Cascade of Descriptors to Detect and Track Objects Across Any Network of Cameras

Alexandre Alahi[1,2], Pierre Vandergheynst[1], Michel Bierlaire[2], Murat Kunt[1]

Ecole Polytechnique Federale de Lausanne
[1]*Signal Processing Laboratory,* [2]*Transportation and Mobility Laboratory*
CH-1015 Lausanne - Switzerland

Abstract

Most multi-camera systems assume a well structured environment to detect and track objects across cameras. Cameras need to be fixed and calibrated, or only objects within a training data can be detected (e.g. pedestrians only). In this work, a master-slave system is presented to detect and track any objects in a network of uncalibrated fixed and mobile cameras. Cameras can have non-overlapping field-of-views. Objects are detected with the mobile cameras (the slaves) given only observations from the fixed cameras (the masters). No training stage and data are used. Detected objects are correctly tracked across cameras leading to a better understanding of the scene.

A cascade of grids of region descriptors is proposed to describe any object of interest. To lend insight on the addressed problem, most state-of-the-art region descriptors are evaluated given various schemes. The covariance matrix of various features, the histogram of colors, the histogram of oriented gradients, the scale invariant feature transform (SIFT), the speeded up robust features (SURF) descriptors, and the color interest points [1] are evaluated. A sparse scan of the cameras'image plane is also presented to reduce the search space of the localization process, approaching nearly real-time performance. The proposed approach outperforms existing works such as scale invariant feature transform (SIFT), or the speeded-up robust features (SURF). The approach is robust to some changes in illumination, viewpoint, color distribution, image quality, and object deformation. Objects with partial occlusion are also detected and tracked.

Key words: Object Detection, Object Tracking, Region Descriptors, Cascade of descriptors, Multi-View, Mobile cameras, Pedestrian Recognition

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 framework to automatically detect and track objects in such a mixed network of uncalibrated cameras. Since cameras can be moving, their views are often not overlapping. Objects need to be localized in each camera view, and tracked (i.e. re-identify) across views.

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. cameras held by pedestrians or placed in cars) 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 [3]. For that purpose they have mounted cameras on cars. Those cameras could collaborate with the fixed cameras installed in the cities to better detect pedestrians or any moving objects (such as animals).

[1]In 2002, approximately four million just for the UK [2]

Figure 1: Left column: objects of interest highlighted in a fixed (master) camera. Right column: Corresponding objects detected and tracked in a mobile (slave) camera by our proposed approach

Most multi-camera systems assume a well structured environment, cameras need to be fixed and calibrated [4, 5, 6]. Moving objects are detected by modeling the background of the scene [7]. The foreground points extracted by each camera are projected in a common reference given a homography or a fundamental matrix estimated during calibration steps [4]. Then, objects are detected and matched in a common reference plane

[8]. However, these systems fail to detect and match objects across uncalibrated and moving cameras.

Object detection with mobile cameras is usually solved with pattern recognition techniques [9, 10, 11]. A set of features such as Haar wavelet coefficients [12, 13], histogram of oriented gradient [9, 14] or covariance matrices of a set of features [10, 15], are extracted from a large number of training samples to train a classifier with a support vector machine [12, 16], or boosting approaches [17, 10]. Thousands of observations of the objects of interest are desired, and only objects present in the training data can be detected.

In this work, a multi-camera system is proposed based on a master-slave approach. Objects are detected with mobile cameras (from now on called slaves) given only observations from fixed cameras (masters). Detected objects are correctly matched across cameras. The proposed framework can be applied to any pair of uncalibrated cameras. It only assumes that objects are correctly detected in at least one view, the master view. Either simple processing can be achieved in that view (i.e. foreground extraction with a fixed camera) [7], or a user can manually select an object (object query). Cameras do not require overlapping field-of-views. A validation step [18] is proposed to evaluate the presence of an object in the views of the cameras. No calibration, neither training process or data is used. The detection and matching process is only based on the appearance of the objects across cameras.

An Object Descriptor (OD) is proposed to handle deformations occurring in the presented applications (e.g. safety, surveillance, or robot navigation) such as: *(i)* photometric deformation, *(ii)* viewpoint changes (i.e. rotation around the vertical or horizontal axes), *(iii)* object deformation (e.g. walking pedestrians), and *(iv)* partial occlusions. The OD is made of a cascade of grids of region descriptors. The detection and matching process is speeded-up by the interest point locations leading to a sparse search space. The state-of-the-art regions descriptors such as the covariance matrix [19] of various features, the histogram of colors [20], the histogram of oriented gradients [14], the scale invariant feature transform (SIFT) [21], the speeded-up robust features (SURF) descriptors [22], and the color interest points [1] are evaluated to provide insight on the object detection and tracking problem.

The rest of the paper is structured as follows. First, a review of existing region descriptors is given. Then, the proposed OD is presented in Section 3 followed by its localization strategy. The master-slave object detection and matching approach is described in Section 5. Finally, the detailed experimentation is presented in Section 7. The proposed system is evaluated given various schemes and strategies. 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 [23]. 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 [20] quantize 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 log-RGB [24], YCrCb [25], or Lab color spaces can also be used. Color histogram can be sufficient for monocular tracking [26] 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 [27]. Nevertheless, in many applications, color histograms are not discriminative enough to match or detect objects.

Another efficient-to-compute descriptor is the Histogram of Oriented Gradient (HOG). It is 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 derivatives, a gradient field is computed assigning to each pixel a magnitude $mg(x, y)$ and an angle $o(x, y)$:

$$mg(x, y) = \sqrt{I_x^2(x, y) + I_y^2(x, y)} \tag{1}$$

$$o(x, y) = \arctan\left(\frac{I_y(x, y)}{I_x(x, y)}\right) \tag{2}$$

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

HOG has been extensively used to detect pedestrians in static images [9, 14, 28]. It is also the key component of the descriptor proposed by Lowe in [21].

Wang *et al.* in [29] use HOGs that incorporate detailed spatial distribution of objects color across different body parts. Likewise, Gheissari *et al.* in [30] segment the body into salient parts and combine color and edgel histograms for appearance representation and person re-identification.

To compare the histograms, any distance measure can be used: Correlation, ℓ_1-norm , ℓ_2-norm , intersection, Chi-square, Bhattacharyya. We compare all those distances and conclude that Bhattacharyya [26] distance is performing either better or similar than other distances:

$$\sigma_r(H_1, H_2) = \sqrt{1 - \sum_i \frac{H_1(i).H_2(i)}{\sqrt{\sum_i H_1(i). \sum_i H_2(i)}}} \tag{3}$$

where $H(i)$ is the histogram value of bin i.

