21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Appearance Modeling for Person Re-Identification using Weighted Brightness Transfer Functions

Ankur Datta Lisa M. Brown Rogerio Feris Sharathchandra Pankanti
IBM Research
{*ankurd, lisabr, rsferis, sharat*}*@us.ibm.com*

Abstract

Appearance of individuals across multiple cameras varies a lot due to illumination and viewpoint changes making person re-identification a challenging problem. In this paper, we describe how to model this appearance variation by using a novel Weighted Brightness Transfer Function (WBTF). In combination with powerful low-level features, we show that WBTF leads to large performance improvements by assigning different weights to different BTFs and combining them accordingly. We have compared our algorithm on two public benhmark datasets: VIPeR and CAVIAR4REID dataset, achieving new state-of-the art performance on both datasets.

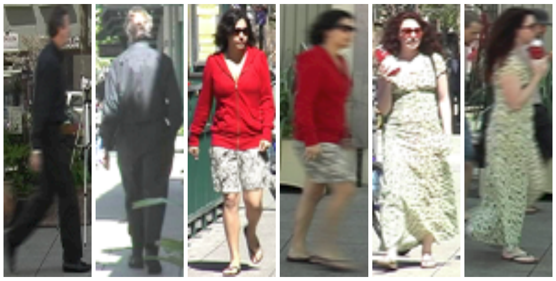

Figure 1. Person re-identification is a challenging problem due to the large variability in appearance across cameras due to illumination, viewpoint and posture changes (images from the VIPeR dataset [3]).

1. Introduction

Person re-identification refers to the problem of identifying the same person across multiple cameras or across the same camera if the person has previously exited the camera field-of-view and then reenters it again. This capability is of immense value in surveillance situations where the objective is to model long-term activities of people to uncover suspicious or anomalous behavior. However, appearance of the same person can change drastically across two cameras making re-identification a challenging problem.

There exists several indirect and direct approaches to the problem of modeling appearance variability of objects in general, and more specifically for person re-identification. Within the context of person re-identification, several authors have advocated for design of better image matching features and distance learning functions to indirectly model the appearance variability. A high-dimensional signature of texture, gradient and color features is learnt in [10] which is then projected into a latent space using Partial Least Squares. Cheng *et al.* in [1] propose to use a detailed body parts model using pictorial structures to localize body parts and then use their visual characteristics for re-identification. In contrast to feature design approaches, several authors have proposed the use of feature selection and distance learning functions to indirectly model appearance variability. Zheng *et al.* in [12] describe a Probabilistic Relative Distance Comparison (PRDC) model that aims to learn the optimal distance to maximize the matching accuracy. Authors in [4] proposed to discriminatvely select an ensemble of localized color and texture features using boosting. RankSVM [9] formulated the problem of re-identification as a ranking problem within the context of SVM classification paradigm.

As opposed to the above mentioned indirect approaches, direct approaches model the Brightness Transfer Functions (BTF) between cameras to transfer appearance. Porikli [7] and Javed *et al.* [6] estimate BTF to transfer appearance between cameras for the purpose of object tracking. These methods first compute multiple BTFs with equal weight and then rely on Mean BTF (MBTF) directly or compute BTF subspace respectively. As opposed to MBTF, Cumulative BTF

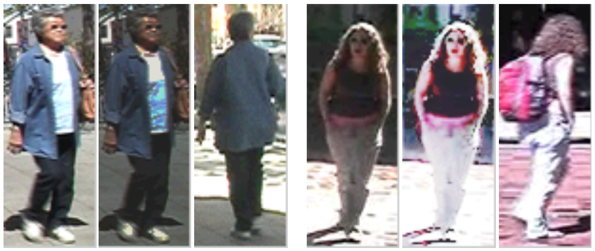

Figure 2. Two examples from VIPeR [3] that show strong change in appearance of the same person from camera 1 (first image) to camera 2 (third image) due to large changes in illumination conditions. WBTF-based appearance transfer (middle image) mitigates these strong illumination effects.

