A Master-Slave Approach to Detect and Match Objects Across Several Uncalibrated Moving Cameras

Alexandre Alahi [*] [†] Pierre Vandergheynst [†]
Michel Bierlaire [*] Murat Kunt [†]

March 10, 2009

Report TRANSP-OR 090309
Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne
transp-or.epfl.ch

[*]TRANSP-OR, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland
[†]LTS, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

Abstract

Most multi-camera systems assume a well structured environment to detect and match objects across cameras. Cameras need to be fixed and calibrated. In this work, a novel system is presented to detect and match any objects in a network of uncalibrated fixed and mobile cameras. A master-slave system is presented. Objects are detected with the mobile cameras (the slaves) given only their observations from the fixed cameras (the masters). No training stage and data are used. Detected objects are correctly matched across cameras leading to a better understanding of the scene.

A cascade of dense region descriptors is proposed to describe any object of interest. Various region descriptors are studied such as color histogram, histogram of oriented gradients, Haar-wavelet responses, and covariance matrices of various features. The proposed approach outperforms existing work such as scale invariant feature transform (SIFT), or the speeded up robust features (SURF). Moreover, a sparse scan of the image plane is proposed to reduce the search space of the detection and matching process, approaching nearly real-time performance. The approach is robust to changes in illuminations, viewpoints, color distributions and image quality. Partial occlusions are also handled.

1 Introduction

Visual cameras are now installed in major cities[1] and integrated into many devices such as phones or vehicles. Such deployment of cameras in fixed and moving platforms has promoted the need to develop a novel framework to automatically detect and match objects in such a mixed network of cameras.

In a surveillance application, the use of data provided by all cameras capturing a given scene, leads to a better understanding of the objects of interest. Object identification (*e.g.* face recognition) or behavior analysis (*e.g.* facial expression) need high resolution features. 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 such high resolution features. In a safety context, car manufacturers and institutions are interested in detecting potential collision of cars with pedestrians in urban areas [2]. For that purpose

[1] In 2002, approximately four millions just for the UK [1]

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

they have mounted cameras on cars. Those cameras could collaborate with the fixed cameras installed in the cities to better detect pedestrians. Finally, the proposed system can also be used to help the navigation of a mobile robot.

Most multi-camera systems assume a well structured environment. Cameras need to be fixed and calibrated [3, 4, 5]. Moving objects are detected by modeling the background of the scene [6]. The foreground points extracted by each camera are projected in a common reference given a homography or a fundamental matrix estimated at calibration step [3]. Then, objects are detected and matched in a common reference plane. However, these systems fail for uncalibrated and moving cameras.

Object detection with mobile cameras addresses the problem from the view point of pattern classification [7, 8, 9]. A set of features is extracted from a large number of training samples to train a classifier. Thousand of observations of the objects of interest are required. However, only objects

present in the training data can be detected.

In this work, a novel multi-camera system is proposed based on a master-slave approach. Objects are detected with a mobile camera (from now on called slave) given observations from a fixed camera (also called master in the rest of the paper). Detected objects are correctly matched across cameras. The proposed framework can be applied to any pair of uncalibrated cameras. It only supposes that objects are correctly detected in at least one view, the master view. Either a simple processing can be achieved in that view (*i.e.* foreground extraction with a fixed camera) [6], or an user can manually select an object (object query).

The proposed approach consists of two steps. First, objects of interest observed by a master are assumed to be present in the slave. Then, the best candidates are validated by our second step, the validation stage [10]. The presence of an object is detected in the field of view of the mobile camera without any training process or data. No calibration between the cameras is used. The detection and matching process is only based on the appearance of the objects across cameras.

In the next section, a brief review of existing region descriptors is given. After formulating the problem, sections 4 and 5 describe the proposed approach to robustly detect and match objects across cameras. An object descriptor is presented considering deformations occurring in the presented applications (*e.g.* safety, surveillance, or robot navigation), such as photometric deformation, or viewpoint changes (*i.e.* rotation around the vertical or horizontal axes). It is made of a cascade of region descriptors. In section 6, various region descriptors are evaluated such as the covariance matrices [11] of various features, the histogram of colors [12], the histogram of oriented gradients [13], the scale invariant feature transform (SIFT) [14], the speeded up robust features (SURF) descriptors [15], and the color interest points [16]. Sparse and dense descriptions of the objects are evaluated. Moreover, a sparse scan of the image plane is presented to reduce the search space of the detection and matching process, approaching nearly real-time performance. Experiments show that objects are successfully detected even if the cameras have significant changes in image quality, illumination, and viewpoint as illustrated in Figure 1. Partial occlusions are also handled.

2 Existing Region Descriptors

A wide assortment of region descriptors has been proposed in the literature to address specific goals. From monocular or multi-view tracking problems, to image retrieval, simple and complex descriptors have been used.

The most basic high dimensional descriptor is the vector of pixel intensities [17]. Cross-correlation can be used to compute the distance between the descriptors. Its high dimensionality leads to high computational complexity without being robust to geometric deformation. A natural alternative is to describe the distribution of the pixel intensities by histograms. It copes with translations and rotations. Striker and Orengo [12] quantizes the HSV color space instead of the RGB. They use 16 bins for Hue and 4 for the Saturation and Value to match images. The Bhattacharyya distance [18] or the L_2 norm can be used to compare the histograms. Color histogram can be sufficient for monocular tracking [18] but leads to poor performance in a multi-view system. It is vulnerable to bad camera calibration and illumination changes. The inter-camera illumination change can be modeled to reduce such an effect [19]. Nevertheless, in many applications, color histograms are not discriminative enough to match regions.

The covariance descriptor is presented by Tuzel *et al.* [11] to outperform histogram descriptors. For each pixel, a set of features is extracted. Alahi *et al.* in [20, 21] compare various set of features. The grayscale intensity, the RGB values, the norm of the first and second order derivatives, the gradient magnitude and its angle are considered. The pixel coordinates are integrated in the feature vector to consider the spatial information of the features. 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 is given by the distance proposed by Forstner and Moonen [22] summing the generalized eigenvalues of the covariances. Although, a fast method based on integral images exists to compute the covariance matrices [11], similarity measurement takes time. Therefore, it is interesting to evaluate other low complexity descriptors and compare them with the covariance descriptor.

Histograms of Oriented Gradients (HOG) are efficient to compute descriptors based on the first order derivatives of the image intensity. From these derivatives, a gradient field is computed assigning to each pixel a magnitude and an angle. A histogram is formed where each bin is the sum of all magnitudes with the same orientation in a given region. HOG has been

extensively used to detect pedestrians in static images [7, 13, 23]. It is also the key component of the descriptor proposed by Lowe in [14].

Lowe presents a method to extract feature points invariant to scale, rotation, substantial range of affine distortion, 3D viewpoint, illumination, and addition of noise: scale-invariant feature transform (SIFT) [14]. Scale-space extrema is detected by difference-of-Gaussian function. Histograms of gradient direction are assigned to keypoints and used to create the descriptors. Bay *et al.* propose an interest point detector and descriptor outperforming SIFT in terms of speed and accuracy: speeded up robust features (SURF) [15]. Their descriptor is based on the distribution of the Haar-wavelet responses within the interest point neighborhood. Their detector and descriptor don't use color information. Gabriel *et al.* in [16] consider color interest points. The R,G,B values and first-order derivatives of the (R,G,B) channels are considered to describe each interest point. Similarity between two regions is computed by summing the distance between IPs with shortest mahalanobis distance. However, interest point based matching perform poorly with noisy low resolution images (see section 6).

