Research Article

Impostor Resilient Multimodal Metric Learning for Person Reidentification

Muhamamd Adnan Syed , Zhenjun Han , Zhaoju Li, and Jianbin Jiao

University of Chinese Academy of Sciences, Beijing, China

Correspondence should be addressed to Zhenjun Han; hanzhj@ucas.ac.cn

Received 15 January 2018; Accepted 22 March 2018; Published 3 May 2018

Academic Editor: Deepu Rajan

Copyright © 2018 Muhamamd Adnan Syed et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In person reidentification distance metric learning suffers a great challenge from impostor persons. Mostly, distance metrics are learned by maximizing the similarity between positive pair against impostors that lie on different transform modals. In addition, these impostors are obtained from *Gallery* view for query sample only, while the *Gallery* sample is totally ignored. In real world, a given pair of query and *Gallery* experience different changes in pose, viewpoint, and lighting. Thus, impostors only from *Gallery* view can not optimally maximize their similarity. Therefore, to resolve these issues we have proposed an impostor resilient multimodal metric (IRM3). IRM3 is learned for each modal transform in the image space and uses impostors from both *Probe* and *Gallery* views to effectively restrict large number of impostors. Learned IRM3 is then evaluated on three benchmark datasets, VIPeR, CUHK01, and CUHK03, and shows significant improvement in performance compared to many previous approaches.

1. Introduction

Person reidentification (Re-ID) matches a given person across a large network of nonoverlapping cameras [1], and is fundamentally used for person tracking in camera networks. Despite years of research, reidentification is still a challenging problem as the data space in Re-ID is multimodal (modal in our work is defined as the space which is formed by the joint combination of different changes a given pair images of the same person undergo in different camera views) and the observed images in different views undergo various different changes in poses [2], viewpoints [3], lighting [4], background clutter, and also experience occlusion.

Most approaches in Re-ID can mainly be divided into two categories: robust features extraction [5–13] for representation and globally learning distance metric for matching [14, 15]. These global metrics [16–19] project features into low dimension subspace where they tend to maximize the discrimination among different persons; however, these metrics still suffer a great challenge from impostor (an impostor is a person that belongs to the other person and, however, possess higher similarity with the given query than the right *Gallery* sample) samples [20, 21]. Though, in past some attempts are made to eliminate impostors [14, 20–22], however, all

these attempts have not given due consideration of different transform modals on which the reidentification images lie [23]. This situation is illustrated in Figure 1, where we have shown three transform modals M_1 , M_2 , and M_3 in the image space. M_1 contains a positive pair (query and *Gallery*) enclosed in green rectangles for which a metric is learned, while there are two more pairs lying in modals M_2 and M_3 , respectively. View b images (enclosed in red rectangles) in M_2 and M_3 are similar to query in M_1 and, thus, are impostors for query sample. In conventional approaches [14, 20–22], the metric between query and *Gallery* samples in M_1 is learned using the impostor sample from M_2 (Metric DM_1) or M_3 (Metric DM_2) as a constraint. Therefore, when the similarity for positive pair is learned under the constraint of an impostor person lying on a different transform modal other than the positive pair, then the learned similarity metric would not be the optimal matching function, which can be proved from poor retrieval results in *Ranklist 1* and *Ranklist 2* in Figure 1.

Further, in Figure 1 previous approaches [14, 20–22] have used impostor samples for query sample only from the *Gallery* view, while totally ignoring the *Gallery* sample. Therefore, to resolve the above shortcomings in [14, 20–22] we have proposed an impostor resilient multimodal metric,

FIGURE 1: Three Modals M_1 , M_2 , and M_3 in Image Space. Query and Gallery lie in modal M_1 , while one impostor for query lies in modal M_2 and the other in modal M_3 . Metric DM_1 is learned using the impostor from modal M_2 , and DM_2 is learned using the impostor from modal M_3 . Then, the obtained retrieval results of DM_1 and DM_2 are shown in *Ranklist 1* and *Ranklist 2*, respectively. Correct Match is in green rectangle.

referred as IRM3, which eliminates the impostors largely and attains an optimal matching between positive pair. The objective of IRM3 is to maximize the matching of a positive pair against both the negative gallery samples (NGS) (samples which are not impostors and belong to different persons), as well as against impostors by taking into account the modal a given pair, its negative gallery samples, and its impostors reside. Further, in contrast to [14, 20–22], it also takes into consideration impostor samples for both the query and its respective Gallery sample. This pair of impostors are referred to as *Cross views impostors (CVI)* which are obtained for query and Gallery samples from their opposite views and help in further maximizing the similarity between given query and Gallery samples. The contributions of our impostor resilient multimodal metric IMR3 are as follows:

- (i) Improving impostors resistance by jointly exploiting the transform modals [23], as well as impostor samples from both *Probe* and *Gallery* views;
- (ii) With our IMR3 approach a significant gain in performance is obtained in Multikernel Local Fisher Discriminant Analysis (MK-LFDA) [44].

2. Methodology

Figure 2 shows the framework of our IRM3. In Figure 2, first color and texture features are extracted from each

training sample; then, different modalities are discovered in the image space. These modalities are discovered by using sum of squares clustering which is explained in Section 2.2. Finally, for each modal cross views impostors (CVI) (explained in Section 2.3) and negative gallery samples (NGS) (explained in Section 2.4) are generated to train the modal metric M_k for each transform modal k . In our work, the modal metric M_k is learned using MK-LFDA [44], and the learning procedure is explained in Section 2.6. Finally, in Section 2.7 we have explained how we have performed matching between test query and Gallery.

