Object Detection and Matching with Mobile Cameras Collaborating with Fixed Cameras

Alexandre Alahi, Michel Bierlaire, Murat Kunt

Object Detection and Matching with Mobile Cameras Collaborating with Fixed Cameras

Alexandre Alahi[1,2], 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. A system is presented to detect and match any objects with mobile cameras collaborating with fixed cameras observing the same scene. No training data is needed. Various object descriptors are studied based on grids of region descriptors. Region descriptors such as histograms of oriented gradients and covariance matrices of different set of features are evaluated.
A detection and matching approach is presented based on a cascade of descriptors outperforming previous approaches. The object descriptor is robust to any changes in illuminations, viewpoints, color distributions and image quality. Objects with partial occlusion are also detected. The dynamic of the system is taken into consideration to better detect moving objects. Qualitative and quantitative results are presented in indoor and outdoor urban scenes.

1 Introduction

Low-cost digital cameras and progress in processing large data sets such as digital videos 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 from all cameras capturing a given scene, 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 of interest to capture high resolution features.

Many car manufactures and institutions 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.* using a single low resolution camera (for example 320×240), are not performing well enough in such application. In a classification framework, only a restricted number of features such as shape [2], histogram of oriented gradient [3, 1], or covariance matrices [4], can be used to detect a pedestrian in a single image. These systems suffer from high false positive rates and from the restriction to only detect objects present in their training data.

Alahi *et al.* in [5] propose a system to cope with the limitations of the previous systems. Features extracted from fixed cameras are used to detect objects

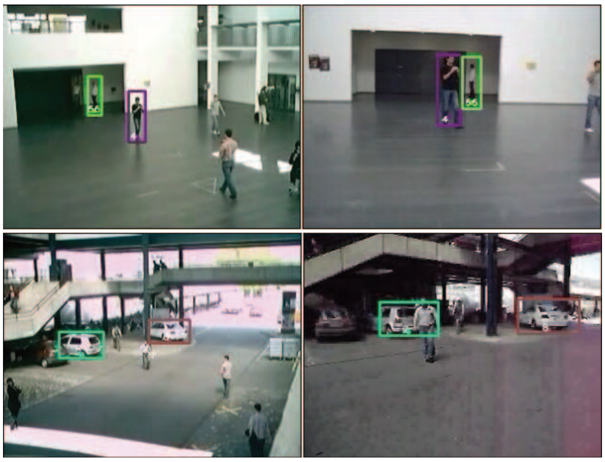

Fig. 1. Left column: objects of interest highlighted in a fixed camera.Right column: objects detected and matched by the proposed approach in a mobile

in mobile cameras[1]. An object descriptor based on a cascade of grids of region descriptors is presented. They obtain higher performance than approaches considering a region with a single descriptor. They compare the covariance descriptor proposed by [7] with histograms of color information.

In this paper, we compare several object descriptors with the proposed cascade of descriptors. We study the impact of describing an object with various grids of region descriptors. The covariance descriptor is compared with the extensively used histograms of oriented gradients [1]. The covariance of various feature sets are considered. In addition, a preprocessing step is added comparing the proportion of edges between two regions to enhance the performance of the approach both qualitatively and quantitatively. Finally, the dynamics of the objects are considered. Several observations are kept to describe a moving object instead of a single observation as in [5].

Experiments show that objects are successfully detected even if the cameras have significant change in image quality, illumination, and viewpoint. Figure 1 presents two examples of the detection and matching of the approach presented in this paper. Partial occlusions are also handled.

The paper is structured as follows: first, an overview of the two region descriptors used in this work is given. Then, after a presentation of the proposed object descriptor, the detection and matching process is presented. In section 5, the performances of different descriptors are evaluated on challenging data sets. Quantitative and qualitative results are given.

[1] Indeed, a very large number of fixed cameras have been installed in major cities (*e.g.* in 2002, approximately four millions just for the UK [6]).