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) [21]. 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) [22]. Their descriptor is based on the distribution of the Haar-wavelet responses within the interest point neighborhood. Their detector and descriptor do not use color information. Gabriel *et al.* in [1] 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 7).

A more complex descriptor is the covariance descriptor. It is first presented by Tuzel *et al.* [19] to outperform histogram descriptors. For each pixel, a feature vector $\mathbf{f}_n$ is extracted. Alahi *et al.* in [31, 15] 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. Typically,

$$\mathbf{f}_n = (x, y, I, I_x, I_y)^T \tag{4}$$

with I the grayscale intensity, I_x and I_y the norm of the first order derivatives.

The pixel coordinates (x, y) are integrated in the feature vector to consider the spatial information of the features. The covariance of a region is computed as:

$$C_i = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{f}_n - \mathbf{m})(\mathbf{f}_n - \mathbf{m})^T \tag{5}$$

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 is given by the distance proposed by Forstner and Moonen [32] summing the generalized eigenvalues of the covariances:

$$\sigma_r(C_1, C_2) = \sqrt{\sum_i \ln^2 \lambda_i(C_1, C_2)} \tag{6}$$

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

Although, a fast method based on integral images exists to compute the covariance matrices [19], similarity measurement takes time.

Other descriptors exist such as steerable filters [33], gaussian derivatives [34], complex filters [35], phase-based local features [36], and moment invariants [37]. However, according to Mikolajczyk and Schmid [38], 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 7, the performances of the best presented descriptors are compared.

3. Object Descriptor

3.1. A Collection of Grids of Descriptors

An 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 subrectangles of equal sizes (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 distances σ_i between corresponding blobs segmenting the grids. Since, many objects can be deformable, and some partially occluded, only the most similar blobs are kept, the best β percent. In this way, blobs belonging to the background can potentially be discarded as well(see Figure 2).

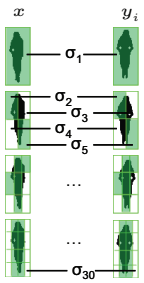


Figure 2: A collection of grids of descriptors. Left column is the object of interest. Right column is a region to compute similarity. Colored blobs are kept to compute the global distance ($\beta = 50\%$)

Let σ be the set of all distances:

$$\sigma = \{\sigma_1, \sigma_2, ..., \sigma_N\} \tag{7}$$

where N is the sum of all the blobs segmenting the grids.

We define the sorted set:

$$\sigma_s = \{\sigma_a, \sigma_b, \sigma_c, ...\} \tag{8}$$

where $\sigma_a < \sigma_b < \sigma_c < ...$ and $| \sigma_s | = N$.

Hence, the final similarity measurement is :

$$\phi(x, y_i) = \frac{1}{p}\sum_{k=1}^{p}\sigma_s\{k\} \tag{9}$$

where $\sigma_s\{k\}$ is the k^{th} distance of the sorted set σ_s, and $p = \lceil \beta N \rceil$.

Therefore, the final similarity measurement is the average of a sparse measurement of isolated blobs of various size and position.

3.2. Several Observations

An object of interest can have several observations. Typically, the appearance of moving objects can change across time from the same view-point. 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 3).

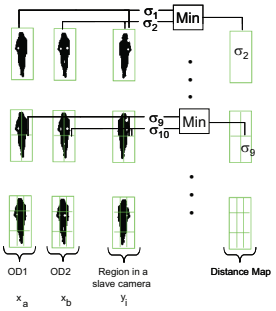


Figure 3: Generation of the distance map between a set of observations of an object from a master camera and a region in the slave camera.

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.

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

$$D = \{OD_i, OD_2, ..., OD_m\}. \tag{10}$$

We define the "similarity" distance σ_{set} as the sum of all distances between the ODs of a set:

$$\sigma_{set}(D) = \sum_{\forall k,l \in D} \sigma(OD_k, OD_l) \tag{11}$$

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:

$$D_p = \{D_1, ..., D_{m+1}\} = \begin{array}{c} \{\{OD_n, OD_2, ..., OD_m\}, \\ \{OD_1, OD_n, ..., OD_m\}, \\ ..., \\ \{OD_1, OD_2, ..., OD_n\}, \\ \{OD_1, OD_2, ..., OD_m\}\} \end{array} \tag{12}$$

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

$$D_u = \arg \max_{\forall D_i \in D_p} \sigma_{set}(D_i) \tag{13}$$

where D_u is the new updated set of observations.

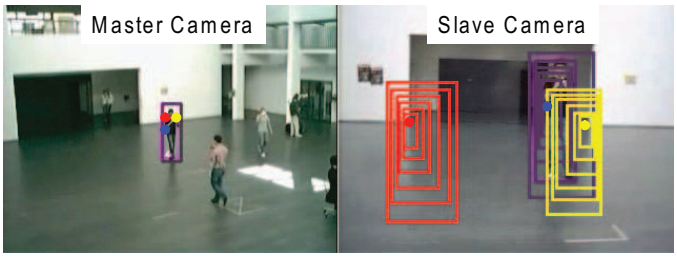


Figure 4: Illustration of an object described by 3 IP. The most similar IPs in the slave camera leads to 3*6 candidate regions only

4. Object Localization

4.1. Preliminary remarks

Given the proposed OD, i.e. the collection of grids of region descriptors, we are interested in localizing the most similar objects in the target image plane. Two strategies are evaluated to select the candidate regions in the target image: a dense or sparse approach.

4.2. Dense scan

All possible regions in the target image are compared with the OD (similar to a brute force search): a window of size proportional to the object bounding box scans the image plane at different scales. Six scales are used with a 25% scaling factor between two consecutive scales and a jumping step equivalent to 15% of the window size.

A basic pruning technique is applied to discard regions with very low similarity: the difference between the proportion of edges in two regions is used to 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. However, this pruning technique does not reduce the search space sufficiently, a further reduction is needed.

4.3. Sparse selection

A dense scan of the target image leads to thousands 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 [22] 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[2]). Figure 4 illustrates the approach.

In Section 7, both strategies, i.e. dense and sparse selection of the candidate regions are compared.

[2]Six different scales are also used.

4.4. Cascade of Coarse to Fine Descriptors

Comparing the collection of grids of region descriptors is computationally costly. Some regions can be easily discarded without knowing the local information. Therefore, an approach similar to a cascade of classifiers 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 (see Figure 5). 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:

$$N_r \times \rho^N = 1 \tag{14}$$

$$\rho = N_r^{-1/N} \tag{15}$$

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.

