more then 100%, whereas in comparable cities like Hamburg and Berlin the number of incidences decreased. In the year 2007, the Viennese police documented 77 bank robberies [1], whereas only 20 Berlin and 12 Hamburg banks where robbed in the same period [1].

CCTV cameras and DVRs are part of the basic equipment in Austrian bank branches. However, due to the Austrian privacy laws, the bank is required to delete any data within a certain amount of time unless it contains the recording of suspicious activity or criminal offences.

Interviews with security experts of our partner bank additionally revealed the following information, which they gath-

[1] http://www.kripo-online.at/portal/0805-bankraub.asp

ered partly by internal statistics, partly by interviewing convicted felons:

- In over 95% of the robberies, a silent alarm was raised by the staff.
- The average robbery lasts not more than one minute.
- Virtually all robbers explored the bank branch they robbed beforehand.
- A majority of robbers stated that they would have chosen another branch if a member of staff had addressed them in person since they feared to be recognised more easily.
- The vast majority (94%) of the robbers are not detered by the presence of CCTV cameras. The publication of low quality images in newspapers even encouraged some of the robbers to commit the felony.

Based on these facts, the main focus of the *tripleB-ID* project was set on preventing bank robberies. However, also other crimes committed in bank foyers will be investigated during the course of the project. This includes detection of manipulation of the foyer's cash machines as well as the detection of people spying out other people's PIN numbers at these machines.

2.2 Main Objectives

The aforementioned provides the basis for the projects main goals: to help preventing, to detect and to support the investigation of crimes committed in bank foyers, where the special focus of *tripleB-ID* lies on the prevention and on the support of investigation.

Behaviours of interest Together with the project partners, several scenarios were defined as suspicious. These scenarios include amongst others:

- wandering around in the foyer without using a machine (e.g. the ATM) or contacting a member of staff over an extended period of time.
- operating a machine over an unusual long period of time.
- more than one person operating a machine at the same time.

In addition to these suspicious activities, we are interested in detecting aggressive behaviour. Beside these well defined scenarios, we also intend to detect unusual behaviour, i.e. behaviour that cannot be described by the data previously labelled as normal.

The detection of suspicious behaviour may have several implications, such as informing a member of staff to judge the situation and take appropriate action or labelling the video data as suspicious. Video sequences labelled as suspicious can subsequently be stored for a longer period of time.

Forensics In case a crime cannot be prevented, the system should be able to provide information usable by the police, such as a detailed high-quality view of the face [17] and an accurate estimation of the size of a suspect.

In order to collect higher quality facial images, standard face detection as proposed by Viola and Jones [23] is performed on the static camera. The dynamic camera is subsequently zoomed to the detected area and collects facial images. The face detection is coupled to a pedestrian detector [20] to reduce the number of false positives. The higher resolution facial images are then connected to the trajectory of the person (see Figure 1).

The height of a person is estimated by applying visual metrology [7]. The height measurement of a person is updated and consolidated during the tracking. Furthermore, an interface for performing manual measurements is provided as well.

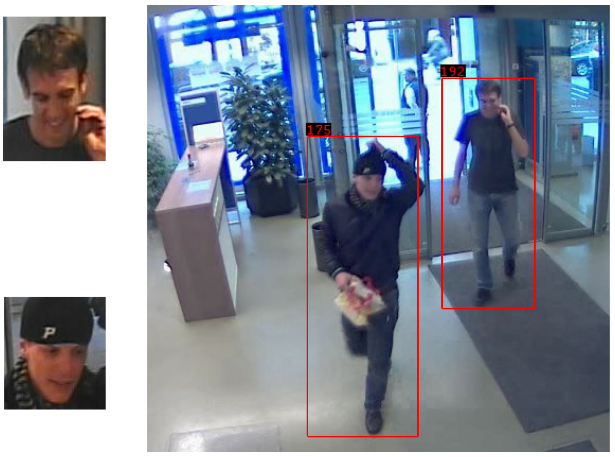

Figure 1. Detected persons with the estimated height and the higher quality facial images. The facial images were recorded by the dynamic camera.

2.3 Technical Challenges

A computer vision system addressing the aforementioned high level objectives faces several challenges.

