Person Re-identification Using Haar-based and DCD-based Signature

Sławomir Bąk, Etienne Corvee, Francois Brémond, Monique Thonnat
INRIA Sophia Antipolis, PULSAR group
2004, route des Lucioles, BP93
06902 Sophia Antipolis Cedex - France
firstname.surname@sophia.inria.fr

Abstract

In many surveillance systems there is a requirement to determine whether a given person of interest has already been observed over a network of cameras. This paper presents two approaches for this person re-identification problem. In general the human appearance obtained in one camera is usually different from the ones obtained in another camera. In order to re-identify people the human signature should handle difference in illumination, pose and camera parameters. Our appearance models are based on haar-like features and dominant color descriptors. The AdaBoost scheme is applied to both descriptors to achieve the most invariant and discriminative signature. The methods are evaluated using benchmark video sequences with different camera views where people are automatically detected using Histograms of Oriented Gradients (HOG). The re-identification performance is presented using the cumulative matching characteristic (CMC) curve.

1. Introduction

Detection and tracking of moving objects constitute the main problem of video surveillance applications. The number of targets and occlusions produce ambiguity which introduces a requirement for reacquiring objects which have been lost during tracking. However, the ultimate goal of any surveillance system is not to track and reacquire targets, but to understand the scene and to determine whether a given person of interest has already been observed over a network of cameras. This issue is called the person re-identification problem.

The person re-identification presents a number of challenges beyond tracking and object detection. The overall appearance of an individual as well as biometrics (e.g. face or gait) are used to differentiate individuals. In this work we consider appearance-based approaches which build a specific human appearance model to re-identify a given individual. This model has to handle differences in illumination, pose and camera parameters. Nevertheless, our approach follow a classical scheme. First, a human detection algorithm is used to find out people in video sequences. Then, the individual is tracked through few frames to generate a human signature. The signature has to be based on discriminative features to allow browsing the most similar signatures over a network of cameras to determine where the person of interest has been observed. It can be achieved by signature matching which has to handle differences in illumination, pose and camera parameters.

The human signature generation is the main subject of this paper. We develop two person re-identification approaches which use haar-like features and dominant color descriptor (DCD), respectively. The haar-based approach uses the AdaBoost scheme to find out the most discriminative haar-like feature set for each individual. This set of haar-like features combined through a cascade represents a human signature. The DCD signature is built by extracting the dominant colors of upper and lower body parts. These two sets of dominant colors are also combined using the AdaBoost scheme to catch different appearance corresponding to one individual.

The outline of the paper is the following. Related work is presented in Section 2. Section 1 describes the overview of the approach. Signature generation is presented in Sections 4 and 5. Section 6 describes experimental results and Section 7 contains some concluding remarks and future work.

2. Related work

Several approaches have been developed where invariant appearance models represent signatures of human. If the system considers only a frontal viewpoint then the triangular graph model [4] or shape and appearance context model [18] can be used. Otherwise, if multiple overlapping cameras are available, it is possible to build a panoramic appearance map [3]. In [7] the authors build a model based on interest-point descriptors using views from different cam-