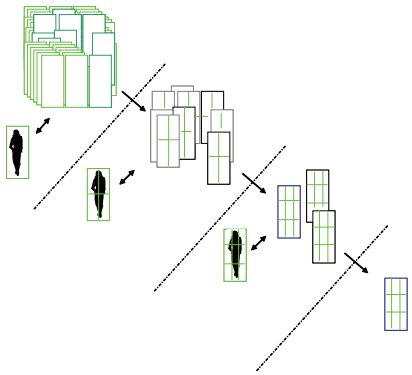


Figure 5: A three stages cascade of coarse to fine descriptors

Figure 6 illustrates the remaining regions with their similarity after each stage.

5. A Master-Slave Object Detection and Matching Approach

5.1. 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. An "Object localization" operator is defined, Φ, which maps a region x to the N most similar regions in a given image I:

$$\Phi(x, I, N) = \{y_1, y_2, ..., y_N\} = Y \tag{16}$$

First, the operator Φ is used to match an observation x from the master to the most similar regions in the slave:

$$\Phi(x, I_s, N_s) = \{y_1, y_2, ..., y_{N_s}\} = Y_x \tag{17}$$

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

$$\Phi(y_i, I_m, N_m) = \{\hat{x}_1, ..., \hat{x}_{N_m}\} = \hat{X}_i \tag{18}$$

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 7). 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:

$$\vartheta(y_i|x, \Phi(y_i, I_m, N_m)) = \vartheta(y_i|x, \hat{x}_1, ..., \hat{x}_j) \in [0, 1]. \tag{19}$$

Moreover, the dynamic of the system can be considered to increase the performance. If results from previous frames are available, they 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 object localization operator becomes:

$$\Phi(\{x^t, x^{t-i}, ..., x^{t-j}\}, I_s, N_s) = Y_x^t \tag{20}$$

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:

$$\Phi(\{y^{t-1}, y^{t-2}, ..., y^{t-k}\}, I_s, N_s) = Y_{y^{t-1}}^t \tag{21}$$

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

$$y_x^t = \arg \max_{y_i^t \in \{Y_x^t, Y_{y^{t-1}}^t\}} \vartheta(y_i^t|x^t, \Phi(y_i^t, I_m, N_m)) \tag{22}$$

If such a y_x^t does not exist (all $\vartheta < T$), it means that the object is not present in the image plane of the slave camera. The choice of the threshold T will be discussed in Section 7.5.

5.2. 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 22 illustrates an example of a single object detected and tracked in the slave camera.

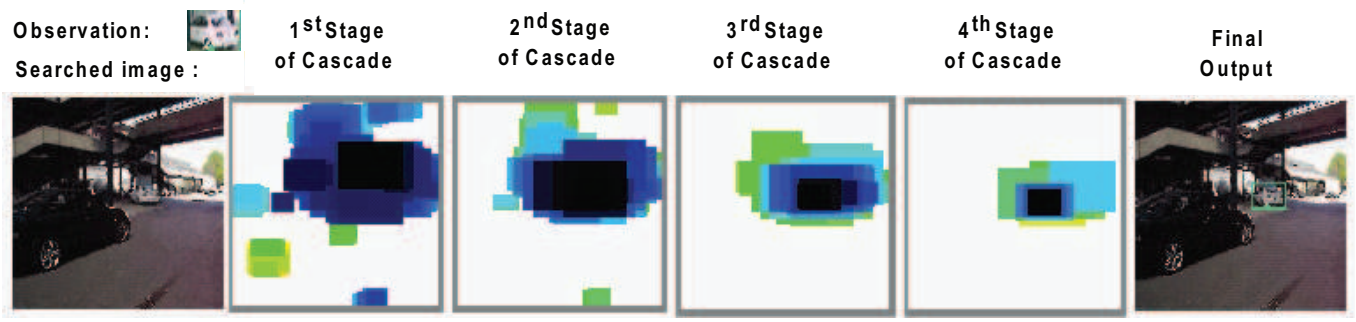


Figure 6: Illustration of the most similar regions after each stage of the algorithm (in Jet format, white regions are the least similar and black ones the most)

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

for *each object x in the master* **do**

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

$$y_x^1 = \arg \max_{y_i \in \{\Phi(x^1, I_s, N_s)\}} \vartheta(y_i | x^1, \Phi(y_i, I_m, N_m)) \quad (23)$$

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

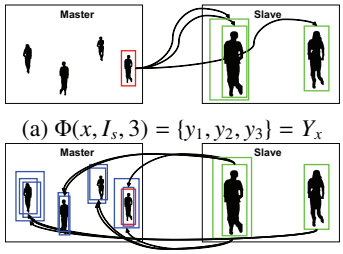
$$y_x^2 = \arg \max_{y_i \in \{\Phi(y_x^1, I_s, N_s)\}} \vartheta(y_i | x^2, \Phi(y_i, I_m, N_m)) \quad (24)$$

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

$$y_x^2 = \arg \max_{y_i \in \{\Phi(x^1, x^2, I_s, N_s)\}} \vartheta(y_i | x^2, \Phi(y_i, I_m, N_m)) \quad (25)$$

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

end



(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 7: 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.

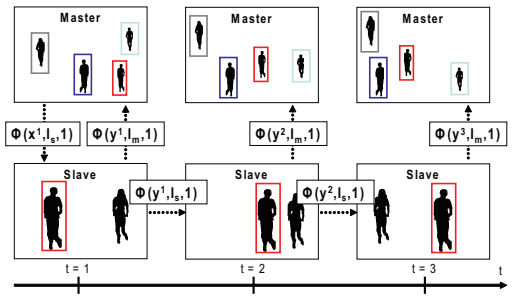


Figure 8: Illustration of the detect, track, and validate process. Only one object is validated and tracked across frames

6. 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:

$$\varsigma_l(x, \hat{x}_i) = 1 - \left(\frac{1 - O}{1 - t_1} w_o + \frac{1 - C}{1 - t_1} w_c + \frac{D_c}{t_2} w_d\right) \quad (26)$$

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 of $\hat{x}_i$ is covered by x. (see Figure 9)
- D_c measures the similarity of the center of two bounding boxes. The smallest euclidian distance between the center, the highest D_c.
- t_1 is the minimum amount of O and C required, and t_2 the minimum distance D_c.

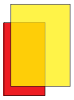


Figure 9: Illustration of the bounding boxes x (in red) and $\hat{x}_i$ where $C \approx 0.75$, $O \approx 0.4$

Note that we choose $\varsigma(x, \hat{x}_i) > 0$ if and only if C and $O > t_1$ and $D_c < t_2$: $t_1 = 0.3$ and $t_2 = 0.75 * max(width_x, height_x)$ leads to satisfactory 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)[3].

