2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance

Person Re-identification Using Spatial Covariance Regions of Human Body Parts

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 is the person re-identification problem. 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. We propose a new appearance model based on spatial covariance regions extracted from human body parts. The new spatial pyramid scheme is applied to capture the correlation between human body parts in order to obtain a discriminative human signature. The human body parts are automatically detected using Histograms of Oriented Gradients (HOG). The method is evaluated using benchmark video sequences from i-LIDS Multiple-Camera Tracking Scenario data set. The re-identification performance is presented using the cumulative matching characteristic (CMC) curve. Finally, we show that the proposed approach outperforms state of the art methods.

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.

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 approach which build a specific human signature model to re-identify a given individual. This model has to handle differences in illumination, pose and camera parameters. In our approach a human detection algorithm is used to find out people in video sequences. Then, we generate a human signature using the image of the individual. 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 computation is the main subject of this paper. We develop a person re-identification approach using spatial covariance regions of human body parts. The body parts detector based on Histogram of Oriented Gradients (HOG) is applied to establish the correspondence between body parts. Then, we offer the covariance descriptor to find out the similarity between corresponding body parts. Finally, we take an advantage of the idea of spatial pyramid matching to design a new dissimilarity measure between human signatures. This dissimilarity measure is able to capture the correlation between body parts.

The outline of the paper is the following. Related work is presented in Section 2. Section 3 describes the overview of the approach. Signature generation is presented in Section 4. Section 5 describes experimental results and Section 6 contains some concluding remarks and future work.

2. Related work

Several approaches have been developed where invariant appearance model represents the signature of a 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 [9] the authors build a model based on interest-point descriptors using views from different cam-

978-0-7695-4264-5/10 $26.00 © 2010 IEEE
DOI 10.1109/AVSS.2010.34

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 [14], the clothing color histograms taken over the head, shirt and pants regions together with the approximated height of the person have been used as the discriminative feature. Recently, the ensemble of localized features (ELF) [8] 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 seem promising but they 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 [12] 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

In this paper we develop a person re-identification approach which uses covariance matrices. For establishing spatial correspondences between individuals we use human detector and human body parts detector (see Section 3.1). The invariant signatures for comparing different regions are generated by combining color and structural information. The color information is captured by covariance descriptors based on colors channels and their gradients. Invariance to differences in ambient illumination is achieved by color normalization (see Section 4.1). The structural information is represented by body parts. Moreover, the spatial information is also employed by using spatial pyramid matching (see Section 4.3) in the two-dimensional image space of the region of interest.

We now describe the human detection algorithm and the human body parts detection approach. Both approaches are based on Histogram of Oriented Gradients (HOG) descriptor.

3.1. Human and body parts detector

We use a Histogram of Oriented Gradient (HOG) based technique to automatically detect humans in different scenes before their visual signatures are extracted for re-identification purposes. Our HOG technique is adapted from the face detection technique [1] to detect human silhouette. The detection algorithm extracts histograms of gradient orientation, using a Sobel convolution kernel, in a multi-resolution framework. The technique was originally designed to detect faces using the assumption that facial features remain approximatively at the same location. A set

Figure 1. Mean human image with corresponding edge magnitudes and the 15 most dominant cells. From the left: 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 this later image superposed with the 15 most dominant cells of size 8×8 pixels.

Figure 2. Examples of detected people.

of 9 non-overlapping square cells (*e.g.* the right eye is located in the top left corner of a square window) were used to represent a face. However, location of human silhouette features do not remain constant in template with the varying poses (*e.g.* knees are constantly changing position when walking; a shoulder changes position from walking to slightly bending when pushing a trolley).

The modified algorithm detects humans using 15 cells located at specific locations around the human silhouette as shown in Figure 1. The system is trained using $10,000$ positive and $20,000$ negative image samples from the NICTA database [13]. Edge orientations were sampled into 8 bins of HOG. Figure 2 shows an example of several detected persons in dynamically occluded scenario.

In order to compute visual signatures of particular body parts of a person, we have applied our people detector to body parts detectors. A body part detector is the same HOG detector described above for people detection but trained on various areas of a person. We have trained 5 body parts: the top, the torso, legs, the left arm and the right arm. Their detected positions in a detected person allow us to extract covariance matrices in their corresponding locations (see Figure 3).

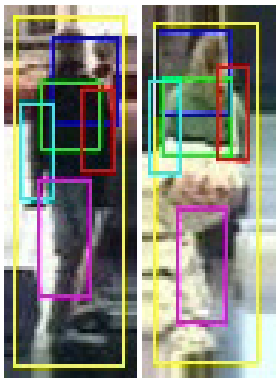

Figure 3. Illustration of the human and body parts detection results. Detections are indicated by 2D bounding boxes. Colors correspond to different body parts: the full body (yellow), the top (dark blue), the torso (green), legs (violet), the left arm (light blue) and the right arm (red).

