Object Detection and Matching in a Mixed Network of Fixed and Mobile Cameras

Alexandre Alahi[1,2], Pierre Vandergheynst[1], Michel Bierlaire[2], Murat Kunt[1]
Swiss Federal Institute of Technology
[1]Signal Processing Laboratory, [2]Transportation and Mobility Laboratory
CH-1015 Lausanne - Switzerland
firstname.lastname@epfl.ch

ABSTRACT

This work tackles the challenge of detecting and matching objects in scenes observed simultaneously by fixed and mobile cameras. No calibration between the cameras is needed, and no training data is used. A fully automated system is presented to detect if an object, observed by a fixed camera, is seen by a mobile camera and where it is localized in its image plane. Only the observations from the fixed camera are used.

An object descriptor based on grids of region descriptors is used in a cascade manner. Fixed and mobile cameras collaborate to confirm detection. Detected regions in the mobile camera are validated by analyzing the dual problem: analyzing their corresponding most similar regions in the fixed camera to check if they coincide with the object of interest.

Experiments show that objects are successfully detected even if the cameras have significant change in image quality, illumination, and viewpoint. Qualitative and quantitative results are presented in indoor and outdoor urban scenes.

Categories and Subject Descriptors

I.4.8 [**Image Processing And Computer Vision**]: Scene Analysis

General Terms

Algorithms, Performance

Keywords

Object Detection, Object Matching, Region Descriptor

1. INTRODUCTION

Detection of objects of interest in digital videos has been in the focus of the research community over the past decade. Low-cost digital cameras and progress in processing large data sets such as video streams have promoted the installation of cameras on fixed and moving platforms. Cameras are now integrated into many devices such as phones or vehicles. The use of data provided concurrently from several cameras, leads to a better understanding of the objects of interest. Mobile cameras (e.g. a camera held by a pedestrian or placed in a car) benefit from their proximity to the objects to capture high resolution features. Pan-tilt-zoom cameras can also be used to optically zoom on an object of interest to capture relevant features.

Most of the systems assume a well structured environment. Cameras need to be fixed and calibrated [5, 11, 7], or only a given object can be detected, *e.g.* a pedestrian [6, 18, 20].

This work presents a system where any object can be detected (*e.g.* animals, cars, urban signs, pedestrians, etc.) in the image plane of any camera (fixed or mobile) given only its single observation from another viewpoint. To evaluate the performance of the approach, a system similar to [2] is studied: any object observed by a fixed camera is searched within the image plane of a mobile camera to find a match. Only observations from the fixed camera are used.

The main drawback of the system presented by Alahi *et*

Figure 1: Two challenging scenes. Left column shows the objects of interest highlighted in the fixed camera. Right column presents the matched objects detected by our algorithm in the mobile camera

AREA'08, October 31, 2008, Vancouver, British Columbia, Canada.

al. in [2] is its failure to classify if an object is really present in the image plane of a mobile camera: they suppose that any object observed by a fixed camera has a correspondence in the mobile camera. Therefore, the best match is chosen as the object of interest. Moreover, the mobile camera does not communicate with the fixed camera to validate the detection process.

In this paper, all the objects observed by a fixed camera are searched within the mobile camera and only those present in the mobile camera are detected. A full collaboration between the fixed and mobile cameras is proposed to validate the detections and help rejecting false positives.

Experiments show that objects are successfully detected even if the cameras have significant change in image quality, illumination, and viewpoint as illustrated in Figure 1. Partial occlusions are also handled.

The paper is structured as follows: the related work is briefly presented in the next section. Then, a formulation of the problem is given. Section 4 presents the matching algorithm. Section 5 describes the validation scheme. Finally, the performance of the approach is evaluated on challenging data sets. Quantitative and qualitative results are given. The paper ends with concluding remarks.

2. RELATED WORK

2.1 Object Detection with mobile cameras

Most of the systems to detect an object with a mobile camera generate a model given a training data. In a classification framework, a region is classified given a set of extracted features.