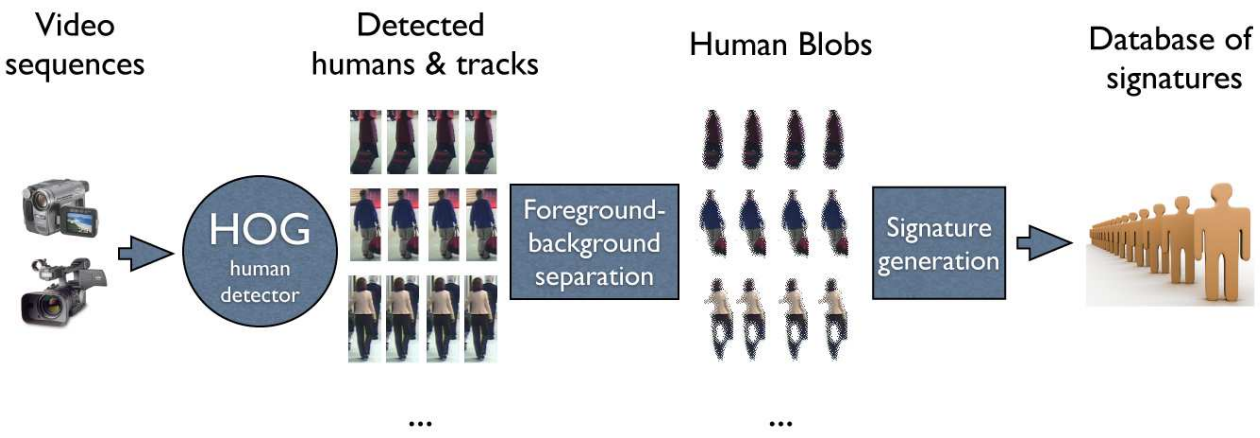


Figure 1. The re-identification system.

eras. Unfortunately, under challenging conditions where the views from different cameras are not given a priori, a local descriptor matching approach performs poorly [4]. In [12] the clothing color histograms taken over the head, shirt and pants regions together with the approximated height of the person has been used as the discriminative feature. Recently, the ensemble of localized features (ELF) [6] has been proposed. Instead of designing a specific feature for characterizing people appearance, a machine learning algorithm constructs a model that provides maximum discriminability by filtering a set of simple features.

Also other more complicated template methods show promise but there are very expensive in both memory and computation cost [15]. Subspace methods together with manifolds are used to model pose and viewpoint changes. However, in [8] the full subspace of nonrigid objects approximated by nonlinear appearance manifold becomes too large to represent a human signature accurately. Thus we propose to study efficient features which are also reliable to build human signature under different camera viewpoints.

3. Overview of the approach

3.1. The re-identification system

A person can be recognized in one camera if his/her visual signature has been previously extracted in another camera. We have tested our re-identification algorithm with manually annotated people in two non-overlapping scenes (see Section 6.1) to validate the method. Nevertheless, in real scenarios the re-identification algorithms work on automatically extracted data. Therefore, the evaluation of the robustness of our method has also been performed on automatically detected humans. Our algorithm uses Histograms of Oriented Gradients (HOG) to automatically detect and track humans (see Figure 1). Each detected human

Figure 2. Examples of tracked people.

Figure 3. Mean human image with corresponding edge magnitudes and the 15 most dominant cells.

is tracked in order to accumulate images with person of interest. From these images we extract human blobs using foreground-background separation technique (see Section 3.3). Finally, sets of human blobs are used by AdaBoost scheme to create a reliable visual signature. The AdaBoost scheme is applied to haar-like features (see Section 4) and to dominant color descriptor (see Section 5).

For each detected and tracked human a visual signature is generated. All such created visual signatures from different scenes are stored in one human signature database. The

performance evaluation of our re-identification algorithms is based on querying the human signature database by extracted signatures. The results are analyzed using cumulative matching characteristic (CMC) curve [5].

3.2. Human detection and tracking

We have adapted the HOG based technique used in [2] to detect and track people. The detection algorithm extracts histogram of gradient orientation, using a Sobel convolution kernel, in a multi-resolution framework to detect human shapes at different scales. The technique was originally designed to detect faces using the assumption that facial features remain approximately at the same location in a 9 non-overlapping cells square window (e.g. the right eye is located in the top left corner of a square window). The modified algorithm detects humans using 15 cells located at specific locations around the human silhouette as shown in Figure 3. The first image shows the mean human image calculated over all positive samples in the database; the second image shows the corresponding mean edge magnitude response; the third image shows the later image superposed with the 15 most dominant cells. The cell bounding boxes are drawn with a color set by their most dominant edge orientation with scheme defined in the last image

The most dominant cells used to characterize human shapes are the 15 most dominant cells selected among 252 cells covering the human sample area. These most dominant cells are the cells having the closest HOG vector to the mean HOG vector calculated over the vectors (of the corresponding cell) from a human database. The NICTA database [10] is used to train the human detection algorithm with 10,000 positive (human) samples and 20,000 negative (background scene) samples. Cells are of size 8x8 pixels and a database sample is of size 64x80 pixels. Figure 2 shows an example of several tracked persons in dynamically occluded scenarios.

