Unsupervised data association for Metric Learning in the context of Multi-shot Person Re-identification

Furqan M. Khan, Francois Bremond
INRIA Sophia Antipolis-Mediterrannee
2004 Route des Lucioles, Sophia Antipolis Cedex, France
{furqan.khan | francois.bremond}@inria.fr

Abstract

Appearance based person re-identification is a challenging task, specially due to difficulty in capturing high intra-person appearance variance across cameras when inter-person similarity is also high. Metric learning is often used to address deficiency of low-level features by learning view specific re-identification models. The models are often acquired using a supervised algorithm. This is not practical for real-world surveillance systems because annotation effort is view dependent. In this paper, we propose a strategy to automatically generate labels for person tracks to learn similarity metric for multi-shot person re-identification task. We demonstrate on multiple challenging datasets that the proposed labeling strategy significantly improves performance of two baseline methods and the extent of improvement is comparable to that of manual annotations in the context of KISSME algorithm [14].

1. Introduction

The task of person re-identification (ReID) has gained significance in the context of visual surveillance as it allows for extended visual tracking and behavior understanding of individuals. In this context, the goal of ReID task is to make association between different tracks (sets of images) of a person, often acquired from different cameras with non-overlapping fields of view. ReID process is generally divided into two phases: *signature* and *similarity* computation. Signature computation refers to acquisition of appearance model from the images that constitute a track. As more than one image is available for signature creation, the signature is often represented as a set of low-level image descriptors and the task is referred to as *multi-shot* ReID.

ReID task is complex as methods have to rely on an individual's global appearance, such as color/texture of clothes, or motion because acquisition of biometric information is not always possible due to unconstrainted movement of individuals. These cues are not unique for every person and are significantly affected by scene illumination as well as camera properties and viewpoint (Fig. 1). Designing low level features with optimal trade-off between discriminative power and invariance to external factors is a significant challenge for ReID methods.

A common solution to address appearance variance is to learn view (scene or camera) specific similarity metrics to deal with inter-view feature transformations. Most metric learning algorithms are supervised by manually labeling each input pair as either matching - *positive* - or non-matching - *negative*. Notable improvement in performance can be achieved this way. However, since the models are view specific, any modification in camera network requires (re)training of a considerable number of models. Therefore, for a real-world surveillance system, supervised metric learning is unattractive as it adds significantly to the maintenance cost. In this paper, we propose a simple, yet effective, automated strategy to label data so that metric learning can proceed without supervision.

Figure 1: Variance in appearance of individuals

In multi-shot scenario, learning algorithms can take advantage of the fact that each track provides multiple images of person, which can be used to create set of positive feature pairs for metric learning. However, when tracks are generated by an automated detection and tracking system, track fragmentation (splitting of one track into multiple) may result in each person having multiple tracks per view. Therefore, images from two arbitrarily selected tracks cannot be used to create negative set as it would mislead the learning algorithm. To address reliable label assignment, we first obtain a distribution of pair-wise distances between tracks from different cameras. We then select a small number of track pairs based on this distribution to generate positive and negative training sets for metric learning. We demonstrate that on multiple benchmarks, PRID2011 [12], iLIDS-VID [28], and iLIDS-AA [3], that the proposed strategy noticably improves performance of a baseline method to achieve new state-of-the-art by learning Mahalanobis metric using KISSME algorithm [14] with automatically generated labels. We also show that the performance of "unsupervised" algorithm is comparable to fully supervised KISSME algorithm.

2. Related Work

A considerable effort has been dedicated in the past to design robust feature descriptors for ReID task ([4, 5, 9, 10, 12, 13, 20, 18, 23, 24, 27, 30, 33, 34, 36]). These methods often try to capture color, texture or shape properties of the target person through inventive feature design. However, conflicting requirements for feature design make it difficult to create a one-for-all solution.

Consequently, recent trend in the literature is to overcome weakness of low-level features in handling complex ReID scenarios by using supervised machine learning techniques to adapt a similarity metric or a ranking function for a set of cameras [1, 3, 6, 7, 8, 11, 14, 17, 19, 25, 26, 28, 31, 32, 35]. Almost all of these approaches require that input samples from different sources be associated with each other, often making pairs, and divided into two sets depending on whether the pair (or group) correspond to a single person or not. The two sets are often referred to as *matching* and *non-matching* sets, or *positive* and *negative* sets, respectively. The goal of learning is to separate the two distributions corresponding to distance between pairs in the positive set and between pairs in the negative set. Labeling of input pairs as positive or negatives is done manually, which is tedious and unattractive for real-world systems because the trained models are tied to specific cameras (or scenes).