Papageorgiou *et al.* in [14] and [15] use haar wavelet coefficients of a set of normalized pedestrian images. They classify the images with a support vector machine (SVM) and a "bootstrapping" method. Recent works have shown that histogram of oriented gradient (HOG) is an efficient and robust shape-based cue [17, 6, 18]. Tuzel *et al.* in [20] use covariance matrices as object descriptors. They have less false positives for the same detection rate as opposed to previous approaches. Gavrila uses a template matching technique based on hierarchical representation of the templates [10, 9]. Shape matching is based on distance transforms (chamfer distance). A reasonable shape extraction is needed.

Broggi *et al.* in [4] and [3] detect pedestrians without any training. Their detection is based on morphological characteristics of pedestrians (size and aspect ratio), vertical linear filter, and the strong vertical symmetry of the human shape. Moreover, an assumption about the region where a pedestrian can be found is done. Thus, their system only operates on flat roads with smoothly varying slope. In addition, multiple detection of the same pedestrian occurs, and pedestrian with monochrome clothing are hardly detected.

Recently, Leibe *et al.* in [12] present a system that is able to detect objects and estimate their 3D localization with stereo cameras mounted on a car.

However, these systems suffer from high false positive rates and from the restriction to only detect objects present in their training data.

2.2 Matching Objects across Cameras

Detecting an object with a camera is challenging, but finding its correspondence in other cameras is an additional defy. Techniques to detect an object with a camera can not be used to match the objects across cameras. By definition, they remove the discriminative parts between two objects of the same category. To find correspondence between two views, most of the systems suppose static and calibrated cameras. A homography matrix is estimated at calibration step to project a point (usually the ground plane points) in the image plane of the cameras to a common reference.

Mueller *et al.* in [13] mark with same label the nearest object with the same size and center of gravity. Caspi *et al.* in [5] match objects by fusing the estimated trajectories obtained by each camera. Those systems fail to match objects if the cameras are moving. The homography matrix is not available anymore.

In this work, objects are matched across fixed and mobile cameras. No calibration is used. The single observation of the objects in another view is sufficient. In the next sections, a system is presented to detect objects of interest with a mobile camera and match them across other cameras based on their appearance in the cameras views.

3. PROBLEM FORMULATION

Given an observation x of an object O in a fixed camera, we wish to detect its presence in the view of a mobile camera, and if present, locate it in its image plane. No additional training data should be used.

Let y_i be a potential region in the mobile camera. x and y_i are subsets of an image bounded by a rectangular bounding box.

We define the "Region Matching" operator, Φ, which maps a region x to the N_y most similar regions in a given image I_m:

$$\Phi(x, I_m, N_y) = \{y_1, y_2, ..., y_{N_y}\} = Y_x \quad (1)$$

with I_m the image plane of the mobile camera.

The precise notion of similarity will be described in section 4.

The same operator Φ can be used to map any y_i to a set of $\hat{x}_i$ referred in this paper as the dual problem:

$$\Phi(y_i, I_f, N_x) = \{\hat{x}_1, ..., \hat{x}_{N_x}\} = \hat{X}_i \quad (2)$$

$\hat{X}_i$ are the regions in the fixed camera similar to y_i.

If a region $\hat{x}_i$ matches x, then the corresponding y_i should be the region bounding object O in the mobile camera (see Figure 2). If none of the $\hat{x}_i$ coincides with x, object O should not be present in the view of the mobile camera.

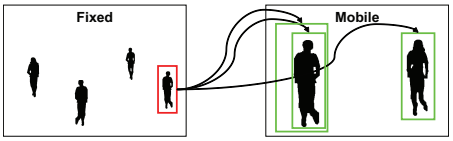
We hence define an operator ϑ to validate if a region y_i matches x:

$$\vartheta(y_i|x, \hat{X}_i) = \vartheta(y_i|x, \hat{x}_1, ..., \hat{x}_j) \in [0, 1] \quad (3)$$

