Robust Descriptors Fusion for Pedestrians' Re-identification and Tracking Across a Camera Network

Ahmed Derbel[1], Yousra Ben Jemaa[2], Sylvie Treuillet[3], Bruno Emile[3], Raphael Canals[3] and Abdelmajid Ben Hamadou[1]

[1]*MIRACL Laboratory, SFAX University, Sfax, Tunisia*

[2]*Signal and System Research Unit, TUNIS University, Tunis, Tunisia*

[3]*PRISME Laboratory, ORLEANS University, Orleans, France*

Keywords: People Identification and Tracking, Multi-camera, Cascade of Descriptors, AdaBoost, Cumulative Matching Characteristic.

Abstract: In this paper, we introduce a new approach to identify people in multi-camera based on AdaBoost descriptors cascade. Given the complexity of this task, we propose a new regional color feature vector based on intra and inter color histograms fusion to characterize a person in multi-camera. This descriptor is then integrated into an extensive comparative study with several existing color, texture and shape feature vectors in order to choose the best ones. We prove through a comparative study with the main existing approaches on the VIPeR dataset and using Cumulative Matching Characteristic measurement that the proposed approach is very suitable to identify a person and provides very satisfactory performances.

1 INTRODUCTION

People identification and tracking is an active research area that can be applied in video surveillance, behavioral analysis and blind guiding. This complex recognition process needs the use of robust descriptors allowing a good pedestrian labeling despite several complex problems depending on the work context.

In a mono-camera context, persons' tracking consists in periodically locating pedestrians who appear in a single camera field of view. This task, although intuitive for the human cognitive system, is very difficult to automate and requires good management of the great shape silhouettes' variations as well as partial occlusions which significantly affect the pedestrian images.

In some cases, the installation of several non-overlapping cameras is necessary to cover wide areas. So, to ensure robust multi-camera tracking, it is necessary to solve the pedestrians' re-identification problem. In fact, pedestrians' re-identification consists in identifying persons who appear in a camera field of view and deciding if they have previously been observed. To address this challenging objective, a robust identification process is required to manage large pose variations, changes in lighting conditions, intrinsic and extrinsic camera parameter variations in the network, scale variations and finally partial occlusions.

To ensure multi-camera people re-identification and tracking, several approaches have been suggested in the literature. In a mono-camera context, some existing approaches use a motion model such as a Kalman filter or a particle filter to predict the position of tracked persons. This type of approach also requires appearance models like color histogram (Tung and Matsuyama, 2008) or a fusion of color histograms and histograms of oriented gradients (Yang et al., 2005) to make the particles weight correction. In a multi-camera context, many works are based on merging priori knowledge on the camera network such as probabilities and inter-camera transition time with appearance models such as color histograms (Nam et al., 2007) or (Gilbert and Bowden, 2006) to ensure multi-camera people tracking and re-identification. Some others works use only appearance models such as color histogram (Cai et al., 2008), regional histograms (Alahi et al., 2010) or spatiogram (Truong Cong et al., 2010) to ensure proper

Derbel A., Jemaa Y., Treuillet S., Emile B., Canals R. and Ben Hamadou A..
Robust Descriptors Fusion for Pedestrians' Re-identification and Tracking Across a Camera Network.
DOI: 10.5220/0004275500920097
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2013), pages 92-97
ISBN: 978-989-8565-48-8

pedestrians' identification. The fusion of color and texture descriptors is also a strategy used in (Gray and Tao, 2008) (Prosser et al., 2010). This last category of approaches is the best suited for people re-identification and mono or multi-camera persons' tracking because of its simplicity and its similarity to the human identification cognitive system. That's why we have adopted this strategy in our work.

In this paper, we propose a new regional color descriptor which will be subsequently integrated into an extensive comparative study with different existing descriptors based on color, texture and shape information applied to people identification. This comparative study will allow us to select the most robust of them which will be merged using Adaboost algorithm to ensure optimal performances.