In case of multi-shot ReID, training data for each person is available as sets of images. The data can be generated by using automated people detection and tracking systems. Li *et al*. [16] benefit from this scenario and construct local metric fields for positive and negative image pairs without manual labeling based on the assumption that each person has only one track per view. In practice, this assumption is often violated for automatically produced data as it contains significant noise and track segmentation. Multiple tracks per person per view adulterate negative set and hence adversely affect learning algorithm. Instead of trivially selecting negative pairs, we first obtain the distribution of a primitive distance measure with respect to all pairs of inputs and then select negative pairs based on this distribution. This reduces the number of false negative pairs.

To the best of our knowledge, our approach to automatically label data for reliable metric learning irrespective of the learning algorithm is the only such effort in the context of multi-shot ReID. Our experiments validate that our approach is successful in producing reliable annotations for metric learning.

3. Mahalanobis distance Learning

Person re-identification is often divided into two main phases: signature extraction and similarity measurement. Due to susceptibility of low level features to illumination and viewpoint changes, learning a similarity metric based on Mahalanobis distance has garnered considerable interest in ReID community. For a pair of vectors $\boldsymbol{x}_{ij} = (\boldsymbol{x}_i, \boldsymbol{x}_j)$, squared Mahalanobis distance is defined as:

$$d^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{M} (\boldsymbol{x}_i - \boldsymbol{x}_j) \qquad (1)$$

where $\boldsymbol{M} \succeq 0$ is a positive semidefinite matrix.

Some of the popular algorithms to learn matrix $\boldsymbol{M}$ from a set of vector pairs $\boldsymbol{X} = \{\boldsymbol{x}_{ij} | i = 1 : m, j = 1 : n\}$ are LMNN [31, 32], ITML [7], DML [11] and KISSME [14]. These algorithms require that the training set be divided into *positive* ($\boldsymbol{X}^+$) and *negative* ($\boldsymbol{X}^-$) subsets. Set $\boldsymbol{X}^+$ consists of vector pairs $\boldsymbol{x}_{ij}$ for which both $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same person, while set $\boldsymbol{X}^-$ consists of non-matching vector pairs. The goal of this work is to automatically find associated (and non-associated) pairs to create sets $\boldsymbol{X}^+$ and $\boldsymbol{X}^-$ given tracks of persons from two different views (or sources) as reliably as possible without human intervention. Once these sets are created, any of the above metric learning algorithms can be used for metric learning.

Our data labeling approach is agnostic to specific metric learning algorithm. However, for our experiments we used KISSME [14] for its simplicity, low computation cost and effectiveness under challenging conditions. KISSME algorithm assumes independent Gaussian generation processes with parameters $\theta^+ = (0, \Sigma^+)$ and $\theta^- = (0, \Sigma^-)$ for positive and negative pairs $(\boldsymbol{x}_i, \boldsymbol{x}_j)$, respectively, based on their difference vector $\boldsymbol{x}_i - \boldsymbol{x}_j$. Given pair associations, the co-

variance matrices Σ^+ and Σ^- can be computed as follows:

$$\Sigma^+ = \sum_{\boldsymbol{x}_{ij} \in \boldsymbol{X}^+} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \quad (2)$$

$$\Sigma^- = \sum_{\boldsymbol{x}_{ij} \in \boldsymbol{X}^-} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \quad (3)$$

Given the covariance matrices, Mahalanobis metric with matrix $\boldsymbol{M}$ that reflects properties of log-likelihood test ratio for a sample pair being a non-match versus otherwise is obtained by clipping the spectrum of $\hat{\boldsymbol{M}}$ by eigenanalysis:

$$\hat{\boldsymbol{M}} = ({\Sigma^+}^{-1} - {\Sigma^-}^{-1}) \quad (4)$$

4. Unsupervised data association for metric learning

Our approach is based on the assumption that multiple images of a person are available and grouped for training. However, it is not necessary that the identity of each person is known apriori or that each person has only one set of images available for a particular view. Therefore, our algorithm can be used to train similarity metrics using data generated by an automated pedestrian detection and tracking system with track fragmentation.

