Eurographics Workshop on 3D Object Retrieval (2013)
S. Biasotti, I. Pratikakis, U. Castellani, T. Schreck, A. Godil, and R. Veltkamp (Editors)

Features Accumulation on a Multiple View Oriented Model for People Re-Identification

J. García[1], C. Kambhamettu[2], A. Gardel[1], I. Bravo[1], J.L. Lázaro[1]

[1] Department of Electronics, University of Alcalá, Alcalá de Henares, 28871 Madrid, Spain.
[2] VIMS Laboratory, Dept. of Computer and Information Sciences, University of Delaware, Newark, DE 19716, United States.

Abstract

People re-identification process provides relevant information in order to understand the scene. In this paper it is proposed a multiple view oriented model for performing people re-identification in a camera network. An appearance model for different perspectives is generated from people trajectories. Global and local features besides path orientation are extracted from each person image given a short-term tracking. Experimental results based on different sets of features demonstrate the effectiveness of our proposal.

Categories and Subject Descriptors (according to ACM CCS): I.4.8 [Image Processing and Computer Vision]: Scene Analysis—

1. Introduction

People re-identification is the visual recognition for the same person in disjoint camera views, considering a certain set of different identities. The knowledge about the identities of people enables a system to fully understand of the scene [BCBT11]. Surveillance systems where a camera network is monitoring different scenarios represent a perfect example to perform these tasks. Typically, this type of surveillance network is composed by different cameras with non-overlapping field of views located in a large area. To address this goal, a feature set is extracted from each person detected on a captured image, process known as identification. The features to model the appearance are classified into two groups: global and local. The difference between them is the area where the feature is applied. Local features focus on information where there are interest points while global features are present in a large extent of a person. Typically, a set of features that define a person is referred to as signature. Re-identification process is carried out through a similarity measure comparing signatures between different views. When using non-overlapping camera views, several problems are added to the re-identification process. Different perspectives are captured when a person crosses different cameras; normally are defined four poses: front, back, left and right side. This aspect is needed to take into account when using different views to create a signature. The people appearance

Figure 1: *Examples of VIPeR dataset [GT08]. Each column is one pair of images corresponding to the same person.*

may be affected by lighting changes between different locations. The fusion of global and local features help create more robust signatures. Temporal and spatial constraints can reduce effects of transition time between disjoint cameras. Each camera captures an uncontrolled environment from a distance, this means that the recognition of biometric aspects such as face, eyes or gait detection, provide low reliability due to difficult segmentation, using low resolution and low framerate video. In the Figure 1, some examples about the problems earlier described are shown. Other methods extract

features from multiple images and use learning process in order to obtain a signature more suitable respect to perspective changes. Finally, there are contributions more focused in a robust distance measure trying quantify and differentiate features by learning the distance measure that is most likely to get correct matches [ZGX12].

In this paper, we propose a multiple view oriented model to perform people re-identification in a camera network where an appearance model for different perspectives is generated from the people trajectory. The paper is organized as follows. In Section 2 previous related works are reviewed respect to the subject under discussion. In Section 3 we present our contribution in the people re-identification field, and finally, in Section 4 an analysis of the results processing public datasets is provided.

2. Related Work

Most contributions generate a signature based on appearance model from a single or multiple images corresponding to same person from a set of features. These can be classified into two groups: global and local features. Typically, global features are obtained via chromatic histogram from different color spaces [CAK10]. Texture feature extraction is other option to characterize the people appearance model [GT08]. These features are more stable than color features so the signature is more independent respect to viewpoint changes. Other type of signatures are complemented by local features. [OSP09] use local histograms from *HSV* color space. They are determined in areas around specific points of interest. Furthermore, each region of interest is used by a *SURF* (Speeded Up Robust Features) descriptor. Other local features like a *Haar-like* features are extracted to define a signature as in [BCBT10]. [BS11] use gradient location and orientation histogram (*GLOH*), combining ideas from both *SIFT* and shape context.