This paper is organized as follows: Section 2 describes the different descriptors including the one we propose. In order to select the best descriptors, we present in Sections 3 and 4 two performed tests used to evaluate the representative power of these feature vectors in terms of people re-identification and tracking. Section 5 illustrates the proposed approach consisting in merging descriptors already selected and presents a comparative study with several existing works. Conclusions and future works end the paper.

2 FEATURE VECTORS

In this section, we introduce our new descriptor and several existing color, texture and shape descriptors used to represent a pedestrian in multi-camera.

2.1 Proposed Descriptor

Regional color histograms are frequently applied in multi-camera people identification according to their robustness against large pose variations and partial occlusions (Alahi et al., 2010). Despite the application of the colors normalization step, the most color descriptors remain very sensitive to lighting conditions changes that affect pedestrians' silhouettes in multi-camera. For this reason, we propose a new color descriptor that integrates intra-people regional histograms to ameliorate the pedestrian characterization and reduce the impact of lighting conditions variations across any network of camera.

The first step in implementing this descriptor consists in quantizing the persons' images to reduce the effect of large color variation and then dividing horizontally each quantized image into four blocks in order to calcule a regional histogram for each block (Figure 1). This division allows considering the articulated nature and color differences between human body parts. For each person matching, two types of feature vectors are calculated. The first feature vector type of size 6 contains Bhattacharyya distances between the four bands color histograms existing in a person's image: it characterizes the intra-class similarity (Figure 1.a). A second feature vector type of size 4 contains the Bhattacharyya distances between each two color histograms of the same band existing in two different persons' images (Figure 1.b): it characterizes the inter-class similarity. Before performing the persons' matching, intra and inter feature vectors are normalized by dividing each element by the sum of the corresponding vector.

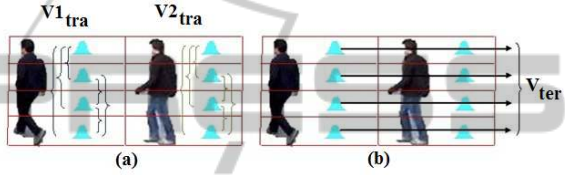

Figure 1: Intra-class (a) and inter-class (b) vector of two people.

The similarity measure between two persons' images depends on intra and inter feature vectors according to Equation 1:

$$similarity = w_1[1 - std(V_{ter})] + w_2[1 - (std(V1_{tra} - V2_{tra}))] \quad (1)$$

where $V1_{tra}$, $V2_{tra}$ are the feature vectors for both intra-persons, V_{ter} is the inter-person characteristic vector, *std* is the standard deviation which represents the difference whatsoever intra or inter people, and w_1 and w_2 are the weights assigned respectively to intra and inter-person similarity. According to an experimental study, parameters w_1 and w_2 are fixed to 0.5 to ensure optimal performances.

2.2 Existing Descriptors

To represent persons in multi-camera context, we have tested as color descriptors, the color histogram and the spatiogram which is a color histogram extension incorporating spatial intensity distribution. As texture descriptor, we have used the co-occurrence matrix and the Schmid and Gabor filters. Also, we have tested the most popular shape descriptors used in literature like the histogram of oriented gradients (HOG) and the Zernike and Hu moments. All already introduced descriptors are more explained in (Derbel et al., 2012).

3 PEOPLE RE-IDENTIFICATION

3.1 VIPeR Dataset

To evaluate the performance of each descriptor in terms of re-identifying a person in a camera network, we have used the VIPeR image database (Gray and Tao, 2008) that contains *632* pedestrians' image pairs taken from different points of view with large lighting conditions variations. All images are spatially normalized as 128×48 pixels.

3.2 Evaluation Protocol and Results

Before extracting color and texture feature vectors, a Greyworld normalization is applied to all the pedestrians' images. This normalization reduces the effect of large color variations caused by lighting condition changes between cameras.