2.1. Feature Extraction. RGB, HSV, LAB, YCbCr, and SCNCd histograms are extracted according to similar settings in [45] using 32 bins per channel, and settings in [12], respectively. Then, all five features are concatenated together. Similarly, DenseSIFT, SILTP, and HOG are extracted according to the settings in [46], [11], and [47], respectively, and are concatenated together. Dimension of color and texture features after concatenation become large, and since Re-ID data is multiview we have used CCA [48] to reduce dimension. However, to keep the local discriminative information of each type of feature we have applied CCA to color and texture features individually. By cross validation on VIPeR and CUHK03 we obtained optimal dimension for color feature to be 900, and texture feature to be 700. Finally, the

FIGURE 2: Impostor resilient multimodal metric learning (IRM3) for person reidentification.

reduced color and texture features are concatenated to form a feature vector F of size 1600.

2.2. Partition Image Space. Let X be the image space of a camera view; then X is

$$X = \{x_i\}_{i=1, \dots, n}, \quad (1)$$

where x_i is the feature representation F_i of person i and n are the number of persons in X . Since images in X lie on different transform modals, therefore, there exist distinct clusters of different modals in X . Each of these modal clusters has its own unique transformation and visual patterns; thus, all the persons belonging to a modal k can be obtained using sum of squares clustering as

$$S_w = \frac{1}{n} \sum_{k=1}^K \sum_{i=1}^n z_{k,i} (x_i - m_k)(x_i - m_k)^T, \quad (2)$$

where K are the number of modals in X , S_w is scatter matrix of within transform modals, $z_{k,i}$ is the association of x_i with transform modal k , and m_k is the center of the k th transform modal.

In (2), each modal center m_k is critical in discovering distinct, stable, and nonempty modals in X . Thus, choosing any sample (x_i^a, x_i^b) as center m_k of any given modal k , it is necessary to make sure it is a right choice. In order to make sure a chosen modal center is right it has to fulfill two conditions: First, (i) if the chosen sample (x_i^a, x_i^b) is a center of modal cluster k , then, all the persons in modal k will be its neighbors, and it has the highest number of nearest neighbors. Second, (ii) center m_k and all its nearest neighbors lie on the same modal; therefore, these neighbors will share similar patterns with the center m_k in both *Probe* and *Gallery* views.

Now, we compute the number of nearest neighbors for each person in training set by taking into consideration the above two conditions. For this purpose, we have used both *Probe* (x_i^a) and *Gallery* (x_i^b) samples of each person to obtain four lists of neighbors, which are computed from both camera views. To acquire most reliable neighbors we then select only top@40 (here, the reason to choose top@40 neighbors is to maintain maximum reliability with minimum time and memory cost in large datasets. For instance, when we have $k = 16$ modals in CUHK03, then, in each modal there will be at least 78 training persons. Now, to obtain a center sample x_i of any modal it must have at least 51% neighbors in that modal, and thus, we take top@40 neighbors which is in actual 52% proportion of the training persons in a modal to find out whether x_i is a center or not) (top@20 for VIPeR) neighbors from each list and then perform an intersection operation among all the four lists to obtain the cardinality value, as well as IDs of the neighbors which are common in both *Probe* and *Gallery* views of a given person. This cardinality value and the IDs of the obtained neighbors are then stored in a matrix. Further, this procedure is repeated for the rest of the remaining $n - 1$ persons in the training set, and then their cardinality values, as well as IDs of the neighbors, are also stored in the same matrix.

Using this matrix we will now obtain our K initial centers for K modal transforms. These K centers are chosen as the K top persons with highest number of neighbors. However, it could be possible that two or more persons can have the same cardinality value, as well as share the same nearest neighbors IDs. In that condition simply choosing top K persons will not be the best solution; instead, we chose only those top K persons that do not have any person IDs common in their neighbors lists. In addition, for situations where more than two persons have the same cardinality and share same neighbors IDs, we randomly chose any one person from

them to represent that modal center. Finally, getting the K modal centers the optimal partitioning of the image space X is obtained by minimizing the trace of within transform modals scatter matrix as

$$\arg \min \text{tr} (S_w). \quad (3)$$

Though the image space is partitioned into K modals, however, to ensure the obtained modals are distinct and stable (in our work a stable modal is formed when it contains at least 15% training persons) we have updated the modal centers and repartitioned the space for further $t = 3$ times. The modal centers are updated as

$$m_k = \frac{1}{N'_k} \sum_{i=1}^n z_{k,i} * x_i, \quad (4)$$

where N'_k is the number of persons in modal k and given as

$$N'_k = \sum_{i=1}^n z_{k,i}. \quad (5)$$

Computing the initial modal centers is computationally tedious in our work, however, it has still moderate computational burden. For the training size of n persons the complexity is about $\mathcal{O}(t \times K \times n)$, where t is the number of iterations, and K is the number of modals.