The objective of ReID algorithm design is to make conditional distributions of distance between pairs of signatures conditioned on the label (positive or negative) of the pair to be mutually exclusive. That is, one conditional distribution is zero at every point where the other is not. In a probabilistic framework, labels for signature pairs, and hence vector pairs, can be acquired given both conditional distributions, even if they are not mutually exclusive. The problem is to acquire these distributions without manual labeling. On the other hand, it is possible to find the empirical marginal distribution of distance for a ReID algorithm without manual annotation. Our approach uses empirical marginal distribution of distance to label data pairs.

To learn Mahalanobis metric for a multi-shot signature representation, such as [9, 16, 26, 2], which represents each signature $\boldsymbol{S}_u$ for track u as a set $\boldsymbol{S}_u = \{\boldsymbol{x}_i | i = 1..K_u\}$ of feature descriptors $\boldsymbol{x}_i$ to capture multi-modality of appearance, one can obtain empirical distribution of distance between pairs of tracks in a training set using the selected representation with Euclidean distance, by assuming $\boldsymbol{M}$ to be an identity matrix. It is reasonable to assume that in any collection of training examples, the number of negative pairs would, in general, considerably outnumber the positive pairs. Therefore, the empirical marginal distribution is heavily influenced by the number of negative pairs, and closely resembles the conditional distribution of negative pairs. Furthermore, one can expect a common scenario for ReID methods is when the true conditional distributions are not completely mutually exclusive, *i.e*. there is some overlap between the two distributions, however, the distributions have distinct modes. This means that the positive pairs are not distributed uniformly on both sides of the mode of the empirical distribution. That is the further the distance of a pair is away from the mode of the distribution, the probability that pair is positive increases or decreases depending on whether the distance is smaller than the mode or greater. Therefore, the signature pair with the highest distance in the population is very unlikely to be a positive pair. Consequently, we sort all the signature pairs in the training set based on their distance in decreasing order and select top N pairs $\{p_k = (\boldsymbol{S}_u^{p_k}, \boldsymbol{S}_v^{p_k}) | k = 1, 2..N\}$ to generate set of negative vector pairs $\boldsymbol{X}^-$:

$$\boldsymbol{X}^- = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) | \boldsymbol{x}_i \in \boldsymbol{S}_u^{p_k}, \boldsymbol{x}_j \in \boldsymbol{S}_v^{p_k}, k = 1..N\} \quad (5)$$

However, we cannot use similar strategy to generate set of positive vector pairs $\boldsymbol{X}^+$ using a number of bottom pairs because the probability that a signature pair with the smallest distance is positive isn't necessarily high enough due to large number of negative signature pairs. Therefore, we use the constraint that each signature consists of feature descriptors that belong to the same person to generate positive set. Nevertheless, to keep positive and negative sets balanced and corresponding, we use the same signatures used to generate negative vector set. The positive set is then randomly sub-sampled to have same size as the negative set. Precisely, the set $\boldsymbol{X}^+$ is defined as a random sub-sample of:

$$\tilde{\boldsymbol{X}}^+ = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) | \quad \boldsymbol{x}_i \in \boldsymbol{S}_t^{p_k}, \boldsymbol{x}_j \in \boldsymbol{S}_t^{p_k} \setminus \boldsymbol{x}_i, \quad k = 1..N, t \in \{u, v\}\} \quad (6)$$

4.1. Signature representation

To validate our hypothesis about empirical marginal distribution, let's consider two approaches based on the representation in [2], which has shown promising performance on multiple multi-shot ReID benchmarks. The approach represents each signature using a set of multi-modal feature distributions, precisely GMMs, and the similarity between two signatures is partially based on computing Mahalanobis distance between means of GMM components. Hence it provides a suitable test case for evaluation of metric learning with our data association approach.