3.3. Foreground-background separation

The output of the detection algorithm is the set of the 2D bounding boxes with tracked individuals. The color-based foreground-background separation technique [1] is used to obtain the mask which allows to separate the person region from the background region (see Figure 4 (a) and (b)). In this technique the color-features inside the target are labeled as 'foreground class' and color-features outside the target are labeled as 'background class'. The probability density function of the color-feature in the target region and in its local background are obtained to find the log-likelihood ratio of the sample belonging to the 'foreground class'.

In this paper we assume that the person mask (i.e. blob) contains all the information to represent the human signature. We present two approaches to generate the human signature: haar-based approach which separates a space of humans extracted from a video sequence and an approach based on the dominant color descriptor. Both approaches are described in the following sections.

(a) original image (b) mask (c) haar features

Figure 4. The color-based foreground-background separation technique [1] and haar signature: a) original image; b) foreground-background separation mask; c) discriminative haar-like features.

4. Haar-based Signature

4.1. Haar-like features

In this work we use an extended set of haar-like features [9] which significantly enrich the basic set proposed by Viola and Jones [17]. Similarly to [17] we develop a cascade of classifiers generated by a boosting scheme. In contrary to Tieu and Viola [16], the threshold of the weak classifiers separating positives and negatives is not computed using Gaussian models. In our approach the threshold computation is based on the concept of information entropy used as a heuristic to produce the smallest cascade. This idea is originated from Quinlan [14]. A human signature is based on 20×40 pixel sub-window which leads to the tremendous number of haar-like features needed to be considered (435,750). Even though haar-like features can be calculated efficiently using *an integral image*, this number of features makes the feature selection phase more time consuming.

We have decided to use the color-based foreground-background separation technique (see Section 3.3) to obtain the mask which allows us to filter out meaningless haar-like features (see Figure 4). Only features inside the mask are considered. This step decreases the feature set from 435,750 to about 20,000 features depending on the person mask, significantly speeding-up the learning process. The huge decrease of the feature set is obtained by ignoring all patterns which intersect the background area.

We use AdaBoost to select the most discriminative feature set for each individual. The most discriminative feature set forms a strong classifier. We have applied the method [17] for combining increasingly more complex clas-

sifiers into a cascade structure. Finally, a signature is represented by this cascade of strong classifiers. In general 5 to 10 strong classifiers are sufficient to discriminate the individual. During signature learning we choose the features which are the most discriminative for the specific individual. Hence, we assume that few frames of the object of interest are given for the learning process. These few frames may carry the information about different poses and can help to catch all pose variety. Nevertheless, if only one image of the person is given, we generate different view points of the person by sliding a window over the image in different directions. These image samples are used as positive samples presented to the boosting algorithm. Negative samples can be obtained by gathering anything which is not an object of interest. In our approach we use the other detected people appearing in the video sequence to get negative samples. This mechanism allows us to find out the most discriminative features which separate one individual from the rest of the detected people. Hence, the person is represented by the set of wavelet functions which differentiate several spatial orientations. This signature can be used as a detector [17] to find out where the person of interest appeared in the another camera. In our case, we have defined a distance similarity function between two signatures to retrieve human signatures in the most efficient way. For each person detected in a video sequence the signature is generated. By indexing the human signatures the system can browse and retrieve the most similar signatures over a network of cameras to determine where the person of interest has been observed.

4.2. Haar-distance computation

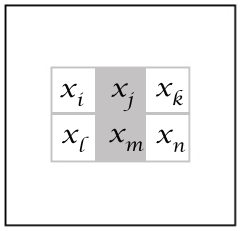


Figure 5. Illustration of a haar-like feature. The sum of the white pixels is subtracted from the sum of the grey pixels.