In the other hand, local features are descriptors applied over local interest point as Harris detector, Hessian-Laplace, etc. Other type of descriptors used to create a signature are covariance descriptors as proposed in [BCBT11]. Using the dense descriptors philosophy, an overlapping grid structure is applied to the image. A cell is defined in each point of the grid where is calculated the covariance descriptor, so the signature is composed by a large vector of covariance descriptors. Researches as [CAK10] determine the silhouette from background/foreground update process in order to only include information belonging to the person. Some authors propose the body segmentation in parts. In [FBP*10] the body is divided in three parts: legs, trunk and head. Each extracted feature is weighted according to the body part and respect to the distance between the half body (vertical orientation) and the feature location. Another possible classification of appearance models takes into account the number of person images in order to create the signature. Single shot methods [GT08] only use one image where appears the person to extract features, while an image set is required in multiple shots methods [BCBT11]. Select a single or multiple method depends on the availability/use of tracking information. Multiple shots methods form a signature more independent that single methods respect to captured viewpoint or lighting changes. In contrast fusion techniques are necessary to combine information from multiple images. The major problem in the re-identification methods are variations in appearance when a person is captured from different perspectives. Given a pair of signatures corresponding to a remarkable different viewpoint, the match between local features decreases and only the influence respect to global features provides similarity in the matching process. To mitigate this problem, we propose a multiple view oriented model which represents a signature composed of different images. Each image of the model represents a different updated viewpoint.

3. Multiple View Oriented Model

3.1. Model Overview

Since a camera provides a short-term tracking of people crossing the scene, multiple images may be used to model the people appearance. These methods, referred to as multiple shots, merge the information extracted from all the images to create a signature more suitable to perspective or people orientation changes. We propose a multiple view oriented model (*MVOM*) which represents a signature composed of different images where each one provides an updated view of the person. The model is composed by different feature vectors, each one extracted from an update view. Two events can occur to add an image to the model, these are explained below:

- *Direction changes*: The trajectory generated by a person across the scene captured by a camera may contain direction changes due to static objects which are located in the scene, crosses between people or his/her own trajectory. These situations are exploited by our *MVOM* to obtain different perspectives of the person and are quantified using an orientation parameter according to camera location. An image is added to the model when it is detected a strong change in a short period of time or a weak change in a large period of time.
- *Periodic acquisition*: Complementary, every certain period of time an image is incorporated into the model. Thus, several images with the same orientation are collected to obtain an appearance model less dependent. This type of data acquisition for appearance model leads to a larger database of possible similar images that could be later refined to reduce amount of data to be processed.

3.2. MVOM Paramerization

Formally, the common re-identification process is defined as follows. Let $\mathcal{C} = \{1, 2, ..., N_C\}$ a set of non-overlapping camera views where N_C is the total number of camera views

involved in the re-identification process. Assuming that a set of people images $\mathcal{I} = \{i_n \mid n = 1,2,...,N_I\}$, where N_I is the total number of images obtained from C and each $i_n \in \mathbb{R}^{W \times H}$ where $W \times H$ is the size of an image, it is defined $\mathcal{P} = \{1,2,...,N_P\}$ as a set of labels where N_P is the total number of labels. Each element $\mathcal{P}_n$ represents a single person across the field of view captured by a camera network. The signature of an image i_n belonging to an individual is represented by a features vector $\mathbf{x}_n \equiv \mathbf{f}(i_n)$. The person which corresponds to the image i_n must satisfy $y_n \equiv y(i_n)$ to correspond to a label y_n. So that the set of images, representing to the same person, is defined as $\mathcal{S}_z \subseteq \mathcal{I}$ where each image $i_n = \{i \mid y(i) = z\}$. Given a image from the database and a new candidate image in order to share the same label, their feature vectors should match, which is given by a distance function expressed as:

$$M(y_{n1}, y_{n2}) = m(\mathbf{f}(i_{n1}), \mathbf{f}(i_{n2}))$$

where $m(,)$ is a certain specific function to measure the similarity between two signatures. Thus, depending on the descriptors chosen, the function $m(,)$, should be different.

