A MASTER-SLAVE APPROACH FOR OBJECT DETECTION AND MATCHING WITH FIXED AND MOBILE CAMERAS

Alexandre Alahi[1,2]*, David Marimon*[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

ABSTRACT

Typical object detection algorithms on mobile cameras suffer from the lack of a-priori knowledge on the object to be detected. The variability in the shape, pose, color distribution, and behavior affect the robustness of the detection process. In general, such variability is addressed by using a large training data. However, only objects present in the training data can be detected. This paper introduces a vision-based system to address such problem.

A master-slave approach is presented where a mobile camera (the slave) can match any object detected by a fixed camera (the master). Features extracted by the master camera are used to detect the object of interest in the slave camera without the use of any training data. A single observation is enough regardless of the changes in illumination, viewpoint, color distribution and image quality.

A coarse to fine description of the object is presented built upon image statistics robust to partial occlusions. Qualitative and quantitative results are presented in an indoor and an outdoor urban scene.

Index Terms— Object detection, Object matching, Covariance descriptor

1. INTRODUCTION

Governmental agencies, car manufacturers, and many institutes are interested in detecting potential collision of cars with pedestrians in urban areas. For that purpose, they have mounted cameras in cars. Expensive systems exist such as stereo cameras combined with other sensors [1]. Low-cost systems, *e.g.* a single low resolution camera (320 x 240), is not performing well enough in such applications. Only a restricted number of features such as shape [2], or histogram of oriented gradient [3, 1] can be used to detect a pedestrian in a single image. Tuzel *et al.* in [4] use covariance matrices as object descriptors. They have less false positives for the same detection rate as opposed to previous approaches. However, the variability in the appearance of pedestrians (*e.g.* clothing), their articulated structure, and the non-rigid kinematics affect the performance of the system. This effect can be reduced if additional priors are integrated. Such priors can be stable features extracted from other sensors. Few years ago, such priors were not available whereas nowadays, they are. Indeed, very large number of fixed cameras have been installed in major cities (*e.g.* in 2002, approximately four millions just for the UK [5]). Therefore, features extracted from those fixed cameras can help the detection in mobile cameras. Moving objects are more easily detected with fixed cameras. For instance, background subtraction is a natural approach to detect a moving object [6].

In this work, a master-slave system is introduced. The goal of the system is to find a match of any object of interest in the image plane of a slave camera (from now on called slave) given its observation in a master camera (also called master in the rest of the paper). In our setup, the master is fixed and the slave is mobile. Moreover, a coarse to fine region descriptor is proposed combining the strength of existing descriptors to best address the proposed goal. Experiments show that objects are successfully detected even if the cameras have a meaningful change of image quality, illumination, and viewing point. Figure 1 presents two examples of the detection and matching of the approach presented in this paper. Partial occlusions are also handled.

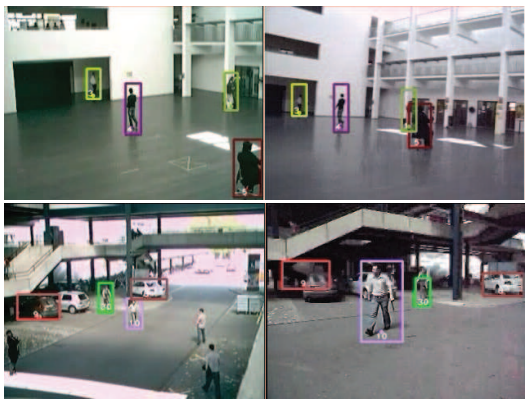


Fig. 1. Left column: Object detected by a fixed (master) camera — right hand-side: result of the proposed approach in a mobile (slave) camera

The rest of the paper is structured as follows: after a formulation the problem, the region descriptor is described with its similarity measurement. Then, the performance of the system is evaluated on different data sets. Quantitative and qualitative results are given. The paper ends with concluding remarks.

2. A MASTER-SLAVE OBJECT DETECTION AND MATCHING APPROACH

2.1. Problem Formulation

Given an observation of an object O in a master camera, we wish to locate it in the image plane of a slave camera. Only features extracted from the region bounding the object in the master are used. In this paper, only a single observation is considered. Future work will consider several observations and take into account the dynamics of the objects.

