EXPLOITING MULTIPLE DETECTIONS TO LEARN ROBUST BRIGHTNESS TRANSFER FUNCTIONS IN RE-IDENTIFICATION SYSTEMS

Amran Bhuiyan, Alessandro Perina and Vittorio Murino

Pattern Analysis and Computer Vision (PAVIS)
Istituto Italiano di Tecnologia, Genova, Italy.

ABSTRACT

Re-identification systems aim at recognizing the same individuals in multiple cameras and one of the most relevant problems is that the appearance of same individual varies across cameras due to illumination and viewpoint changes. This paper proposes the use of *Cumulative Weighted Brightness Transfer Functions* to model this appearance variations. It is multiple frame-based learning approach which leverages consecutive detections of each individual to transfer the appearance, rather than learning brightness transfer function from pairs of images. We tested our approach on standard multi-camera surveillance datasets showing consistent and significant improvements over existing methods on three different datasets without any other additional cost. Our approach is general and can be applied to any appearance-based method.

Index Terms— Re-identification, Brightness Transfer Function, Video surveillance

1. INTRODUCTION

Person re-identification (ReID) refers to the problem of recognizing individuals at different times and locations. A schematic illustration of the problem is given in Fig. 1a, where the task is to match detections of the same person acquired by the two cameras. Re-identification involves different cameras, views, poses and illuminations and it has recently drawn a lot of attention due to its significant role in visual surveillance systems, including person search and tracking across disjoint cameras.

The core assumption in re-identification is that individuals do not change their clothing so that appearances in the several views are similar, nevertheless it still consists in a very challenging task due to the non-rigid structure of the human body, the different perspectives with which a pedestrian can be observed, and the highly variable illumination conditions (as an example see the images of the same lady on the top of Fig. 1a).

Re-identification approaches can mainly be organized in two classes of algorithms: direct and learning-based methods. In the former group, algorithms search for the most discriminant features to form a powerful descriptor for each individual

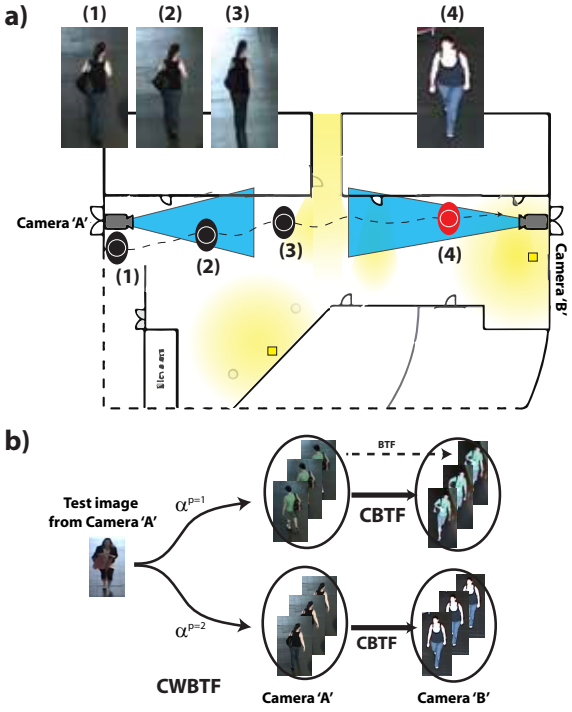

Fig. 1. a) Typical indoor system b) Overview of our approach.

[1, 2, 3, 4, 5]. In contrast, learning-based methods have techniques that learn metric spaces where to compare pedestrians, in order to guarantee a high re-identification rates [6, 7, 8, 9, 10, 11]. Finally, we find methods that learn the transformation that the appearance of a person undergoes when passing from one domain to another [12, 13, 14, 15, 16].

This work lies in the latter, camera-specific, category, which is very relevant in large video surveillance networks where individuals are observed using various cameras across a large environment.