2 Region Descriptors

2.1 Covariance matrix

Covariance matrices are a very attractive descriptor first used by Tuzel *et al.* [7], [8], [4]. For each pixel, a set of features is extracted. Alahi *et al.* in [5] use the grayscale intensity, I, and the norm of the first order derivatives with respect to x and y, I_x and I_y:

$$\mathbf{f}_n = (x, y, I, I_x, I_y) \tag{1}$$

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:

$$C_i = \frac{1}{N-1} \sum_{n=1}^{N} (\mathbf{f}_n - \mathbf{m})(\mathbf{f}_n - \mathbf{m})^T \tag{2}$$

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 [9]:

$$\sigma_r(B_1, B_2) = \sqrt{\sum_i \ln^2 \lambda_i(C_1, C_2)} \tag{3}$$

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

2.2 Histogram of Oriented Gradients

Histograms of Oriented Gradients (HOG) are efficient to compute descriptors based on the first order derivatives with respect to x and y of the image intensity (denoted by I_x and I_y). From these two derivatives, a gradient field is computed assigning to each pixel a magnitude $mg(x, y)$ and an angle $\theta(x, y)$:

$$mg(x, y) = \sqrt{I_x^2(x, y) + I_y^2(x, y)} \tag{4}$$

$$\theta(x, y) = \arctan\left(\frac{I_y(x, y)}{I_x(x, y)}\right) \tag{5}$$

The angle values $\theta \in [0, 360[$ are quantized to N discrete levels θ_i. A histogram is formed where each bin is the sum of all magnitudes with the same orientation θ_n in a given region.

The Bhattacharyya distance [10] or the L_2 norm can be used to compute the distance, σ_r, between the histograms.

3 Object Descriptor

3.1 A Collection of Grids of Descriptors

In this work, a descriptor is created for each observation x_i of an object, referred to as an object descriptor (OD). An observation correspond to the rectangular region bounding the object of interest in the fixed camera at a given frame.

An object descriptor (OD) is used taking into account local and global information. It is a collection of grids of region descriptors (see figure 2). 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.

Fig. 2. A collection of grids of descriptors

3.2 Similarity Measurement

Similarity between grids of descriptors is computed by generating a distance map representing the distances between corresponding blobs. If several observations are available for the same object, the minimum distance, σ_r, between each blob is selected among all observations (see figure 3).

Finally, since many objects do not have a rectangular shape and some can be partially occluded, only the most similar distances are kept. Thereupon, blobs belonging to the background can also be discarded. The sum of the smallest distances in the distance map is the final similarity measurement σ.

4 Object Detection and Matching

4.1 Problem Formulation

Given a set of observations $\{x_1, x_2, ..., x_n\}$ of an object O in a fixed camera, we wish to locate it in the image plane of a mobile camera. No additional training data should be used.

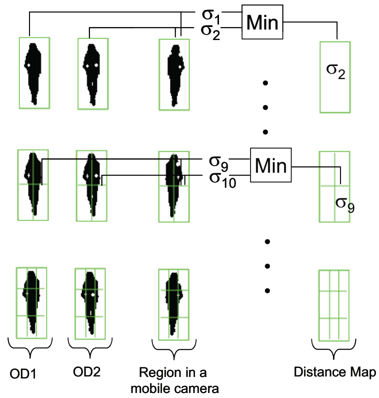

Fig. 3. Generation of the distance map between a set of observations of an object from a fixed camera and a region in the mobile camera.

4.2 The Approach

Given the ODs of an object, all possible regions in the image plane of a mobile camera are compared with the ODs. 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 ODs is computed to find the region with highest similarity. Each region is translated by 15% of its width or height, and scaled by 25% (6 scales are considered in this work).

4.3 Preprocessing: 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 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. "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 best candidates, *i.e.* regions with highest

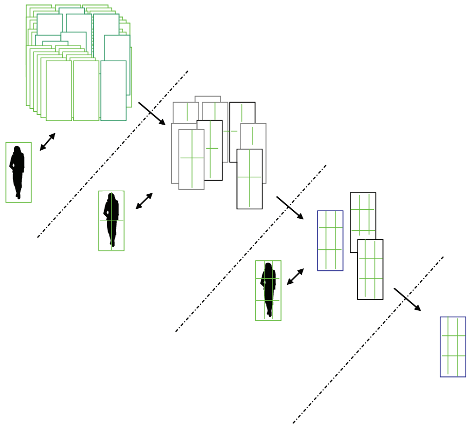

Fig. 4. A three stages cascade of coarse to fine descriptors. At each stage, a finer grid is filtering out remaining regions