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:

$$\varsigma(x, \hat{x}_i) = \frac{1}{1 + c_1 e^{-c_2 \cdot O}} w_o + \frac{1}{1 + c_1 e^{-c_2 \cdot C}} w_c + \frac{1}{1 + c_1 e^{-c_2 \cdot D_c}} w_d \quad (27)$$

c_1 and c_2 are the parameters of the logistic function.

Figure plots the behavior of the logistic operator as opposed to the linear one. Figure 11 presents an example of the value obtained with ς and ς_l.

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

$$\vartheta(y_i|x, \Phi(y_i)) = \max_{\hat{x}_i \in \Phi(y_i)} \varsigma(x, \hat{x}_i) \times w(y_i) \quad (28)$$

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

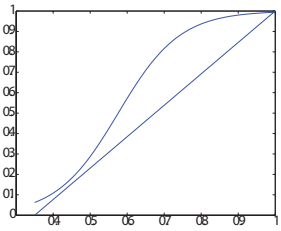


Figure 10: x-axis represents C or O; y-axis represents its contribution to ς and ς_l. It can be seen that for values of C or O close to 1, the contribution remains also almost 1 (full) for the logistic operator.

Figure 11: The linear ς_l gave 0.63 and the proposed ς gives 0.86

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 3.1):

$$w(y_i) = \frac{\phi(x, y_i)}{\max_{y_j \in \Phi(x)} \phi(x, y_j)} \quad (29)$$

where $\phi(x, y_i)$ is the similarity measurement defined in Section 3.1.

7. Performance Evaluation

7.1. Data Sets

Indoor and outdoor data sets have been used to evaluate the proposed master-slave approach. Each data set is composed of video sequences captured concurrently by a master and a slave camera from the same scene[4]. Masters are located at a height equivalent to the first floor of a building. Slaves 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 master and slave cameras. Sensing devices are also different. Indeed, slave cameras have a low-quality 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.

To further evaluate the performance of the proposed descriptor to track object identity, the VIPeR data set is used [5]. It contains hundreds of pedestrian image pairs taken from arbitrary viewpoints (45 to 180 degree view difference) under varying illumination conditions[39]. Hence, pedestrian recognition or re-identification can be evaluated.

7.2. Experiments

To evaluate the proposed master-slave framework, thousands of objects are selected within the masters, to find correspondence in 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.

[4] The video sequences with their ground truth data (in xml format) are available at: http://lts2www.epfl.ch/~alahi/data.htm

[5] The VIPeR dataset is available at: http://vision.soe.ucsc.edu/node/178

First, only objects present in the view of the slaves are searched, i.e. cameras have overlapping field-of-views (Section 7.4). Hence, the number of false positives is equal to the number of false negatives, leading to a similar recall and precision measures. Then, to compute the performance of the full approach (detect, track, and validate), all the objects of interest in the master are selected, i.e. all moving objects and some static objects such as signs and cars. All the objects are searched in the slaves even if they are not present in the field-of-view of the cameras (Section 7.5). The proposed approach should detect only objects present in the slave and locate them.

Finally, to measure the performance of the proposed descriptors to recognize or re-identify pedestrians given the VIPeR data set, we measure the cumulative matching characteristic (CMC) curve and similar to Gray *et al.* in [25]. The CMC curve is calculated by selecting pedestrians in the first (master) camera view and finding the ranking of their match in the collection of pedestrians observed by the second (slave) camera. All the pedestrians observed by the slave are sorted given their distance to the probe pedestrian in the master.

7.3. Region Descriptors Evaluation

After studying the literature and considering their relevant results, the state-of-the-art region descriptors are compared for the object descriptor.

First, the color histogram is evaluated as a benchmark of the simplest low cost descriptor. Various color spaces and bin partitions are studied. The RGB, log-RGB, HSV, YCrCb, Lab and opponent color space are evaluated. The Bhattacharyya distance is used to compare histograms since it performs better than other distances (such as ℓ_1-norm , ℓ_2-norm , intersection, Chi-square).

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 [33], gaussian derivatives [34], complex features [35], phase-based local features [36], and moment invariants [37]. Eight to sixteen bin partitions are compared.

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.* are tested. Then, by changing the parameter to a finer grid size and a bigger filter size, we reach better performance (referred to as Haar SURF tuned).

Finally, the covariance descriptor is exhaustively evaluated for various feature vectors since Tuzel *et al.* [19] introduced such a descriptor to outperform histogram descriptors. All the presented color channels, the first and second order derivatives, the magnitude and angle of the gradient, are used to form the feature vector. Many combinations of features are tested.

All these descriptors are intensively studied for various schemes and parameters. Table 1 illustrates the best performing ones.

Region Descriptors	
Histogram of Color	64 bins for RGB: H(64R,64G,64B) 32 bins for RGB: H(32R,32G,32B) 64 bins for log-RGB: H(log 64R,64G,64B) 32 bins for log-RGB: H(log 32R,32G,32B) 64 bins for YCrCb: H(64Y,64Cr,64Cb) 8 bins for YCrCb: H(8Y,8Cr,8Cb) 64 bins for HSV: H(64H,64S,64V) 32 bins for H, 8 bins for S, V: H(32H,8S,8V) 16 bins for H, 4 bins for S, V: H(16H,4S,4V) 64 bins for Lab: H(64L,64a,64b)
HOG	8 bins: HOG 8 12 bins: HOG 12 16 bins: HOG 16
Haar-wavelet responses	SURF distribution [22]: Haar(SURF) SURF distribution tuned: Haar(SURFtuned)
Covariance	$C(x, y, I_x, I_y)$ $C(x, y, I_{xx}, I_{yy})$ $C(x, y, mg, o)$ $C(x, y, I, I_x, I_y)$ $C(x, y, I, I_x, I_y, I_{xx}, I_{yy})$ $C(x, y, I, I_x, I_y, mg, o)$ $C(x, y, I, I_{xx}, I_{yy}, mg, o)$ $C(x, y, I, I_x, I_y, I_{xx}, I_{yy}, mg, o)$ $C(x, y, R, G, B, I_x, I_y, I_{xx}, I_{yy})$ $C(x, y, R, G, B, I_x, I_y, mg, o)$ $C(x, y, H, S, V, I_x, I_y, I_{xx}, I_{yy})$ $C(x, y, Y, Cr, Cb, I_x, I_y, I_{xx}, I_{yy})$ $C(x, y, Y, Cr, Cb, I_x, I_y, mg, o)$

Table 1: Summary of the best performing region descriptors evaluated for the object descriptor. 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 o the gradient magnitude and angle.