Scalable multi-camera surveillance The data of several statically mounted cameras, with partially overlapping as well as non-overlapping views, has to be processed for the higher level algorithms. Amongst other topics, this task includes: Background maintenance and motion detection [4], pedestrian detection [11], single and multi-camera tracking [25] and data fusion in order to get a single representation of the current scene.

Robust vision algorithms Visual surveillance in a bank foyer is a 24/7 application and therefore it is inevitable to use robust algorithms. This requirement has a severe impact on the selection of algorithms. For example, the use of online learning algorithms is undeniable advantageous since the amount of training data is reduced and the system is able to adapt to changes over the time. However, if not taken care, classifiers

may drift over time due to wrong updates and end up in an unreliable state [11]. Classifier grids [20] and conservative learning algorithms [19] have shown promising results in this area.

Action recognition Trajectory analysis is a simple way to detect events of interest and for some of the depicted scenarios this may be sufficient (e.g. for the detection of people wandering around). However, for other scenarios such as the detection of aggressive behaviour (or unusual behaviour in general) the analysis of trajectories is of little help. Here more sophisticated models for action recognition - a very active research area in computer vision [22] - have to be applied. A promising research direction in this field are spatio-temporal interest points [15] and their use in a bag-of-words scheme [24].

2.4 Setup

In order to create a real-world testbed for bank surveillance, six static cameras and a dynamic PTZ camera were installed in a bank branch in Vienna, providing overlapping views of an area of approximately 10×10 m. The dynamic camera has a 35x optical zoom and is collocated with a static camera directed towards the bank's entrance area. These two cameras are intended to be used in a master-slave scenario [3, 9] to acquire high-resolution facial images of people entering the bank (see Section 3).

The large overlap in the field-of-views of the cameras is essential to allow for a robust long-term tracking of people in the presence of occlusions. A major issue in multi-camera tracking is to know the relationship between the cameras to obtain geometric information like object position [6] and camera fields-of-view (FOVs) [14] in world coordinates. Therefore, we recovered the homographies between every single camera and the ground plane by laying a calibration pattern on the floor and manually matching points between camera views. As a result, from every camera view image points assumed to lie on the ground plane can be mapped to a common world coordinate system.

Figure 2 shows the views of two cameras installed in the bank branch. Figure 2(a) shows the bank's entrance view of the static camera collocated with the dynamic camera. The yellow grid represents the established world coordinate system on the ground plane whereas one square corresponds to 0.5×0.5 m. In Figure 2(b) the view of a second camera, showing a large part of the overall bank branch area, can be seen. Here additionally the FOVs of two other cameras are visualized (yellow and red region, respectively) which were also computed from the estimated homographies.

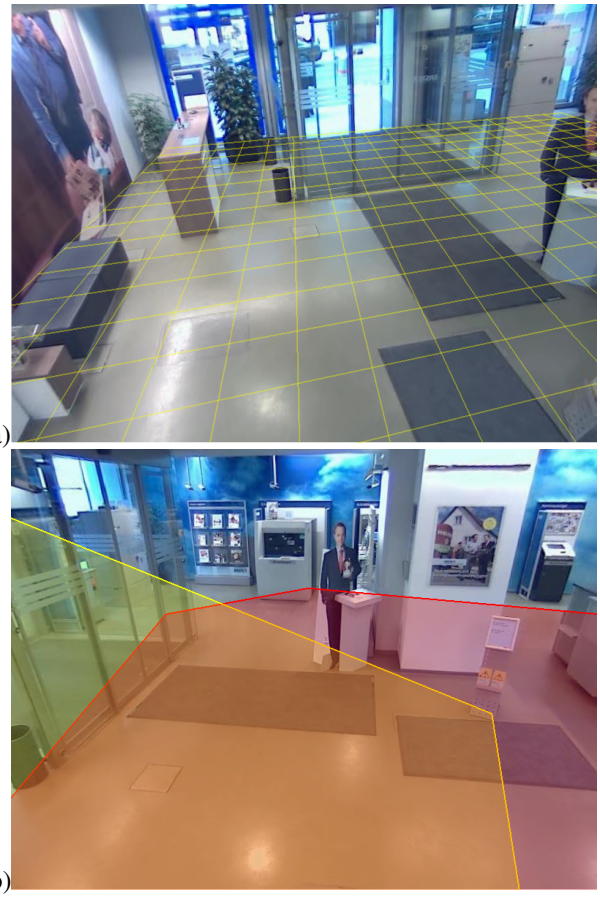

Figure 2. View of (a) the bank branch's entrance with superimposed world coordinate system, and (b) the overall branch with superimposed FOVs of two other installed cameras.