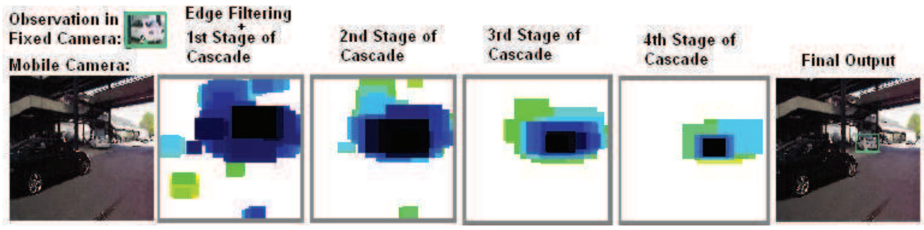

Fig. 5. Illustration of the most similar regions after each stage of the algorithm (in Jet format, white regions are the least similar and black ones the most)

similarity (top ρ% of the evaluated regions), remain (see figure 5). The percentage (ρ) of regions to keep after each stage can be fixed, or adaptive such that after each stage, the same percentage is kept:

$$N_r \times \rho^{N_s} = 1 \tag{6}$$

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.

$$\rho = N_r^{-1/N_s} \tag{7}$$

4.5 Updating the Observations

An object moving in a scene can have different appearances across time even from a fixed viewpoint. A set of relevant observations should be kept to detect the same object with a mobile camera at all time.

In order to cover the most different appearances of an object, the most dissimilar observations are kept. As a result, if an object does not have a similar appearance in the mobile camera with the current observation, it might have a better similarity with an older observation.

Let D be the set of observations of an object, and m the number of observations to keep:

$$D = \{OD_i, OD_2, ..., OD_m\} \tag{8}$$

We define the "set similarity" operator as the sum of all distances between the ODs of a set:

$$\sigma_{set}(D) = \sum_{\forall k,l \in D} \sigma(OD_k, OD_l) \tag{9}$$

Initially, the set D correspond to the m first observations of the object. Then, given a new observation OD_n, $m+1$ choices of the set D are possible, referred to as D_p:

$$D_p = \{D_1, ..., D_{m+1}\} = \begin{array}{c} \{\{OD_n, OD_2, ..., OD_m\}, \\ \{OD_1, OD_n, ..., OD_m\}, \\ ..., \\ \{OD_1, OD_2, ..., OD_n\}, \\ \{OD_1, OD_2, ..., OD_m\}\} \end{array} \tag{10}$$

The set with the most dissimilarity (highest σ_{set}) is kept:

$$D_u = \arg \max_{\forall D_i \in D_p} \sigma_{set}(D_i) \tag{11}$$

where D_u is the new updated set of observations.

5 Performance Evaluation

5.1 Data Sets

Indoor and outdoor data sets have been used [11]. Each data set is composed of video sequences captured concurrently by a fixed and a mobile camera from 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 images are recorded at 25fps with a resolution of 320×240. Figures 1 and 11 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.

5.2 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. In the second one, random rigid objects in the scene are selected to prove generalization of the approach to any objects of interest. Only objects present in the view of the mobile camera are selected in the fixed camera. Hence, the performance of the system is quantitatively measured by computing the percentage of correct detected and matched objects in the mobile camera, %TP. It is clear that the false positive rate is simply its complementary, hence does not need to be reported.

5.3 Region Descriptors

First, each object is described with a single region descriptor to select the best ones for the remaining evaluation. The following region descriptors are evaluated:

- Covariance of:
  - (x, y, I, I_x, I_y): as in [5]
  - (x, y, I_x, I_y): to measure impact of the intensity value
  - (x, y, I_{xx}, I_{yy}): the 2^{nd} order derivatives
  - (x, y, mg, θ): gradient magnitude and angle
  - $(x, y, I, I_x, I_y, I_{xx}, I_{yy})$: the grayscale values, 1^{st} and 2^{nd} order derivatives
  - $(x, y, I, I_x, I_y, mg, \theta)$: the grayscale values, 1^{st} order derivatives and gradient magnitude with angle
  - $(x, y, I, I_{xx}, I_{yy}, mg, \theta)$: the grayscale values, 2^{nd} order derivatives and gradient magnitude with angle
  - $(x, y, I, I_x, I_y, I_{xx}, I_{yy}, mg, \theta)$: set of all features
- Histogram of Oriented Gradient with 8, 12, and 16 bins

Color features are not considered since Alahi *et al.* in [5] show that it degrades the performance.

Figure 6 presents the %TP over all the descriptors on both data sets. Interestingly, when pedestrians are considered, the HOGs perform as good as the best covariance descriptors and even slightly better with 16 number of bins. On the second data set, for rigid objects, it is not the case.