Other descriptors exist such as steerable filters [24], gaussian derivatives [25], complex filters [26], phase-based local features [27], and moment invariants [28]. However, according to Mikolajczyk and Schmid [29], their proposed descriptor, called gradient location-orientation histogram (GLOH), as well as SIFT descriptor, outperforms these descriptors. GLOH is a variant of SIFT computing the HOGs in a log-polar location grid and reducing the final descriptor size with principal component analysis (PCA). Nevertheless, it is computationally more demanding than SIFT.

In section 6, the performance of the best presented descriptors are compared. It can be seen that each of the presented descriptors performs poorly if our proposed scheme is not used.

3 A Master-Slave Object Detection and Matching Approach

3.1 Problem Formulation

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

Let y_i be a potential region in the slave. x and y_i are rectangular subsets of an image. A "Region Matching" operator is defined, Φ, which maps a region x to the N most similar regions in a given image I:

$$\Phi(x, I, N) = \{y_1, y_2, ..., y_N\} = Y \tag{1}$$

The precise notion of similarity will be described in section 4. The operator Φ is used to match an observation x from the master to the most similar regions in the slave:

$$\Phi(x, I_s, N_s) = \{y_1, y_2, ..., y_{N_s}\} = Y_x \tag{2}$$

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:

$$\Phi(y_i, I_m, N_m) = \{\hat{x}_1, ..., \hat{x}_{N_m}\} = \hat{X}_i \tag{3}$$

where I_m is the image plane of the master.

In order to validate if a detected region in the slave really matches the same object in the master, the dual problem is evaluated. If a region $\hat{x}_i$ coincides with x, then the corresponding y_i should be the region bounding object O in the slave (see Figure 2). If none of the $\hat{x}_i$ coincides with x, object O is probably not present in the view of the slave. Hence, an operator ϑ validates if a region y_i matches x:

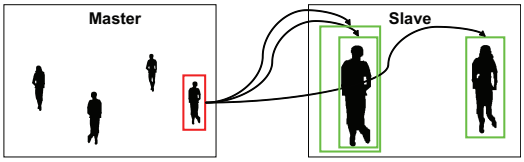
$$\vartheta(y_i|x, \Phi(y_i, I_m, N_m)) = \vartheta(y_i|x, \hat{x}_1, ..., \hat{x}_j) \in [0, 1]. \tag{4}$$

Moreover, the dynamic of the system can be considered to increase the performance. If results from previous frames are available, they can help the decision at the current frame. Two types of prior are useful. First, an object moving in a scene can have different appearances across time even from a fixed viewpoint. A set of relevant observations, $\{x^t, x^{t-i}, ..., x^{t-j}\}$, can be kept to detect the same object with a slave camera. The region matching operator becomes:

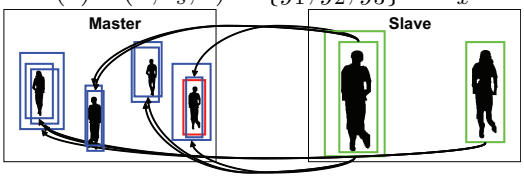
$$\Phi(\{x^t, x^{t-i}, ..., x^{t-j}\}, I_s, N_s) = Y_x^t \tag{5}$$

Second, the results of a detected object in the slave at previous frames, $\{y^{t-1}, y^{t-2}, ..., y^{t-k}\}$, can be used to detect the same object at the current frame, corresponding to a tracking approach:

$$\Phi(\{y^{t-1}, y^{t-2}, ..., y^{t-k}\}, I_s, N_s) = Y_{y^{t-1}}^t \tag{6}$$



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



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

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

As a result, the problem can be formulated as follows: find the region y_x^t in the mobile camera that maximizes $\vartheta(y_i^t|x^t, \Phi(y_i^t, I_m, N_m))$ for all $y_i^t \in \{Y_x^t, Y_{y^{t-1}}^t\}$:

$$y_x^t = \arg \max_{y_i^t \in \{Y_x^t, Y_{y^{t-1}}^t\}} \vartheta(y_i^t|x^t, \Phi(y_i^t, I_m, N_m)) \tag{7}$$

If such a y_x^t does not exist (all $\vartheta = 0$), it means that the object is not present in the image plane of the slave camera.

3.2 Detect, Track, and Validate

In order to solve the formulated problem, the approach can be summarized as follows. First, an object observed by a master is searched in the image plane of the slave with the Φ operator. The dual problem is evaluated to validate the candidates. Then, at the next frames, prior from the slave is first used to search the new frames. If the tracked region validates the dual problem, then the corresponding object is not searched given observation from the master. However, If none of the candidates match the initial object, the process is repeated without considering the prior from the slave. Algorithm 1 summarizes the approach and figure 3 illustrates an example of a single object detected and tracked in the slave camera.

Algorithm 1: Overview of the approach "detect, track, and validate"

Input: A set of objects $\{x_1, x_2, ..., x_p\}$ observed in the master camera

Output: Location $\{y_{x_1}, ..., y_{x_q}\}$ of the corresponding objects in the image plane of the slave camera

foreach *object x in the master* **do**

1. At $t = 1$, detect and validate:

$$y_x^1 = \arg \max_{y_i \in \{\Phi(x^1, I_s, N_s)\}} \vartheta(y_i | x^1, \Phi(y_i, I_m, N_m)) \tag{8}$$

2. At $t = 2$,
If y_x^1 exists, track and validate:

$$y_x^2 = \arg \max_{y_i \in \{\Phi(y_x^1, I_s, N_s)\}} \vartheta(y_i | x^2, \Phi(y_i, I_m, N_m)) \tag{9}$$

If y_x^2 or y_x^1 do not exist, detect given prior from the master and validate:

$$y_x^2 = \arg \max_{y_i \in \{\Phi(x^1, x^2, I_s, N_s)\}} \vartheta(y_i | x^2, \Phi(y_i, I_m, N_m)) \tag{10}$$

3. Repeat step 2 till object x is present in the master

end

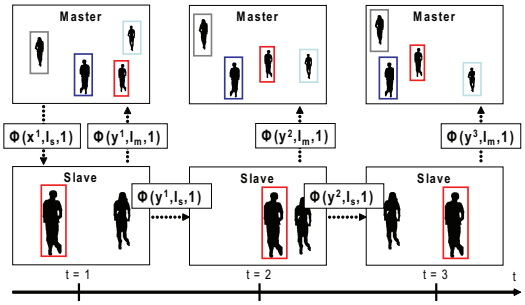


Figure 3: Illustration of the detect, track, and validate process. Only one object is validated and tracked across frames

4 Region Matching

4.1 Preliminary remarks

The region matching operator matches a region bounding an object of interest to the most similar regions in a different image plane. In this work, an object descriptor (OD) made of several region descriptors is created from the region bounding the object of interest. Then, a set of candidate regions in the given image are compared with the computed OD. Two strategies are evaluated to select the candidate regions in the image: a dense or sparse approach.

4.2 Dense selection

All possible regions in the given image are compared with the OD using a brute force search. A window of size proportional to the object bounding box scans the image plane at different scales[2].

A greedy pruning technique is applied to discard regions with very low similarity. 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.

[2]Six scales are used with a 25% scaling factor between two consecutive scales and a jumping step equivalent to 15% of the window size.