3 Automatic Mapping Between a Static and Dynamic Camera

For a master-slave system composed of a static and dynamic camera as described above, the relation between the two camera views has to be known. Thus, for each point in the static camera's frame the pan and tilt angles to center the PTZ camera's frame on this point are established. To solve this problem, in the past basically two groups of approaches have been proposed. The first one derives the geometric relation between the two cameras through a complete camera calibration, i.e. the computation of the intrinsic and extrinsic camera parameters [8]. Approaches of this kind were presented by Horaud et al. [12] and Jain et al. [13] whereas the former uses a stereo calibration and the latter calibrates the cameras separately. However, although by these methods an accurate model for the camera relation is obtained, a drawback of camera calibration is the need for a skilled person handling the calibration patterns and marks. The second group of approaches relies on the computation of a look-up-table (LUT) for the image points. The LUT stores for every image point in the static camera's frame the pan and tilt parameters such that this point is in the center of the dynamic camera's frame. During the creation of the LUT, only certain points are learned and the remaining LUT entries are interpolated. One such approach was presented by Zhou et al. [26]. It needs no calibration pattern but manual definition of correspondences. The approach presented by Senior et al. [21] aims at a more automatic solution by learning transformations from unlabelled training data. However, for a new scene a considerable amount of training data is needed to learn the

new LUT, thus an immediate use after camera installation is not possible.

In order to develop a fast, fully-automatic and simple approach that can be applied by a non-expert as well we follow the method presented by Badri et al. [2]. The approach is based on the matching of SIFT features [16] between the views of two collocated cameras. However, our approach operates in the other direction: instead of moving the dynamic camera to well-defined points in the scene we are moving the dynamic camera in regular rotation steps and obtain a LUT entry for each step. Thus, no convergence criterion for aligning both views has to be defined.

3.1 Feature-Based Mapping Method

In our method the dynamic camera is moved in regular steps and for every step SIFT keypoints are matched between the frames of the static and dynamic camera. SIFT features were introduced by Lowe [16] as a method for extracting local image descriptors that are highly discriminative for object recognition. SIFT features are invariant to changes in image translation, scaling, and rotation and partially invariant to changes in illumination and affine distortion and proofed to be very robust compared to other local image descriptors [18]. In the next step, the matched keypoints are used for a robust homography estimation between the two views by means of RANSAC [10]. Please note that the cameras are collocated, i.e. their centers of projection are nearly identical, hence a homography is sufficient to describe the transformation between the two camera views. After a homography has been estimated, the center of the dynamic camera's frame can be transformed to the coordinate system of the static camera's frame. Finally, the current pan and tilt angles are stored in the LUT at the transformed image point. This is done for every view of the dynamic camera. To keep the method simple, we neglect the aspect of having more zoom levels. Usually, for different zoom levels different LUTs are needed, although the discrepancies are low. We leave this issue open for further research. The final step of the method is to interpolate the empty entries in the LUT. For this purpose we use Thin-Plate-Splines (TPS) interpolation [5] which produces a smooth representation of the pan and tilt angles in the LUT. The TPS interpolation is also essential for a robust LUT generation when homogeneous regions occur in the scene which preclude the estimation of the homography due to the absence of discriminative visual features.

In the following we give a more formal description of our method. We have a static camera providing the image I_s and a dynamic camera providing the image $I_d(\alpha, \beta)$ at pan angle α and tilt angle β. The goal is to learn a LUT L which stores for every point $p_s = (x_s, y_s)$ of I_s the values α and β such that the point p_s is in the center of I_d:

$$L(x_s, y_s) = (\alpha, \beta) \tag{1}$$

The algorithm is summarised in Alg. 1. To detect cases where I_d is outside of I_s and a homography was estimated from false matches, only homographies with an absolute value of the determinant in the range of 0.2 to 5 are allowed.

```
1  Compute SIFT interest points in I_s
2  for α = α_start to α_end with stepsize α_stepsize do
3      for β = β_start to β_end with stepsize β_stepsize do
4          Compute SIFT interest points in I_d
5          Estimate homography H between I_s and I_d by
           applying RANSAC on matched interest points
6          if 0.2 ≤ |det(H)| ≤ 5 then
7              Determine the point p_s = (x_s, y_s, 1)^T by
               transforming the center point p_c of I_d with
               H: p_s = H p_c
8              Store the current angles in the LUT:
               L(x_s, y_s) = (α, β)
9          end
10     end
11 end
12 Interpolate L for all points using thin-plate-splines
```

Algorithm 1: Algorithm for computing the LUT.