Three regions descriptors based on the covariance matrix and the three HOGs are selected to continue the evaluation process:

- Covariance of:
  - (x, y, I, I_x, I_y): since it gave the best performance in [5].
  - $(x, y, I, I_x, I_y, mg, \theta)$: since it gave the best performance on both data sets when 7 features are selected.
  - $(x, y, I, I_x, I_y, I_{xx}, I_{yy}, mg, \theta)$: since it gave the best performance among other features set.
- Histogram of Oriented Gradient with 8, 12, and 16 bins.

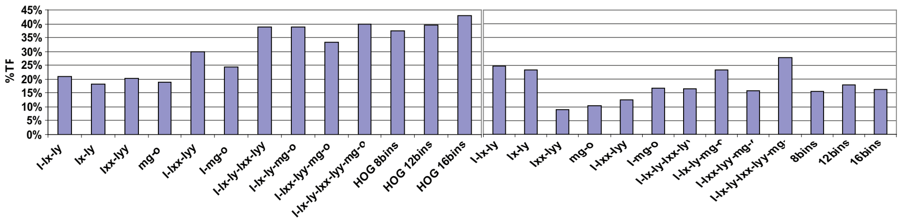

Fig. 6. %TP for various region descriptors on data set 1 (left-side), and data set 2 (right-side).

5.4 Detection and Matching Process

In average, 85% of the regions are filtered out with the proposed preprocessing step without removing the true positive regions. In addition, without such preprocessing, the performance of the descriptors are decreased by 23% in average. Reducing the search space, increases the likelihood to find the right region.

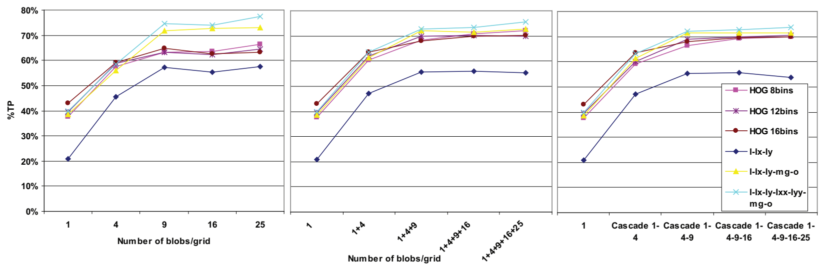

Fig. 7. %TP on the 1st Data set (pedestrians only) for various region descriptors based on 3 Object descriptors: Left hand-side:A single grid of n blobs (here n=1,4,9,16,25). Middle column: collection of grids. Right hand-side: Cascade of grids.

Considering a region with a single descriptor is not enough. Local information is lost in the global behavior. Figures 7 and 8 present 3 strategies: first, an object is described by a single grid (first graph). Various number of blobs per grid are considered. Increasing the number of blobs, increases the performance over all descriptors reaching a limit. Second, an object is described by a collection of grids. The final similarity measurement is the sum of the measurements over all the grids. With covariance matrices, considering several grids does not give better performance than simply considering the grid with higher number of blobs, although, for HOG, it is the case. The proposed cascade of grids leads to very similar performance with much lower computation cost. The number of

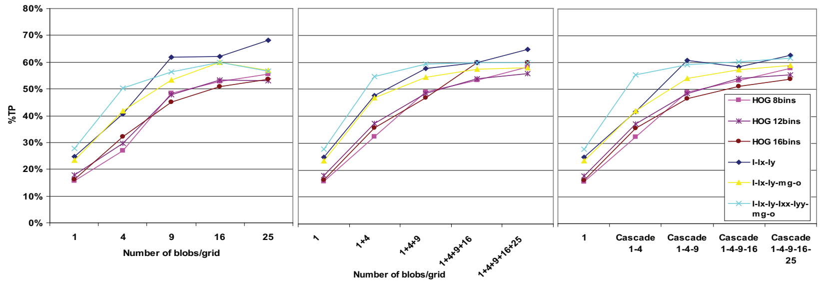

Fig. 8. %TP on the 2nd Data set (rigid objects only) for various region descriptors based on 3 different Object descriptors.

descriptors to compute is much less than the previous two approaches. Figure 9 presents the performance of the cascade of descriptors for various ρ with respect to the number of region descriptor computed.