(a) original (b) original (c) normalized (d) normalized

Figure 4. The first two columns show original images of the same person captured from different cameras in different environments. The last two columns show these images after histogram equalization.

4. Signature computation

In this section we propose a scheme to generate the human signature. The human and body parts detector returns six regions of interest corresponding to body parts: the full body, the top, the torso, legs, the left arm and the right arm. The top part is composed of the torso and the head (see Figure 3).

Once the body parts are detected, the next step is to handle color dissimilarities caused by camera and illumination differences. Thus, we applied a color normalization technique called histogram equalization (see Section 4.1). Then, on such normalized image, the covariance regions (see Section 4.2) of body parts are computed to generate a human signature. Furthermore, the dissimilarities between these regions corresponding to different images are combined using an idea derived from the spatial pyramid match kernels (see Section 4.3). The following sections describe in detail the aforementioned steps.

4.1. Color normalization

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, color normalization procedure has been carried out in order to obtain invariant signature. We use a technique called histogram equalization [10]. This technique is based on the assumption that the rank ordering of sensor responses is preserved across a change in imaging conditions (lighting or device). Histogram equalisation is an image enhancement technique originally developed for a single channel image. The aim was to increase the overall contrast in the image by brighting dark areas of an image, increasing the detail in those regions. Histogram equalisation achieves this aim by stretching range of histogram to be as close as possible to an uniform histogram. The approach is based on the idea that amongst all possible histograms, an uniformly distributed histogram has maximum entropy [5]. Maximizing the entropy of a distribution we maximize its information and thus histogram equalization maximizes the information content of the output image. We apply the histogram equalization to each of the color channels (RGB) to maximize the entropy in each of those channels and obtain the invariant image. Figure 4 illustrates the effect of applying the histogram equalization technique to images of the same individual captured by different cameras. These images highlight the fact that a change in illumination leads to a significant change in the colors captured by the camera. The last two columns show result images after applying histogram equalization procedure. It is clear that the resulting images are much more similar than the two original images.

4.2. Covariance Regions

Once color has been normalized, the covariance descriptor is applied to the regions corresponding to detected body parts. In this section we present an overview of this covariance descriptor [16] and its specialization to our re-identification problem.

Let I be an image. The method can be generalized to any type of image such as a one dimensional intensity image, three channel color image or even other types of images, *e.g.* infrared. Let F be the $W \times H \times d$ dimensional feature image extracted from I

$$F(x, y) = \phi(I, x, y) \tag{1}$$

where the function ϕ can be any mapping such as color,

intensity, gradients, filter responses, *etc*. For a given rectangular region $R \subset F$, let $\{f_k\}_{k=1\ldots n}$ be the d-dimensional feature points inside R. The region R is represented with the $d \times d$ covariance matrix of the feature points

$$C_R = \frac{1}{n-1}\sum_{k=1}^{n}(f_k - \mu)(f_k - \mu)^T \tag{2}$$

where μ is the mean of the points.

In our approach we define the mapping $\phi(I, x, y)$ as

$$\left[x, y, R_{xy}, G_{xy}, B_{xy}, \|\nabla^R_{xy}\|, \theta^R_{xy}, , \|\nabla^G_{xy}\|, \theta^G_{xy}, , \|\nabla^B_{xy}\|, \theta^B_{xy}\right] \tag{3}$$

where x and y are pixel location, R_{xy}, G_{xy}, B_{xy} are RGB channel values and ∇ and θ corresponds to gradient magnitude and orientation in each channel, respectively.

The input image region is mapped to $d = 11$ dimensional feature image. The covariance region descriptor is represented by an 11×11 matrix. We assume that the covariance of a distribution inside the region of interest is enough to discriminate it from other distributions. The descriptor encodes information of the variances of the defined features inside the region, their correlations with each other and spatial layout.

Concerning time consumption, there is an efficient way to compute covariance descriptors using integral images based on an idea proposed by [17] and applied to covariance matrices in [16]. Thus, there is an efficient way to compute a large number of covariance descriptors corresponding to subregions of the image of interest.

4.2.1 Covariance matrix distance

The covariance matrices are not in Euclidean space. Hence, we use the distance definition proposed by [2] to compute the dissimilarity between regions. This dissimilarity of two covariance matrices is defined as

$$\rho(C_i, C_j) = \sqrt{\sum_{k=1}^{d} \ln^2 \lambda_k(C_i, C_j)} \tag{4}$$

where $\lambda_k(C_i, C_j)_{k=1\ldots d}$ are the generalized eigenvalues of C_i and C_j, determined by

$$\lambda_k C_i x_k - C_j x_k = 0, \quad k = 1 \ldots d \tag{5}$$

and $x_k \neq 0$ are the generalized eigenvectors.