3.2 Experiments

For evaluation a *AXIS 223M* was used as static camera and a *Sony SNC-RZ50P* served as dynamic camera. The static camera provides images with a resolution up to 1600×1200. The dynamic camera captures images with a resolution of 640×480 and has a 26x optical zoom. Its mechanical range of the pan and tilt angle is between -170° and +170° and -25° and +90°, respectively. The step size for the mechanical positioning of both the pan and tilt angle lies at 0.07°.

Figure 3(a) shows the image I_s of the static camera of the scene used for the experiments. According to the terminology introduced above, the values $\alpha_{start} = -40°$, $\alpha_{end} = 40°$, $\beta_{start} = -20°$, $\beta_{end} = 30°$ and a stepsize of 5° was used, thus in total 187 images were captured by the dynamic camera. The zoom of the dynamic camera was set to 6.5x. As mentioned before, the consideration of different zooms is part of future research.

In order to evaluate the accuracy of our method in absence of ground truth data from a calibrated scene, we placed five targets in the scene after LUT generation, as shown in Figure 3(b). Targets 1, 2 and 3 were placed on the left wall, on the right wall and on the floor, respectively. Targets 4 and 5 were placed on chairs adding new geometry to the scene.

The obtained errors for the pan and tilt angles are shown in Figure 4. Errors are reported for the static camera's maximum resolution of 1600×1200 (red circles) as well as for a resolution of 640×480 (blue asterisks) to examine the practicability of our approach on cheap cameras with lower resolutions. The dashed lines indicate the average errors for the five targets. They lie in the range of $0.58°$ to $0.62°$ and $0.31°$ to $0.36°$. It can be also seen from Figure 4 that the maximum pan and tilt error lies at $1.81°$ and $0.87°$, respectively. These errors come from target 4, an object which was not apparent in the scene during LUT generation. However, also target 5 adds new geometry to scene and its errors lie between $0.17°$ and $0.27°$ which is lower than the average error. Figure 5 show the indi-

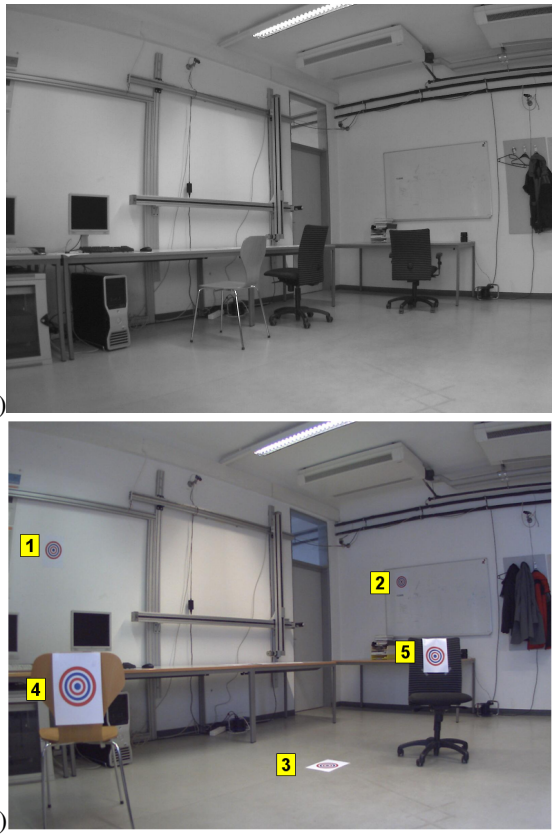

Figure 3. (a) The image I_s of the static camera used for the experiments, (b) targets newly placed in the scene.

vidual results of the method for targets 1 and 4 with a resolution of 640×480. The targets have a diameter of $16cm$ and in the worst result of the experiments (target 4) the target is missed by $\sim 14cm$.