In our proposal, we define subsets $\mathcal{S}_z^{'}$ where all images belong to the same label from short-term tracking. The subset size depends on two factors: time within the field of view of the camera and trajectory changes of the person. If the tracking is robust the same label is directly assigned to the whole subset. We propose extract a features vector and a orientation value $\mathbf{x}_n \equiv [\ \mathbf{f}(i_n) \quad \theta(i_n)\ ]$ from each image which has been added to the model. In this way a subset $\mathcal{S}_z^{'}$ can be expressed as:

$$\mathcal{S}_z^{'} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} \mathbf{f}(i_1) & \theta(i_1) \\ \vdots & \vdots \\ \mathbf{f}(i_N) & \theta(i_N) \end{bmatrix}$$

Given a re-identification process between two subsets $\mathcal{S}_{z1}$ and $\mathcal{S}_{z2}$ as the query subset and a selected subset from available subsets corresponding to a label, respectively. The function $m(,)$ is calculated for each pair of vectors $\mathbf{f}(i_n)$ within a common range of $\theta(i_n)$. In cases where a common range is not possible, an all-to-all matching is calculated. The best match is selected in order to join the query subset with the same label.

3.3. Accumulation of Global and Local Features

Different steps are necessary in order to extract a set of features from a image where multiple people can appear. A background/foreground segmentation algorithm is the first step to detect all non-static objects as proposed in [CAK10] where a method based on an adaptive spatial-colorimetric model is presented. Moreover, an estimator of orientation from the short-term tracking algorithm is responsible to define a orientation scalar value between the person trajectory and the camera. In this paper, we assume masks of silho ette that define only information of the person to compu the set of features and it is established a orientation value f each silhouette that appears in the image. Different featur are accumulated in a vector in order to encode the visual a pearance of a person. A feature vector is constructed fro each image, thus a label of person is composed by multip feature vectors.

4. Experimental Results

In this section, it is shown different experiments to evalua the proposed multiple view oriented model. The results a presented through two common curves used in people r identification: Cumulative Matching Characteristic (CMC curve and Synthetic Recognition Rate (SRR) curve. Tl CMC curve provides the recognition percentage which re resents the expectation of finding the correct label with major matched labels. Furthermore, the SRR curve provid the probability that any label between the best matched l bels is correct. The dataset used is proposed in [MM12]. It composed of three non-overlapping cameras in real scenar where 61 different individuals cross the entire scene. Eac camera capture images with a resolution of 320x240 pixe and a framerate of 1 frames per second. The size of peop images is normalized to 96x48. This dataset include differe viewpoints of the people and lighting changes. A short-ter tracking is collected for all people in each camera. Thu the overall re-identification process is composed of 183 sul sets $\mathcal{S}_z^{'}$. Other benchmark datasets, such as VIPeR, can n be used to show comparisons as it only provides two si gle views of each person. To compare the effectiveness the proposed model respect to other contributions, two se of features are used to implement the multiple view or ented model proposed in [FBP*10] and [MM12]. Also th re-identification functions $m(,)$ proposed by the authors a used, in order to show the advantages from using a mult ple view oriented model as proposed, without improving th features of individual images. The first set of features is con posed by weighted color histograms, maximally stable col regions (*MSCR*) and recurrent high-structured patches. A features are weighted respect to vertical axis and classify i two principal regions. The area corresponding to the head not taken into account to extract features because it normall contains very few pixels providing little information abo the person. The set proposed in [MM12] exploit three loc features: *SIFT* features, pyramid of histograms of orient tion (*PHOG*) and Haralick texture features. The authors in plement the same body partitioning. *SIFT* features are use to extract the chromatic appearance from *HSV* color spac in different interest points. The *PHOG* is calculated in thre levels and they are accumulated into a single oriented hi togram. Haralick texture features are determined in the r gions where it is concentrated more information (legs an trunk regions). Figure 2 represents CMC and SRR curve Our methods is compared with single-shot case propose