2.3. Cross Views Impostors (CVI). After getting the distinct modals in the image space X , we can now obtain the set of CVI for each positive pair (x_i^a, x_i^b) lying in modal k from both of its *Probe* and *Gallery* views. We believe in real world situation (open set) where a positive pair has always limited or few samples; these CVI can be exploited to deliver subtle and differentiating information in metric learning that can differentiate a given pair more efficiently against large number of diverse real world impostors, as well as negative gallery samples. These impostors are obtained by comparing the similarity value of a given person pair against the other persons in *Gallery* and *Probe* views. First, the similarity values for a Probe sample x_i^a are computed with the whole *Gallery* view using metric M_{ini} and CCA reduced feature F as

$$S_{\text{probe}}^i = (x_i^a - x_j^b)^T M_{\text{ini}} (x_i^a - x_j^b), \quad (6)$$

where x_i and x_j are CCA reduced feature F of person i and j , while M_{ini} is a globally learned metric with feature F using K-LFDA [45]. We have used *linear* kernel to save memory and computational time. Similarly, the similarity values for Gallery person x_i^b are obtained with the whole *Probe* view as

$$S_{\text{gallery}}^i = (x_i^b - x_j^a)^T M_{\text{ini}} (x_i^b - x_j^a). \quad (7)$$

These obtained values S_{probe}^i and S_{gallery}^i for person (x_i^a, x_i^b) in modal k are then stored into two sets as

$$\begin{aligned} \text{Sim}_{x_i^a} &= [S_{\text{probe}, i'}^i]_{i'=1, \dots, N'_k}, \\ \text{Sim}_{x_i^b} &= [S_{\text{gallery}, i'}^i]_{i'=1, \dots, N'_k}, \end{aligned} \quad (8)$$

where N'_k refers to the number of persons in a modal k . Now, we compare each similarity value in these sets with the reference similarity value $S_{(x_i^a, x_i^b)}^{\text{ref}}$ of a given pair (x_i^a, x_i^b) to obtain its CVI set $\text{Set}_{(x_i^a, x_i^b)}^{\text{C.V.I.}}$ as

$$\text{Set}_{(x_i^a, x_i^b)}^{\text{C.V.I.}} = [x_p], \quad (9)$$

here $x_p = S_{\text{probe}, p}^i < S_{(x_i^a, x_i^b)}^{\text{ref}}$, or $x_p = S_{\text{gallery}, p}^i < S_{(x_i^a, x_i^b)}^{\text{ref}}$,

and p is the index of impostor person, and $S_{(x_i^a, x_i^b)}^{\text{ref}}$ is computed as

$$S_{(x_i^a, x_i^b)}^{\text{ref}} = (x_i^a - x_i^b)^T M_{\text{ini}} (x_i^a - x_i^b). \quad (10)$$

Further, using (6)–(10), CVI set $\text{Set}_{(x_i^a, x_i^b)}^{\text{C.V.I.}}$ for all the N'_k persons in the modal k are computed. The computational cost of generating cross views impostors for a modal k is about $\mathcal{O}(3 \times N'_k)$, where $N'_k \ll n$.

2.4. Negative Gallery Samples (NGS). We have also used negative gallery samples (NGS) to learn metric M_k . Set of NGS, denoted as $\text{Set}_{(x_i^a, x_i^b)}^{\text{Ng}}$, for person pair (x_i^a, x_i^b) are obtained from Gallery view only as

$$\text{Set}_{(x_i^a, x_i^b)}^{\text{Ng}} = [x_q], \quad (11)$$

where $q \neq p$ in $\text{Set}_{(x_i^a, x_i^b)}^{\text{C.V.I.}}$, $q \neq i$ for probe i ,

where q is the index of NGS. Further, the set of NGS $\text{Set}_{(x_i^a, x_i^b)}^{\text{Ng}}$ for all N'_k persons in modal k is then obtained using (11).

2.5. Triplet Formation. Getting the set of CVI $\text{Set}_{(x_i^a, x_i^b)}^{\text{C.V.I.}}$ and NGS $\text{Set}_{(x_i^a, x_i^b)}^{\text{Ng}}$ for all N'_k persons in modal k we will now generate triplet samples to learn metric M_k . Since the positive samples for each person x_i are too scarce compared to the number of negative samples, therefore, following the protocol of data augmentation in [49] we augment each person pair five times. Similarly, following the protocol in [39] we generate 20 triplets for each positive pair. Now, the triplet samples T_i^{imp} and T_i^{Ng} for person x_i using impostor p and negative Gallery q are given as

$$\begin{aligned} T_i^{\text{imp}} &= [\langle x_i^a, x_i^b, p \rangle], \\ T_i^{\text{Ng}} &= [\langle x_i^a, x_i^b, q \rangle], \end{aligned} \quad (12)$$

where p and q are taken from respective sets $\text{Set}_{(x_i^a, x_i^b)}^{\text{C.V.I.}}$ and $\text{Set}_{(x_i^a, x_i^b)}^{\text{Ng}}$ of person x_i .

2.6. Impostor Resilient Multimodal Metric (IRM3). Taking triplets from T_i^{imp} and T_i^{Ng} , metric IRM3 for modal k is learned using MK-LFDA [44]; however, to save both the computational time and memory requirements we adapted

[44] and use three RBF kernels and one χ^2 kernel. The weights for these kernels are learned globally for once for each dataset in our work using the similar method in [44]. The reason to learn weights globally is to save both time and computational burden. Further, there is considerably minor effect on kernel weights; even the weights are learned globally. This is due to the fact that the global space is comprised of all the existing modals, and thus, all the modals contribute in learning the global weights. For learning weights of kernels all the extracted features are used individually, and the dimensions of these features are also individually reduced to 450 by CCA before learning weights. In all our experiments the obtained weights for VIPeR are 0.3, 0.22, and 0.22 for RBF kernels, while weight for χ^2 kernel is 0.26. For CUHK01 and CUHK03 the obtained weights for RBF kernels are 0.28, 0.24, and 0.24, while weight for χ^2 kernel is 0.24. The σ values in all the datasets for the three RBF kernels are set to the mean value of modal k , as well as (mean value + mean/2) and (mean value – mean/2). These values for σ are chosen to model all the different variations in the modal k , while the σ value for χ^2 kernel is also set to mean value of modal k . The mean value in our work is the similarity value between Probe and Gallery samples of center m_k . Finally, the metric M_k is learned as