(CBTF) [8] first accumulates all the observations and then computes a CBTF to preserve observations that may be under-represented. However, CBTF approach, like MBTF approaches, assigns equal weight to all the observations.

In contrast to MBTF and CBTF, we present a novel approach that uses a Weighted Brightness Transfer Function (WBTF) that assigns unequal weights to observations based on how close they are to test observations. Observations from the training dataset that are close in feature space to the test observation are assigned higher weights for the purpose of BTF computation compared to observations that are distant. We also describe the use of a high-dimensional signature of color and texture features for the purpose of image matching. The main contributions of our work are thus two-fold: 1) A novel WBTF that models appearance BTF using weights of observations proportional to their distance in the feature space. 2) A high-dimensional color and texture signature for image matching. Together, these two contributions have lead to a new state-of-the art performance on two public benchmark dataset: VIPeR [3] and CAVIAR4REID [1]. In the next section, we describe our approach for constructing WBTF and image matching signature.

2. Weighted Brightness Transfer Function

Given a pair of observation O_i and O_j corresponding to the same observation from two cameras C_i and C_j, a BTF function H_i^j transfers a brightness value B_i in O_i to its corresponding brightness value B_j in O_j,

$$B_j = H_i^j(B_i). \tag{1}$$

In order to calculate H_i^j, we would need pixel to pixel correspondences between O_i and O_j, however this is not possible for person re-identification due to self-occlusions and viewpoint changes. Therefore, we employ normalized cumulative histograms of object brightness values for the computation of H_i^j under the assumption that the percentage of brightness values less than or equal to B_i in O_i is equal to the percentage of brightness values less than or equal to B_j in O_j [6, 5]. Note that object observations O_i and O_j correspond only to the areas of the image that represent the person. Let H_i and H_j be the normalized cumulative brightness histograms of observations O_i and O_j respectively then,

$$H_i(B_i) = H_j(B_j) = H_j(H_i^j(B_i)). \tag{2}$$

The BTF function H_i^j can thus be computed as follows,

$$H_i^j(B_i) = H_j^{-1}(H_i(B_i)), \tag{3}$$

where H^{-1} is the inverted cumulative histogram. In case of a color image, each color channel needs to be transformed separately. Associated with every training pair of images, $\{O_i, O_j\}$, we now have its BTF H_i^j.

Let O_i^T be a test image from cameras C_i. Additionally let $O_i^T(p)$ and $O_i^T(\tilde{p})$ represent a segmentation of O_i^T into person(foreground) and non-person(background) image regions respectively (we defer the description of how we segment an image into person/non-person regions until Section 2.1). Then let,

$$\begin{aligned} D^H = \{{}^kH_i^j | \alpha_k = \psi(O_i^T(\tilde{p}), O_i(\tilde{p})), \alpha_k \leq \delta\}, \\ |D^H| = K, \end{aligned} \tag{4}$$

be a set of K BTFs associated with K training images $O_i(\tilde{p})$ whose background areas are at most δ distance away in the feature-space from the background areas of the test image $O_i^T(\tilde{p})$, ψ is the bhattacharyya distance between the feature representations described in Section 2.2 and α_k is the matching cost. Then a Weighted BTF (WBTF) is defined as a linear combination of all the BTFs in D^H,

$$H_{WBTF} = \sum_{k=1}^{K} \alpha_k {}^kH_i^j. \tag{5}$$

The principal advantage of H_{WBTF} is that it assigns more weight to the BTF of those images that are closer to the test image as opposed to assigning equal weight to all BTFs. We use H_{WBTF} to map illumination from C_i to C_j and then the rank-1 (other ranks follow analogously) re-identification problem is defined as follows,

$$\arg\min_j \eta\psi(\tilde{O}_i^T(p), O_j^T(p)) + \psi(\tilde{O}_i^T(\tilde{p}), O_j^T(\tilde{p})), \tag{6}$$

