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 benchmark datasets: VIPeR and CAVIAR4REID dataset, achieving new state-of-the-art performance on both datasets.

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 discriminatively 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 $O_j$,

$$B_j = H_i^j(B_i).$$ (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,

$$H_i(B_i) = H_j(B_j) = H_j^i(B_i)).$$ (2)

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

$$H_i^j(B_i) = H_j^{-1}(H_i(B_i)),$$ (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 $O_i^T(\tilde{p})$ and $O_j^T(\tilde{p})$ 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,

$$D^H = \{k \mid H_i^j(\alpha_k = \psi(O_i^T(\tilde{p}), O_i(\tilde{p})), \alpha_k \leq \delta\},$$ (4)

$$|D^H| = K,$$

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

$$H_{WBTF} = \sum_{k=1}^{K} \alpha_k H_i^j.$$ (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,

$$\arg \min_j \eta \psi(O_i^T(p), O_j^T(p)) + \psi(\tilde{O}_i^T(\tilde{p}), O_j^T(\tilde{p})),$$ (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 \( O^T_i \) against all the test images \( O^T_j \), \( j = \{1, \ldots, N\} \) 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,

\[
\Delta = (1 - \eta) \frac{1}{E(S_i, F^{S_i})} + \eta \frac{1}{\rho(S_i, F^{S_i})},
\]

\[
P(S_i = F) = 1 \text{ if } \Delta \geq \epsilon, 0 \text{ otherwise}, \tag{7}
\]

\[
P(S_i = B) = 1 - P(S_i = F), \tag{8}
\]

where \( E(S_i, F^{S_i}) \) and \( \rho(S_i, F^{S_i}) \) are the minimum Euclidean and Bhattacharyya 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

VlPeR dataset contains 632 corresponding pairs of images of pedestrians from different viewpoints, illumination and posture conditions [3]. Each image in the dataset is of \( 48 \times 128 \) size. CAVIAR4REID [1] contains 72 different individuals in the database, with image sizes ranging from \( 17 \times 39 \) to \( 72 \times 144 \), 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-
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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>( p = 316 )</th>
<th>( p = 432 )</th>
<th>( p = 532 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r = 1 )</td>
<td>( r = 5 )</td>
<td>( r = 10 )</td>
<td>( r = 20 )</td>
</tr>
<tr>
<td>Ours (WBTF)</td>
<td>21.99</td>
<td>35.76</td>
<td>31.77</td>
</tr>
<tr>
<td>CPS [1]</td>
<td>21.84</td>
<td>35.76</td>
<td>31.72</td>
</tr>
<tr>
<td>SDALF [2]</td>
<td>20.00</td>
<td>35.76</td>
<td>31.72</td>
</tr>
<tr>
<td>PRDC [12]</td>
<td>15.66</td>
<td>29.75</td>
<td>20.21</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>8.16</td>
<td>29.75</td>
<td>20.21</td>
</tr>
<tr>
<td>Li-Norm</td>
<td>4.18</td>
<td>29.75</td>
<td>20.21</td>
</tr>
<tr>
<td>Bhattacharyya</td>
<td>4.65</td>
<td>29.75</td>
<td>20.21</td>
</tr>
</tbody>
</table>

Figure 5 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.

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