$$\max_{M_k} \text{tr} \left(\frac{M_k^T S_B M_k}{M_k^T S_W M_k} \right), \quad (13)$$

where matrices S_B and S_W are obtained with similar method in [44]. Now, (13) is then solved using generalized eigenvalue problem [50] in (14) to obtain first $r' = 300$ eigenvectors corresponding to eigenvalues with largest magnitude as

$$S_B \varphi = \lambda S_W \varphi. \quad (14)$$

2.7. Reidentification. From Figure 2, reidentification between test pair (x_i^a, x_j^b) is performed by first determining the transform modal the test pair belongs to using K -NN classifier. In K -NN classifier, the parameter K is set to the number of modals in the image space; that is, in VIPeR the value of K is set to the number of modals $k = 7$. Then, the features of (x_i^a, x_j^b) are projected into the weighted multikernel space of the respective modal, followed by the respective modal metric M_k to perform matching as

$$d_{(x_i^a, x_j^b)} = (x_i^a - x_j^b)^T M_k (x_i^a - x_j^b)^T. \quad (15)$$

3. Experiments

Our IRM3 metric is evaluated on three benchmark datasets: VIPeR, CUHK01, and CUHK03. We follow the evaluation protocol of [33] for test/train split for VIPeR, CUHK01, and CUHK03 datasets. However, in our work we have tested CUHK01 for $P = 486$ only, while CUHK03 is tested for both *Labelled* and *Detected* settings. All the experiments are conducted in *single-shot* mode, and all the reported Cumulative Matching Curves (CMC) are obtained by averaging the results over 20 trials.

3.1. Experiment Protocols. To thoroughly analyze the performance of IRM3 we have devised three evaluation strategies. These strategies evaluate IRM3 performance with different number of discovered modals K in X , with *Gallery view* impostors (GVI) (GVI are the impostors from Gallery view only and are obtained in similar way as in previous conventional metrics [14, 20–22]), as well as *Cross views* impostors (CVI).

- (i) IRM3 only: it is basic multimodal metric, learned with only *Negative Gallery Samples* (NGS).
- (ii) IRM3 + GVI (p'): IRM3 is learned with impostors from *Gallery view* (GVI), as well as with NGS. Here p' refers to the number of impostors taken from *Gallery view* to form triplet samples and have values $p' = 5, 10, \text{ and } 15$, while the remaining triplets are formed using NGS.
- (iii) IRM3 + CVI (p'): IRM3 is learned with CVI, as well as with NGS. Here p' refers to number of CVI samples used to form triplets and have values $p' = 5, 10, \text{ and } 15$, while the remaining triplets are formed using NGS.

All the samples from NGS, GVI, and CVI contain most difficult instances for a person and are randomly sampled offline, before training metric. In all the three strategies above, we have partitioned image space into $k = 3, 5, \text{ and } 7$ for VIPeR, while, for CUHK01 we have used $k = 6, 7, \text{ and } 10$ partitions, and for CUHK03 $k = 13, 14, \text{ and } 16$ partitions are used, respectively.

3.2. Results on VIPeR

Comparison with State-of-the-Art Features. Results of IRM3 metric are compared with three state-of-the-art features LOMO [11], GoG [25], and mom_f^{LE} [24] in Table 1. All the results in Table 1 are obtained for $K = 7$ modals, and our IRM3 + CVI ($p' = 15$) has attained rank@1 **52.81%** and has outperformed all the three features of reidentification, providing evidence that if the metric can address multimodal transform variations well as well as have strong resistance against impostors then the matching accuracy can be improved. Our learned IRM3 + CVI ($p' = 15$) considers optimizing all the rank orders simultaneously and, thus, has large improvement at rank@5 and rank@10.

Comparison with Metric Learning. We also compared metric IRM3 with 7 metrics. From Table 1 IRM3 + CVI ($p' = 15$) has outperformed both multimodal metric LAFT [23] and impostor resistance metric LISTEN [21]. The prime difference between IRM3 and [21, 23] is its capability of addressing both the person modal transform, as well as capability of further maximizing the matching against joint constraint of cross views impostors. All these are the causes of great challenge in matching pedestrians. In Table 1 only SS-SVM [16] is a metric that tries to model the transform modal for each individual person; however, it never paid attention to acquire resistance against impostors and thus has **19.21%** lower rank@1 accuracy than IRM3 + CVI ($p' = 15$). Though IRM3 has successful results, still it has **1.36%** lower rank@1 than SCSP [38].

TABLE 1: Top matching comparison on VIPeR.

		Single-shot, $P = 316$		
	Method	$r = 1$	$r = 5$	$r = 10$
F	LOMO [11]	40.0	-	80.51
	moM _f ^{LE} [24]	48.0	76.8	85.4
	GoG [25]	49.7	79.7	88.7
DF	SIR-CIR [26]	35.76	68.38	82.9
DMN	GS-CNN [27]	37.8	66.9	77.4
	DGD [28]	38.6	-	-
	LSTM [29]	42.4	68.7	79.4
	MuDeep [30]	43.03	74.36	85.76
	E2E-CAN [31]	47.2	79.2	89.2
	DLPA [32]	48.7	74.7	85.1
	Quadruplet-net [33]	49.05	73.10	81.96
	JLML [34]	50.2	74.2	84.3
	ITL [35]	15.2	34.2	45.9
M	LAFT [23]	29.6	-	69.3
	WARCA [36]	37.47	70.78	-
	LISTEN [21]	39.62	69.97	82.9
	L-1 graph [37]	41.5	-	-
	SS-SVM [16]	42.66	70.1	84.27
	SCSP [38]	53.54	82.59	91.49
	IRM3 only	45.92	82.90	91.63
	IRM3 + GVI ($p' = 15$)	50.39	85.79	95.73
	IRM3 + CVI ($p' = 15$)	52.81	87.95	97.29

Obviously, VIPeR has large pose, misalignment, and body parts displacement issues which are specifically not addressed in our work and, thus, is necessarily needed to improve the matching and results largely.