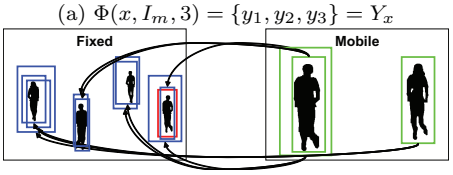
As a result, the problem can be formulated as follows: for a given x, find the region y_x in the mobile camera that maximizes $\vartheta(y_i|x, \hat{X}_i)$ for all $y_i \in Y_i$:

$$y_x = \arg\max_{y_i \in Y_i} \vartheta(y_i|x, \hat{X}_i) \quad (4)$$

If such a y_x does not exist, it means that the object is not present in the image plane of the mobile camera.



(a) $\Phi(x, I_m, 3) = \{y_1, y_2, y_3\} = Y_x$



(b)For $i = 1:3$: $\Phi(y_i, I_f, 3) = \{\hat{x}_1, \hat{x}_2, \hat{x}_3\} = \hat{X}_i$

Figure 2: Illustration of the Φ operator. (a) An object x, highlighted in the fixed camera, is mapped to the best 3 regions in the mobile camera. (b) Then, each region y_i is mapped back to 3 regions in the fixed camera. If those regions coincide with x, there is a match.

4. REGION MATCHING

4.1 The Approach

An object descriptor (OD) is created from the region bounding the object of interest in the fixed camera. Then, all possible regions in the image plane of the mobile camera are compared with the OD. A window of size proportional to the object bounding box scans the image plane of the mobile camera at different scales. For each region, its similarity with the OD is computed to find the region with highest similarity. Therefore, a discriminative region descriptor is needed.

4.2 Covariance matrix

Covariance matrices are a very attractive descriptor first used by Tuzel *et al.* [19], [16], [20]. For each pixel, a set of features are extracted. Alahi *et al.* in [2] used the grayscale intensity, I, and the norm of the first order derivatives with respect to x and y, I_x and I_y:

$$f_n = (x, y, I, I_x, I_y) \tag{5}$$

Other features such as the R,G,B values or the second order derivatives, the gradient magnitude, mg, and its angle, θ, can also be used. The pixel coordinates, x and y, are integrated in the feature vector to consider the spatial information of the features. Finally, the covariance of a region is computed as:

$$C_i = \frac{1}{N-1}\sum_{n=1}^{N}(f_n - m)(f_n - m)^T \tag{6}$$

where N is the number of points in the region, and m the mean vector of all the feature vectors.

With covariance matrices, several features can be fused in a lower dimensionality without any weighting or normalization. They describe how features vary together.

Similarity between two regions B_1 and B_2 is given by the following distance proposed by [8]:

$$\sigma_1(B_1, B_2) = \sqrt{\sum_i \ln^2 \lambda_i(C_1, C_2)} \tag{7}$$

where $\lambda_i(C_1, C_2)$ are the generalized eigenvalues of the covariance matrices C_i

4.3 A Collection of Grids of Descriptors

An object descriptor (OD) is used taking into account local and global information. It is a collection of grids of region descriptors (see figure 3). Each grid segments the object into different number of blobs of equal sizes. Grids of finer blob size describe local information whereas grids of coarse blob size describe a more global behavior.

Figure 3: A collection of grids of descriptors

Similarity between two objects, $\phi(x, y_i)$, is computed by summing distance between corresponding blobs segmenting the grids. Since, many objects do not have a rectangular shape and some can be partially occluded, only the most similar blobs are kept. In this way, blobs belonging to the background can also be discarded (see figure 4).

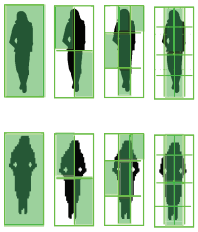


Figure 4: Top row is an object detected by the fixed camera. Bottom row is a region within the mobile camera. At least half of the blobs are kept to compute the global distance

4.4 Matching Process

4.4.1 Preprocessing step: Edge Filtering

Some regions in the mobile camera do not need to be compared with the ODs. They can be discarded with a simple preprocessing. The difference between the proportion of edges in two regions can give a quick indication about their similarity. If the proportion of edges is not similar, the region is discarded. As a result, fewer regions remain to be analyzed and it increases the likelihood to detect the right object by reducing the search space.