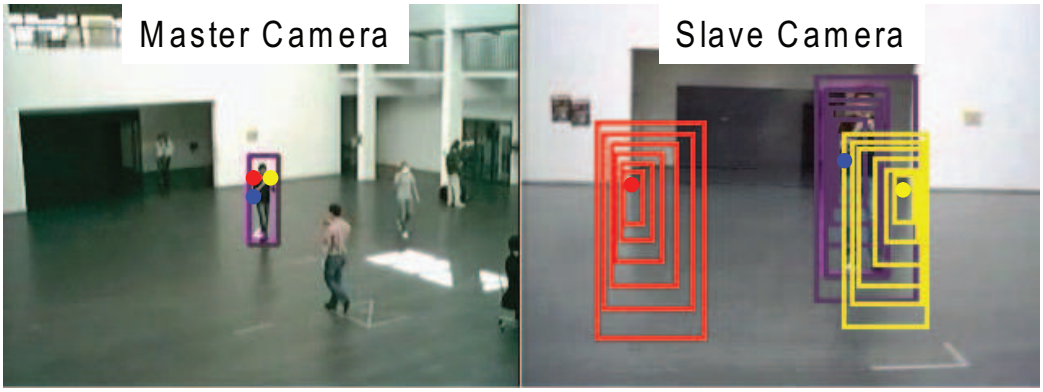


Figure 4: Illustration of an object described by 3 IP. The most similar IPs in the slave camera leads to 3*6 candidate regions

4.3 Sparse selection

A dense selection of candidate regions leads to thousand of regions to evaluate. In order to reduce the cardinality of such a set, a sparse selection given by the interest point (IP) extracted from the object of interest is proposed. All the interest points found on the object are matched to the most similar IPs in the image. Any existing detector and descriptor can be used. In this work, SURF [15] is used to detect and describe the IPs due to its low computational cost.

Each IP extracted from the object is represented by its coordinates with respect to the center of the bounding box. Therefore, a matched IP corresponds to a bounding box with the same spatial coordinates with respect to the center of the candidate region (up to a scale[3]). Figure 4 illustrates the approach.

In section 6, both strategies, *i.e.* dense and sparse selection of the candidate regions are compared.

4.4 A Collection of Grids of Descriptors

An object descriptor (OD) is proposed taking into account local and global information. It is a collection of grids of region descriptors. Each grid segments the object into a different number of sub-rectangles of equal sizes

[3]Six different scales are also used.

(referred to as blobs in the rest of the paper). Grids of finer blob size describe local information whereas grids of coarse blob size describe a more global behavior.

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, the best β percent. In this way, blobs belonging to the background can also be discarded (see figure 5).

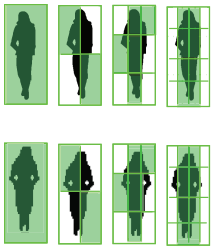


Figure 5: A collection of grids of descriptors. Top row is the object of interest. Bottom row is a region to compute similarity. Colored blobs are kept to compute the global distance ($\beta = 0.5$)

4.5 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.* larger 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 remain, *i.e.* regions with highest similarity, top ρ% of the evaluated regions.

The parameter ρ can be fixed (typically 30%)or chosen such that after each stage the same percentage is kept and one region remains after N stages:

$$N_r \times \rho^N = 1 \tag{11}$$

$$\rho = N_r^{-1/N} \tag{12}$$

where N_r is the total number of regions in the image plane to compare with the object descriptor, and N is the total number of stages to use.

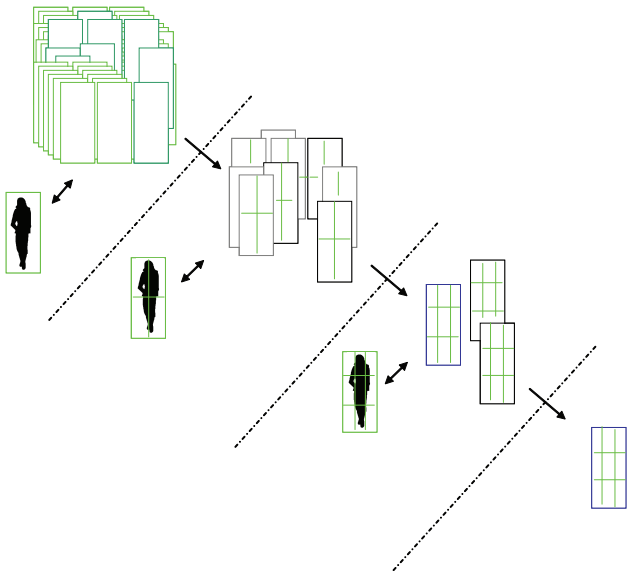


Figure 6: A three stages cascade of coarse to fine descriptors

Figure 7 illustrates the remaining regions with their similarity after each stage.

4.6 Several Observations

The master can consider several observations from the object of interest. In fact, only moving objects are treated since their appearance can change across time. Therefore, the ϕ operator can use several observations of an object in the matching process. Each observation leads to an OD. To compute the similarity of a region in the given image, the minimum distance, σ_r, between each blob of the grids is selected among all ODs leading to a distance map (see figure 8).

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 with the current observation, it might have a better similarity with an older observation.

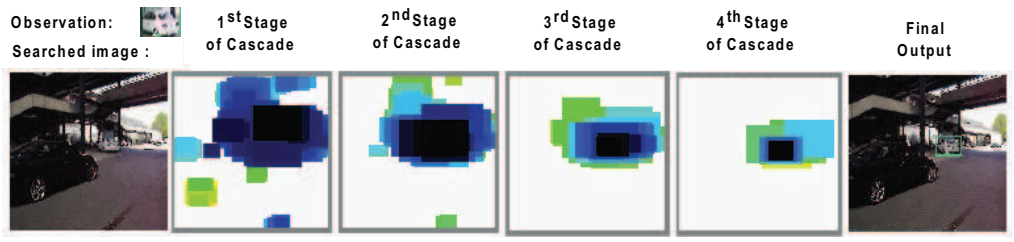


Figure 7: Illustration of the most similar regions after each stage of the algorithm (in Jet format, white regions are the least similar and black ones the most)

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

$$D = \{OD_i, OD_2, ..., OD_m\}. \tag{13}$$

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

$$\sigma_{set}(D) = \sum_{\forall k,l \in D} \sigma(OD_k, OD_l) \tag{14}$$

Initially, the set D corresponds 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:

$$D_p = \{D_1, ..., D_{m+1}\} = \begin{array}{c} \{\{OD_n, OD_2, ..., OD_m\}, \\ \{OD_1, OD_n, ..., OD_m\}, \\ ..., \\ \{OD_1, OD_2, ..., OD_n\}, \\ \{OD_1, OD_2, ..., OD_m\}\} \end{array} \tag{15}$$

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

$$D_u = \arg \max_{\forall D_i \in D_p} \sigma_{set}(D_i) \tag{16}$$

where D_u is the new updated set of observations.

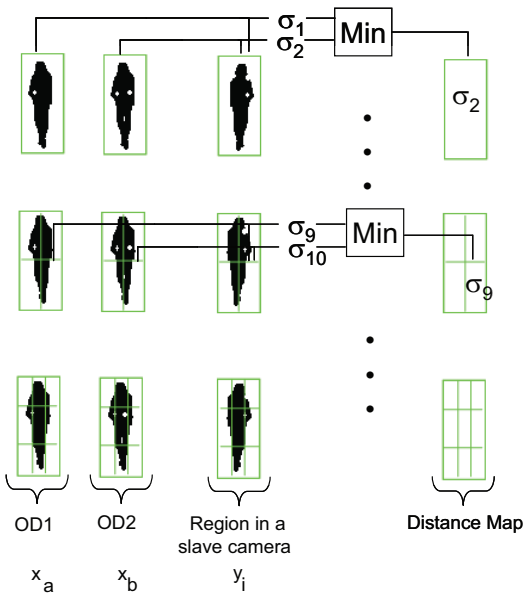


Figure 8: Generation of the distance map between a set of observations of an object from a master camera and a region in the slave camera.

5 Region Validation

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

$$\varsigma_l(x, \hat{x}_i) = 1 - (\frac{1-O}{1-c_1}w_o + \frac{1-C}{1-c_2}w_c + \frac{D_c}{c_3}w_d) \tag{17}$$

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 9)

- D_c measures the similarity of the center of two bounding boxes. The smallest the euclidian distance between the center, the highest D_c.