Comparison with Deep Methods. Though, deep features (DF) and deep matching networks (DMN) have no match with conventional metric learning methods, however, from the results in Table 1 it is clearly evident if two major issues of reidentification (i.e., multimodal transforms, and strong rejection capability against impostors) can be well handled simultaneously, then comparable or even higher performance than deep methods can be attained. Our IRM3 + CVI ($p' = 15$) has **7.1%** and **4.94%** higher rank@1 than Quadruplet-Net [33] and JLML [34], respectively. These obtained results demonstrate the fact that for smaller dataset like VIPeR deep matching networks have insufficient training samples to learn a discriminative network.

At last, Figure 3 shows the comparison of retrieval results of two queries from VIPeR dataset for XQDA [11] and our IRM3 + CVI ($p' = 15$) when $K = 7$ modals are used. Retrieval results of *Query 1* for XQDA find the correct match at *rank* = 4 enclosed in green rectangle (b), while IMR3 finds the match at *rank* = 2 enclosed in green rectangle (e). Similarly, for *Query 2* our IMR3 finds the match at *rank* = 1 enclosed in green rectangle (j); in contrast, XQDA finds the correct match at *rank* = 3 enclosed in green rectangle (h). Thus, our

IRM3 approach improves matching, and consequently rank gets higher.

3.3. Results on CUHK01

Comparison with State-of-the-Art Features. Table 2 summarizes results of IRM3 for $K = 10$ modals and compares the obtained results with LOMO [11], GoG [25], and moM_f^{LE} [24]. Though the three features are discriminative, however, our IRM3 approach is better than the three features in solving the two big challenges of Re-ID, that is, multimodal pedestrians matching and impostors resistance. Since CUHK01 has larger training set than VIPeR, thus, modal transforms can be well learned, and therefore, IRM3 + CVI ($p' = 15$) attains larger discrimination than moM_f^{LE} [24]. Our IRM3 + CVI ($p' = 15$) has **15.15%** higher rank@1 accuracy than moM_f^{LE} due to inherent virtue of handling different modals, person specific variations, and rejecting large number of impostors, all simultaneously.

Comparison with Metric Learning. In Table 2 three most recently proposed metrics CVAML [40], WARCA [36], and L-1 Graph [37] are compared with our IRM3 approach. All the three metrics have assumption of unimodal intercamera transform, rather than multimodal image space. Though WARCA [36] employed hard negative samples as learning constraint, however, ignoring other negative samples from

FIGURE 3: Two queries are shown, *Query 1* and *Query 2*, and their retrieval results using XQDA [11] and our IRM3. Correct match is shown in green rectangle, while blue rectangle shows impostors.

TABLE 2: Top matching comparison on CUHK01.

		Single-shot, $P = 486$		
<i>Method</i>		$r = 1$	$r = 5$	$r = 10$
F	LOMO [11]	49.2	75.7	84.2
	moM _f ^{LE} [24]	64.6	84.9	90.6
	GoG [25]	57.8	79.1	86.2
DMN	MCP-CNN [39]	53.7	84.3	91.0
	DGD [28]	66.6	-	-
	E2E-CAN [31]	67.2	87.3	92.5
	DLPA [32]	75.0	93.5	95.7
	Quadruplet-net [33]	62.55	83.44	89.71
	JLML [34]	69.8	88.4	93.3
M	CVAML [40]	57.3	81.2	86.5
	WARCA [36]	58.34	79.76	-
	L-1 graph [37]	50.1	-	-
	IRM3 only	68.24	88.71	94.31
	IRM3 + GVI ($p' = 15$)	72.91	94.61	97.6
	IRM3 + CVI ($p' = 15$)	76.14	96.90	98.35

Gallery view and not taking into consideration a person modal during learning have made it suffer greatly to attain higher accuracy. On the other hand, IRM3 + CVI ($p' = 15$) has capability to deal all these challenges and, thus, has attained **76.14%** rank@1 accuracy.

Comparison with Deep Methods. In Table 2, we can see several deep matching networks (DMN) have performed much well than conventional metrics on CUHK01. Only K-LFDA when trained with moM_f^{LE} [24] feature attains comparable performance than DMN. However, motivated to resolve the challenges for reidentification in real world (i.e., multimodal image space, and diverse impostors) IRM3 + CVI ($p' = 15$) has much better results than MCP-CNN [39], E2E-CAN [31], Quadruplet-Net [33], and JLML [34], while our IRM3 + CVI ($p' = 15$) has **1.49%** rank@1 than DLPA [32]. DLPA

extracts deep features by semantically aligning body parts, as well as rectifying pose variations. We believe if semantic body parts alignment and rectification of poses variations are included in our IRM3 then the results can be further improved.

3.4. Results on CUHK03

Comparison with State-of-the-Art Features. Table 3 compares LOMO [11] and GoG [25] features with our IRM3 metric in both *Labelled* and *Detected* settings. All the results in Table 3 are obtained for $K = 16$ modals. In Table 3, obtained results are much higher than the two features. The primary reason of gain in performance for IRM3 against the features [11, 25] is mainly due to the difference in their approaches. In [11, 25] a universal feature representation is proposed for

TABLE 3: Top matching comparison on CUHK03.