The descriptors extraction phase is performed on each image according to different settings. For each category of descriptors, different configurations are tested by varying the color space, the quantization step (number of bins) for color descriptors, only the quantization step for texture descriptors and the block size for the Histogram of oriented gradients which is the only local shape descriptor used in this work.

Taking into account all configurations, a set of *41* feature vectors are extracted from each image. To minimize the computation time, we have used only the first half of the VIPeR dataset (that is to say *N=316* pedestrians' images pairs). Next, we calculate the inter-images similarity by computing the Bhattacharyya distance between each descriptor variant: $41 \times N^2$ distance measurements. To have distance measurements between 0 and 1, all extracted feature vectors are normalised by dividing each element by the sum of the corresponding vector.

The comparison is based on the CMC precision measurement (Alahi et al., 2010) (Gray and Tao, 2008) (Prosser et al., 2010) that represents the correct identification rate according to a rank. To do this, each descriptor's distance measurements are stored in a matrix of size $N \times N$. The diagonal contains the similarity of the same pedestrian pair that should be the maximum in each matrix line. For each matrix row noted *i*, *d(i)* represents the number of matches showing greater similarity compared to the diagonal. For a given rank *n*, the recognition rate (normalized between 0 and 1) cumulates the results of all lines boolean test (when $d(i) \leq n$). Figure 2 shows the performance of the best descriptors of each category.

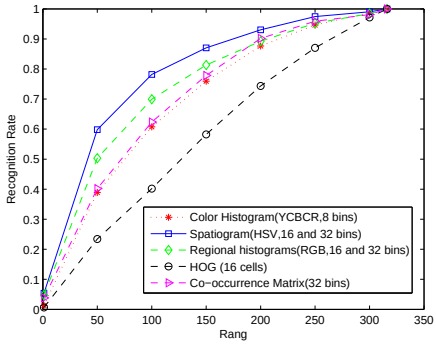

Figure 2: Best descriptors of each category.

4 PEOPLE TRACKING AND RE-IDENTIFICATION

4.1 PETS 2009 and CVLAB Databases

To evaluate simultaneously the descriptors performances in terms of multi-camera people tracking and re-identification, we have used the two pedestrians' images databases PETS 2009 (Sharma et al., 2009) and CVLAB (Fleuret et al., 2008). These two databases contain many peoples' sequences filmed by different cameras and including large pose and lighting conditions changes, scale variation and partial occlusions.

We have selected *606* person's images from the PETS 2009 database representing four pedestrians and *900* person's images from the CVLAB database representing three pedestrians taken from three different cameras. For each person filmed by a camera, all the extracted images are temporally successive allowing us to evaluate the multi-camera pedestrians' tracking and re-identification performances.

Preprocessing includes the following steps: (1) background subtraction, (2) morphological filtering, (3) people detection and bounding box tracing and finally (4) spatial and color image normalization. In both databases, all persons are in standing position. For this reason, a foreground detected region is considered a pedestrian when its height is twice greater than its width.

4.2 Evaluation Protocol and Results

This second test consists in evaluating the performance of the best descriptors for each category selected from the previous comparative study (Figure 2) simultaneously in terms of people tracking and re-identification. In fact, the representative power of

each descriptor can be measured by the average similarity for the same person $\hat{SP}$, the average similarity for different persons $\hat{DP}$, and the number of false alarms that represents the number of times where the same person's similarity is less than the different persons' similarity ($S_{m,i|m,j} < S_{m,i|n,j}$). All these parameters are clarified by Equation 2.

$$\hat{SP}_{m,i|n,j} = \frac{1}{Zp}\sum S_{m,i|n,j} \quad if \;\; m = n \qquad (2)$$

$$\hat{DP}_{m,i|n,j} = \frac{1}{Zn}\sum S_{m,i|n,j} \quad if \;\; m \neq n$$

$$S_{m,i|n,j} = \frac{1}{T}\sum_{t=1}^{T} d_B(V_{m,i,t}, V_{n,j,t+1})$$

$$T = min(nb_{m,i}, nb_{n,j})$$

where m and n represent the person's identities, i and j represent the different cameras used, Z_p and Z_m represent respectively the number of same-person's (when m=n) and different-persons' (when $m \neq n$) matching, S represents the average similarity between two pedestrians' sequences, t is the chronological order of the person's image in the sequence, T is the number of matches made to compare two persons' sequences, d_B is the Bhattacharyya distance between two feature vectors, V represent the feature vector and nb is the number of frames in a sequence.