Haar-like features inspired by [11] are the weighted sums of pixels composing rectangle patterns. For example, in Figure 5 one of the haar-like features (*line feature*) is presented. Mathematically, this line feature can be expressed as the constraint:

$$p_I(\alpha_I(x_i+x_l)+\beta_I(x_j+x_m)+\gamma_I(x_k+x_n)) > p_I\theta_I, \quad (1)$$

where a parity p_I indicates the direction of the inequality sign, the coefficients α_I, β_I and γ_I are chosen arbitrarily, θ_I is the threshold and x represents a pixel value. In general, each haar-like feature can be expressed in this way forming a weak classifier. Therefore a strong classifier as well as a cascade of them can be represented by a set of linear constraints like formula (1). In our approach a detected blob corresponding to an individual is reduced to a $W \times H$ pixel sub-window (let it be 20×40). Hence, $n = W * H$ ($20 * 40 = 800$) dimensions are used to model all haar-like features. If we assume that a pixel value $x \in \Psi$ and $|\Psi| = \tau$ (in general a range of intensity is $\tau = 256$), then the whole signature space is τ^n. The set of constraints defined by haar-like features forms a set of hyper-planes in this space. Moreover, if we assume additionally that we have $2n$ default hyper-planes: $x = \min(\Psi)$ and $x = \max(\Psi)$ for each of the n dimensions, the whole set of hyper-planes produces a hypercube. Therefore, each human signature is represented by different hypercube obtained during learning process. The volume of hypercube generated by the signature s can be denoted as $\mathcal{V}_s$.

Let us assume that two signatures s_i and s_j are given. We already know that both signatures can be represented by the set of constraints. The fusion of these signatures can be obtained by merging these sets of constraints into a new set. The result set may also be expressed as a new hypercube. The volume of the new hypercube produced after fusion is $\mathcal{V}_{s_i s_j}$. We define the distance between two signatures s_i and s_j as:

$$D(s_i, s_j) = 1 - \frac{\mathcal{V}_{s_i s_j}}{\min(\mathcal{V}_{s_i}, \mathcal{V}_{s_j})}. \quad (2)$$

Nevertheless, if we want to build the hypercube in n-dimensions a calculation of 2^n vertices is required. In our approach we use 20×40 pixel sub-window which leads to $n = 800$ dimensions, bringing an unattainable memory and time requirement. For that reason our volume computation does not consider the whole space τ^n. However, the volume is computed using a spare space which is build dynamically by images arriving to the system. Each image is a point in the n-dimensional space and it can be inside or outside hypercube created by a set of haar-like features given as a set of constraints. If image meet all set of constraints then it is inside hypercube. Otherwise the image is outside hypercube. The volume of hypercube is computed using the number of images which satisfy a set of haar-like features (the number of images which are inside the hypercube).

5. DCD Signature

Dominant color descriptor (DCD) has been proposed by MPEG-7 and it is extensively used for image retrieval [19]. DCD is defined as:

$$F = \{\{c_i, p_i\}, i = 1, \ldots, N\}, \quad (3)$$

(a) original image (b) upper body part (c) lower body part

Figure 6. The dominant color separation technique: a) original image; b) upper body dominant color mask; c) lower body dominant color mask.

where N is the total number of dominant colors in the image, c_i is a 3-D color vector, p_i is its percentage, and $\sum_i p_i = 1$.

In our approach we use dominant colors to create a discriminative signature of the people. First, the color-based foreground-background separation technique (see Section 3.3) is used to obtain the region where the person is present. Then, similar to [12], a person is divided into basic body parts. The human body is separated into two parts: the upper body part and the lower body part. The separation is obtained by maximizing the distance between sets of dominant colors of the upper and the lower body (see Figure 3). The combination of the dominant color descriptor of upper and lower body is considered as a meaningful feature to discriminate people. As far as we assume that a few frames of the person of interest are given, the AdaBoost scheme is applied to find out simultaneously the pose invariant and the most discriminative appearance model. For example, the AdaBoost scheme allows to find a different appearance corresponding to the person of interest (back appearance of an individual can be different than his/her frontal appearance). In general to represent the appearance of an individual, two or three weak classifiers are sufficient. A weak classifier is represented by two sets of dominant colors. The first set corresponds to the upper body part and the second set corresponds to the lower body part. The classification is performed using the threshold on the distance between corresponding body parts. For each body part the threshold is computed using the distance between two corresponding sets of dominant colors considered to belong to the same person. The distance between two sets of dominant colors is defined using the improved similarity measure [19]:

$$D^2(F_1, F_2) = 1 - \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} a_{ij} S_{ij}, \tag{4}$$

where $S_{ij} = [1 - |p_i - p_j|] * \min(p_i, p_j)$ and

$$a_{ij} = \begin{cases} 1 - d_{ij}/d_{max}, & d_{ij} \leq T_d \\ 0, & d_{ij} > T_d \end{cases} \tag{5}$$

where d_{ij} is the Euclidean distance between two colors. Threshold T_d is the maximum distance for two colors to be considered as similar and $d_{max} = \alpha T_d$ and α has been set experimentally to 1.2.

One of the most challenging problem using the color as a feature is that images of the same object acquired under different cameras show color dissimilarities. Even identical cameras, which have the same optical properties and are working under the same lighting conditions, may not match in their color responses. Hence, inter-camera color calibration using cross-correlation model function method [13] is used to handle color dissimilarities.

6. Experimental results

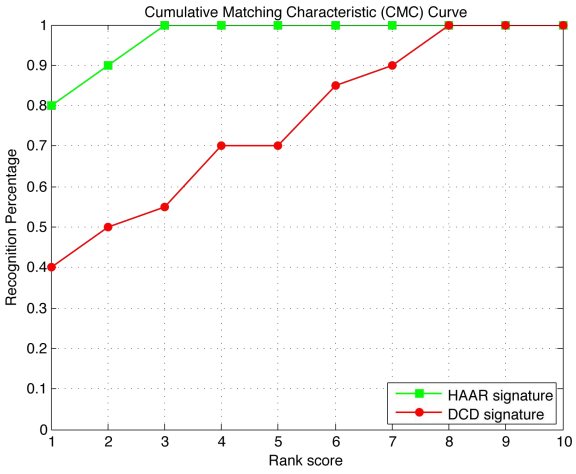
In this section the evaluation of our approach is presented. Given a single human signature, the chance of choosing the correct match is inversely proportional to the number of considered signatures. Hence, we believe the cumulative matching characteristic (CMC) curve is a proper performance evaluation metric [5]. In our approach the signature is computed for each detected person. Let us denote a signature as s_i^c, where i encodes the person id and c encodes the camera id. Then, the signature s_i^c is used as a query to the database of signatures $s_j^{c'} \in \Omega$ such that $c \neq c'$. This evaluation scheme is analogous to a standard surveillance scenario where an operator queries a system with multiple images of the same individual captured over a short period of time from a particular camera to find him/her in a network of cameras.

The experiments were performed on the publicly available data recorded for CAVIAR project (IST 2001 37540, http://homepages.inf.ed.ac.uk/rbf/CAVIAR) and on TREC Video Retrieval Evaluation data (organized by NIST, TRECVID 2008) obtained from Gatwick Airport surveillance system.