4.3. Spatial Pyramid Matching

This section describes how the human signature is computed and how the dissimilarity between two signatures is obtained. First, we introduce the notion of pyramid matching kernel. Then, the signature levels are described. Finally, the dissimilarity functions between two signatures at each level are defined.

The original formulation of pyramid matching has been proposed in [6]. The pyramid matching allows for precise matching of two collections of features in a high dimensional appearance space. Nevertheless, it discards all spatial information. Hence, in [11] an orthogonal approach (pyramid matching in the two-dimennsional image space) has been proposed. Let us assume that X and Y are two sets of vectors in a d-dimensional feature space. The pyramid matching finds an approximate correspondence between these two sets. First, a sequence of grids at resolutions $0, \ldots, L$ is constructed. The grid at level l has 2^l cells. Then, the number of matches that occur at each level of resolution are combined using weighted sums. Matches found at finer resolutions are weighted more highly than matches found at coarser resolutions. The pyramid match kernel is defined as

$$\kappa^L(X, Y) = \mathcal{I}^L + \sum_{l=0}^{L-1} \frac{1}{2^{L-l}}(\mathcal{I}^l - \mathcal{I}^{l+1}) \tag{6}$$

$$= \frac{1}{2^L}\mathcal{I}^0 + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}}\mathcal{I}^l \tag{7}$$

where $\mathcal{I}^l$ is the matching function on level l (in original formulation the matching function was represented by histogram intersection).

In our approach the matching function is based on a similarity defined using the covariance matrix distance. In Figure 5 grid levels are presented. Level 0 corresponds to full body part region. Level 1 is represented by a set of remaining body parts. Finally, level 2 is described by grid cells inside the regions of interest. The human signature is represented by the set of covariance matrices computed at all levels.

The matching function at level 0 is defined as

$$\mathcal{I}^0 = \frac{1}{\rho(C_i, C_j)} \tag{8}$$

where C_i and C_j are a covariance matrices computed on regions corresponding to full body part of individual i and individual j, respectively.

Concerning following levels, we define the matching function as

$$\mathcal{I}^l = \frac{1}{\mathcal{D}^l} \tag{9}$$

where $\mathcal{D}^l$ is a dissimilarity function at level l between two sets of covariance matrices $\mathfrak{C}^i$ and $\mathfrak{C}^j$ computed inside regions of interest. Let us assume that $\mathcal{M}_z$ is the set of z largest ρ distances between corresponding covariance matrices. Then, the dissimilarity functions is defined as

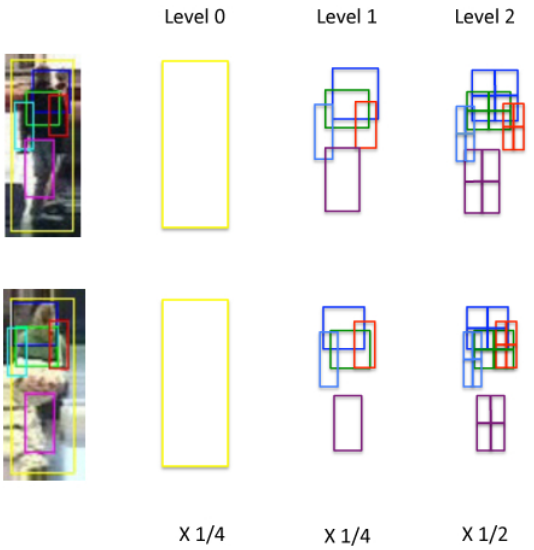

Figure 5. Example of constructing a three-level pyramid. Level 0 corresponds to the full body part. Level 1 and level 2 correspond to the rest of detected body parts and grids inside body parts, respectively. Finally, we weight each level according to eq. (7).

$$\mathcal{D}^l = \frac{\sum_{k=1}^{n(l)} \rho(C_k^i, C_k^j) - \sum_{m \in \mathcal{M}_z} \rho_m}{n(l) - z} \qquad (10)$$

where i and j correspond to individuals and $n(l)$ is the number of compared covariance matrices. The introduction of $\mathcal{M}_z$ increases robustness towards outliers coming from possible occlusions. In evaluation we set $z = \frac{n(l)}{2}$.

Finally, the dissimilarity between two signatures extracted from images I_i and I_j, is defined as

$$\mathfrak{D}(I_i, I_j) = \kappa^L(I_i, I_j). \qquad (11)$$

5. 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 [7].

The experiments are performed on images from 2008 i-LIDS Multiple-Camera Tracking Scenario (MCTS) data set with multiple camera views. The evaluation data set contains 476 images with 119 individuals automatically extracted by [19]. This data set is very challenging since there is a lot of occlusions and often only top part of the person is presented.