For each descriptor, $\hat{SP}$ and $\hat{DP}$ are calculated by making temporally successive matches between the same camera pedestrians' images (tracking case) and between the different cameras pedestrians' images (re-identification case). A robust descriptor must accentuate the gap between $\hat{SP}$ and $\hat{DP}$ and minimize the number of false alarms. Table 1 summarizes all the obtained results.

4.3 Performances Analysis and Interpretation

The results presented in Table 1 confirm those obtained from the first comparative study (Figure 2). In fact, it is very clear that the descriptor ranking according to the people representation power is as follows: 1) Spatiogram (HSV 16 bins), 2) Proposed descriptor (RGB, 16 bins), 3) Co-occurrence matrix (32 bins), 4) Color histogram (YCbCr, 8 bins) and 5) HOG (16 cells).

Since color and texture descriptors are the best feature vectors, we propose to combine them using the AdaBoost algorithm in order to find the best configuration that ensures good performances in terms of pedestrians' re-identification in multi-camera.

5 PROPOSED APPROACH AND COMPARATIVE STUDY

In this section, we introduce firstly the proposed approach intended to ensure a proper multi-camera people identification and secondly a comparative study with the main existing works in literature.

5.1 Proposed Approach

In the first stage, we have divided randomly the VIPeR dataset into two equivalents portions (the first one is for learning and the second one is for testing). Both portions contain *N=316* pairs of pedestrians, each taken from two different cameras. During the learning phase, we have calculated an $N \times N$ matrix *"D"* containing Bhattacharyya distances between two cameras learning images for the most robust descriptors that are shown in Figure 3. Each matrix contains N positive training examples (positive distribution in diagonal) and $N \times (N-1)$ negative training examples (negative distribution otherwise). Three distributions can be applied on each matrix *"D"* such as Gaussian, Exponential or Gamma distribution (Gray and Tao, 2008). According to an experimental study, we choose the Exponential distribution that maximizes the number of *1* in the diagonal and the number of *-1* otherwise for all selected descriptors.

In fact, an Exponential matrix *"E"* is obtained from each descriptor's matrix *"D"* as follows (Equation 3).

$$E(i,j) = \begin{cases} 1 \;\; if \; a \times D(i,j) + b \succ 0 \\ -1 \;\; otherwise \end{cases} \qquad (3)$$

The coefficients a and b can be expressed in term of the estimated parameters of positive distribution (in diagonal) and negative distribution (otherwise) calculated from each matrix *"D"* as Equation 4.

$$a = \lambda_n - \lambda_p, \;\; b = \ln(\lambda_p) - \ln(\lambda_n) \qquad (4)$$

$$\lambda_n = \frac{1}{\mu_n}, \;\; \lambda_p = \frac{1}{\mu_p}$$

where μ_p and μ_n are respectively the mean of positive and negative distributions in matrix *"D"*.

In the second stage, the Adaboost algorithm is applied for selected descriptors on matrices *"E"* in order to calculate a weight for each one. In this work, we have performed two experiments: (a) cascade of color descriptors and (b) cascade of color and texture descriptors. We do not use shape descriptors in the proposed cascade because they give poor performances (Figure 2). Figure 3 shows all selected feature vectors and their respective weights for each experiment.

Table 1: People re-identification and tracking performances analysis.