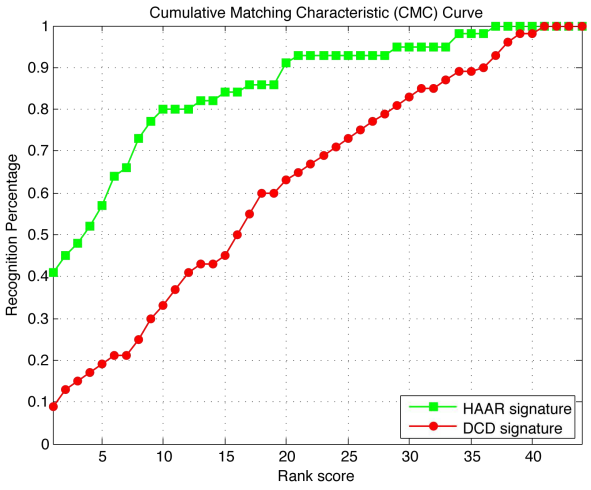
6.1. CAVIAR data

Similarly as in [7] first we have evaluated our approach using CAVIAR data. We have used the set where clips from shopping center in Portugal are given. Each clip was recorded from two different points of view that allows us to evaluate the re-identification algorithms. The low resolution in one of the two cameras makes the data challenging (see Figure 7). We have selected 10 persons to evaluate our approach. Each signature in both cameras has been created using 45 positives obtained by manually annotating data. In the contrary to [7] the human signatures have been created

Figure 7. The sample images from CAVIAR data set. Top and bottom lines correspond to different cameras.



(a) CAVIAR



(b) TRECVID

Figure 8. Cumulative matching characteristic (CMC) curve for haar-based and DCD-based signature with CAVIAR and TRECVID database.

independently for each camera. Our goal was to generate a human signatures using one camera and re-identify an individual in another camera. In [7] the signatures are built by collecting the interest-point descriptors using both cameras which simplifies the re-identification problem. Their goal was to re-identify an individual by signature generated using descriptors obtained from both cameras. In our case the extraction of signatures does not require to match the description of the same individual in different cameras. In haar-based approach 80% of the queries have achieved a top ranking true match which is comparable to [7] (82% precision for a recall of 78%). We have obtained similar results without the strong assumption of observing people by both cameras.

Figure 8(a) shows CMC curves for haar-based and DCD-based signatures. CMC metric shows how performance improves as the number of requested images increases (Rank Score). DCD signature performs poorly because of significant color dissimilarities between both cameras (see Figure 7). Color calibration improves performance but the dissimilarities are so strong that the color transformation function remains an issue. For example the woman at the fifth position appeared white on the first camera and blue on the second camera. Moreover, DCD signature depends strongly on resolution which can produce an ambiguities in the dominant color extraction.

6.2. TRECVID data

The evaluation of the re-identifiaction algorithm has to take into account that the chance of choosing the correct match depends on segmentation results and on the number of considered signatures. Therefore, for evaluation purposes, 44 individuals were detected using the human detector based on HOG. DCD signature performs poorly again. In addition to the strong color dissimilarities, the new issue in the data appears. The people often carry the luggage which can occlude almost half of the person in one of the cameras. This problem also produce some challenges for