		<i>Labelled, P = 100</i>		
	<i>Method</i>	<i>r = 1</i>	<i>r = 5</i>	<i>r = 10</i>
F	LOMO [11]	52.2	82.23	92.14
	GoG [25]	67.3	91.0	96.0
DF	DCAF [41]	74.21	94.33	97.54
DFM	DGD [28]	75.3	-	-
	Quadruplet-net [33]	75.53	95.15	99.16
	MuDeep [30]	76.87	96.12	98.41
	E2E-CAN [31]	77.6	95.2	99.3
	JLML [34]	83.2	98.0	99.4
	DLPA [32]	85.4	97.6	99.4
M	SS-SVM [16]	57.0	85.7	94.3
	Null Sp. [42]	62.55	90.05	94.80
	SSM [43]	76.6	94.6	98.0
	WARCA [36]	78.38	94.55	-
	IRM3 only	78.83	95.97	98.37
	IRM3 + GVI ($p' = 15$)	83.32	98.70	99.54
	IRM3 + CVI ($p' = 15$)	86.17	99.02	99.68
		<i>Detected, P = 100</i>		
	<i>Method</i>	<i>r = 1</i>	<i>r = 5</i>	<i>r = 10</i>
F	LOMO [11]	46.25	78.9	88.55
	GoG [25]	65.5	88.4	93.7
DF	SIR-CIR [26]	52.17	83.7	90.4
	DCAF [41]	67.99	91.04	95.36
DFM	LSTM [29]	57.3	80.10	88.3
	GS-CNN [27]	68.1	88.1	94.6
	E2E-CAN [31]	69.2	88.5	94.1
	MuDeep [30]	75.64	94.36	97.46
	JLML [34]	80.6	96.9	98.7
	DLPA [32]	81.6	97.3	98.4
M	L-1 graph [37]	39.0	-	-
	SS-SVM [16]	51.2	80.8	89.6
	Null Sp. [42]	54.70	84.75	94.80
	SSM [43]	72.7	92.4	96.1
	IRM3 only	72.98	91.7	93.02
	IRM3 + GVI ($p' = 15$)	78.68	95.60	98.09
	IRM3 + CVI ($p' = 15$)	80.77	96.94	98.67

all the different persons, which may not be optimal for all the persons at the same time residing on different modals; in contrast, our motivation is based on discovering distinct modals in the image space and then addressing each modal specifically with empowerment of large number of impostors rejection. Therefore, our IRM3 + CVI ($p' = 15$) (in Labelled setting) has rank@1 accuracy of about **86.17%**.

Comparison with Metric Learning. In Table 3, recently proposed WARCA [36] and SSM [43] are compared with our IRM3 approach. WARCA [36] differs with our IRM3 approach in a way that it only addresses hard negative samples, while SSM [43] differs in a way that it has no measure to account for different modal transforms, as well as having

no resistance against impostors. Our IRM3 + CVI ($p' = 15$) (in Labelled setting) has surpassed [36] and [43] and has attained **9.04%** and **11.1%** rise at rank@1 accuracy, respectively.

Comparison with Deep Methods. Interestingly, in Table 3 all the deep methods in *Labelled* and *Detected* settings have very high performance on CUHK03. These high results demonstrate the fact that CUHK03 is the largest dataset among all and, thus, can help in learning a more discriminative DMN. Even though both JLML [34] and DLPA [32] learn deep body features with global and local body parts alignment, as well as, pose alignment, however, our IMR3 approach benefitted with transform specific metrics empowered with impostors rejection still maintained to attain better results. Our IRM3

TABLE 4: Effect of multimodal transforms + impostor resistance (VIPeR, $P = 316$).

Method	$r = 1$	$r = 5$	$r = 10$
$k = 5, p' = 0$	45.27	82.16	91.1
$k = 7, p' = 0$	45.92	82.90	91.63
$k = 5, \text{GVI} (p' = 15)$	48.88	85.33	95.37
$k = 7, \text{GVI} (p' = 15)$	50.39	85.79	95.73
$k = 5, \text{CVI} (p' = 15)$	52.10	87.14	96.87
$k = 7, \text{CVI} (p' = 15)$	52.81	87.95	97.29

considers optimizing all the rank orders simultaneously and, thus, have large gain at rank@5 and rank@10 in *Labelled* setting.

3.5. Analysis. In Table 4, we analyzed the effect of number of modals K in testing for VIPeR. Initially, we have partitioned image space into $K = 5$ and then tested it without using any impostor sample ($K = 5, p' = 0$) to obtain rank@1 results of about 45.27%. As the more modals are discovered in the image space, such as $K = 7$, then the results get further improved even without using any impostor sample ($K = 7, p' = 0$), and rank@1 becomes 45.92%. The main reason behind this increment is the fact that now we could match more test samples correctly by using their actual modal transforms which were lost when the modals are less discovered in $K = 5$.

In addition, we could also see a positive increment in results when impostors from *Gallery* view are also added in learning metric. Both ($K = 5, \text{GVI} (p' = 15)$) and ($K = 7, \text{GVI} (p' = 15)$) have attained more higher differentiating capability than [14, 20–22], as now, they can restrict impostors by taking into care transform modals a positive pair and impostors undergo.