Performances / Descriptors	$\widehat{SP}$		$\widehat{DP}$		Number of false alarms	
	PETS 2009	CVLAB	PETS 2009	CVLAB	PETS 2009	CVLAB
Color histogram (YCbCr,8 bins)	98.91	99.54	97.73	98.24	2	2
Spatiogram (HSV,16 bins)	87.42	93.87	76.54	87.37	2	2
Proposed descriptor (RGB,16 bins)	97.38	95.67	91.94	91.33	0	1
Co-occurrence matrix (32 bins)	98.00	99.05	93.23	96.65	3	2
HOG (16 cells)	73.49	80.11	69.97	70.65	11	4

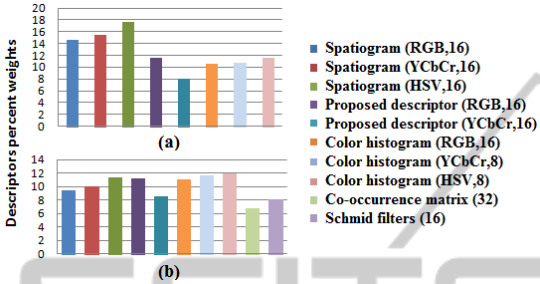

Figure 3: Descriptors percent weights: (a) color descriptors cascade (b) color and texture descriptors cascade.

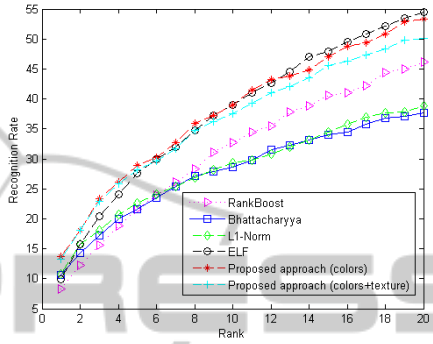

Figure 4: CMC curve for different approaches.

To reduce the search space and more separate different people, these relative descriptors weights will be applied on matrix E_T where E_T is the Exponential matrix calculated from testing images (Equation 5).

$$CSM = \sum_{i=1}^{n} \alpha_i \times E_{Ti} \quad (5)$$

where CSM is the Cascade Similarity Matrix, E_{Ti} and α_i represent respectively the Exponential testing matrix and the relative weight for the i^{th} selected descriptor.

Ambiguity cases (two different matching with the same rank) are managed by Spatiogram (HSV, 16 bins) Bhattacharyya matrix chosen because it is the strongest descriptor (Figure 2). Experimental results are presented in the next section.

5.2 Comparative Study with Existing Works

Here, we compare the two variants of our proposed approach (cascade of color descriptors and cascade of color and texture descriptors) with two categories of approaches which are tested in (Prosser et al., 2010): (i) non-learning methods (Bhattacharyya (Prosser et al., 2010), L1-norm (Prosser et al., 2010)) and (ii) learning methods using AdaBoost (ELF (Gray and Tao, 2008), RankBoost (Freund et al., 2003)).

Since we have been interested in evaluating the impact of the descriptors choice on the recognition rate, we have used the Ababoost classifier which is widely applied in literature. In fact all introduced existing approaches use color histograms and Gabor and Schmid filters to represent pedestrians. However our proposed color approach uses the color histograms, the spatiograms and the proposed descriptors. The proposed color/texture one retain all feature vectors used in the color proposed approach and add the co-occurrence matrix and Schmid filters as explained in Figure 3. Experimental results are shown in Figure 4 representing the performance for lower rank-orders (up to 20) which are the most significant in CMC measure.

Figure 4 indicates that non-learning approaches (Bhattacharyya or L1-norm) are not efficient to identify a person in a multi-camera context. This is due to the high number of possible matches and also to the large variation in poses and lighting conditions in the VIPeR dataset that cannot be bypassed by a simple distance and without using a learning phase. Figure 4 shows too that our two proposed approaches are more efficient compared to ELF since the recognition rate has increased about *3%* (up to rank 8) and RankBoost approaches even if they all use the same classifier. This proves that the used descriptors are more representative in terms of multi-camera people identification. The slight superiority of the ELF approach for ($rank \succ 12$) can be explained by using the absolute difference which has a great separation power compared to the Bhattacharyya distance used in our work.