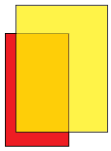


Figure 9: Illustration of the bounding boxes x (in red) and $\hat{x}_i$ where $C \approx 0.75$, $O \approx 0.4$

Note that by choosing $\varsigma(x, \hat{x}_i) > 0$ if and only if C and $O > 30\%$ and $D_c < 0.75 * max(width_x, height_x)$ leads to good enough results.

A weight $w_.$ is associated with each factor to emphasize priority. In this work, focus is first on a high cover of x , then a similar center of mass, finally $\hat{x}_i$ should not be too big with respect to x (decent O)[4].

A linear ς_l may be too sensitive to differences. The logistic operator is used to reduce sensitivity to two regions overlapping with a slight difference:

$$\varsigma(x, \hat{x}_i) = \frac{1}{1 + c_1 e^{-c_2 \cdot O}} w_o + \frac{1}{1 + c_1 e^{-c_2 \cdot C}} w_c + \frac{1}{1 + c_1 e^{-c_2 \cdot D_c}} w_d \tag{18}$$

c_1 and c_2 are the parameters of the logistic function.

Figure 10 presents an example of the value obtained with ς and ς_l.

Figure 10: 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:

$$\vartheta(y_i|x, \Phi(y_i)) = \max_{\hat{x}_i \in \Phi(y_i)} \varsigma(x, \hat{x}_i) \times w(y_i) \tag{19}$$

[4] $w_c = 0.5$, $w_d = 1/3$, and $w_o = 1/6$

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 4.4):

$$w(y_i) = \frac{\phi(x, y_i)}{\max_{y_j \in \Phi(x)} \phi(x, y_j)} \tag{20}$$

where $\phi(x, y_i)$ is the similarity measurement defined in section 4.4.

6 Performance Evaluation

6.1 Data Sets

Indoor and outdoor data sets have been used. Each data set is composed of video sequences captured concurrently by a fixed and a mobile camera from the same scene[5]. 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 25 fps with a resolution of 320×240.

The data sets 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. A rough temporal synchronization of the cameras is used (few frames delay) similar to the delay that can occur in real-world applications.

6.2 Experiments

Thousands of objects are selected within the fixed cameras, *i.e.* the masters, to find correspondence in the mobile cameras, the slaves. Pedestrians and random rigid objects in the scene are selected to prove the generality of the approach.

The performance of the system is quantitatively evaluated by computing the precision (*i.e.* number of true positives divided by the sum of true positives and false positives) and recall (*i.e.* number of true positives divided by the sum of true positives and false negatives) measures. A true positive is an object correctly detected in a slave camera and correctly matched to the corresponding object in the master camera.

[5]The videos sequences with their ground truth data (in xml format) can be found at http://lts2www.epfl.ch/~alahi/data.htm

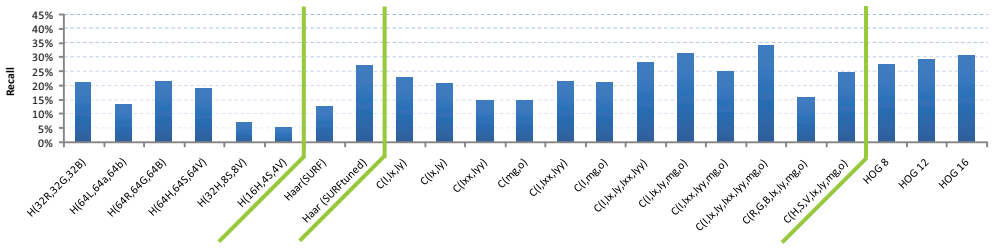


Figure 11: Recall for various region descriptors

To evaluate the performance of the region matching operator, only objects that are present in the view of the mobile camera are searched. Hence, the number of false positives is equal to the number of false negatives, leading to a similar recall and precision measures. However, to compute the performance of the full approach (detect, track, and validate), all the objects of interest in the master camera are selected, *i.e.* all moving objects and some static objects such as signs and cars. All the objects are searched in the slave camera even if they are not present in the field of view of the camera. The proposed approach should detect only objects present in the slave camera and locate them.

6.3 Region Matching

After studying the literature and considering their relevant results, the descriptors presented in table 1 are studied for the region matching operator. First, the color histogram is evaluated as a benchmark of the simplest low cost descriptor. Then, HOG descriptor is considered since Mikolajczyk and Schmid conclude that gradient based descriptors (*i.e.* GLOH, SIFT) outperform other descriptors such as steerable filters [24], gaussian derivatives [25], complex features [26], phase-based local features [27], and moment invariants [28]. Haar-wavelet responses are also analyzed since Bay *et al.* obtained better results with such descriptor than HOG based. Experimental results showed that Haar-wavelet responses are very sensitive to the choice of the filter size and the sampling grid. First, the same choices as Bay *et al.* is tested. Then, by changing the parameter to a finer grid size and a bigger filter size, we reached better performance (referred to as Haar SURF tuned). Finally,

Region Descriptors	
Histogram of Color	64 bins for RGB, HSV, or Lab 32 bins for RGB 32 bins for H, 8 bins for S, V 16 bins for H, 4 bins for S, V
HOG	8 bins 12 bins 16 bins
Haar-wavelet responses	SURF distribution [15] SURF distribution tuned
Covariance	(x, y, I_x, I_y) (x, y, I_{xx}, I_{yy}) (x, y, mg, θ) (x, y, I, I_x, I_y) $(x, y, I, I_x, I_y, I_{xx}, I_{yy})$ $(x, y, I, I_x, I_y, mg, \theta)$ $(x, y, I, I_{xx}, I_{yy}, mg, \theta)$ $(x, y, I, I_x, I_y, I_{xx}, I_{yy}, mg, \theta)$ $(x, y, R, G, B, I_x, I_y, I_{xx}, I_{yy})$ $(x, y, H, S, V, I_x, I_y, I_{xx}, I_{yy})$

Table 1: Summary of the region descriptors evaluated for the region matching operator. x and y are the pixel coordinates, I the grayscale value, I_x and I_y the 1^{nd} order derivatives, I_{xx} and I_{yy} the 2^{nd} order derivatives, mg and θ the gradient magnitude and angle.

the covariance descriptor are exhaustively evaluated for various feature sets since Tuzel *et al.* [11] introduced such descriptor to outperform histogram descriptors. All these descriptors are intensively studied for various schemes.