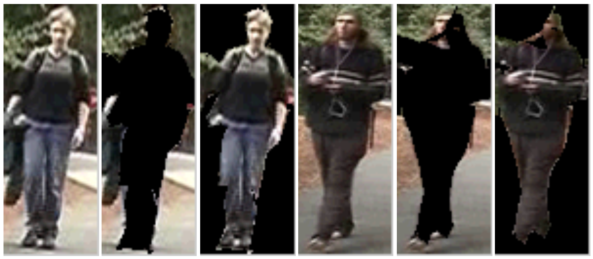

Figure 3. Two examples of segmenting an image into non-person (second image) and person (third image) regions.

where the two terms represent the matching cost for foreground and background of the transformed image $\tilde{O}_i^T$ against all the test images $O_j^T, j = \{1, \ldots, N\}$ respectively and we use $\eta = 3, K = 5$ for our experiments.

2.1. Foreground and Background Estimation

In order to estimate foreground(person) and background(non-person) regions in an image, we over-segment it using Normalized Cuts [11] into S_i segments. We make the assumption that the person will be centered in an image and therefore, we initialize the foreground model (F^{S_i}) using segments that lie in the center and correspondingly we use the segments at four image corners to initialize our background model. A binary label for all the remaining segments can then be determined as follows,

$$\begin{aligned} \Delta &= (1-\eta)\frac{1}{E(S_i, F^{S_i})} + \eta\frac{1}{\rho(S_i, F^{S_i})}, \\ P(S_i = F) &= 1 \; if \; \Delta \geq \epsilon, 0 \; otherwise, \qquad (7) \\ P(S_i = B) &= 1 - P(S_i = F), \qquad (8) \end{aligned}$$

where $E(S_i, F^{S_i})$ and $\rho(S_i, F^{S_i})$ are the minimum Euclidean and Bhattarcharya distance between the center of S_i and any of the segments that lie in the foreground model F^{S_i} and between their color histograms respectively. We use a 10-dimensional histogram per color channel and $\eta = 0.15$ for all our experiments.

2.2. Image Representation and Matching

We use a mixture of low-level color and texture features similar to [4, 12] as our feature representation. Specifically, we divide an image into fifteen horizontal stripes. For each stripe, we compute a 20-dimensional histogram of RGB, HSV, and YCbCr color features. Additionally, we also compute a 405-dimensional HOG feature histogram for each of the RGB color channels for each stripe. Each image, is thus, represented using a 12 channel high dimensional feature vector (20925 dimensions), where each channel is obtained by concatenating features across all stripes. During image matching, $\psi()$ computes an average of 12 bhattacharyya distances between the channels of two images.

3. Experimental Results

We have tested our approach on the two most difficult public benchmark datasets available for testing re-identification: VIPER and CAVIAR4REID. Re-identification performance is measured using the cumulative matching characteristic (CMC) curve which represents the expectation of finding the correct match in the top n matches. So, a top r matching rate indicates the percentage of images correctly identified in the top r from a dataset of p test images. In our experiments, we use an average of 10 trials to report CMC rates.

3.1. Datasets: VIPeR and CAVIAR for re-id

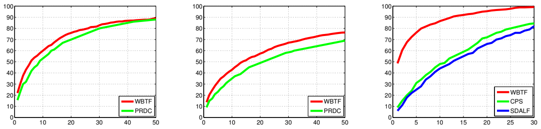

Figure 4. From left to right: (a) CMC curve for VIPeR at p=316 comparing WBTF and PRDC [12] (b) CMC curve for VIPeR at p=532 comparing WBTF and PRDC [12] (c) Comparison of WBTF against CPS [1] and SDALF [2] on the CAVIAR4REID database [1].