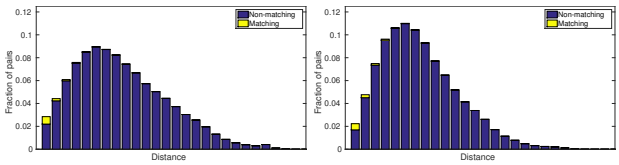
We, however, use a simplified version of the algorithm proposed in [2]. Specifically, we define similarity between two signatures using only the *Mean Pointwise Distance* between component means of two signatures, ignoring component variances altogether. Further, we extend Brownian Covariance (BCov) feature to include gray image intensity, magnitude and orientation of intensity gradient, and Gabor, Laplacian and Gaussian filter responses. In addition, we modify Color Spatio-Histogram (CSH) to include more

color channels, *i.e*. Y, Cr, Cb, H, S, nR, nG, and nB as suggested by [33]. Another difference is that we use Alkaline Information Criterion (AIC) to select optimal number of signature components instead of the regularization function used in [2]. That is, for track u with image set $\boldsymbol{I}_u$, the number of components

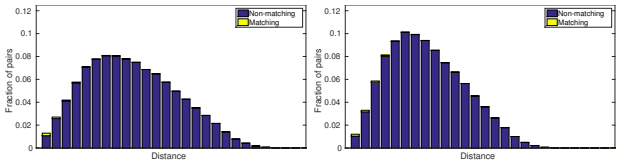
$$K_u = \arg \min_{K=1:K_{MAX}} J(\boldsymbol{I}_u, K) + 2dK, \qquad (7)$$

where d is the dimension of image descriptor and $J(.,.)$ is the standard distortion function for k-means algorithm. We refer to this modified representation as *Multi-Channel Means* (MCM) model. In addition, for generality of our hypothesis, we use a further simplified model that represents a signature using a set of image descriptors corresponding to 10 randomly selected images from the track. We use same three feature descriptors as in MCM for this model and refer to it as *Multi-Channel Random* (MCR) model.

Empirical distance distributions for matching signature pairs in PRID and iLIDS-VID datasets using MCM and MCR representations with Euclidean distance are shown in Figure 2. The distributions are averaged for 10 random trials by selecting half of the datasets. It can be observed that for both datasets and both representations, as the distance increases, the conditional distribution for matching pairs gets close to zero. Therefore, the signature pairs with distance considerably greater than the mode are highly likely to be negative pairs.



(a) PRID



(b) iLIDS-VID

Figure 2: Distributions of distances between pairs of signature of randomly selected half of a) PRID, and b) iLIDS-VID datasets for MCM and MCR representations using Euclidean MPD. The distributions are averaged for 10 trials.

5. Experiments and Results

To show effectiveness of our approach in diverse conditions we experimented with three multi-shot benchmarks. The goal of these experiments is to evaluate: i) sensitivity of learning algorithm to the number of examples selected for training, ii) effect of underlying multi-shot signature representation and iii) effectiveness of unsupervised scheme in comparison to supervised KISSME algorithm and other state-of-the-art approaches. We call KISSME learning with unsupervised annotations as *UnKISSME*. For comparison with state-of-the-art methods, we restrict ourselves to publicly available results on each dataset under consideration.

5.1. Implementation details

We have a limited number of parameters related to signature construction, which we fixed for all our experiments. All images are resized to a fixed windows or 64×192 pixles and are divided into 32×32 overlapping pixels blocks with 16 pixels overlap in each direction. $K_{max} = max(5, 0.1N_t)$, where N_t is the the length of track $\bar{t}$.

5.2. Datasets and experimental setup

For evalulation, we use three challenging datasets:

5.2.1 PRID 2011

PRID 2011 consists of person tracks acquired from two cameras with significant color inconsistency. Although 385 and 749 persons appear in camera 1 and 2, respectively, only 200 appear in both. For evaluation, we followed experimental settings of [28], *i.e*. 178 persons with at least 21 available images are equally and randomly divided into disjoint training and test sets in terms of person IDs. Experiments are repeated 10 times for robust performance estimate using same data splits as in [28].

5.2.2 iLIDS-VID

iLIDS-VID dataset is a subset of iLIDS-MCTS dataset which is collected at a UK airport. The dataset consists of tracks of 300 persons from two different cameras. We follow the same experimental setup of [28] for evaluation by equally and randomly dividing data into disjoint training and test sets. Average performance for 10 trials is reported.

5.2.3 iLIDS-AA

iLIDS-AA is also a subset of iLIDS-MCTS, however, the tracks are produced using automated detection and tracking methods. The dataset has only 100 person tracks each for two cameras. Therefore, following [3] we use a separate set of 40 manually annotated tracks per camera from iLIDS-MCTS. The two datasets are disjoint. We report average performance over 10 trials under these settings.