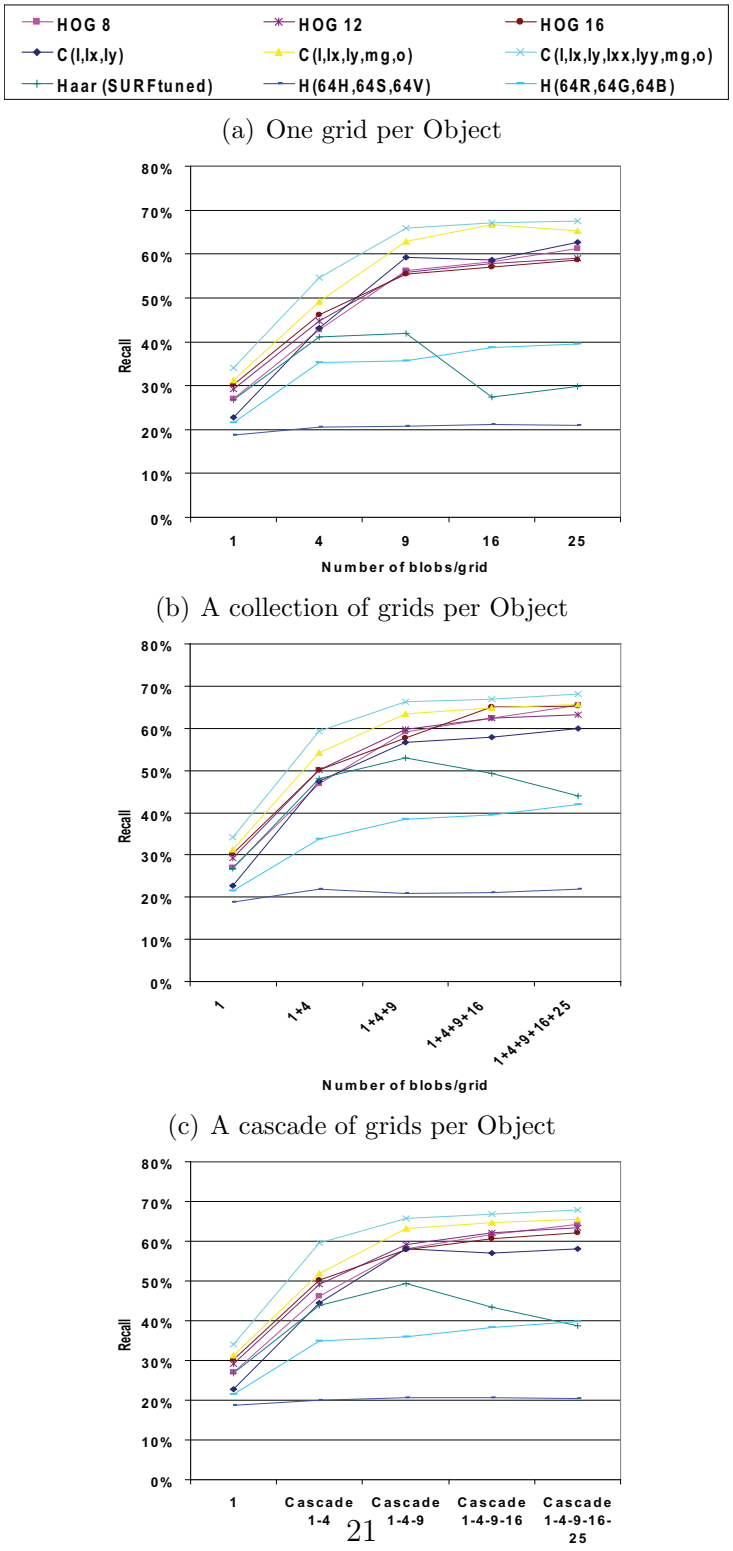
For the sake of clarity, only the best performing descriptors from table 1 are presented in the remaining study. Nevertheless, the performance of all descriptors is presented in figure 11 for the simplest scheme: an object is described by a single descriptor with a brute force search, *i.e.* a dense selection of the candidate regions. Color features perform poorly with histogram and covariance descriptor. Since sensing devices are different, the color distribution is also changed. Hence, color is not the right feature to use. Increasing the number of features increases the performance of the covariance descriptor. The HOGs perform almost as good as the best covariances. However, it is clear that describing an object with a single descriptor leads to very poor performance. Local information is lost in the global behavior. In this work, a cascade of grids of descriptors is proposed to tackle this problem. In order to validate such an approach, the proposed cascade approach is compared with other schemes (figures 12(a) to 12(c)) when a dense search is used.

First, an object is described by a single grid (figure 12(a)). Various numbers of sub-regions per grid are considered. Increasing the number of sub-regions increases the performance with histogram of color, HOG, and covariance descriptors. The color histogram still performs poorly compared to others. Interestingly, the performance of the descriptor based on Haar-wavelet responses increases for a few set of coarse grids and decreases for much finer grids. The filter size and sampling grid are proportional to the sub-rectangle size. As mentioned previously, changing the filter size and sampling grid affects the performance. Hence such a decrease of performance can happen with fine grids (*i.e.* high number of small sub-rectangles).

Second, an object is described by a collection of grids (figure 12(b)). The final similarity measurement is the sum of the distances over all the grids. Considering global and local information increases the performance of all the descriptors reaching a limit.

Finally, figure 12(c) shows that the proposed cascade of grids leads to a very similar performance as the collection of grids but with a much lower computation cost. The number of descriptors to compute is much less than the previous two schemes. Figure 13 presents the performance of the cascade of descriptors for various ρ (refer to section 4.5) with respect to the number of region descriptors needed.

Similarity between two regions is computed by summing the β most simi-



(a) One grid per Object

(b) A collection of grids per Object

(c) A cascade of grids per Object

Figure 12: Recall for various region descriptors with 3 different schemes to describe an object based on a dense search.

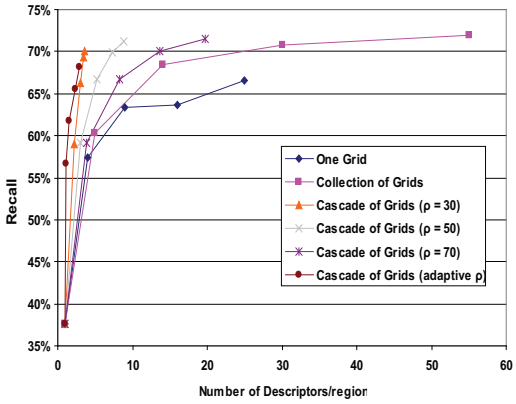


Figure 13: Recall with respect to the number of region descriptors needed

lar blobs (see section 4.4) within the grids of descriptors. Figure 14 illustrates the impact of β on the performance of the cascade of HOG and covariance descriptors. The mean performance between the two descriptors is plotted. The impact of β depends on the percentage of occlusion and photometric changes usually present in the data set. In our application, keeping 75% of the blobs to compute the overall similarity leads to the best performance.

All 3 strategies describe an object in a dense manner. However, an object can be described in a sparse representation obtained by the detected interest points. The state-of-the art interest points detector and descriptor, *i.e.* SIFT ([14]) and SURF([15]), are evaluated for comparison purposes. Figure 16 presents the matched interest points found across cameras with both approaches. The matched interest points do not correspond to the same objects where as our proposed cascade of covariances correctly matched the objects across cameras. Some objects, made of smooth regions, have very few interest points leading to an unfeasible matching process. In addition, the poor image quality affects the detection process. Gabriel *et al.* in [16] compared IP within the region of interest whereas SIFT and SURF matches the IPs over the whole image. By comparing IPs of two regions [16], the performance increases slightly. Various parameters are evaluated for SIFT, SURF, and the color interest points proposed by [16]. They all lead to poor results. The best configuration leads to a recall less than 15%. Therefore, the proposed dense representation of an object outperforms the sparse representation made by interest points. Nevertheless, the matched interest points

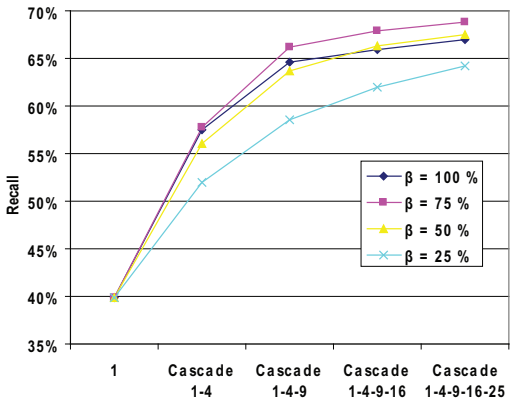
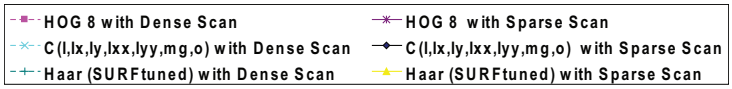


Figure 14: Mean recall of the cascade of HOG and covariance descriptors for various β value

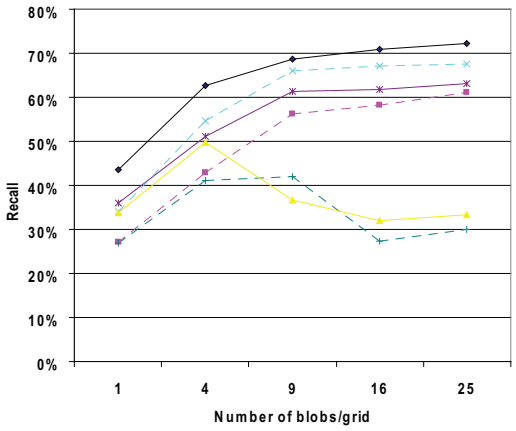
can be used to reduce the search space in the image plane. Hence, a sparse selection of the candidate regions is also evaluated in figure 15.