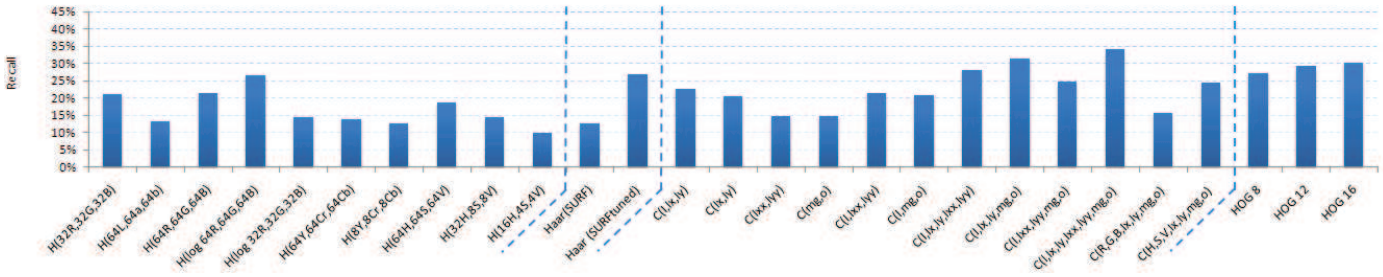


Figure 12: Recall for various region descriptors

7.4. *Object Detection and Tracking across Overlapping Field-Of-Views*

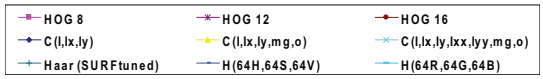
For the sake of clarity, only the best performing descriptors from table 1 are presented in the remaining study. Nevertheless, the performance of some descriptors is presented in Figure 12 for the simplest scheme: an object is described by a single descriptor with a dense scan of the candidate regions. 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 13(a) to 13(c)) when a dense scan is used.

First, an object is described by a single grid (Figure 13(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).

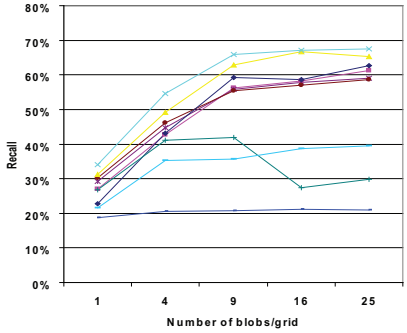
Second, an object is described by a collection of grids (Figure 13(b)). The final similarity measurement is the sum of the distances over all the grids. Considering global and local information increases the performance of all the descriptors reaching a limit.

Finally, Figure 13(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 14 presents the performance of the cascade of descriptors for various ρ (refer to Section 4.4) with respect to the number of region descriptors needed.

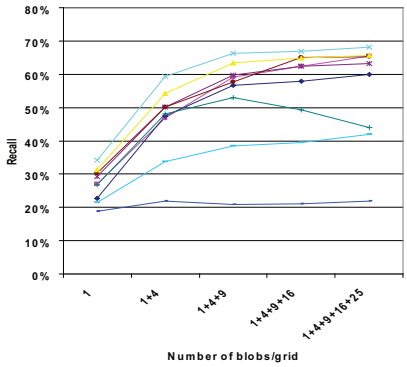
Similarity between two regions is computed by summing the



(a) One grid per Object



(b) A collection of grids per Object



(c) A cascade of grids per Object

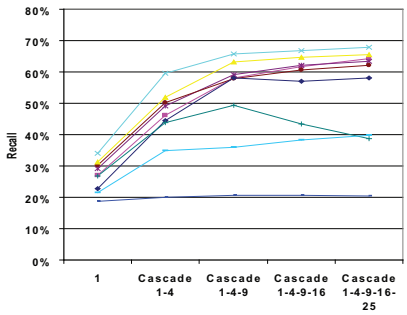


Figure 13: Recall for various region descriptors with 3 different schemes to describe an object based on a dense scan.

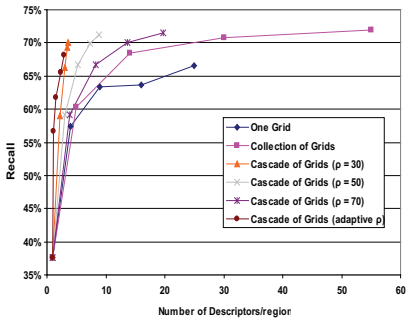


Figure 14: Recall with respect to the number of region descriptors needed

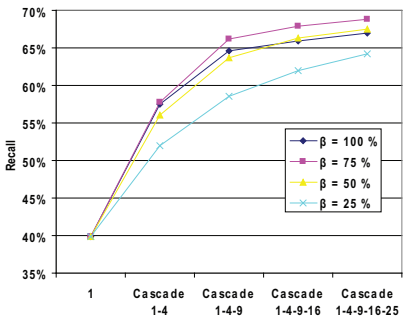


Figure 15: Mean recall of the cascade of HOG and covariance descriptors for various β value

β most similar blobs (see Section 3.1) within the grids of descriptors leading to a sparse similarity measurement. Figure 15 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, object deformation, view-point 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 (grids of region descriptors). 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 ([21]) and SURF([22]), 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 [1] compared IP within the region of interest whereas SIFT and SURF matches the IPs over the whole image. By comparing IPs of two regions [1], the performance increases slightly. Various parameters are evaluated for SIFT, SURF, and the color interest points proposed by [1]. 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 can be used to reduce the search space in the image plane as explained in Section 4.3. A sparse selection of the candidate regions is evaluated in Figure 17.

(a) Matched SIFT interest points

(b) Matched SURF interest points

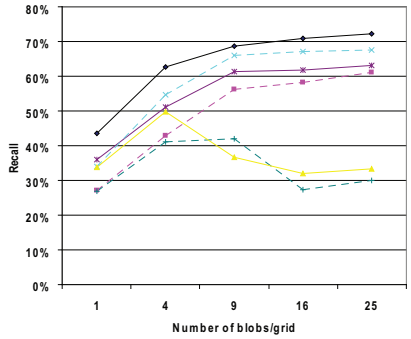
(c) Proposed approach (cascade of covariances)

Figure 16: Left-hand side are the objects observed in the master. Right-hand side are the image plane of the slave to be searched.

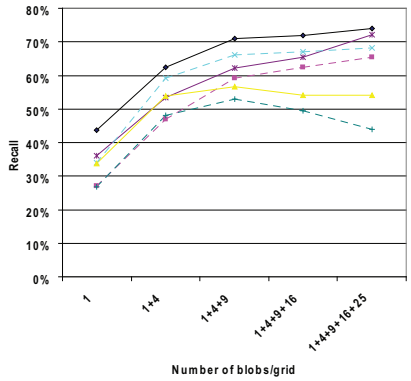
The proposed sparse selection of the candidate regions combined with the dense descriptor outperforms the approach based on a dense selection (see Figure 17). The regions proposed by the interest points are good candidates. The reduced search space increases the likelihood to correctly detect and match 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 with a still low computational cost.