Table 1: Performance comparison for different training set sizes using recognition rates at rank r

Method	r=1	r=5	r=10	r=20
MCM+UnKISSME-25	57.8	82.0	90.4	96.7
MCM+UnKISSME-50	59.2	81.7	90.6	96.1
MCM+UnKISSME-100	58.1	81.9	89.6	96.0
MCM+UnKISSME-150	55.1	81.0	87.8	94.6

(a) PRID

Method	r=1	r=5	r=10	r=20
MCM+UnKISSME-25	36.9	63.9	75.5	83.5
MCM+UnKISSME-50	38.2	65.7	75.9	84.1
MCM+UnKISSME-100	37.9	64.7	75.0	84.3
MCM+UnKISSME-150	35.9	63.3	74.9	83.4

(b) iLIDS-VID

Method	r=1	r=5	r=10	r=20
MCM+UnKISSME-25	61.2	85.1	92.2	95.8
MCM+UnKISSME-50	61.2	85.1	92.8	96.0
MCM+UnKISSME-100	62.2	85.9	92.6	95.9
MCM+UnKISSME-150	59.2	85.4	92.1	95.7

(c) iLIDS-AA

5.3. Sensitivity to training set size

Proposed unsupervised labeling and learning approach is based on utilization of a fraction of track pairs in the training dataset based on their distances. To understand sensitivity of the method, we performed experiments by selecting farthest pairs equal to 25%, 50%, 100%, and 150% of the number of persons (*not pairs*) available for training. Results of our experiments are reported in Table 1, which shows that on iLIDS-VID and PRID, the method is robust to the exact number of selected pairs between 25% and 100% values. On the other hand, performance on iLIDS-AA is not as consistent for low amount of training data. This can be attributed to the fact that the training set is already very small (40 persons). Therefore, training a metric with only 10 or 20 examples is subject to overfitting.

5.4. Representation comparison

To benchmark improvement due to proposed approach, we considered two different signature representations: MCM and MCR. Performance of learned Mahalanobis metric - using both supervised (KISSME) and proposed unsupervised (UnKISSME) schemes - is compared against Euclidean Distance for the two representations in Table 2. Performance of supervised models can be viewed as an upper limit on the performance of the proposed method.

Unsurprisingly, MCM representation significantly outperforms MCR representation on all datasets for both metrics because it is able to retain significantly more information about person appearance and does not suffer from unfortunate sampling. However, the important observation is that the proposed approach improves rank-1 recognition rate for both signature representations over baseline Euclidean metric and removes significant performance gap between supervised learned metric and the baseline with Euclidean distance. Relative improvement in rank-1 recognition rate of unsupervised scheme w.r.t. supervised scheme is $\sim 68\%$ on average for three datasets for both representations, even though absolute improvement using MCR representation is not huge. This leads us to believe that the proposed unsupervised strategy can be used with different representations irrespective of their baseline performance. However, performance gain due to learning for weaker representations may be limited even for supervised training.

Table 2: Performance comparison of different representations using recognition rates at rank r

Method	r=1	r=5	r=10	r=20
MCM+MPD	53.6	83.1	91.0	96.9
MCM+UnKISSME	59.2	81.7	90.6	96.1
MCM+KISSME	64.3	86.1	94.5	98.0
MCR+MPD	48.7	74.0	83.9	93.3
MCR+UnKISSME	50.8	76.6	85.2	93.3
MCR+KISSME	50.7	75.7	85.6	92.6

(a) PRID

Method	r=1	r=5	r=10	r=20
MCM+MPD	34.3	61.5	74.4	83.3
MCM+UnKISSME	38.2	65.7	75.9	84.1
MCM+KISSME	40.3	69.9	79.0	87.5
MCR+MPD	26.9	51.7	64.8	76.7
MCR+UnKISSME	27.9	52.7	65.3	77.5
MCR+KISSME	28.8	53.7	65.9	78.3

(b) iLIDS-VID

Method	r=1	r=5	r=10	r=20
MCM+MPD	56.5	79.7	90.9	95.2
MCM+UnKISSME	61.2	85.1	92.8	96.0
MCM+KISSME	62.9	84.7	93.4	97.0
MCR+MPD	55.6	80.9	88.6	93.7
MCR+UnKISSME	58.1	81.3	89.8	95.6
MCR+KISSME	60.6	83.4	91.5	95.4

(c) iLIDS-AA

It is important to note that labeling data automatically for the whole dataset takes less than few seconds - less than 5 seconds for selected datasets. Therefore, considering the significant amount of manual labeling effort is saved by our strategy, removing two-third of the performance difference gap by the proposed approach is significant.