4.4.2 Cascade of Coarse to Fine Descriptors

Some regions can be easily discarded without knowing the local information. Therefore, an approach similar to a cascade of classifier is proposed as in [2]. "Easy regions" are discarded with coarse grids (*i.e.* grids with small number of blobs). More challenging regions require the use of finer grids (*i.e.* more number of blobs).

The detection process is divided into several stages. At each stage, a finer grid is used. After each stage, only the

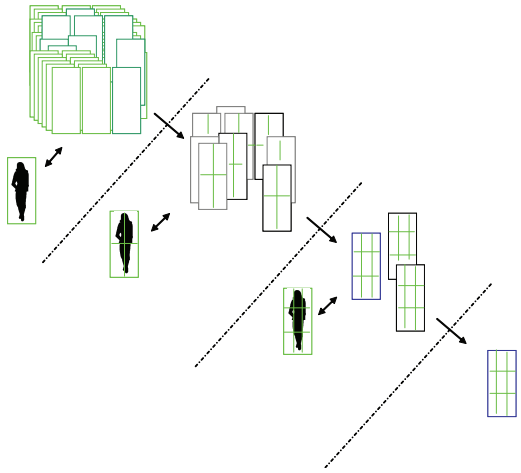


Figure 5: A three stages cascade of coarse to fine descriptors

best candidates, *i.e.* regions with highest similarity (top ρ% of the evaluated regions), remain.

ρ is chosen such that after each stage the same percentage is kept:

$$N_r \times \rho^{N_s} = 1 \tag{8}$$

where N_r is the total number of regions in the mobile camera to compare with the object descriptor, and N_s is the total number of stages to use.

$$\rho = N_r^{-1/N_s} \tag{9}$$

5. REGION VALIDATION

The validation operator, ϑ, evaluates the likelihood that object x matches region y_i in the mobile camera. It considers the dual problem by analyzing the set obtained by $\Phi(y_i, I_f, N_x) = \{\hat{x}_1, \hat{x}_2, ..., \hat{x}_{N_x}\}$. In the next section, the choice of N_x will be studied.

A similarity measure σ between the original x and each $\hat{x}_i$ is estimated based on the spatial arrangement of their bounding boxes:

$$\sigma(x, \hat{x}_i) = e^{-(\frac{1-O}{c_1})^4} w_o + e^{-(\frac{1-C}{c_1})^4} w_c + e^{-(\frac{D_c}{c_2})^4} w_d \tag{10}$$

where

- C is a percentage which represents how much of the original bounding box of x is covered by the bounding box of $\hat{x}_i$. Likewise, O is the percentage which represents how much $\hat{x}_i$ is covered by x. (see figure 6)
- D_c is the euclidian distance between the center of two bounding boxes.

c_1 and c_2 are constants adjusted so that $\sigma = 0$ if C, O and D_c are less than their thresholds.

$\sigma(x, \hat{x}_i) > 0$ if and only if C and $O > 30\%$ and $D_c < 0.75 * max(width_x, height_x)$.

A weight $w_.$ is given to each factor to emphasize priority. In this work, focus is first on a high cover of x, then a similar

Figure 6: An example of the bounding box of x (in red) and $\hat{x}_i$ where $C \approx 0.75$, $O \approx 0.4$

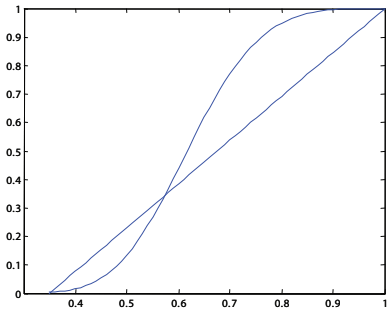


Figure 7: x-axis represents C or O; y-axis represents its contribution to σ and σ_l. It can be seen that for values of C or O close to 1, the contribution remains 1 (full) for the non-linear operator.

center of mass, finally $\hat{x}_i$ should not be too big with respect to x (decent O).