A thorough review of the state-of-the-art shows how these approaches mainly aim to model a function to transfer "appearance" cues between cameras. For example, [12] estimate the brightness transfer function, in short BTF, to transfer the appearance for object tracking. By employing a set of N labeled pairs $\{(I_A^p, I_B^p)\}$, where I_X^p represents an observation of pedestrian p acquired by camera X, they learn multiple BTFs,

978-1-4799-8339-1/15/$31.00 ©2015 IEEE

one for each pedestrian and then rely on a *Mean-BTF*(MBTF) [12]. Porikli [17] used the same setup previously but estimate the transfer function in the form of color. Later, Prosser [13] proposed the *Cumulative-BTF*(CBTF) amalgamating the training pairs before computing the transfer function. In contrast to MBTF and CBTF which end up with a single transfer function, Datta [14] proposed a *Weighted-BTF* (WBTF) that assigns different weights to test observations based on their proximity to training observations. The latter approach showed a remarkable improvement over [13, 17] and therefore we will consider it as our main comparison.

This paper makes another step forward and taking inspiration from real scenarios it bridges the works of [12] and [13].Actually, most of the state-of-art methods learn a single BTF from a pair of images, but nowadays we have powerful tools for person tracking and we robustly have access to at least 5-10 detections of the same individual in consecutive frames [18]. A mild criticism of single pair-based method lies in how they choose labeled pairs. Fig. 1a shows 3 detections from camera A and the detection from camera B. It is easy to figure out how a transfer function learned from the pair (1)-(4) would behave differently from the one learned from the pair (3)-(4). The question we pose here is that if and how these very similar sets of images can be exploited to learn more robust and principled transfer functions. Examples of detections for the same pedestrian are shown in Fig.1b and, although at a first glance it may appear they do not add anything, we will show that considering all of them is indeed useful. More specifically, we propose here the use of the *Cumulative Weighted Brightness Transfer Function* (CWBTF). Our approach assigns unequal weights to each CBTF which, exploiting multiple detections, is more robust of the previous approaches based on single pairs. Our technique is general and strongly outperforms previous appearance transfer function based methods [14, 13, 17] and the basic framework upon which we built it. Unlike previous work, we also considered the effect of increasing the number of pedestrian in the validation set.

2. CUMULATIVE WEIGHTED BRIGHTNESS TRANSFER FUNCTION

Our goal is to find the correspondence between multiple observations of an pedestrian across a camera pair C_i and C_j. As in previous work, we assume a limited validation set of labeled detections that can be used to calculate an inter-camera CBTF. Subsequently the CBTF of each pedestrian in the validation set are weighted according to their distance with respect to the test pedestrian to form the final CWBTF.
We assume to have $N \leq 10$ subsequent frames for each of the P pedestrians in the validation set which we used to learn the transfer functions. To obtain such images, we assume the reliability of a tracking algorithm able to detect single pedestrians for less then a second[1], or alternatively one could simply propagate the detected bounding box for $\frac{N}{2}$ before and after the "labeled" detection, as illustrated by Fig. 2a. In this sense, our approach does not increase the amount of labeled data needed.

To compute the CWBTF, it is necessary to understand the extraction procedure of brightness transfer function, proposed by Javed et al. [12]. In principle, it would be necessary to estimate the pixel-to-pixel correspondences between the pedestrian images in the two camera views, however this is not possible due to self-occlusion and pose difference. Thus, to be robust to occlusions and pose differences, normalized histograms of object brightness values are employed for the BTF calculation under the assumption that the percentage of the image pixels on the observed image I_i with brightness less than or equal to ρ_i is equal to the percentage of image points in the observation I_j with brightness less than or equal to ρ_j. Now, let H_i and H_j be the normalized cumulative histograms of observations I_i and I_j respectively. More specifically, for H_i each bin of brightness value $B_1, \ldots, B_m, \ldots, B_M$ related to one of the three color channels is obtained from the color image I_i as follows:

$$H_i\left(B_m\right) = \sum_{k=1}^{m} I_i\left(B_k\right) \tag{1}$$

Where $I_i\left(B_k\right)$ is the pixel count of brightness value B_k in I_i. $H_i\left(B_i\right)$ represents the proportion of H_i less than or equal to B_i, then $H_i\left(B_i\right) = H_j\left(B_j\right)$ and the BTF function $H_{i\to j}$ can be defined:

$$H_{i\to j}\left(B_i\right) = H_j^{-1}\left(H_i\left(B_i\right)\right) \tag{2}$$

with H^{-1} representing the inverted cumulative histogram.