5.5. Comparison with state-of-the-art

We compare performance of MCM and both unsupervised (UnKISSME) and supervised (KISSME) algorithms

against competing unsupervised and supervised ReID methods, respectively. Table 3 shows recognition rates at different ranks on the datasets where proposed learning strategy with MCM representation outperforms all other unsupervised approaches. The performance improvement is quite significant on PRID dataset. This can be attributed to both underlying MCM representation and the learned metric. However, further improvement is possible if data is labeled manually for KISSME algorithm. In this case, 98% recognition rate can be achieved using MCM at rank-20.

Similarly, on iLIDS-VID dataset, MCM+UnKISSME outperforms all competing unsupervised approaches for ranks less than 15. However, note that without metric learning MCM+MPD significantly underperforms in comparison to the current state-of-the-art algorithm STFV3D [19]. It can further be noted that performance of fully supervised MCM+KISSME is inferior to STFV3D+KISSME. Therefore, performance of unsupervised learning is limited by the performance of supervised method.

Likewise, on iLIDS-AA dataset, MCM+UnKISSME performs considerably better than other methods, including LBDM [16], which uses unsupervised learning. Furthermore, the performance difference between UnKISSME and KISSME metrics with MCM is quite low. This is due to the fact that the training set is quite different (generated manually) from the test set (generated automatically) and is quite small for supervised learning to be maximally efficient.

6. Conclusion

Person re-identification is challenging due to opposing requirements for low-level features. Metric learning has proven successful to create view specific ReID models from data. However, most metric learning algorithms need manual supervision to annotate signatures pairs as either matching or non-matching. The proposed work addresses problem of automatic association of training data as matching or non-matching pairs for metric learning in the context of multi-shot ReID. The approach uses empirical marginal distribution of pair-wise distances from the training data to automatically construct the matching and non-matching training pairs. This approach is independent of the underlying signature representation and the metric learning method, hence it is widely applicable in multi-shot ReID scenarios. We successfully demonstrate effectiveness of proposed approach on multiple data benchmarks using two signature representations and a metric learning algorithm.

Acknowledgement: The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement No. 324359.

Table 3: Performance comparison with state-of-the-art using recognition rates at rank r

Unsupervised Methods	r=1	r=5	r=10	r=20
Color+LFDA[22]	43.0	73.1	82.9	90.3
SDALF[9]	5.2	20.7	32.0	47.9
Salience[34]	25.8	43.6	52.6	62.0
FV2D[21]	33.6	64.0	76.3	86.0
FV3D[19]	38.7	71.0	80.6	90.3
DVDL[13]	40.6	69.7	77.8	85.6
STFV3D[19]	42.1	71.9	84.4	91.6
MCM+UnKISSME	**59.2**	**81.7**	**90.6**	**96.1**
Supervised Methods	r=1	r=5	r=10	r=20
Color+DVR[28]	41.8	63.8	76.7	88.3
ColorLBP+DVR[28]	37.6	63.9	75.3	89.4
ColorLBP+RSVM[28]	34.3	56.0	65.5	77.3
DVR[28]	28.9	55.3	65.5	82.8
DSVR[29]	40.0	71.7	84.5	92.2
Salience+DVR[28]	41.7	64.5	77.5	88.8
SDALF+DVR[28]	31.6	58.0	70.3	85.3
STFV3D+KISSME[19]	64.1	**87.3**	89.9	92.0
MCM+KISSME	**64.3**	86.1	**94.5**	**98.0**