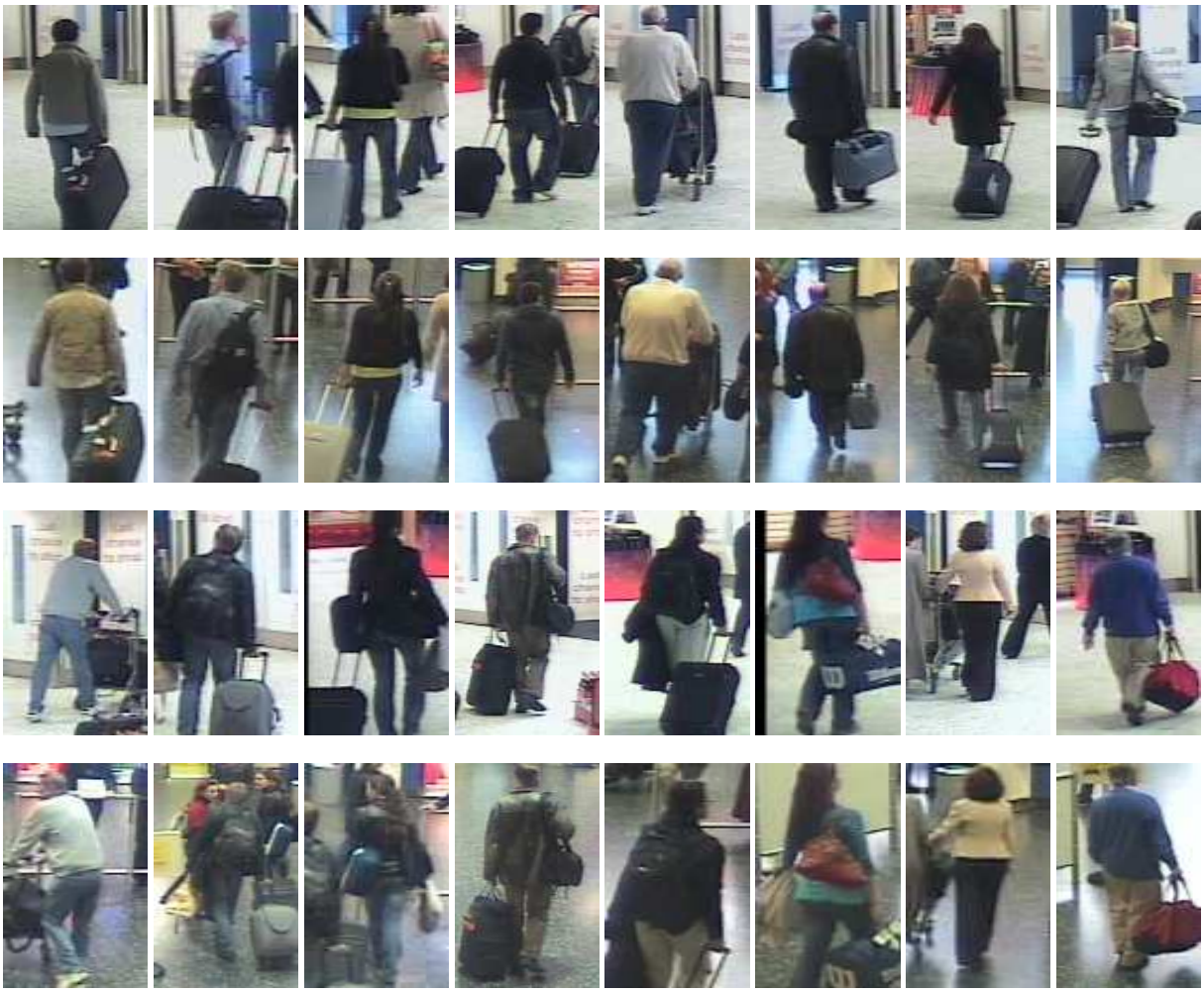


Figure 9. The sample images from TRECVID data set. First and third lines correspond to camera one and second and forth lines correspond to camera two.

re-identification algorithms (see Figure 9). In haar-based approach 41% of the queries have achieved a top ranking true match (see Figure 8(b)) and 80% of the queries have generated a true match in the top ten.

For comparison, in [4] the evaluation were also performed on 44 individuals. Their model fitting approach results in the best performance with approximately 60% of the queries achieving a top ranking true match and over 90% of the queries generating a true match in the top ten. Nevertheless, it is worth noting that [4] is based on a strong assumption: frontal view of the person has to be given. Moreover, in their approach individuals are detected manually. In fact, according to our knowledge, we are the first trying a re-identification approach on real world videos with automatically extracted humans. We try to find the most representative features to match different poses (illustrated by training videos) without any assumption concerning viewpoint. In our approach an individual is detected automatically from the real environment where people are occluding each other and carry luggage which produce ambiguities. The segmentation becomes an issue because not always full body of the person is detected. Often only the half of the person is detected making the person re-identification problem more challenging.

Furthermore, our results seem reasonable when compared to [18]. The obtained accuracy in [18] is 82% in first match for a simpler database (containing mostly frontal human views). In [20] the authors applied [18] on challenging i-LIDS database which is our TRECVID data. The highest performance obtained by [20] was 10% of accuracy in first match, which is significantly less than our 41%. Even though their evaluation process was more challenging (they tested on 119 pedestrians and we tested on 44) our results look better. [20] explains their results by the fact that i-LIDS data is very challenging since it was captured from non-overlapping multiple camera views subject to signifi-

cant occlusion and large variations in both view angle and illumination. Therefore, even if [18] is known for best performance we believe that on TRECVID they would achieve smaller accuracy than ours 41%. The data set of humans gathered for the re-identificaiton approaches can be published if the authorization of i-LIDS is obtained.