As first step of our approach, we first compute a normalized version of the cumulative-BTF. The cumulative histogram cH_i^p considering the N detection of a pedestrian p in camera view i, can be computed from the brightness values as:

$$cH_i^p\left(B_m\right) = \frac{1}{M \cdot N} \sum_{m=1}^{M} \sum_{n=1}^{N} I_n^p\left(B_m\right) \tag{3}$$

being M the number of different brightness levels. The normalization is necessary as bounding boxes can have different sizes. Then, similarly to Eq. 2, its CBTF is computed as follows:

$$CH_{i\to j}^p\left(B_i\right) = cH_j^{-1}\left(cH_i\left(B_i\right)\right) \tag{4}$$

At this point, given a test image and its color histogram, we compute a weighted average of cumulative-BTFs. To be robust to occlusions, camera noise or tracking errors, typical problem when considering a set of detections, we compute the root mean squared distances - ψ - as specified in the Bhattacharya distance between the test image and the set of N detections of each pedestrian p in the validation set.

$$d_\psi^p = \sqrt{\frac{\sum_{n=1}^{N} \psi(I^{test}, I_n^p)^2}{N}} \tag{5}$$

[1] In standard conditions trackers run at 25FPS

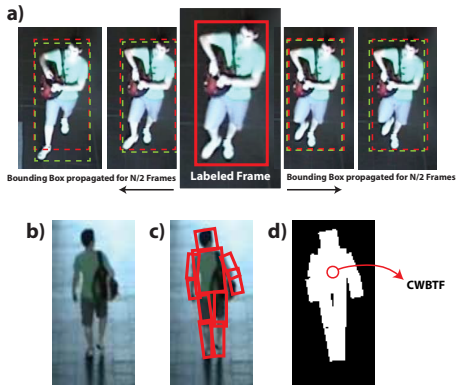

Fig. 2. a) Bounding Box propagation (red) and actual tracking result (green). b) One image from the SAIVAT-SoftBio c) Custom Pictorial Structure. d) Segmentation mask M_s derived its CPS.

It is important to note that while in [14], a weighted BTF is calculated using background only, here we are considering the full set of images.
Observations in the validation set that are close in the feature space to the test observation must be assigned a higher weights for the purpose of CWBTF calculation therefore we compute normalized weights α^p and we average the cumulative-BTFs as follows:

$$\alpha^p = \frac{e^{-\gamma \cdot d_\psi^p}}{\sum e^{-\gamma \cdot d_\psi^p}} \qquad H_{i\to j}^{CWBTF} = \sum_{p=1}^{P} \alpha^p \cdot CH_{i\to j}^p \qquad (6)$$

The parameter γ is a weight factor that can be employed to reduce αs peakier thus averaging between less images. This is akin to what done in [14], where the estimated transfer function is averaged only using the top K matches.

3. RE-IDENTIFICATION WITH CWBTF

The aim of this section is to summarize the framework we used for re-identification, nevertheless our method is clearly independent from any appearance direct method employed. The goal of re-identification is to assign to a test image seen in camera C_i an "identity" choosing among the G identities present in the gallery at camera C_j which acts as training set. We summarize our approach by the following three steps: *i)* first, we isolate the actual body appearance from the rest of the scene and we extract a feature signature from its foreground, *ii)* second, we calculate $H_{i\to j}^{CWBTF}$ to transfer the appearance from C_i to C_j using validation set as explained in the previous section, and *iii)* third, we match the transformed signatures with the gallery and select top matching identities. In the following we detail the first and the last steps.