Interestingly, in our work this impostor resistance can be further enhanced. This is done by using *Cross views impostors* (CVI). From Table 4, it is clear that even for same number of modals say $K = 7$, when (CVI) are used then the differentiation capability of ($K = 7, \text{CVI} (p' = 15)$) gets further enhanced than ($K = 7, \text{GVI} (p' = 15)$) and rank@1 becomes 52.81%. This increment in rank@1 provides a strong evidence that CVI have ability to maximize the similarity of positive pair more than GVI by taking into care both the transform modal, as well as various different changes a given query and *Gallery* samples undergo in different views.

At last, in Figure 4 we have provided a performance comparison at rank@1 when the modal centers are chosen randomly, as well as when the centers are obtained using our method in Section 2.2. Obtained rank@1 accuracy for random centers is poor, because these random centers are obtained just by simply choosing the top- K persons without taking into care their reliability, stability, and IDs.

3.6. Efficiency. We computed the run time of our IRM3 approach using MK-LFDA [44], XQDA [11], and K-LFDA [45] (with χ^2 kernel) on CUHK03. There are 1260 training persons and 100 testing identities. All the algorithms are implemented in MATLAB and run on server machine having 6 CPUs (Xeon(r)e5-2620) with each CPU having 6 cores and

FIGURE 4: Performance at rank@1 when centers m_k are selected randomly, and when centers are selected with our approach provided in Section 2.2.

TABLE 5: Run time comparison on CUHK03 (in seconds).

Method	MK-LFDA [44]	XQDA [11]	K-LFDA [45]
Training	171.26	174.7	160.03
Testing	28.8	30.23	32.45

total memory size of 256 GB. In Table 5, training time of MK-LFDA [44] is faster than XQDA [11] but lower than K-LFDA [45]. However, in testing when the weights of kernels are not learned MK-LFDA [44] is faster than both XQDA and K-LFDA. These timing results support the fact that our proposed method is well applicable in real time applications and in public spaces.

4. Conclusion

This paper presents a metric learning approach that exploits both multimodal transforms and Cross views impostors to improve the capability of metric to differentiate among different persons, as well as enhance rejection capability to decline large number of real world diverse impostors. In real world mostly pedestrian images are multimodal, and in public

spaces several persons share similar clothing; therefore, our IRM3 is learned to tackle such issues of reidentification and person tracking in public spaces. Extensive experiments on three challenging datasets (VIPeR, CUHK01, and CUHK03) demonstrate the effectiveness of our IRM3 metric which has outperformed many previous state-of-the-art metrics. In addition, we further intend to extend our approach for testing in real world scenario and intend to solve various other issues for real time implementation.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- [1] W.-S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *Proceedings of the 2009 20th British Machine Vision Conference, BMVC 2009*, UK, September 2009.
- [2] R. Zhao, W. Oyang, and X. Wang, "Person Re-Identification by Saliency Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 2, pp. 356–370, 2017.
- [3] S. Bık, S. Zaidenberg, B. Boulay, and F. Brmond, "Improving person re-identification by viewpoint cues," in *Proceedings of the 11th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS 2014*, pp. 175–180, Republic of Korea, August 2014.
- [4] R. Rama Varior, G. Wang, J. Lu, and T. Liu, "Learning invariant color features for person reidentification," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3395–3410, 2016.
- [5] C.-H. Kuo, S. Khamis, and V. Shet, "Person re-identification using semantic color names and RankBoost," in *Proceedings of the 2013 IEEE Workshop on Applications of Computer Vision, WACV 2013*, pp. 281–287, January 2013.
- [6] Y. Hu, S. Liao, Z. Lei, D. Yi, and S. Z. Li, "Exploring structural information and fusing multiple features for person re-identification," in *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2013*, pp. 794–799, June 2013.
- [7] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 2360–2367, IEEE, San Francisco, Ca, USA, June 2010.
- [8] A. Bhuiyan, A. Perina, and V. Murino, "Person re-identification by discriminatively selecting parts and features," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 8927, pp. 147–161, 2015.
- [9] S. Khamis, C.-H. Kuo, V. K. Singh, V. D. Shet, and L. S. Davis, "Joint learning for attribute-consistent person re-identification," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 8927, pp. 134–146, 2015.
- [10] J. Roth and X. Liu, "On the exploration of joint attribute learning for person re-identification," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 9003, pp. 673–688, 2015.
- [11] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by Local Maximal Occurrence representation and metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 2197–2206, June 2015.
- [12] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 8689, no. 1, pp. 536–551, 2014.
- [13] Z. Mingyong, Z. Wu, C. Tian, Z. Lei, and H. Lei, "Efficient person re-identification by hybrid spatiogram and covariance descriptor," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2015*, pp. 48–56, June 2015.
- [14] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.
- [15] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '12)*, pp. 2288–2295, IEEE, Providence, RI, USA, June 2012.
- [16] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan, "Sample-specific SVM learning for person re-identification," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pp. 1278–1287, July 2016.
- [17] H. Shi, Y. Yang, X. Zhu et al., "Embedding deep metric for person Re-identification: A study against large variations," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 9905, pp. 732–748, 2016.
- [18] Y.-C. Chen, W.-S. Zheng, J.-H. Lai, and P. C. Yuen, "An asymmetric distance model for cross-view feature mapping in person reidentification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 8, pp. 1661–1675, 2017.
- [19] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang, "Similarity learning on an explicit polynomial kernel feature map for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 1565–1573, June 2015.
- [20] M. Hirzer, P. M. Roth, and H. Bischof, "Person re-identification by efficient impostor-based metric learning," in *Proceedings of the 2012 IEEE 9th International Conference on Advanced Video and Signal-Based Surveillance, AVSS 2012*, pp. 203–208, China, September 2012.
- [21] X. Zhu, X.-Y. Jing, F. Wu et al., "Distance learning by treating negative samples differently and exploiting impostors with symmetric triplet constraint for person re-identification," in *Proceedings of the 2016 IEEE International Conference on Multimedia and Expo, ICME 2016*, July 2016.
- [22] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 6495, no. 4, pp. 501–512, 2011.
- [23] W. Li and X. Wang, "Locally aligned feature transforms across views," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013*, pp. 3594–3601, June 2013.
- [24] M. Gou, O. Camps, and M. Sznaiier, "moM: Mean of Moments Feature for Person Re-identification," in *Proceedings of the 2017*