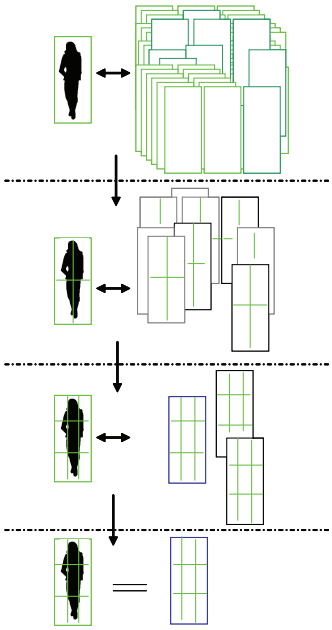


Fig. 2. A three stages coarse to fine approach: Left hand-side is the observed image described with increasing number of blobs. Right hand-side represents all the candidate regions to measure similarity. At each stage, only a few remain (10%)

2.2. The Approach

Since there is not any estimation in the location of the object in the image plane of the slave, all possible regions are evaluated. A window is scanning the image plane at different scales to find the region with highest similarity. For each region, a distance is computed to quantify its similarity with the observation. Therefore, a discriminative region descriptor is needed.

2.3. A Coarse to Fine Region Descriptor

Tuzel *et al.* use a very attractive descriptor: the covariance matrix of image statistics within a region [7, 8, 4]. They obtain higher performance with covariance descriptor rather than commonly used histograms [9].

To detect an object given an observation, the tree-dimensional color vector (R,G,B), and the norm of first and second derivatives of intensity with respect to x and y are used as a feature vector [7]. To track an object across frames, only the intensity and first derivatives are used [8]. The covariance of a region is computed as:

$$C_i = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{f}_n - \mathbf{m})(\mathbf{f}_n - \mathbf{m})^T \qquad (1)$$

where N is the number of points in the region, $\mathbf{f}_n$ the feature vector for a given point, and $\mathbf{m}$ 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. However, experimental results show that if a large region is described, local variations are lost in the global behavior (see section 3). In this work, a coarse to fine approach is presented to address such issue.

The detection process is divided into several stages. At each stage, a region is segmented into blobs (sub-regions) of sizes smaller than in the previous stage. With more blobs a finer description of the object is used. After each stage, only the best candidates, *i.e.* regions with highest similarity (top 10% of the regions evaluated), remain. The process ends when only a single candidate remains.

2.4. Similarity Measurement

Similarity between two regions is computed by summing distance between corresponding blobs segmenting the region. Since, many objects do not have a rectangular shape and some can be partially occluded, only 50% of the blobs are kept, the most similar ones. In this way, blobs belonging to the background can be discarded.

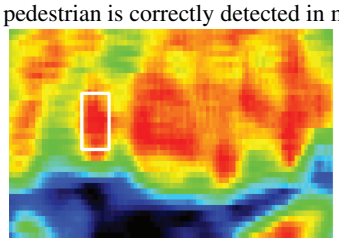
Similarity between two blobs B_1 and B_2 is given by the following distance [8]:

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

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

(a) A pedestrian detected by a fixed camera

(b) The same pedestrian is correctly detected in mobile camera



(c) Similarity map of the image plane of the mobile camera

Fig. 3. A pedestrian detected by a fixed camera is matched in the mobile camera based on above similarity map

Figure 3 presents the similarity map obtained at the last iteration. The white bounding box corresponds to the region selected as the best match.

Fig. 4. Examples with partial occlusions. Left column: detected object in fixed camera. Middle column: output of method described by [7]. Right column: output of proposed approach in this paper

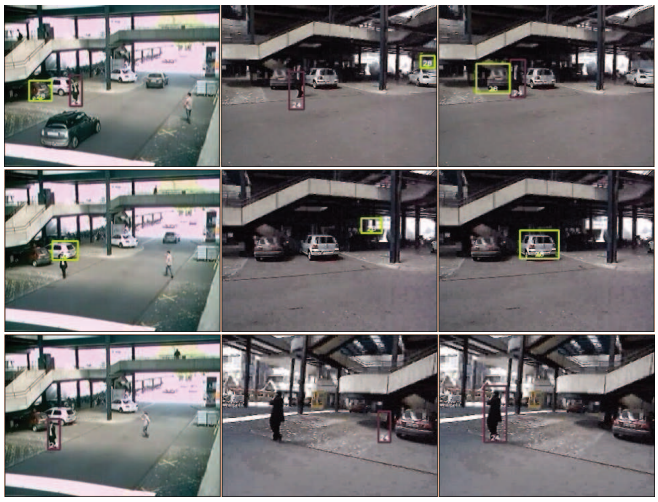


Fig. 5. Examples with severe changes in viewpoints. Left column: detected object in fixed camera. Middle column: output of method described by [7]. Right column: output of proposed approach in this paper

3. PERFORMANCE EVALUATION

3.1. 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 [10]. 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.