3.3 Discussion

The results show that the proposed method is able to learn an accurate LUT for a given scene. The whole process of image acquisition and LUT generation takes about 10 minutes. The main benefit is that it needs no camera calibration, correction of lens distortion or any manual interaction by using a simple, fast and easy-to-use method.

However, it must be noted that these initial experiments serve only as a proof of concept. The main drawback of the method is that it requires a certain amount of strong visual features in the scene. Although the TPS interpolation is able to fill the missing data the method might fail in the presence of large homogeneous regions. Another problem of the method might be a strong radial distortion of the static camera's lens which can corrupt the homography estimation if the zoom level of the dynamic camera is too low.

A conclusion of the experiments is that, for the given setup, the influence of the resolution is negligible and the lower resolution of 640×480 does not deteriorate the results. However, in practice the resolution needed depends on the zoom level

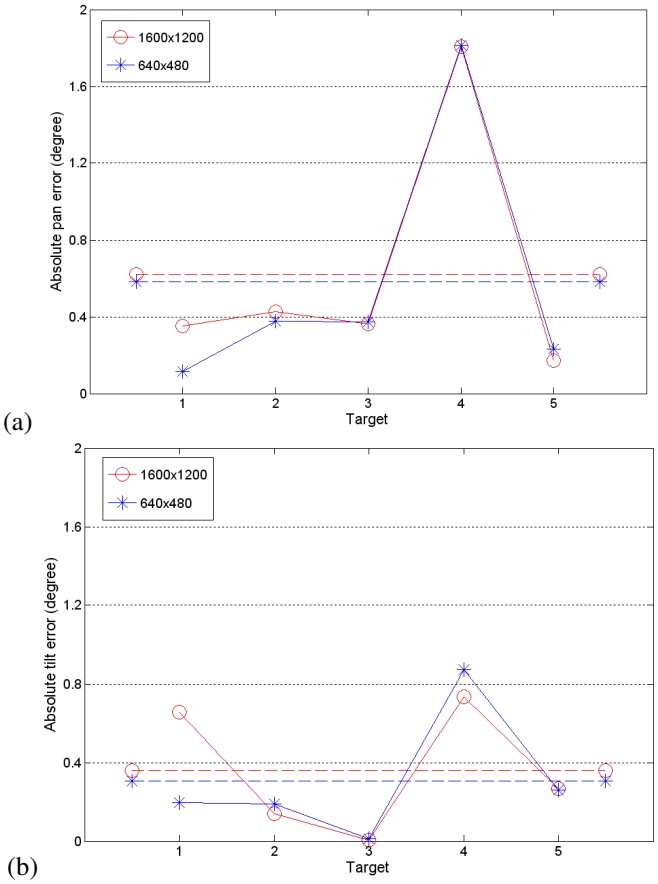

Figure 4. (a) Pan and (b) tilt error for the five targets shown in Figure 3(b).

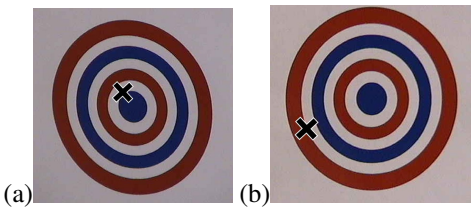

Figure 5. Error of (a) target 1 and (b) target 4 with a resolution of 640×480.

of the dynamic camera and the distance of the visual features from the camera. In the future, the impact of all factors (i.e. presence of visual features, radial distortion, image resolution and zoom level) will be part of a more comprehensive evaluation by means of a comparison to point mapping between calibrated cameras.

4 Conclusion

In this paper an overview of the recently granted *tripleB-ID* project was given. The aim of the project is to establish a multi-camera surveillance system with a twofold intention: first, crime prevention by the detection of suspicious behaviour of

people exploring the bank for robbery planning, and second, supporting the crime solving by providing forensic data like high-resolution facial images and person's height.

In the second part of the paper a method for the automatic mapping between a static camera and a collocated dynamic camera has been proposed. The reported results show the general ability of the method to compute an accurate LUT for a given scene. However, in the future a more detailed evaluation - including an evaluation of its applicability in the testbed - will be conducted.