The non-linear operator, $e^{-(.)^4}$, is used to reduce sensitivity to two regions overlapping with a slight difference. A linear σ_l such as:

$$\sigma_l(x, \hat{x}_i) = 1 - (\frac{1-O}{1-c_1} w_o + \frac{1-C}{1-c_2} w_c + \frac{D_c}{c_3} w_d) \tag{11}$$

was too sensitive to differences. Figure 7 plots the two operators, and figure 8 presents an example of the value obtained with σ and σ_l.

Figure 8: The linear σ_l gave 0.63% and the proposed σ gives 0.86%

Finally, $\vartheta(y_i|x, \Phi(y_i))$ is computed as follows:

$$\vartheta(y_i|x, \Phi(y_i)) = \max_{\hat{x}_i \in \Phi(y_i)} \sigma(x, \hat{x}_i) \times w(y_i) \tag{12}$$

where $w(y_i)$ weights region y_i with respect to other y_j based on the similarity measurement computed by $\Phi(x)$ (in section 4.3):

$$w(y_i) = \frac{\phi(x, y_i)}{\max_{y_j \in \Phi(x)} \phi(x, y_j)} \tag{13}$$

6. PERFORMANCE EVALUATION

6.1 Data Sets

Indoor and outdoor data sets have been used. Each data set is composed of the video sequences captured by a fixed and a mobile camera in the same scene. Fixed cameras are located at a height equivalent to the first floor of a building. Mobile cameras are held by pedestrians walking in the scene. The videos sequences with their ground truth data (in xml format) can be found in [**?**]. The images are recorded at 25fps with a resolution of 320×240. Figure 1 presents an example of images captured by the cameras.

The data sets used have meaningful changes in viewpoint, illumination, and color distribution between fixed and mobile cameras. Sensing devices are also different. Indeed, mobile cameras have a cheap capturing device and hence provide noisy images.

6.2 Experiments

Thousands of frames and objects are selected within the fixed cameras to find correspondence in mobile cameras. In the first data set, only pedestrians are of interest (see figure 12). In the second one, random rigid objects in the scene are selected to prove generalization of the approach to any objects of interest (see figure 13).

6.3 Performance of the Validation Scheme

In this section, focus is on evaluating the validation approach. The relevance of the object descriptor and the matching process (Φ) is extensively studied in [1]. The cascade of grids of covariances matrices outperforms other descriptors based on histograms of oriented gradients, colors, or interest points. The features used in this work are the following:

$$f_n = (x, y, I, I_x, I_y, mg, \theta) \tag{14}$$

They outperform the one used in [2].

In the validation process, two parameters are of interest: the number of regions N_y and N_x to keep within Y_x and $\hat{X}_i$. Figure 9 presents the detection rate and the corresponding number of false positives generated for various $N_.$. They are compared with the previous approach considering the best match of the region matching operator as the matched object (labeled as "best match") without any validation process. With the proposed approach, setting $N_x = N_y = 2$, the false positive (FP) rate is decreased by 70 % while the True Positive (TP) rate decreases by only $\sim 2\%$. For $N_x = N_y = 3$, the number of FP is reduced by half while the percentage of TP is reduced by less than 1%. Having high values for N_x and N_y will not necessarily lead to high performance. Considering $N_y = 2$ and $N_x = 1$ is the best tradeoff for our application in terms of performance.

In addition, a possible approach to reduce even more the false positives rate will be to threshold the similarity measurements ϕ. However, if the validation scheme is not used, it is not interesting to threshold $\phi(x, y_i)$, obtained between the object descriptor and the regions in the mobile camera. Figure 10 illustrates the histogram of the values obtained when the regions are correctly matched (TP) and the ones for the false positives (FP). There is not a clear decision boundary. Typically, setting the threshold to 4.4 will reduce the FP rate by 9% and reduce the TP rate by 11%.