- IEEE International Conference on Computer Vision Workshop (ICCVW)*, pp. 1294–1303, Venice, Italy, October 2017.
- [25] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, “Hierarchical Gaussian descriptor for person re-identification,” in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pp. 1363–1372, July 2016.
 - [26] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, “Joint learning of single-image and cross-image representations for person re-identification,” in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pp. 1288–1296, July 2016.
 - [27] R. R. Variator, M. Haloi, and G. Wang, “Gated siamese convolutional neural network architecture for human re-identification,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*: Preface, vol. 9912, pp. 791–808, 2016.
 - [28] T. Xiao, H. Li, W. Ouyang, and X. Wang, “Learning deep feature representations with Domain Guided Dropout for person re-identification,” in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pp. 1249–1258, July 2016.
 - [29] R. R. Variator, B. Shuai, J. Lu, D. Xu, and G. Wang, “A siamese long short-term memory architecture for human re-identification,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*: Preface, vol. 9911, pp. 135–153, 2016.
 - [30] X. Qian, Y. Fu, Y. Jiang, T. Xiang, and X. Xue, “Multi-scale Deep Learning Architectures for Person Re-identification,” in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5409–5418, Venice, October 2017.
 - [31] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, “End-to-end comparative attention networks for person re-identification,” *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3492–3506, 2017.
 - [32] L. Zhao, X. Li, Y. Zhuang, and J. Wang, “Deeply-Learned Part-Aligned Representations for Person Re-identification,” in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3239–3248, Venice, October 2017.
 - [33] W. Chen, X. Chen, J. Zhang, and K. Huang, “Beyond Triplet Loss: A Deep Quadruplet Network for Person Re-identification,” in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1320–1329, Honolulu, HI, July 2017.
 - [34] W. Li, X. Zhu, and S. Gong, “Person Re-Identification by Deep Joint Learning of Multi-Loss Classification,” in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 2194–2200, Melbourne, Australia, August 2017.
 - [35] W. Liao, M. Y. Yang, N. Zhan, and B. Rosenhahn, “Triplet-Based Deep Similarity Learning for Person Re-Identification,” in *Proceedings of the 2017 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pp. 385–393, Venice, October 2017.
 - [36] C. Jose and F. Fleuret, “Scalable metric learning via weighted approximate rank component analysis,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*: Preface, vol. 9909, pp. 875–890, 2016.
 - [37] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, “Person re-identification by unsupervised ℓ_1 graph learning,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*: Preface, vol. 9905, pp. 178–195, 2016.
 - [38] D. Chen, Z. Yuan, B. Chen, and N. Zheng, “Similarity learning with spatial constraints for person re-identification,” in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pp. 1268–1277, July 2016.
 - [39] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, “Person re-identification by multi-channel parts-based CNN with improved triplet loss function,” in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pp. 1335–1344, July 2016.
 - [40] H. Yu, A. Wu, and W. Zheng, “Cross-View Asymmetric Metric Learning for Unsupervised Person Re-Identification,” in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 994–1002, Venice, October 2017.
 - [41] D. Li, X. Chen, Z. Zhang, and K. Huang, “Learning Deep Context-Aware Features over Body and Latent Parts for Person Re-identification,” in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7398–7407, Honolulu, HI, July 2017.
 - [42] L. Zhang, T. Xiang, and S. Gong, “Learning a discriminative null space for person re-identification,” in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pp. 1239–1248, July 2016.
 - [43] S. Bai, X. Bai, and Q. Tian, “Scalable Person Re-identification on Supervised Smoothed Manifold,” in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3356–3365, Honolulu, HI, July 2017.
 - [44] M. A. Syed and J. Jiao, “Multi-kernel metric learning for person re-identification,” in *Proceedings of the 23rd IEEE International Conference on Image Processing, ICIP 2016*, pp. 784–788, September 2016.
 - [45] F. Xiong, M. Gou, O. Camps, and M. Sznajder, “Person re-identification using kernel-based metric learning methods,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*: Preface, vol. 8695, no. 7, pp. 1–16, 2014.
 - [46] R. Zhao, W. Ouyang, and X. Wang, “Unsupervised salience learning for person re-identification,” in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 3586–3593, IEEE, Portland, Ore, USA, June 2013.
 - [47] G. Lisanti, I. Masi, and A. Del Bimbo, “Matching people across camera views using kernel canonical correlation analysis,” in *Proceedings of the 8th ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC 2014*, Italy, November 2014.
 - [48] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, “Canonical correlation analysis: an overview with application to learning methods,” *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.
 - [49] W. Li, R. Zhao, T. Xiao, and X. Wang, “DeepReID: Deep filter pairing neural network for person re-identification,” in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014*, pp. 152–159, June 2014.
 - [50] Y. Ying and P. Li, “Distance metric learning with eigenvalue optimization,” *Journal of Machine Learning Research*, vol. 13, pp. 1–26, 2012.