Pedestrian Segmentation: In previous section, we showed how transfer functions are learned from the whole image, however to increase the robustness, we apply the transfer function to the foreground normalized histogram only.
We performed this separation by exploiting the Custom Pictorial Structure (CPS) [1]. CPS is based on the framework of [19] where the parts are initially located by general part detectors, and then a full body pose is inferred by solving their kinematic constraints. In CPS [1], the HOG [20] feature based part detector and a linear discriminant analysis (LDA) [21] classifier is used. Moreover, CPS [1] also used the belief propagation algorithm to infer MAP body configurations from the kinematic constraints, isolated as a tree-shaped factor graph. An example of CPS is shown in Fig. 2b-c.
Finally, we generate a segmentation mask M_s employing the following rule $M_s = 1$ if s belongs to at least one of the foreground parts, otherwise $M_s = 0$; this is shown in Fig. 2d.

Feature Extraction and Matching: The feature extraction stage consists in distilling complementary aspects from each body part in order to encode heterogeneous information, so capturing distinctive characteristics of the individuals. There are many possible cues useful for a fine visual characterization. We use standard ReID features: color histogram and maximally stable color regions (MSCR) already considered in [1, 2, 22, 4]. The transfer functions are clearly only applied to the color histogram.
As for feature matching to calculate re-id score, we use the combination of Bhattacharyya distance for histogram signature and the MSCR distance for MSCR feature, as previously done in [4].

4. EXPERIMENTAL RESULTS

In this section, we compare the performance of CWBTFs with the base framework upon which they are applied, e.g., [1] and the state-of-the-art in transfer functions [14, 13, 17]. It is important to note that *i)* all the transfer functions are applied to the same framework and *ii)* CWBTF can in principle be applied to any other appearance-based, direct, approach so comparison with other methods make little sense.
The main performance evaluation figure of merit for ReID is the Cumulative Matching Characteristic (CMC) curve which is a plot of the recognition performance vs. the ReID ranking score. It represents the expectation of finding the correct match in the top k matches. To compare the results numerically, we relied on the normalized area under the CMC (*nAUC*).

As first dataset, we considered SAIVT-SoftBio [23]. It includes annotated sequences (704 $\times$ 576 pixels, 25 frames per second) of 150 people, each of which is captured by a subset of eight different cameras placed inside an institute, providing various viewing angles and varying illumination conditions. A coarse box indicating the location of the annotated person in each frame is provided. We chose this dataset because it provides consecutive frames of same person which is suitable to evaluate the performance of our approach. We

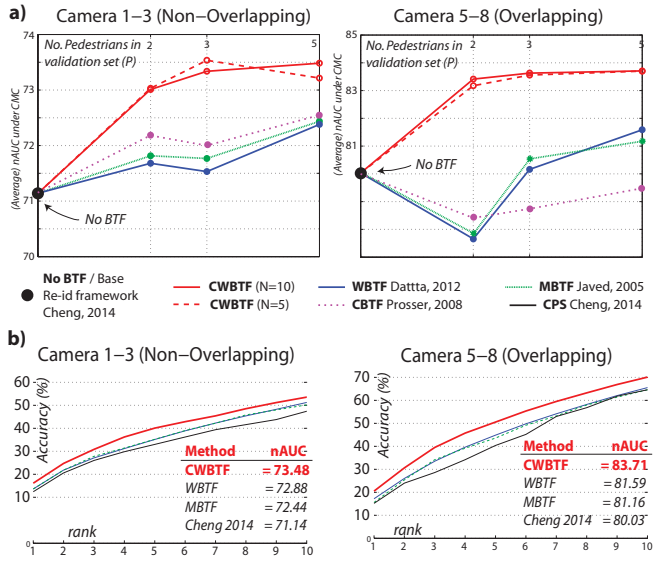

Fig. 3. a) *P*-vs-*nAUC* for two camera pairs of SAIVT-SoftBio. b) CMC and nAUC. For clarity we did not report CMC curve relative to CBTF [13] which performed more like MBTF in the lower ranks.