However, it is possible to threshold the similarity measurement $\phi(y_i, \hat{x_i})$, or the sum $\phi(x, y_i) + \phi(y_i, \hat{x_i})$ obtained in the validation process. Figure 11 shows the histograms for the two cases. Now, an interesting decision boundary exists: if we keep y_i such that $\phi(y_i, \hat{x_i}) < 4.1$ or $\phi(x, y_i) + \phi(y_i, \hat{x_i}) < 8.2$, the remaining FP will be reduced by 50% while reducing the TP rate by 5% only. Therefore, the proposed approach can globally reduces the $\#FP$ by $75 - 85\%$ for a decrease of 5-7% of the TP rate. This is feasible only because of the validation approach considering the dual problem. Without the validation scheme proposed in this work, to reduce the false positive rate by 80%, the TP rate will be reduced by 50%.

Qualitative results are given on both data sets in figures 12 and 13. It can be seen that objects are successfully detected even if the cameras have significant change in image quality, illumination, and viewpoint. In addition, highlighted objects in the fixed camera which are not present in the view of the mobile camera are not generating false positives. Figure 14 presents some missed detection and few false positives.

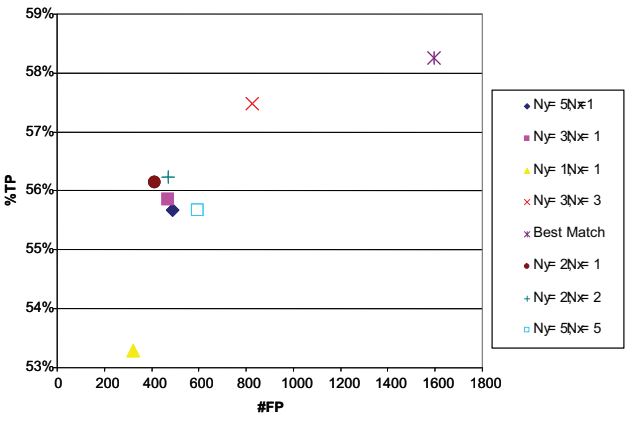


Figure 9: Graph of the percentage of TP with respect to the $\#FP$ for various N_y and N_x.

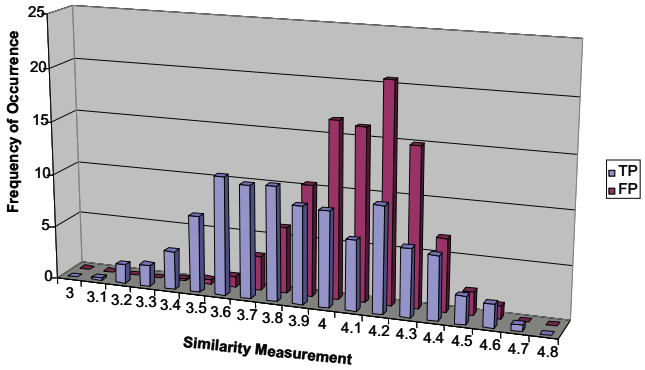
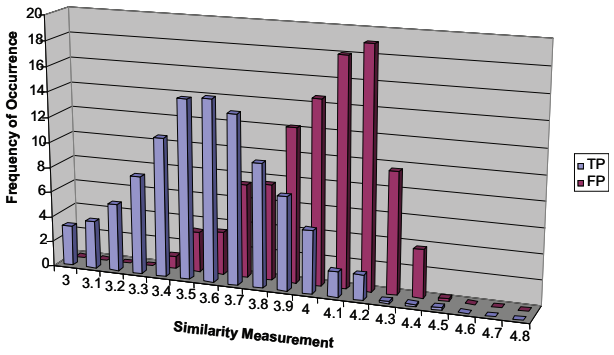
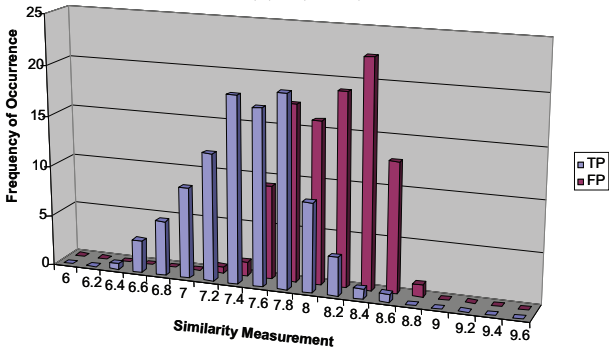