VIPeR dataset contains 632 corresponding pairs of images of pedestrians from different viewpoints, illumination and posture conditions [3]. Each image in the dataset is of 48 × 128 size. CAVIAR4REID [1] contains 72 different individuals in the database, with image sizes ranging from 17 × 39 to 72 × 144, with low image resolution being the primary challenge for re-identification. Table 1 and Figures 4 (a, b) presents the results of our approach on VIPeR dataset as compared against several existing state-of-the art approaches. It can be observed that our approach outperforms all the existing literature that we have tested against. Moreover, the improvement from the previous benchmark becomes more marked as the number of training samples decreases from 316(p=316, p is number of test im-

Methods	p(# of test classes) = 316				p = 432				p = 532			
	r = 1	r = 5	r = 10	r = 20	r = 1	r = 5	r = 10	r = 20	r = 1	r = 5	r = 10	r = 20
Ours(WBTF)	**21.99**	**46.84**	**59.97**	**75.95**	**15.05**	**35.76**	**50.81**	**64.24**	**13.72**	**31.77**	**42.86**	**57.42**
CPS[1]	21.84	46.00	57.21	71.50	-	-	-	-	-	-	-	-
SDALF[2]	20.00	38.00	48.5	65.00	-	-	-	-	-	-	-	-
PRDC[12]	15.66	38.42	53.86	70.09	12.64	31.97	44.28	59.95	9.12	24.19	34.40	48.55
AdaBoost	8.16	24.15	36.58	52.12	6.83	19.81	29.75	43.06	4.19	12.95	20.21	30.73
L1-Norm	4.18	11.65	16.52	22.37	3.80	9.81	13.94	19.44	3.55	8.29	12.27	17.59
Bhattacharyya	4.65	11.49	16.55	23.83	4.19	10.35	14.19	20.19	3.82	9.08	12.42	17.88

Table 1. Top ranked matching rate (%) on VIPeR. p is the number of classes in the testing set; r is the rank. Bold numbers represent the best-score in every column.

ages) to 100(p=532). For instance, we report a 20% and 50% improvement at $r = 1$ for $p = 432$ and $p = 532$ respectively as compared to the previous best matching performance. Figure 4(c) present our results on the CAVIAR4REID dataset compared against CPS [1] and SDALF [2]. Our approach achieves a significant improvement in performance over existing approaches despite the low-resolution of CAVIAR4REID dataset.

3.2. WBTF compared to no appearance modeling and MBTF

Figure 5(a) shows re-identification performance with and without using Weighted BTF (WBTF) for appearance modeling and Figure 5(b) compares WBTF and Mean BTF (MBTF) at different rank positions on the VIPeR dataset. It can be observed that there is a significant performance drop-off if we don't use appearance modeling and that WBTF outperforms MBTF.

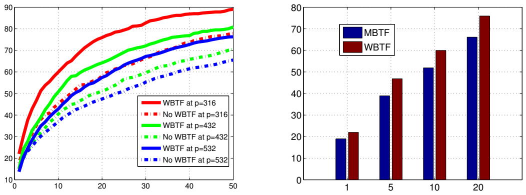

Figure 5. (a) Difference in performance between WBTF and no appearance modeling. (b) Difference in performance between WBTF and MBTF for rank=1, 5, 10, and 20 on the VIPeR dataset with $p = 316$.

4. Conclusions

We have presented a novel approach that uses a Weighted BTF to transfer appearance information between camera views. The key advantage of WBTF is that it assigns higher weight to BTFs of training observations that are closer to the test observation, as opposed to assigning equal weight to all the BTFs. Our approach has achieved new state-of-the art performance on VIPeR and CAVIAR4REID datasets.

References

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

[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] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS*, 2007.

[4] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.

[5] M. Grossberg and S. Nayar. Determining the Camera Response from Images: What is Knowable? *PAMI*, 2003.

[6] O. Javed, K. Shafique, and M. Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *CVPR*, 2005.

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

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

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

[10] W. R. Schwartz and L. S. Davis. Learning discriminative appearance-based models using partial least squares. In *SIBGRAPI*, 2009.

[11] J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*, 2000.

[12] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011.