considered one pair of non-overlapping cameras (1-3) and one pair of overlapping (5-8) (see [23] for more details). For each camera pair, we fixed the number of identities in the gallery to $G = 50$. In all our experiments the gallery and the validation set are kept disjoint and we repeated each task 10 times by randomly picking the identities in validation and gallery. We did not investigate the effect of γ which is set to 1. In the first experiment, we varied the number of pedestrian $P = [0, 2, 3, 5]$ in the validation set, picking $N = [5, 10]$ consecutive frames each (we used the bounding boxes provided by a tracking algorithm). When $P = 0$, no transfer function is used and our framework becomes the same as [1]. In Fig.3a we report the graphs with *P*-vs-*nAUC* for both camera pairs. In both cases, transferring the brightness helped the re-identification and CWBTF significantly outperformed [14, 13, 17]. For the second experiment, we fixed $N = 5$ and, consistently with [14], we report the CMC curves setting $P = 5$. Results are shown in Fig. 3b, where the improvement is again significant especially in low ranks.

As further datasets we considered CAVIAR4REID [24] and PRID 2011 [25]. The former contains images of pedestrians extracted from outdoor sequences and it provides a challenging real world setup. From the 72 identified different individuals, 50 are captured by two cameras with 20 images for each of them and 22 from only one camera with 10 images for each them. In our experiments, we focused on only the former 50, first 5 of which are used as validation set and next 40 are used for evaluating the performance. The latter consists of images extracted from trajectories recorded from two static outdoor cameras. Images from these cameras contain a viewpoint change and a stark difference in illumination, background and camera characteristics. Results for these datasets are reported in Fig. 4a, where the improvement of our technique is again clearly visible.

As final test we evaluated the robustness of our approach to the transfer function; in the specific, we considered the *Inter-Camera Color Calibration* - ICC of Porikli [17]. Working in the exact same way of Sec. 2, we adapted (for the first time) [12, 13, 14] to transfer the color (this yield to *Mean-ICC* etc) and we compared with our *Cumulative-Weighted-ICC*. In the case of SAIVT-SoftBio, we observed a consistent degradation of the performances (for all the methods) and we did not report the results. This can be explained by the illumination conditions in indoor environment. The results of color calibration for CAVIAR4REID and PRID 2011 are instead shown in Fig.4b. It is evident how all the methods generally worked with ICC and outperformed the base method ([1], Black dot), the cumulative-weighted transfer showed again the best performance and was the only one to robustly performed in all both datasets.

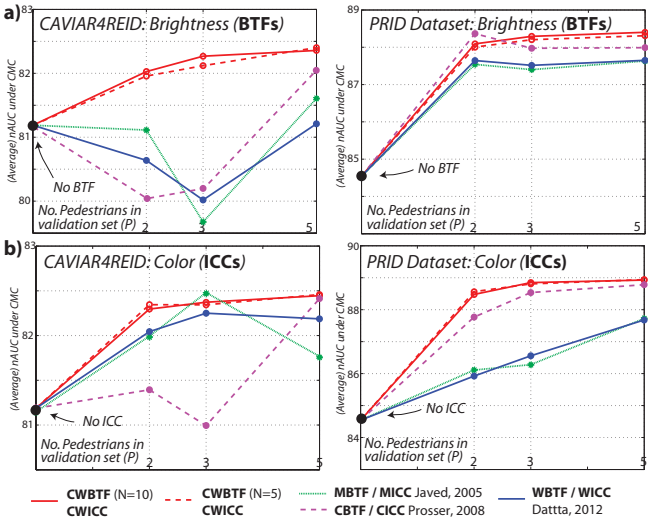

Fig. 4. *P*-vs-*nAUC* for Caviar4REID dataset and PRID 2011 datasets. On the top we transferred brightness and on the bottom we transferred color.

Analyzing all the experimental findings, it is obvious to say that our proposed method works consistently for all the datasets outperforming all the state-of-the-art.