7. Conclusion and Future work

We have presented Haar-based and DCD-based approaches for the person re-identification problem. The evaluation has been performed on automatically detected humans using Histograms of Oriented Gradients. The results indicate that the haar-like features are reliable to handle viewpoint and pose changes.

In the future work, we plan to apply the graph-cut optimization method in order to improve the foreground-background separation technique. The extraction of foreground seems to be a bottleneck in the re-identification approaches. Furthermore, the DCD and haar-like features can be combined to form a robust human signature. Also, additional research has to be carried out in order to handle color dissimilarities. Finally, consideration of different features like shape, 3D size or silhouette might be beneficial.

Acknowledgements

This work has been supported by Agence National de la Recherche (ANR) and VIDEO-ID project.

References

[1] S. Bąk, S. Suresh, F. Brémond, and M. Thonnat. Fusion of motion segmentation with online adaptive neural classifier for robust tracking. In A. Ranchordas and H. Araújo, editors, *VISSAPP (2)*, pages 410–416. INSTICC Press, 2009. 3

[2] E. Corvee and F. Bremond. Combining face detection and people tracking in video sequences. In *3rd International Conference on Imaging for Crime Detection and Prevention - ICDP09*, December 2009. 3

[3] T. Gandhi and M. M. Trivedi. Person tracking and reidentification: Introducing panoramic appearance map (pam) for feature representation. *Mach. Vision Appl.*, 18(3):207–220, 2007. 1

[4] N. Gheissari, T. B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1528–1535, Washington, DC, USA, 2006. IEEE Computer Society. 1, 2, 7

[5] D. Gray, S. Brennan, and H. Tao. Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, 2007. 3, 5

[6] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*, pages 262–275, Berlin, Heidelberg, 2008. Springer-Verlag. 2

[7] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *Distributed Smart Cameras, 2008. ICDSC 2008. Second ACM/IEEE International Conference*, pages 1–6, Sept. 2008. 1, 5, 6

[8] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *CVPR '03: Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–313–I–320 vol.1, June 2003. 2

[9] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–900–I–903 vol.1, 2002. 3

[10] G. Overett, L. Petersson, N. Brewer, L. Andersson, and N. Pettersson. A new pedestrian dataset for supervised learning. *In IEEE Intelligent Vehicles Symposium*, 2008. 3

[11] C. P. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. *Computer Vision, 1998. Sixth International Conference on*, pages 555–562, 1998. 4

[12] U. Park, A. Jain, I. Kitahara, K. Kogure, and N. Hagita. Vise: Visual search engine using multiple networked cameras. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference*, volume 3, pages 1204–1207, 0-0 2006. 2, 5

[13] F. Porikli. Inter-camera color calibration by correlation model function. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 2, pages II–133–6 vol.3, Sept. 2003. 5

[14] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, March 1986. 3

[15] C. Stauffer and E. Grimson. Similarity templates for detection and recognition. In *CVPR '01: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–221–I–228 vol.1, 2001. 2

[16] K. Tieu and P. Viola. Boosting image retrieval. *International Journal of Computer Vision*, 56(1):17–36, 2004. 3

[17] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR '01: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 511–518, 2001. 3, 4

[18] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *ICCV '07: Proceedings of the 2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, Oct. 2007. 1, 7, 8

[19] N.-C. Yang, W.-H. Chang, C.-M. Kuo, and T.-H. Li. A fast mpeg-7 dominant color extraction with new similarity measure for image retrieval. *J. Vis. Comun. Image Represent.*, 19(2):92–105, 2008. 4, 5

[20] W.-S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *British Machine Vision Conference, BMVC*, London, 2009. 7