The computational cost of the different approaches to detect

(a) One grid per Object



(b) A collection of grids per Object



(c) A cascade of grids per Object

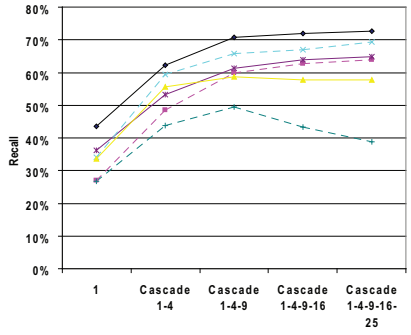


Figure 17: Recall for various region descriptors with 3 different schemes to describe an object based on a sparse search.

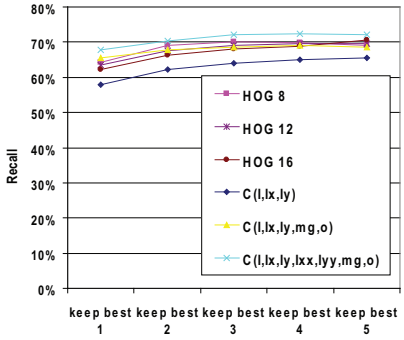


Figure 18: Recall with respect to the number of best match kept

and match objects is also a crucial point. Table 2 summarizes the performance of the presented approaches to search for one object in the slave camera. 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 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 19, 20, 21, and 22. 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 object localization operator on challenging images (see Figure 23). 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 object identification. In Section 7.6, we explicitly address the intra-category recognition problem showing how the cascade of grids of descriptors outperform other schemes.

Figure 18 presents the performance of the approach if several regions in the slave are kept to locate the object of interest. Considering two or three regions is enough to increase the performance. The validation operator classifies those candidate regions as either matching or not the object of interest by evaluating the dual problem.

7.5. Object Detection and tracking across Non Overlapping Field-Of-Views

Since we are dealing with mobile slave cameras, some of them can have non overlapping field-of-views with the master

Region Descriptors	Recall	Cost
SIFT detector and descriptor [21]	< 0.15	250 ms
SURF detector and descriptor [22]	< 0.15	31 ms
Covariance descriptor [19]	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

Figure 19: Examples of correctly detected and matched objects in indoor scene. 1st column: objects of interest seen in a master. 2nd column: corresponding detected objects in a slave

Figure 20: Examples of correctly detected and matched objects in outdoor scene. 1st column: objects of interest seen in a master. 2nd column: corresponding detected objects in a slave

Figure 21: First column: objects detected by a master. Second column: corresponding objects detected with the slave given only the observation from the master.

Figure 22: Given the left-hand side image, the same pedestrian is detected regardless of changes in viewpoint and scale.

Figure 23: 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

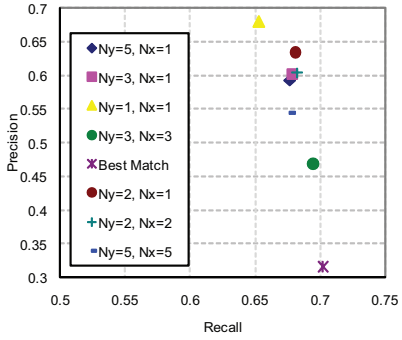


Figure 24: Recall/precision graph for various N_s and N_m.

cameras. An object observed by a master is searched for in all the slaves at all time frames even if the object might not be present in the views of the slaves. The validation operator classifies the presence of the object in the views.

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 5.1). Figure 24 presents the recall/precision graph for various $N_.$. They are compared with the greedy approach considering the best match proposed by the object localization 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 slave 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 25 illustrates the histogram of the values obtained when the regions are correctly matched (TP) and the ones for the false positives (FP). There is no clear decision boundary. Typically, setting the threshold to 4.4 reduces the FP rate by 9% and reduces the TP rate by 11%. However, it is possible to threshold the similarity measurement $\phi(y_i, \hat{x}_i)$, or the sum $\phi(x, y_i) + \phi(y_i, \hat{x}_i)$ obtained in the validation process. Figure 26 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

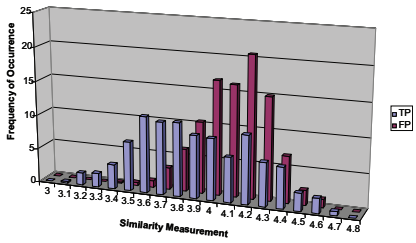
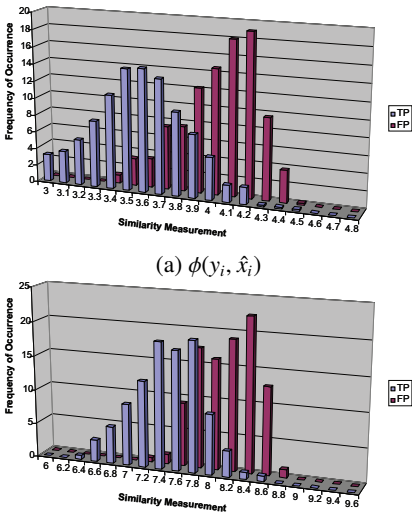


Figure 25: Histogram of the similarity measurements $\phi(x, y_i)$ for a set of TP and FP



(a) $\phi(y_i, \hat{x}_i)$

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

Figure 26: Histogram of the similarity measurements in the validation process

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 27 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 decreases the precision rate by 6%.

Qualitative results are given in figures 28 and 29. 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 master which are not present in the view of the slave do not generate false positives. Figure 30 presents some missed detections and few false positives.