5. CONCLUSION

This paper proposes the use of cumulative histograms from multiple consecutive detections to learn better and more robust transfer functions. Augmenting the pool of labeled data *within* a camera can be easily carried out by relying on tracking algorithms or simply by propagating the label for few frames. Our results clearly demonstrate a significant improvement over previous ways to model appearance variations.

6. REFERENCES

[1] D. Cheng and M. Cristani, "Person re-identification by articulated appearance matching. in person re-identification, isbn 978-1-4471-6295-7," *Springer*, 2014.

[2] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," *In CVPR*, 2010.

[3] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino, "Multiple-shot person re-identification by hpe signature," *In ICPR*, pp. 1413–1416, 2010.

[4] A. Bhuiyan, A. Perina, and V . Murino, "Person re-identification by discriminatively selecting parts and features," *Workshop on Visual Surveillance and Re-identification in ECCV*, 2014.

[5] I. Kviatkovsky, A. Adam, and E Rivlin, "Color invariants for person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 99, 2012.

[6] M. Dikmen, E. Akbas, T.S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," *In ACCV*, pp. 501–512, 2010.

[7] D. Gray and H. Tao, "Viewpoint invariant perdestrian recognition with an ensemble of localized features," *In ECCV*, pp. 262–275, 2008.

[8] C. Liu, S. Gong, C.C. Loy, and X. Liu, "Person re-identification: What features are important," *In The International Workshop on Re-identification in conjunction with ECCV*, vol. LNCS 7583, pp. 391–401, 2012.

[9] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," *In The International Workshop on Re-identification in conjunction with ECCV LNCS*, vol. 7583, pp. 413–422, 2012.

[10] B.Prosser, W.S.Zheng, S.Gong, and T.Xiang, "Person re-identification by support vector ranking," *In BMVC*, 2010.

[11] W. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," *In CVPR*, pp. 649–656, 2011.

[12] O. Javed, K. Shafiq, and M.Shah, "Appearance modeling for tracking in multple non-overlapping cameras," *In CVPR*, 2005.

[13] B.Prosser, S.Gong, and T.Xiang, "Multi-camera matching using bi-directional cumulative brightness transfer functions," *In BMVC*, 2008.

[14] A. Datta, L.M. Brown, R. Feris, and S. Pankanti, "Appearance modeling for person re-identification using weighted brightness transfer functions," *In ICPR*, 2012.

[15] Y. Brand, T. Avraham, and M. Lindenbaum, "Transitive re-identification," *In BMVC*, 2013.

[16] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch, "Learning implicit transfer for person re-identification," *In The International Workshop on Re-identification in conjunction with ECCV LNCS*, vol. 7583, pp. 381–390, 2012.

[17] F.Porikli, "Inter-camera color calibration using cross correlation model function," *In ICIP*, 2003.

[18] X. Chen and B. Bhanu, "Soft biometrics integrated multi-target tracking," *In ICPR*, 2014.

[19] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," *In CVPR*, 2009.

[20] N. Dalal and B. Triggs, "Histogram of oriented gradients for human detection," *In CVPR*, 2005.

[21] P.E. Forssen, "Maximally stable color regions for recognition and matching," *In CVPR*, 2007.

[22] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification.," *In Computer Vision and Image Understanding*, vol. 117:130-144, 2013.

[23] A. Bialkowski, S. Denman, P. Lucey, S. Sridharan, and C.C. Fookes, "A database for person re-identification in multi-camera surveillance networks," *In Digital Image Computing: Techniques and Application (DICTA 2012)*, pp. 1–8, 2012.

[24] D. Cheng, M. Cristani, M.Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," *In BMVC*, 2011.

[25] Martin Hirzer, Csaba Beleznai, Peter M. Roth, and Horst Bischof, "Person re-identification by descriptive and discriminative classification," in *Proc. Scandinavian Conference on Image Analysis (SCIA)*, 2011, The original publication is available at www.springerlink.com.