The proposed sparse selection of the candidate regions combined with the dense descriptor outperforms the approach based on a dense selection (see figure 15). The regions proposed by the interest points are good candidates. The reduced search space increases the likelihood to correctly detect and match the objects. The number of regions to keep after each stage of the cascade approach, ρ, can be increased with the sparse selection since few candidates are examined. With both selection, dense and sparse, 30 % of the regions are kept after each stage. Yet, increasing ρ can lead to better recall measures for a still low computational cost.

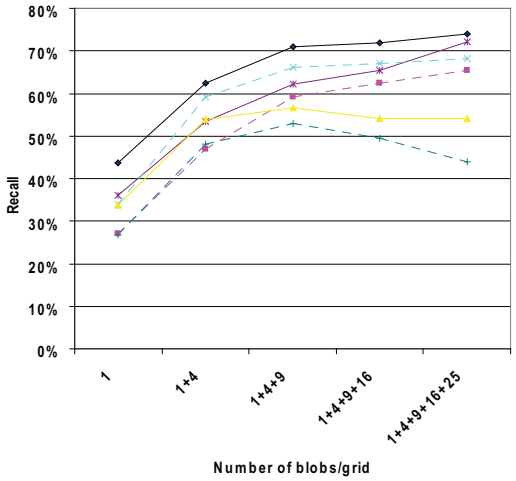
The computational cost of the different approaches to detect and match objects is also a crucial point. Table 2 summarizes the performance of the presented approaches. Note that the full cost of the approaches is measured, *i.e.* the cost of allocating memories, computing descriptors, comparing them, and creating and sorting lists of distances. The implementation is written in C/C++, without any optimization, and running on a Intel core 2 duo (2.8 GHZ with 4 GB RAM). Therefore, the absolute cost of an approach is not informative since it can be reduced, but the relative costs are interesting. The proposed sparse selection combined with the cascade of dense descriptors outperforms other approaches in terms of recall rate and computation cost. The cascade of covariances has the best recall rate closely followed by the cascade



(a) One grid per Object



(b) A collection of grids per Object



(c) A cascade of grids per Object

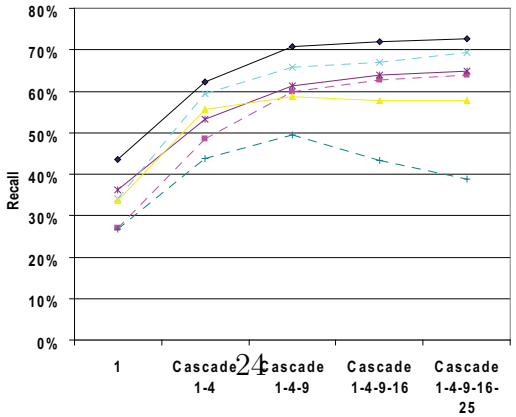


Figure 15: Recall for various region descriptors with 3 different schemes to describe an object based on a sparse search.

Region Descriptors	Recall	Cost
SIFT detector and descriptor [14]	< 0.15	250 ms
SURF detector and descriptor [15]	< 0.15	31 ms
Covariance descriptor [11]	0.20	4350 ms
Dense selection combined with		
Collection of HOGs	0.65	5588 ms
Cascade of HOGs	0.64	520 ms
Collection of Covariances	0.68	30 703 ms
Cascade of Covariances	0.69	2324 ms
Sparse selection combined with		
Collection of HOGs	0.72	558 ms
Cascade of HOGs	0.66	75 ms
Collection of Covariances	0.74	1042 ms
Cascade of Covariances	0.74	291 ms

Table 2: Recall rate and computation cost of various approaches

of HOG. However, HOG has a lower computational complexity. Although, integral images are not used to compute the HOG descriptors as opposed to the covariances, they still run faster. Hence, if computational complexity is an issue, the proposed cascade of HOG might be a viable alternative.

Qualitative results are given in figures 18 and 19. Objects with severe change of viewpoint or partial occlusion are correctly detected and match. Furthermore, a set of images has been randomly selected from a data set to illustrate the strength of the region matching operator on challenging images (see figure 20). It can be seen that very low resolution images made of smooth areas can also be detected and matched. Also, faces are correctly matched across images encouraging the use of the descriptor for other applications such as face identification.

Figure 17 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. The region validation scheme classifies those candidate regions as matching or not the object of interest by evaluating the dual problem.

(a) Matched SIFT interest points

(b) Matched SURF interest points

(c) Proposed approach (cascade of covariances)

Figure 16: Left-hand side are the objects observed in the fixed camera. Right-hand side are the image plane of the mobile cameras to be searched.

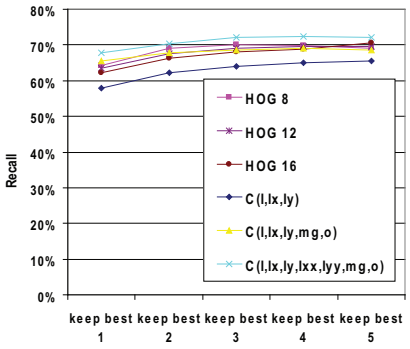


Figure 17: Recall with respect to the number of best match kept

6.4 Region Validation

The performance of the validation operator depends on two parameters: the number of regions to keep in the searched image plane, N_s, and the number of regions to keep in the dual problem, N_m (see section 3.1). Figure 21 presents the recall/precision graph for various $N_.$. They are compared with the greedy approach considering the best match proposed by the region matching operator as the matched object (labeled as "best match") without any validation process. With the proposed validation operator, setting $N_m = N_s = 2$, the number of false positives is decreased by 70 % while the true positive rate decreases by only $\sim 2\%$. In other words, it means that almost all the objects present in the view of the mobile camera are correctly classified as present while the others are correctly discarded with a success rate of 70 %. For $N_m = N_s = 3$, the number of false positives is reduced by half while the precision is reduced by less than 1%. Higher values for N_m and N_s do not necessarily lead to higher performance. Considering $N_s = 2$ and $N_m = 1$ is the best tradeoff for our application in terms of cost and precision rate.