7.6. People Recognition

To further evaluate the strength of the proposed cascade of region descriptors, we evaluate its performance to solve the people recognition problem. Note that no tuning, neither training

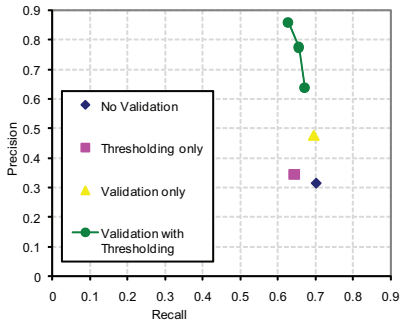


Figure 27: Overview of the recall/precision graph for various thresholding strategies

Figure 28: Correct detections and no false positives. First column: objects detected by a master. Second column: corresponding objects detected and matched with a slave

Figure 29: Correct detections and no false positives. First column: objects detected by a master. Second column: corresponding objects detected and matched with a slave

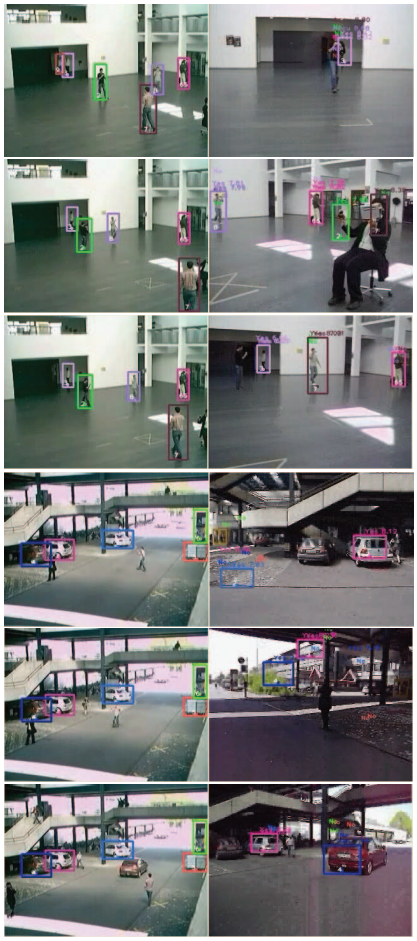


Figure 30: Some false positives and missed true positives. First column: objects detected by a master. Second column: corresponding objects detected and matched with a slave

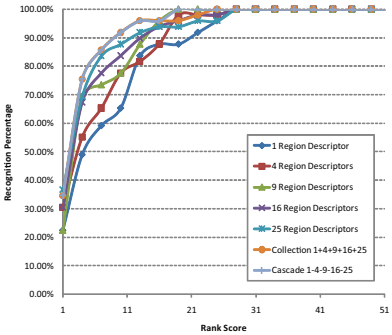


Figure 31: The CMC curve for color histogram given all 3 schemes. YCrCb color space with 64 bins per channel is used

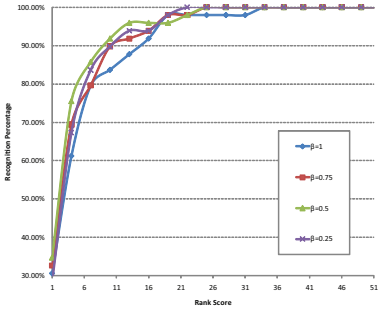


Figure 32: The CMC curve for color histogram given various β value

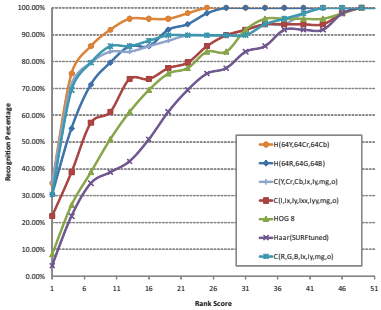


Figure 33: The CMC curve given 50 pedestrians present in the scene

is performed for such specific problem. The cascade of grids of descriptors (1-4-9-16-25) is compared with the collection of grids (5 grids of 1,4,9,16, and 25 region descriptors), and the single grids (1, 4, 9, 16, or 25 region descriptors) of color histograms in Figure 31. All various color spaces and bin partition are compared. The YCrCb color space with 64 bins per channel performed best with the Bhattacharyya distance. Using finer grid size increases the performance. For such specific recognition problem, the collection and cascade schemes also outperform the single grid of descriptors. Although the cascade and collection of descriptors perform similarly, the cascade of descriptors outperforms other schemes in term of computation cost since a much smaller number of region descriptors are computed and compared (referred to Figure 14). Figure 34 illustrates the most similar pedestrians found in a camera given an observation from another camera.

The sparse similarity measurement also influences the recognition rate (see Figure 32). If all the blobs are kept ($\beta = 1$), the worst performance is achieved. However, keeping half of the blobs, the overall performance increases by roughly 25%. Note that even if beta is low (e.g. $\beta = 0.25$), we have a gain in performance.

In addition, all 4 region descriptors are compared in Figure 33 given the cascade of grids of descriptors (1-4-9-16-25). Interestingly, the color feature is performing better than other features given such problem. Indeed, in order to identify the pedestrians across a set of pedestrian images (i.e. when only pedestrians are present in the search space), the color feature is

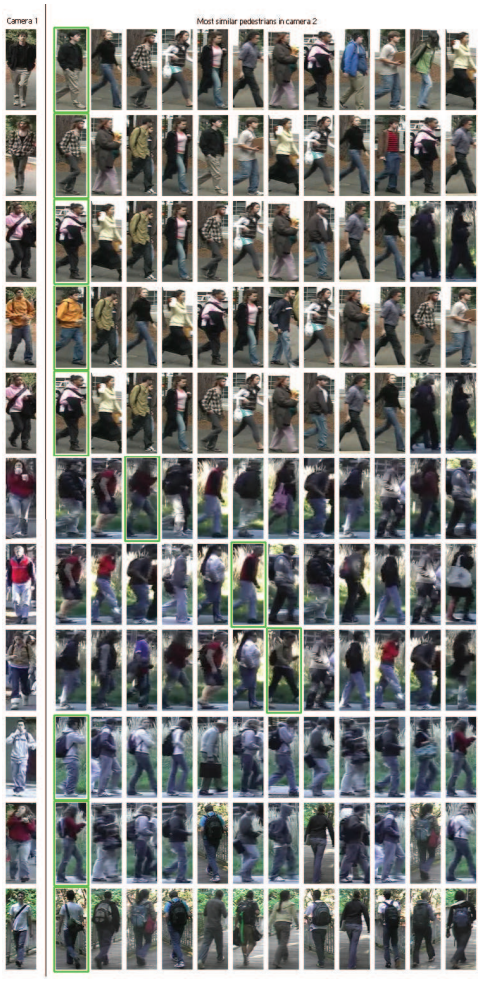


Figure 34: Examples of objects observed by a camera (left side), and the corresponding most similar pedestrians found by the cascade of grids of color histograms (YCrCb space)within another camera (right side). The images are sorted (most similar are the left ones) and the green bounding box is the correct match.

a relevant cue. It coincides our intuition since all the objects have similar shapes (all pedestrians), and only their clothes, hence the color feature is the most discriminative feature to match pedestrians across other pedestrians. However, when a set of random objects are present in the search space similar to our presented application (object localization across camera views), the features considering the shape of the objects (e.g. gradients) are more relevant and still perform well given the re-identification problem (e.g. the covariance descriptor).