Acknowledgement

This paper was supported by the Austrian Research Promotion Agency, KIRAS initiative.

References

[1] *Polizeiliche Kriminalstatistik 2007*. Republik Österreich, Bundesministerium für Inneres, 2008.

[2] J. Badri, C. Tilmant, J. M. Lavest, Q. C. Pham, and P. Sayd. Camera-to-camera mapping for hybrid pan-tilt-zoom sensors calibration. In *Proc. of SCIA*, pages 132–141, 2007.

[3] J. Batista, P. Peixoto, and H. Araujo. Real-time active visual surveillance by integrating peripheral motion detection with foveated tracking. In *Proc. of VS*, pages 18–25, 1998.

[4] P. Blauensteiner, H. Wildenauer, A. Hanbury, and M. Kampel. Motion and Shadow Detection with an Improved Colour Model. In *Proc. of ICSIP*, pages 627–632, 2006.

[5] F.L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *PAMI*, 11(6):567–585, 1989.

[6] R.T. Collins, A.J. Lipton, H. Fujiyoshi, and T. Kanade. Algorithms for cooperative multisensor surveillance. *Proceedings of the IEEE*, 89(10):1456–1477, 2001.

[7] A. Criminisi. *Accurate Visual Metrology from Single and Multiple Uncalibrated Images*. Springer, 2001.

[8] A. Del Bimbo, F. Dini, A. Grifoni, and F. Pernici. Exploiting single view geometry in pan-tilt-zoom camera networks. In *Proc. of ECCV - M2SFA2 Workshop*, 2008.

[9] A. Del Bimbo, F. Dini, A. Grifoni, and F. Pernici. Uncalibrated framework for on-line camera cooperation to acquire human head imagery in wide areas. In *Proc. of AVSS*, pages 252–258, 2008.

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

[11] H. Grabner, P.M. Roth, and Bischof H. Is pedestrian detection really a hard task? In *Proc. of PETS*, pages 1–8, 2007.

[12] R. Horaud, D. Knossow, and M. Michaelis. Camera cooperation for achieving visual attention. *MVA*, 16(6):1–2, 2006.

[13] A. Jain, D. Kopell, K. Kakligian, and Y.-F. Wang. Using stationary-dynamic camera assemblies for wide-area video surveillance and selective attention. In *Proc. of CVPR*, pages 537–544, 2006.

[14] S. Khan and M. Shah. Consistent labeling of tracked objects in multiple cameras with overlapping fields of view. *PAMI*, 25(10):1355–1360, 2003.

[15] I. Laptev. On space-time interest points. *IJCV*, 64(2):107–123, 2005.

[16] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[17] G. Medioni, J. Choi, C.-H. Kuo, and D. Fidaleo. Identifying noncooperative subjects at a distance using face images and inferred three-dimensional face models. *SMC-A*, 39(1):12–24, 2009.

[18] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, 2005.

[19] P. M. Roth, H. Grabner, D. Skočaj, H. Bischof, and A. Leonardis. On-line conservative learning for person detection. In Rama Chellappa, James Ferryman, and Tieniu Tan, editors, *Proc. of PETS*, pages 223–230, 2005.

[20] P. M. Roth, Sternig S., Grabner H., and Bischof H. Classifier grids for robust adaptive object detection. In *Proc. of CVPR*, 2009.

[21] A. W. Senior, A. Hampapur, and M. Lu. Acquiring multi-scale images by pan-tilt-zoom control and automatic multi-camera calibration. In *Proc. of WACV-MOTION - Volume 1*, pages 433–438, 2005.

[22] P. Turaga, R. Chellappa, V.S. Subrahmanian, and O. Udrea. Machine Recognition of Human Activities: A Survey. *CSVT*, 18(11):1473–1488, 2008.

[23] P. Viola and M.J. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.

[24] Y. Wang and G. Mori. Human action recognition by semi-latent topic models. *PAMI*, 31(10):1762–1774, 2009.

[25] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys*, 38(4), 2006.

[26] X. Zhou, R. T. Collins, T. Kanade, and P. Metes. A master-slave system to acquire biometric imagery of humans at distance. In *Proc. of IWVS*, pages 113–120, 2003.