(a) PRID

Unsupervised Methods	r=1	r=5	r=10	r=20
SDALF[9]	5.1	19.0	27.1	37.9
Salience[34]	10.2	24.8	35.5	52.9
FV2D[21]	18.2	35.6	49.2	63.8
FV3D[19]	25.3	54.0	68.3	**87.7**
DVDL[13]	25.9	48.2	57.3	68.9
STFV3D[19]	37.0	64.3	**77.0**	86.9
MCM+UnKISSME	**38.2**	**65.7**	75.9	84.1
Supervised Methods	r=1	r=5	r=10	r=20
MLF[35]	11.7	29.1	40.3	53.4
Color+RSVM[28]	16.4	37.3	48.5	62.6
ColorLBP+DVR[28]	32.7	56.5	67.0	77.4
ColorLBP+RSVM[28]	20.0	44.0	52.7	68.0
DVR[28]	23.3	42.4	55.3	68.6
DSVR[29]	39.5	61.1	71.7	81.0
MTL-LORAE[26]	43.0	60.1	70.3	85.3
STFV3D+KISSME[19]	**43.8**	69.3	**80.0**	**90.0**
MCM+KISSME	40.3	**69.9**	79.0	87.5

(b) iLIDS-VID

Unsupervised Methods	r=1	r=5	r=10	r=20
MRCG-B [4]	47.1	68.8	75.9	87.1
RSCNN [15]	50.1	73.8	83.8	91.5
LBDM [16]	57.8	80.5	88.9	**96.3**
MCM+UnKISSME	**61.2**	**85.1**	**92.8**	96.0
Supervised Methods	r=1	r=5	r=10	r=20
COSMATI [3]	33.8	59.2	71.2	82.7
MCM+KISSME	**62.9**	**84.7**	**93.4**	**97.0**

(c) iLIDS-AA

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015.

[2] Anonymous. Person re-identication for real-world surveillance systems. In *ECCV - under review*, 2016.

[3] S. Bak, G. Charpiat, E. Corvee, F. Bremond, and M. Thonnat. Learning to match appearances by correlations in a covariance metric space. In *ECCV*, 2012.

[4] S. Bak, R. Kumar, and F. Bremond. Brownian descriptor: a rich meta-feature for appearance matching. In *WACV*, 2014.

[5] L. Bazzani, M. Cristani, A. Perina, M. Farenzena, and V. Murino. Multiple-shot person re-identification by hpe signature. In *ICPR*, 2010.

[6] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang. Similarity learning on an explicit polynomial kernel feature map for person. In *CVPR*, 2015.

[7] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, 2007.

[8] M. Dikmen, E. Akbas, T. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *ACCV*, 2010.

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

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

[11] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, 2009.

[12] M. Hirzer, P. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, 2012.

[13] S. Karanam, Y. Li, and R. J. Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *ICCV*, 2015.

[14] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[15] W. Li, Y. Wu, Y. Kawanishi, M. Mukunoki, and M. Minoh. Riemannian set-level common-near-neighbor analaysis for multiple-shot person re-identification. In *International Conference on Machine Vision Applications*, 2013.

[16] W. Li, Y. Wu, M. Mukunoki, Y. Kuang, and M. Minoh. Locality based discriminative measure for multiple-shot human re-identification. *Neurocomputing*, 2015.

[17] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015.

[18] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: What features are important? In *ECCV Workshops and Demonstrations*, 2012.

[19] K. Liu, W. Zhang, and R. Huang. A spatio-temporal appearance representation for video-based pedestrian re-identification. In *ICCV*, 2015.

[20] X. Liu, M. Song, Q. Zhao, D. Tao, C. Chen, and J. Bu. Attribute-restricted latent topic model for person re-identification. *Pattern Recognition*, pages 4204–4213, 2012.

[21] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher descriptors for person re-identification. In *ECCV Workshops*, 2012.

[22] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*, 2013.

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

[24] W. R. Schwartz and L. S. Davis. Learning discriminative apperance-based models using partial least squares. In *Brazilian Symposium on Computer Graphics and Image Processing*, 2009.

[25] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang. Person re-identification with correspondence structure learning. In *ICCV*, 2015.

[26] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao. Multi-task learning with low aank attribute embedding for person re-identification. In *CVPR*, 2015.

[27] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *ECCV*, 2006.

[28] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*, 2014.

[29] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by discriminative selection in video ranking. *T-PAMI*, 2016.

[30] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *ICCV*, 2007.

[31] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2006.

[32] K. Q. Weinberger and L. K. Saul. Fast solvers and efficient implementations for distance metric learning. In *ICML*, 2008.

[33] M. Zeng, Z. Wu, C. Tian, L. Zhang, and L. Hu. Efficient person re-identification by hybrid spatiogram and covariance descriptor. In *CVPR Workshop*, 2015.

[34] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *ICCV*, 2013.

[35] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identfiation. In *CVPR*, 2014.

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