In addition, a possible approach to reduce the false positives rate is 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 from the master and the regions in the slave camera. Figure 22 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 reduces the FP rate by 9% and reduce the TP rate by 11%. However, it is possible to

Figure 18: Examples of correctly detected and matched objects in indoor scene. 1st column: objects of interest seen in a fixed camera. 2nd column: corresponding detected objects in a mobile camera (output of the proposed approach)

Figure 19: Examples of correctly detected and matched objects in outdoor scene. 1st column: objects of interest seen in a fixed camera. 2nd column: corresponding detected objects in a mobile camera (output of the proposed approach)

Figure 20: Examples of images randomly selected from a data set. Left column are manually selected regions, and right column are the corresponding regions detected and matched by our proposed approach

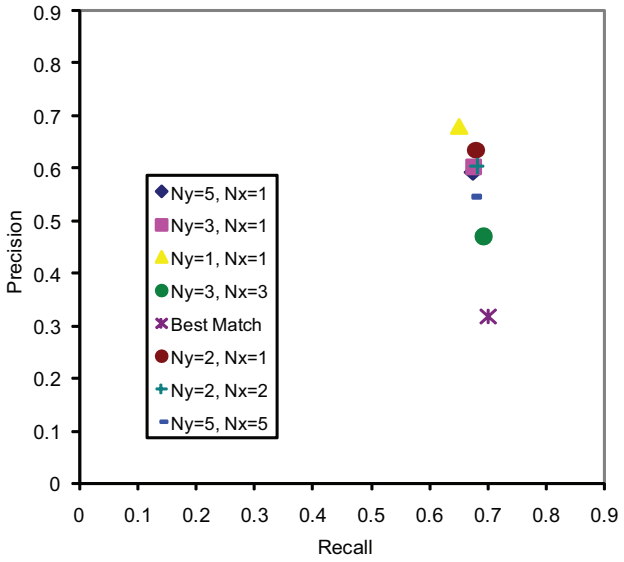


Figure 21: Recall/precision graph for various N_s and N_m.

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 23 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 is reduced by 50% while reducing the TP rate by 5% only. Therefore, the proposed approach can globally reduce the number of false positives by $75\% - 85\%$ for a decrease of 5-7% of the precision rate. This is feasible only because of the validation approach considering the dual problem. Without the validation scheme proposed in this work, a reduction of the false positive rate by 80% (with thresholding), would require a reduction of the precision rate by 50%. Figure 24 summarizes the overall performance with the different thresholding strategies.

When priors are available, the performance of the system increases. The gain in performance depends on the behavior of the objects. By keeping three observations from the master, the global performance increases by 7%. Moving objects are much better detected. Considering the prior from the slave increases the recall rate by 12% and the decreases the precision rate by 6%.

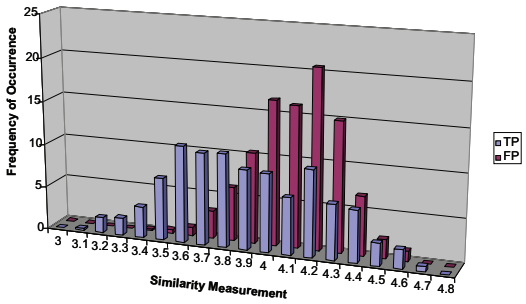
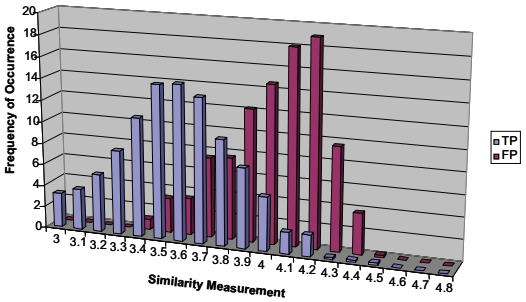
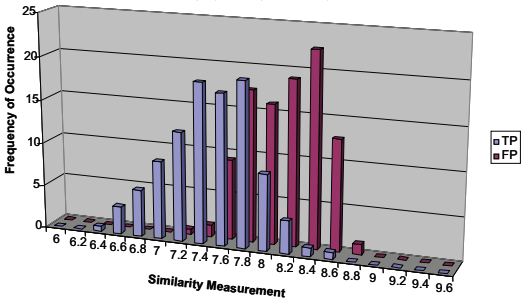


Figure 22: Histogram of the similarity measurements $\phi(x, y_i)$ for a set of TP and FP



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



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

Figure 23: Histogram of the similarity measurements in the validation process

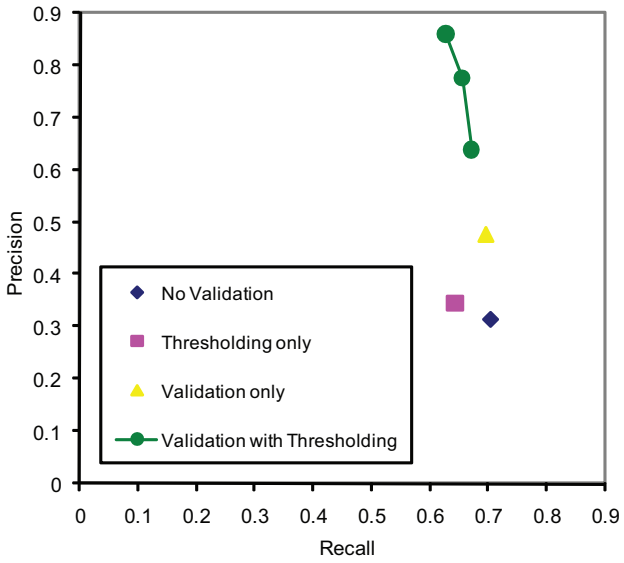


Figure 24: Overview of the recall/precision graph for various thresholding strategies

Qualitative results are given in figures 25 and 26. 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 27 presents some missed detections and few false positives.

7 Conclusions

A novel framework is presented to detect and match any objects across multiple uncalibrated cameras. It only supposes that objects are correctly detected in at least one camera, the master. Objects are successfully detected and matched with slave cameras even if the cameras have significant changes in image quality, illumination, and viewpoint. Partial occlusions are also handled. The proposed cascade of descriptors outperforms current state-of-the art approaches both qualitatively and quantitatively. It is generic to any region descriptors. Its strength has been proven for covariance and HOG descriptors. Furthermore, no training is necessary to detect the presence of any class of objects in the view of a mobile camera. The proposed validation process overcomes the use of a training data. Future work can evaluate the proposed cascade of descriptors for monocular tracking. Objects can be matched across frames based on the similarity measure obtained with the cascade of grids of descriptors. Moreover, the descriptor can also be used as a cascade of features for classification purposes.

Acknowledgments

We are very grateful to Pascal Frossard for useful comments. We also thank Samuel Egli and Hind Hannouch for practical discussions.

References

[1] M. McCahill, C. Norris, Cctv in london (2002).

[2] www.watchover-eu.org.

Figure 25: Correct detections and no false positives. First column: objects detected by a fixed camera. Second column: corresponding objects detected and matched with a mobile camera

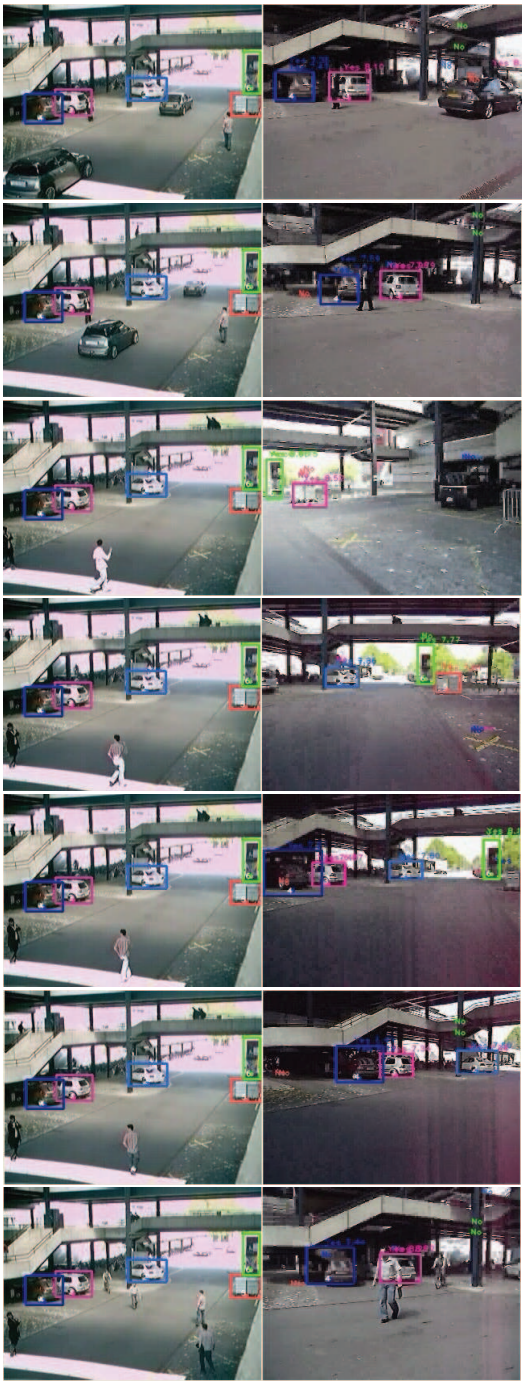


Figure 26: Correct detections and no false positives. First column: objects detected by a fixed camera. Second column: corresponding objects detected and matched with a mobile camera