Looking at the performances of the two proposed approaches, we can conclude that the integration of texture descriptors slightly deteriorates the performance proving that the fusion of only robust color descriptors (color histogram, spatiogram and proposed

regional histograms) represents the best strategy to identify pedestrians in a multi-camera context.

6 CONCLUSIONS AND FUTURE WORKS

In this paper, we have introduced a new regional color histograms feature vector to characterize a person which is integrated into an extensive comparative study between different existing descriptors based on color, texture and shape information applied to people re-identification and tracking in multi-camera. To ensure this objective, two separate tests have been performed. The first one consists in evaluating the performances of already introduced feature vectors in terms of people re-identification as CMC curves on VIPER pedestrians dataset. The second test, that is more generic, allows us to evaluate simultaneously the discriminatory power of these descriptors in terms of persons tracking and re-identification.

Given the complexity of the multi-camera pedestrians re-identification and the number of constraints to manage, a new approach based on a fusion of descriptors selected from two performed comparative studies is presented in this paper. Two variants of the proposed approach (cascade of color descriptors and cascade of color and texture descriptors) are tested and compared with several existing approaches. Experimental results show that the proposed color-based approach provides very satisfactory performances despite the highly articulated human body, lighting conditions changes and large pose variations.

Future work will focus on developing a robust behavioral analysis module and merging it with the proposed cascade of color descriptors to improve the multi-camera people tracking and identification performances.

REFERENCES

Alahi, A., Vandergheynst, P., and Bierlaire, M. (2010). Cascade of descriptors to detect and track objects across any network of cameras. *Computer Vision and Image Understanding*, 114(6):624–640.

Cai, Y., Huang, K., and Tan, T. (2008). Matching tracking sequences across widely separated cameras. *International Conference on Image Processing*, pages 765–768.

Derbel, A., Ben Jemaa, Y., Canals, R., Emile, B., Treuillet, S., and Ben Hamadou, A. (2012). Comparative study between color texture and shape descriptors for multi-camera pedestrians identification. *IEEE International Conference on Image Processing Theory Tools and Applications*, pages 313–318.

Fleuret, F., Berclaz, J., Lengagne, R., and Fua, P. (2008). Multi-camera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):267–282.

Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *Journal ofMachine Learning Research*, 4:933–969.

Gilbert, A. and Bowden, R. (2006). Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity. *European Conference on Computer Vision*, pages 125–136.

Gray, D. and Tao, H. (2008). Viewpoint invariant pedestrian recognition with an ensemble of localized features. *European Conference on Computer Vision*, pages 262–275.

Nam, Y., Ryu, J., Choi, Y., and Cho, W. (2007). Learning spatio-temporal topology of a multi-camera network by tracking multiple people. *World Academy of Science Engineering and Technology*, 24:175–180.

Prosser, B., Zheng, W. S., Gong, S., and Xiang, T. (2010). Person re-identification by support vector ranking. *British Machine Vision Conference*, pages 1–11.

Sharma, P. K., Huang, C., and Nevatia, R. (2009). Evaluation of people tracking, counting and density estimation in crowded environments. *IEEE Int'l Workshop Perf. Eval. of Tracking and Surveillance*, pages 39–46.

Truong Cong, D. N., Meurie, C., Khoudour, L., Achard, C., and Lezoray, O. (2010). People re-identification by spectral classification of silhouettes. *Signal Processing*, 90(8):2362–2374.

Tung, T. and Matsuyama, T. (2008). Human motion tracking using a color-based particle filter driven by optical flow. *InternationalWorkshop on Machine Learning for Vision-based Motion Analysis*.

Yang, C., Duraiswami, R., and Davis, L. (2005). Fast multiple object tracking via a hierarchical particle filter. *IEEE International Conference on Computer Vision*, pages 212–219.