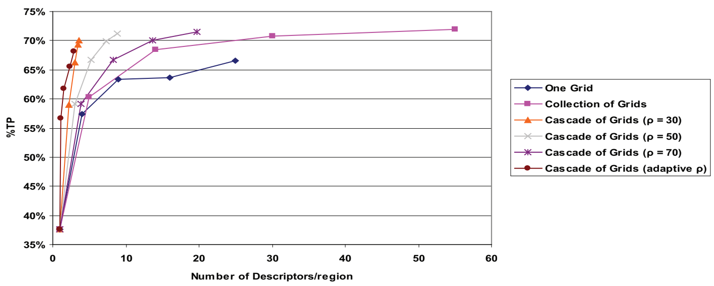

Fig. 9. %TP with respect to the number of region descriptors needed

Figure 10 presents the performance of the approach if several regions in the mobile camera are kept as matching the object of interest. Considering two or three regions is enough to increase the performance. Further work could classify those candidate regions as matching or not the object of interest by evaluating the posterior probability: checking if those regions match the same object in the fixed camera.

Finally, considering several observations increases the performance of the system. It is not relevant to give a quantitative results since it depends on the behavior of the objects. Nevertheless, by keeping three observations the performance increases by 7%. Moving objects are much better detected.

Figure 11 presents the output of our proposed system (5 stages cascade of HOG of 8 bins) on both data sets.

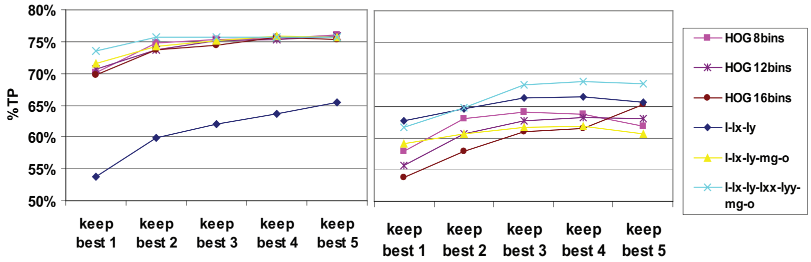

Fig. 10. %TP with respect to the number of best match kept on data set 1 (left-side), and data set 2 (right-side).

Fig. 11. Random Examples of the proposed system. 1st and 3rd column: objects of interest seen in fixed camera. 2nd and 4th column: corresponding detected objects in the mobile camera (output of the proposed approach)

6 Conclusions

A system is presented to detect and match any objects detected by a fixed camera in the image plane of a mobile camera. The performance of various object descriptors are measured both quantitatively and qualitatively. The proposed approach based on edge filtering, and the cascade of region descriptors outperform other approaches or lead to similar performance with much less computation cost, making it feasible for a real-time application. Although, the object descriptor based on HOG have slightly lower performance than those based on the covariances, they are much faster to compute and compare. Indeed, even with the fast implementation proposed by [7] to compute the covariance matrices, they are still expensive to measure similarity if many features are used. Future work will combine those descriptors to evaluate the posterior probability of the detected regions.

References

1. Suard, F., Rakotomamonjy, A., Bensrhair, A., Broggi, A.: Pedestrian Detection using Infrared images and Histograms of Oriented Gradients. In: Procs. IEEE Intelligent Vehicles Symposium 2006, Tokyo, Japan (2006) 206–212
2. Gavrila, D.: Pedestrian detection from a moving vehicle. (2000) II: 37–49
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. (2005) I: 886–893
4. Tuzel, O., Porikli, F., Meer, P.: Human Detection via Classification on Riemannian Manifolds. Proc. CVPR (2007) 1–8
5. Alahi, A., Marimon, D., Bierlaire, M., Kunt, M.: 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)
6. McCahill, M., Norris, C.: Cctv in london. (2002)
7. Tuzel, O., Porikli, F., Meer, P.: Region covariance: A fast descriptor for detection and classification. Proc. 9th European Conf. on Computer Vision (2006)
8. Porikli, F., Tuzel, O., Meer, P.: Covariance Tracking using Model Update Based on Lie Algebra. IEEE Conf. on Computer Vision and Pattern Recognition (2006)
9. Forstner, W., Moonen, B.: A metric for covariance matrices. Qua vadis geodesia (1999) 113–128
10. Comaniciu, D., Ramesh, V.: Real-time tracking of non-rigid objects using mean shift (2003) US Patent 6,590,999.
11. Videos available at: http://ltswww.epfl.ch/ alahi/data.htm