Figure 10: Histogram of the similarity measurements $\phi(x, y_i)$ for a set of TP and FP



(a) $\phi(y_i, \hat{x_i})$



(b) $\phi(x, y_i) + \phi(y_i, \hat{x_i})$

Figure 11: Histogram of the similarity measurements in the validation process

7. CONCLUSIONS

A system is presented to detect and match in the image plane of a mobile camera objects observed by a fixed camera. No calibration between the camera is needed. No training data is used. The proposed system is able to classify the presence of an object in the mobile camera by analyzing the dual problem. The presented validation process reduces the false positives rate considerably without significantly affecting the detection rate (%TP).

Further work will evaluate the impact of fusing several object descriptors such as a histogram of oriented gradients with the covariance matrices. In addition, considering the dynamic of the system will increase the overall performance of the system.

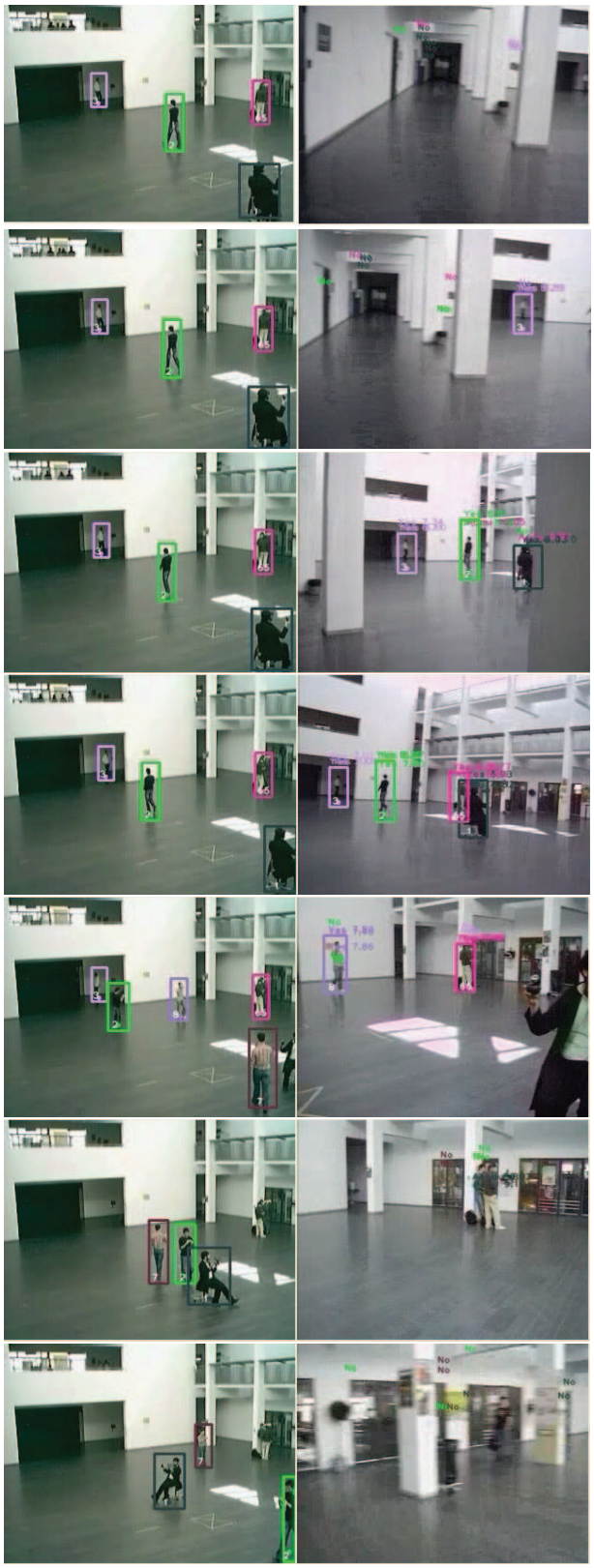


Figure 12: Correct detections and no FP. First column: objects detected by fixed camera. Second column: corresponding objects detected and matched with the mobile camera