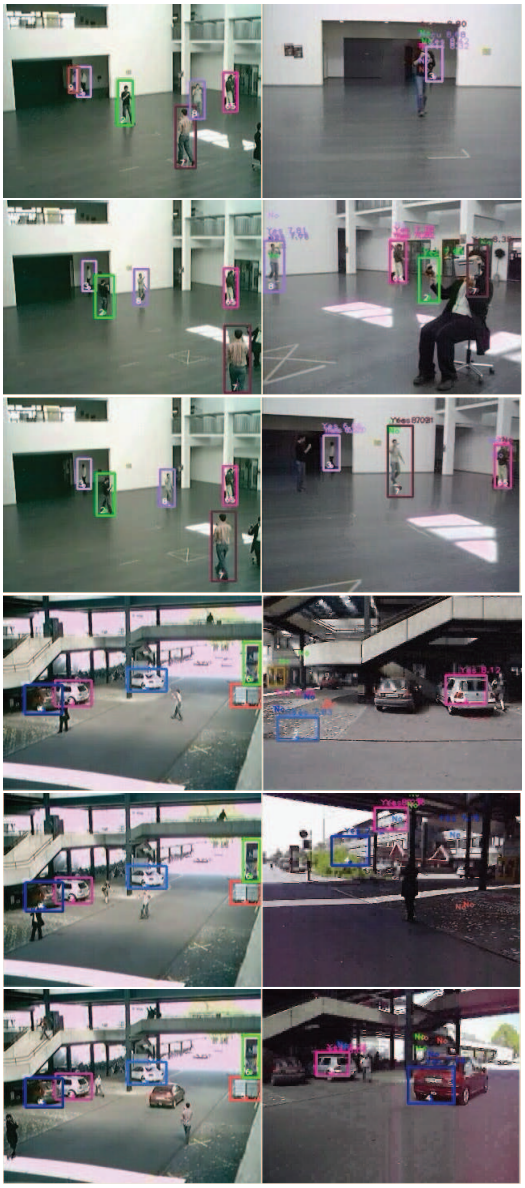


Figure 27: Some false positives and missed true positives. First column: objects detected by a fixed camera. Second column: corresponding objects detected and matched with a mobile camera

[3] Y. Caspi, D. Simakov, M. Irani, Feature-based sequence-to-sequence matching, International Journal of Computer Vision 68 (1) (2006) 53–64.

[4] S. Khan, M. Shah, A multiview approach to tracking people in crowded scenes using a planar homography constraint, in: European Conference on Computer Vision 2006, 2006, pp. IV: 133–146.

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

[6] F. Porikli, Achieving real-time object detection and tracking under extreme conditions, Journal of Real-Time Image Processing 1 (1) (2006) 33–40.

[7] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: CVPR05, 2005, pp. I: 886–893.

[8] O. Tuzel, F. Porikli, P. Meer, Pedestrian detection via classification on riemannian manifolds, IEEE Transactions on Pattern Analysis and Machine Intelligence 20.

[9] B. Leibe, N. Cornelis, K. Cornelis, L. Van Gool, E. Zurich, Dynamic 3D scene analysis from a moving vehicle, CVPR07.

[10] A. Alahi, P. Vandergheynst, M. Bierlaire, M. Kunt, Object Detection and Matching in a Mixed Network of Fixed and Mobile Cameras, in: The ACM International Conference on Multimedia, Vancouver, 2008, pp. 152–160.

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

[12] M. Stricker, M. Orengo, Similarity of color images, in: Proc. SPIE Storage and Retrieval for Image and Video Databases, Vol. 2420, San Jose CA USA, 1995, pp. 381–392.

[13] F. Suard, A. Rakotomamonjy, A. Bensrhair, A. Broggi, Pedestrian detection using infrared images and histograms of oriented gradients, in: Procs. IEEE Intelligent Vehicles Symposium 2006, Tokyo, Japan, 2006, pp. 206–212.

[14] D. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[15] H. Bay, T. Tuytelaars, L. Van Gool, SURF: Speeded Up Robust Features, Lecture Notes in Computer Science 3951 (2006) 404.

[16] P. Gabriel, J. Hayet, J. Piater, J. Verly, Object tracking using color interest points, IEEE AVSS (2005) 159–164.

[17] A. Rosenfeld, G. Vanderbrug, Coarse-fine template matching, IEEE Transactions on Systems, Man and Cybernetics 7 (1977) 104–107.

[18] D. Comaniciu, V. Ramesh, Real-time tracking of non-rigid objects using mean shift, uS Patent 6,590,999 (Jul. 8 2003).

[19] B. Prosser, S. Gong, T. Xiang, Multi-camera matching under illumination change over time, The 10th European Conference on Computer Vision (ECCV).

[20] A. Alahi, D. Marimon, M. Bierlaire, M. Kunt, A Master-Slave Approach for Object Detection and Matching with Fixed and Mobile Cameras, in: 15th IEEE International Conference on Image Processing, San Diego, 2008, pp. 1712–1715.

[21] A. Alahi, M. Bierlaire, M. Kunt, Object Detection and Matching with Mobile Cameras Collaborating with Fixed Cameras, in: The 10th European Conference on Computer Vision (ECCV), Marseilles, France, 2008, pp. 1542–1550.

[22] W. Forstner, B. Moonen, A metric for covariance matrices, Qua vadis geodesia (1999) 113–128.

[23] A. Shashua, Y. Gdalyahu, G. Hayun, Pedestrian detection for driving assistance systems: Single-frame classification and system level performance, in: IVS04, 2004, pp. 1–6.

[24] W. Freeman, E. Adelson, The design and use of steerable filters, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (9) (1991) 891–906.

[25] L. Florack, B. Ter Haar Romeny, J. Koenderink, M. Viergever, General intensity transformations and differential invariants, Journal of Mathematical Imaging and Vision 4 (2) (1994) 171–187.

[26] A. Baumberg, Reliable feature matching across widely separated views, in: Computer Vision and Pattern Recognition, Vol. 1, IEEE Computer Society; 1999, 2000, pp. 131–137.

[27] G. Carneiro, A. Jepson, Multi-scale phase-based local features, in: Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, Vol. 1, 2003, pp. 171–187.

[28] F. Mindru, T. Tuytelaars, L. Gool, T. Moons, Moment invariants for recognition under changing viewpoint and illumination, Computer Vision and Image Understanding 94 (1-3) (2004) 3–27.

[29] K. Mikolajczyk, C. Schmid, A Performance Evaluation of Local Descriptors, IEEE Transactions on Pattern Analysis and Machine Intelligence (2005) 1615–1630.