3.2. 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 top row of figure 1). In the second one, random objects in the scene are selected to prove generalization of the approach to any object of interest (see figures 4 to 6).

Table 1 compares the performance of the proposed approach with other methods. The work described by Tuzel *et al.* [7] is used as baseline. Since they obtained higher performance than previous approaches based on histograms (on RGB or HSV space), the performance of such histograms is not extensively studied. Beside the covariance descriptor, two typical histogram descriptors are used to evaluate their impact on a coarse to fine approach. The first one is a 2D histogram of 64 bins on Hue and Saturation values. The second one, is a simple 1D histogram of 16 bins on grayscale values. Histogram similarity is computed through the Bhattacharyya distance [9].

Interestingly, the coarse to fine approach is not always increasing the performance of the system. If the right feature descriptor is not used, the performance can even decrease, *e.g.* histogram on H-S (see table 1). With an 1D histogram, the coarse to fine approach increases the performance in some cases. Indeed, in the first data set, objects are not very similar to each other, thus the performance increased. In the second one, objects have less difference in grayscale, hence the performance is not changed. When a more sophisticated descriptor is used, *i.e.* the covariance matrix, the performance is quadrupled in both data sets with the coarse to fine approach. Therefore, the covariance descriptors with the coarse to fine approach is the right combination for such application. Note that when computing the covariance matrix, the color vector is not increasing the performance. Indeed, color is not a reliable feature to use in our system since the same object viewed by two cameras can have a drastically change of color distribution.

Comparing the performance of a 1D histogram and the proposed method by [7], it can be seen that for our application, even if a sophisticated descriptor is used (*i.e.* covariance matrix), it is not always performing better than a simpler descriptor (1D histogram). Integrating the covariance descriptor in a coarse to fine framework reveals its strength. Qualitative results are given in figure 6. Figures 5 and 4 illustrates robustness to change of viewpoint and partial occlusions, respectively.

4. CONCLUSIONS

A novel master-slave system is presented to detect objects without the use of any training data. Any sort of object can be matched in a slave camera as long as a single observation in the master camera is available. A coarse to fine region descriptor is proposed robust to changes of illumination, viewpoint, and image quality. Qualitative

Fig. 6. Random examples. Left column: detected object in fixed camera. Middle column: output of method described by [7]. Right column: output of proposed approach in this paper

Method based on	$TP1$	$TP2$
Full Region described by		
Histogram on H, S	10.0	15.3
Histogram on I	32	28.2
Covariance of features		
$I, \|I_x\|, \|I_y\|$ [8]	21.2	17.8
$R, G, B, \|I_x\|, \|I_y\| \ \|I_{xx}\|, \|I_{yy}\|$ [7]	17.5	15.1
Coarse to fine approach made of		
Histogram on H, S	8.8	17
Histogram on I	68.76	28.4
Covariances on I, $\|I_x\|$, $\|I_y\|$ *(proposed method)*	**80.4**	**69.0**

Table 1. Performance measurement. TP1 and TP2 are respectively the percentage of correctly detected and matched objects in the first and second data sets.

and quantitative results have shown very promising results and encourage the development of such systems. Future work will focus on several observations from the same fixed camera (across time). Furthermore, observations from several master cameras (fixed cameras with different viewpoints) can be considered to increase the performance of the system.

5. REFERENCES

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

[2] D. Gavrila, "Pedestrian detection from a moving vehicle," 2000, pp. II: 37–49.

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

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

[5] Michael McCahill and Clive Norris, "Cctv in london," 2002.

[6] F. Porikli, "Achieving real-time object detection and tracking under extreme conditions," *Journal of Real-Time Image Processing*, vol. 1, no. 1, pp. 33–40, 2006.

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

[8] 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.

[9] D. Comaniciu and V. Ramesh, "Real-time tracking of non-rigid objects using mean shift," July 8 2003, US Patent 6,590,999.

[10] ""http://ltswww.epfl.ch/ alahi/datasets/","" .