Figure 13: Correct detections and no FP. First column: objects detected by fixed camera. Second column: corresponding objects detected and matched with the mobile camera

Figure 14: Some FP and missed TP. First column: objects detected by fixed camera. Second column: corresponding objects detected and matched with the mobile camera

8. REFERENCES

[1] A. Alahi, M. Bierlaire, and M. Kunt. Object detection and matching with mobile cameras collaborating with fixed cameras. In *The ECCV, Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*, 2008.

[2] A. Alahi, D. Marimon, M. Bierlaire, and M. Kunt. A master-slave approach for object detection and matching with fixed and mobile cameras. In *Accepted IEEE Int. Conf. on Image Processing (ICIP), San Diego, CA, USA*, 2008.

[3] M. Bertozzi, A. Broggi, R. Chapuis, F. Chausse, A. Fascioli, and A. Tibaldi. Shape-based pedestrian detection and localization. In *Procs. IEEE Intl. Conf. on Intelligent Transportation Systems 2003*, pages 328–333, Shangai, China, Oct. 2003.

[4] A. Broggi, M. Bertozzi, A. Fascioli, and M. Sechi. Shape-based pedestrian detection. *Proc. IEEE Intelligent Vehicles Symp*, pages 215–200, 2000.

[5] Y. Caspi, D. Simakov, and M. Irani. Feature-based sequence-to-sequence matching. 68(1):53–64, June 2006.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. pages I: 886–893, 2005.

[7] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multi-camera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.

[8] W. Forstner and B. Moonen. A metric for covariance matrices. *Qua vadis geodesia*, pages 113–128, 1999.

[9] D. Gavrila. Pedestrian detection from a moving vehicle. pages II: 37–49, 2000.

[10] D. Gavrila and V. Philomin. Real-time object detection for "smart" vehicles. pages 87–93.

[11] S. Khan and M. Shah. A multiview approach to tracking people in crowded scenes using a planar homography constraint. pages IV: 133–146, 2006.

[12] B. Leibe, N. Cornelis, K. Cornelis, L. Van Gool, and E. Zurich. Dynamic 3D Scene Analysis from a Moving Vehicle. *CVPRŠ07*, 2007.

[13] K. Mueller, A. Smolic, M. Droese, P. Voigt, and T. Wienand. Multi-texture modeling of 3d traffic scenes. *icme*, 2:657–660, 2003.

[14] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. *Proc. Computer Vision and Pattern Recognition*, 97:193–199, 1997.

[15] C. Papageorgiou and T. Poggio. Trainable pedestrian detection. *Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on*, 4, 1999.

[16] F. Porikli, O. Tuzel, and P. Meer. Covariance Tracking using Model Update Based on Lie Algebra. *IEEE Conf. on Computer Vision and Pattern Recognition*, 2006.

[17] A. Shashua, Y. Gdalyahu, and G. Hayun. Pedestrian detection for driving assistance systems: Single-frame classification and system level performance. pages 1–6, 2004.

[18] F. Suard, A. Rakotomamonjy, A. Bensrhair, and A. Broggi. Pedestrian Detection using Infrared images and Histograms of Oriented Gradients. In *Procs. IEEE Intelligent Vehicles Symposium 2006*, pages 206–212, Tokyo, Japan, June 2006.

[19] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. *Proc. 9th European Conf. on Computer Vision*, 2006.

[20] O. Tuzel, F. Porikli, and P. Meer. Human Detection via Classification on Riemannian Manifolds. *Proc. CVPR*, pages 1–8, 2007.