We follow the scheme of evaluation presented in [19]. It allows us to be the most comparable with published results. First, one image for each person was randomly selected as the database image. Then, the rest was used as query images. This procedure was repeated 10 times and the average performances are presented in Figure 6.

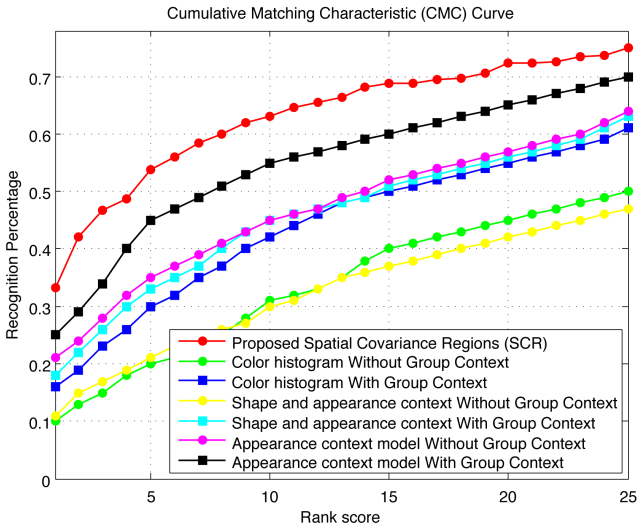

Figure 6. Cumulative matching characteristic (CMC) curves obtained on i-LIDS data set. Our descriptor, noted as Spatial Covariance Regions (SCR) outperforms state of the art methods evaluated in [19]. The methods: Appearance Context model and Shape and Appearance Context model has been proposed in [18].

Three descriptors for the person re-identification problem has been evaluated in [19]. One is the simple color histogram using 16 bins. Remaining two are the appearance context model [18] and the shape and appearance context model [18]. These three descriptors were evaluated in two ways: with and without group context information. Group context information stand for the information about the people around the individual. This information reduces the ambiguity in person re-identification problem (see details in [19]).

Descriptors which used only an image of individual (no group context information) are noted as circle-doted lines. The rest (rectangle-doted lines) took advantage of group context information. The evaluation results obtained in [19] are presented in Figure 6 together with our performance. It is worth of noting that our approach was evaluated on images which contain only one person of interest (without group context information). We can see that our descriptor (noted as Spatial Covariance Regions, SCR) outperforms all descriptors presented in [19].

SCR outperforms significantly all descriptors without group context information. The best performance reported in [19] is that approximately 21% of the queries achieved a top ranking true match. SCR matching rate for top rank is around 32%. Moreover, descriptors which have used context information also performed worse than our SCR. The best rate for these descriptors is approximately 25% which is still less than our 32%.

It is worth of noting that the performance is not very high

because the person images from the i-LIDS data are very challenging since they were captured from non-overlapping multiple camera views subject to significant occlusions and large variations in both view angle and illumination.

6. Conclusions

We have proposed a spatial covariance regions descriptor for the person re-identification problem. The evaluation has been performed on 2008 i-LIDS Multiple-Camera Tracking Scenario (MCTS) data set. It has been shown that this spatial covariance descriptor is effective as a discriminative feature applied to the person re-identification problem. Our SCR descriptor outperforms state of the arts methods evaluated in [19].

In the future work, we will apply our descriptor to video sequence where the information from consecutive frames can be used to create a discriminative human signature. Furthermore, the covariance matrices can be represented as a connected Riemannian manifold where the distance between covariance matrices can be formulated as a geodesic distance what should be also investigated. Moreover, the extraction of foreground seems to be a bottleneck in the re-identification approaches. Future investigation will include methods which remove background from a human signature.

Acknowledgements

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

References

[1] 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. 2

[2] W. Förstner and B. Moonen. A metric for covariance matrices. In *Quo vadis geodesia ...?, Festschrift for Erik W. Grafarend on the occasion of his 60th birthday, TR Dept. of Geodesy and Geoinformatics, Stuttgart University*, 1999. 4

[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

[5] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Addison-Wesley Longman Publishing Co., Boston, MA, USA, 2001. 3

[6] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, pages 1458–1465, 2005. 4

[7] 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. 5

[8] 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

[9] 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

[10] S. D. Hordley, G. D. Finlayson, G. Schaefer, and G. Y. Tian. Illuminant and device invariant colour using histogram equalisation. *Pattern Recognition*, 38:2005, 2005. 3

[11] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society. 4

[12] 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

[13] G. Overett, L. Petersson, N. Brewer, L. Andersson, and N. Pettersson. A new pedestrian dataset for supervised learning. *In IEEE Intelligent Vehicles Symposium*, 2008. 2

[14] 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

[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] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *Proc. 9th European Conf. on Computer Vision*, pages 589–600, 2006. 3, 4

[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. 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, 5

[19] W.-S. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *British Machine Vision Conference, BMVC*, London, 2009. 5, 6