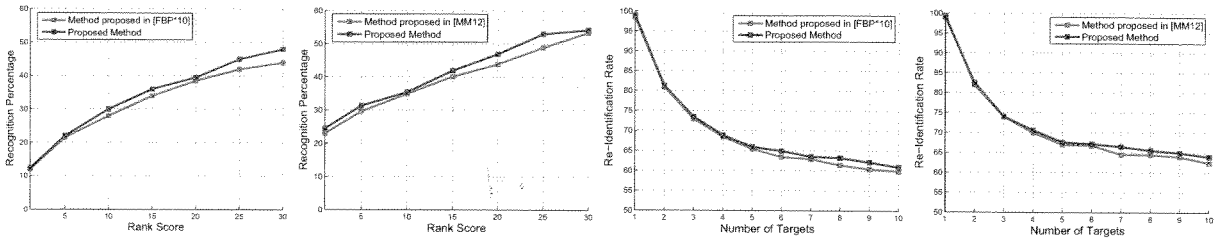

Figure 2: *Performances on dataset proposed in [MM12]. In (a) and (c) the proposed method is compared with [FBP*10]. In (b) and (d) the proposed method is compared with [MM12].*

in [FBP*10] and [MM12], respectively. The multiple view oriented model perform a single-shot matching process using a query image with a database image with similar orientation. The major problems are lighting changes and the low resolution of images. Consequently, poor feature vectors are obtained in each pair of images resulting in a calculated distance not satisfactory. *Rank Score = 1* is the rank score result which represents the correct re-identification result. In figure 2.a, the proposed method provides a recognition percentage of 11.3% while the method proposed in [FBP*10] presents a 10% in *Rank Score = 1*. Moreover, the difference between recognition percentages increases in other rank score. A similar case occurs in figure 2.b, where the proposed method provides a recognition percentage of 22.9% and the method proposed in [MM12] have a 21% in *Rank Score = 1*. The proposed model reduces perspective changes as multiple vectors with different orientations represent a more robust appearance of the person.

5. Conclusions

In this paper it is presented a multiple view oriented model to carry out people re-identification process. Due to variations in appearance from different perspectives of the person, we propose a model composed of different views of the person. Each image is represented by a features vector and a parameter of estimated orientation extracted from person trajectory. The proposed model works without training stages to carry out the re-identification process and is not necessary previously knowledge about the full dataset. Different experiments have been performed with two feature sets proposed by other authors to provide a reliable comparison. The major problems in the re-identification process are due to lighting changes and low image resolution. However, problems related to the appearance variation due to perspective changes have been reduced.

6. Acknowledgements

This research was supported by the University of Alcalá (ref.UAH2011/EXP-001), throught the project Sistema de Arrays de Cámaras Inteligentes (SACI).

References

[BCBT10] BAK S., CORVEE E., BRÉMOND F., THONNAT M.: Person re-identification using haar-based and dcd-based signature. In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on* (29 2010-sept. 1 2010), pp. 1 –8. 2

[BCBT11] BAK S., CORVEE E., BREMOND F., THONNAT M.: Multiple-shot human re-identification by mean riemannian covariance grid. In *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on* (30 2011-sept. 2 2011), pp. 179 –184. 1, 2

[BS11] BAUML M., STIEFELHAGEN R.: Evaluation of local features for person re-identification in image sequences. In *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on* (30 2011-sept. 2 2011), pp. 291 –296. 2

[CAK10] CONG D.-N. T., ACHARD C., KHOUDOUR L.: People re-identification by classification of silhouettes based on sparse representation. In *Image Processing Theory Tools and Applications (IPTA), 2010 2nd International Conference on* (july 2010), pp. 60 –65. 2, 3

[FBP*10] FARENZENA M., BAZZANI L., PERINA A., MURINO V., CRISTANI M.: Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on* (june 2010), pp. 2360 –2367. 2, 3, 4

[GT08] GRAY D., TAO H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Computer Vision âĂŞ ECCV 2008*, Forsyth D., Torr P., Zisserman A., (Eds.), vol. 5302 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2008, pp. 262–275. 1, 2

[MM12] MARTINEL N., MICHELONI C.: Re-identify people in wide area camera network. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on* (june 2012), pp. 31 –36. 3, 4

[OSP09] OLIVEIRA I., SOUZA-PIO J.: People reidentification in a camera network. In *Dependable, Autonomic and Secure Computing, 2009. DASC '09. Eighth IEEE International Conference on* (dec. 2009), pp. 461 –466. 2

[ZGX12] ZHENG W., GONG S., XIANG T.: Re-identification by relative distance comparison. *Pattern Analysis and Machine Intelligence, IEEE Transactions on PP*, 99 (2012), 1. 2