8. Conclusions

A master-slave system is presented to address the challenging problem of detecting and tracking any objects across any network of cameras. Cameras are uncalibrated, moving, and with non-overlapping field-of-views. Only the appearances of the objects are used. Most state-of-the-art region descriptors are evaluated to address such problem. Using them "as is", i.e. a single descriptor per object leads to very poor performance. Using a grid of region descriptors similar to Low [21], or Bay *et al.* [22] increases the performance. However, we propose a cascade of grids of region descriptors outperforming other approaches. The use of coarse to fine grids combined with the sparse similarity measurement allow the detection and tracking of deformable objects with partial occlusion and change of view-points. The reduced search space driven by the matched interest points promotes near real-time performance.

Although many region descriptors have been studied, future work can investigate other descriptors. We notice that when an object needs to be localized in an image plane, color features degrade the performance whereas to re-identify or recognize pedestrians within a data set of pedestrians only (intra category), it performs best. Shape information is a crucial feature to localize objects and color increase the re-identification rate. A descriptor combining both feature can lead to promising performance. The covariance descriptor combined both features but did not perform best for the addressed problem since equivalent weight are given to the features. One can further combine those features given more weight to shape during the object inter-category detection (localization problem) and more to color during the intra-category (recognition problem). Also, generic and easy-to-compute grids of descriptors (a set of rectangular grids made of uniform sub-rectangles) are used. Future work can use more sophisticated grids. Notwithstanding, they demonstrated a gain in performance using the proposed cascade of grids of descriptors given the sparse similarity measurement.

Acknowledgments

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

References

[1] P. Gabriel, J. Hayet, J. Piater, J. Verly, Object tracking using color interest points, IEEE International Conference on Advanced Video and Signal based Surveillance (2005) 159–164.
[2] M. McCahill, C. Norris, Cctv in london (2002).
[3] www.watchover-eu.org.
[4] Y. Caspi, D. Simakov, M. Irani, Feature-based sequence-to-sequence matching, International Journal of Computer Vision 68 (1) (2006) 53–64.
[5] S. M. Khan, M. Shah, Tracking multiple occluding people by localizing on multiple scene planes, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (3) (2009) 505–519.
[6] F. Fleuret, J. Berclaz, R. Lengagne, P. Fua, Multicamera people tracking with a probabilistic occupancy map, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2) (2008) 267–282.
[7] F. Porikli, Achieving real-time object detection and tracking under extreme conditions, Journal of Real-Time Image Processing 1 (1) (2006) 33–40.
[8] A. Alahi, Y. Boursier, L. Jacques, P. Vandergheynst, Sparsity driven people localization with heterogeneous cameras network, Submitted to IEEE transactions on Image Processing.
[9] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. I: 886–893.
[10] O. Tuzel, F. Porikli, P. Meer, Pedestrian detection via classification on riemannian manifolds, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (2008) 1713–1727.

[11] B. Leibe, N. Cornelis, K. Cornelis, L. Van Gool, E. Zurich, Dynamic 3D scene analysis from a moving vehicle, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2007.
[12] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, T. Poggio, Pedestrian detection using wavelet templates, IEEE Conference on Computer Vision and Pattern Recognition 97 (1997) 193–199.
[13] C. Papageorgiou, T. Poggio, Trainable pedestrian detection, in: IEEE International Conference on Image Processing, Vol. 4, 1999.
[14] F. Suard, A. Rakotomamonjy, A. Bensrhair, A. Broggi, Pedestrian detection using infrared images and histograms of oriented gradients, in: Intelligent Vehicles Symposium, Tokyo, Japan, 2006, pp. 206–212.
[15] 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, Marseilles, France, 2008, pp. 1542–1550.
[16] H. Cheng, N. Zheng, J. Qin, Pedestrian detection using sparse gabor filter and support vector machine, Intelligent Vehicles Symposium (2005) 583–587.
[17] P. Viola, M. Jones, D. Snow, Detecting pedestrians using patterns of motion and appearance (2003).
[18] 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.
[19] O. Tuzel, F. Porikli, P. Meer, Region covariance: A fast descriptor for detection and classification, in: European Conference on Computer Vision, 2006.
[20] 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.
[21] D. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.
[22] H. Bay, T. Tuytelaars, L. Van Gool, SURF: Speeded Up Robust Features, Lecture Notes in Computer Science 3951 (2006) 404.
[23] A. Rosenfeld, G. Vanderbrug, Coarse-fine template matching, IEEE Transactions on Systems, Man and Cybernetics 7 (1977) 104–107.
[24] B. Funt, G. Finlayson, Color constant color indexing, IEEE transactions on Pattern analysis and Machine Intelligence (1995) 522529.
[25] D. Gray, H. Tao, Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features, in: Proceedings of the 10th European Conference on Computer Vision: Part I, Springer, 2008, pp. 262–275.
[26] D. Comaniciu, V. Ramesh, Real-time tracking of non-rigid objects using mean shift, uS Patent 6,590,999 (2003).
[27] B. Prosser, S. Gong, T. Xiang, Multi-camera matching under illumination change over time, in: European Conference on Computer Vision, 2008.
[28] A. Shashua, Y. Gdalyahu, G. Hayun, Pedestrian detection for driving assistance systems: Single-frame classification and system level performance, in: International Symposium on Intelligent Vehicles, 2004, pp. 1–6.
[29] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, P. Tu, Shape and appearance context modeling, in: IEEE International Conference on Computer Vision, 2007.
[30] N. Gheissari, T. Sebastian, R. Hartley, Person reidentification using spatiotemporal appearance, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, 2006.
[31] 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.
[32] W. Forstner, B. Moonen, A metric for covariance matrices, Qua vadis geodesia (1999) 113–128.
[33] W. Freeman, E. Adelson, The design and use of steerable filters, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (9) (1991) 891–906.
[34] 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.
[35] A. Baumberg, Reliable feature matching across widely separated views, in: Conference on Computer Vision and Pattern Recognition, Vol. 1, IEEE Computer Society, 2000, pp. 131–137.
[36] G. Carneiro, A. Jepson, Multi-scale phase-based local features, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, 2003, pp. 171–187.
[37] 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.
[38] K. Mikolajczyk, C. Schmid, A Performance Evaluation of Local Descriptors, IEEE Transactions on Pattern Analysis and Machine Intelligence (2005) 1615–1630.
[39] D. Gray, S. Brennan, H. Tao, Evaluating Appearance Models for Recognition, Reacquisition, and Tracking, in: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance., 2007.