Person Re-identification by Local Maximal Occurrence Representation and Metric Learning

Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li
Center for Biometrics and Security Research, National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun East Road, Beijing 100190, China
{scliao,yhu,xiangyu.zhu,szli}@nlpr.ia.ac.cn

Abstract

Person re-identification is an important technique towards automatic search of a person's presence in a surveillance video. Two fundamental problems are critical for person re-identification, feature representation and metric learning. An effective feature representation should be robust to illumination and viewpoint changes, and a discriminant metric should be learned to match various person images. In this paper, we propose an effective feature representation called Local Maximal Occurrence (LOMO), and a subspace and metric learning method called Cross-view Quadratic Discriminant Analysis (XQDA). The LOMO feature analyzes the horizontal occurrence of local features, and maximizes the occurrence to make a stable representation against viewpoint changes. Besides, to handle illumination variations, we apply the Retinex transform and a scale invariant texture operator. To learn a discriminant metric, we propose to learn a discriminant low dimensional subspace by cross-view quadratic discriminant analysis, and simultaneously, a QDA metric is learned on the derived subspace. We also present a practical computation method for XQDA, as well as its regularization. Experiments on four challenging person re-identification databases, VIPeR, QMUL GRID, CUHK Campus, and CUHK03, show that the proposed method improves the state-of-the-art rank-1 identification rates by 2.2%, 4.88%, 28.91%, and 31.55% on the four databases, respectively.

1. Introduction

Person re-identification is a problem of finding a person from a gallery who has the same identity to the probe. This is a challenging problem because of big intra-class variations in illumination, pose or viewpoint, and occlusion. Many approaches have been proposed for person re-identification [40, 8], which greatly advance this field.

Two fundamental problems are critical for person re-identification, feature representation and metric learning. An effective feature representation should be robust to illumination and viewpoint changes, and a discriminant metric should be learned to match various person images. Many efforts have been made along the two directions to tackle the challenge of person re-identification. For feature representation, several effective approaches have been proposed, for example, the ensemble of local features (ELF) [10], SDALF [2], kBiCov [33], fisher vectors (LDFV) [32], salience match [46], and mid-level filter [48]. These hand-crafted or learning based descriptors have made impressive improvements over robust feature representation, and advanced the person re-identification research. However, how to design or learn a robust feature for the person re-identification challenge still remains an open problem.

Another aspect of person re-identification is how to learn a robust distance or similarity function to deal with the complex matching problem. Many metric learning algorithms have been proposed considering this aspect [5, 49, 18, 14, 24]. In practice, many previous metric learning methods [43, 4, 5, 14, 18, 38] show a two-stage processing for metric learning, that is, the Principle Component Analysis (PCA) is first applied for dimension reduction, then metric learning is performed on the PCA subspace. However, this two-stage processing may not be optimal for metric learning in low dimensional space, because samples of different classes may already be cluttered after the first stage.

In this paper, we propose an efficient feature representation called Local Maximal Occurrence (LOMO), and a subspace and metric learning method called Cross-view Quadratic Discriminant Analysis (XQDA). The LOMO feature analyzes the horizontal occurrence of local features, and maximizes the occurrence to make a stable representation against viewpoint changes. Besides, we find that applying the Retinex transform is useful to handle illumination variations in person re-identification. To learn a discriminant metric, we propose to learn a discriminant low dimensional subspace by cross-view quadratic discriminant analysis, and simultaneously, a QDA metric is learned on

the derived subspace. We show that the problem can be formulated as a Generalized Rayleigh Quotient, and a closed-form solution can be obtained by the generalized eigenvalue decomposition. We also present a practical computation method for XQDA, as well as its regularization and dimension selection. The proposed method is shown to be effective and efficient through person re-identification experiments on four public databases, and we also demonstrate how the proposed components lead to improvements.

2. Related Work

Many existing person re-identification approaches try to build a robust feature representation which is both distinctive and robust for describing a person's appearance under various conditions [10, 15, 7, 12, 41, 3]. Gray and Tao [10] proposed to use AdaBoost to select good features out of a set of color and texture features. Farenzena et al. [6] proposed the Symmetry-Driven Accumulation of Local Features (SDALF) method, where the symmetry and asymmetry property is considered to handle viewpoint variations. Ma et al. [32] turned local descriptors into the Fisher Vector to produce a global representation of an image. Cheng et al. [3] utilized the Pictorial Structures where part-based color information and color displacement were considered for person re-identification. Recently, saliency information has been investigated for person re-identification [47, 46, 29], leading to a novel feature representation. In [42], a method called regionlets is proposed, which picks a maximum bin from three random regions for object detection under deformation. In contrast, we propose to maximize the occurrence of each local pattern among all horizontal sub-windows to tackle viewpoint changes.

Besides robust features, metric learning has been widely applied for person re-identification [43, 4, 11, 5, 49, 18, 14, 24]. Zheng et al. [49] proposed the PRDC algorithm, which optimizes the relative distance comparison. Hirzer et al. [14] proposed to relax the PSD constraint required in Mahalanobis metric learning, and obtained a simplified formulation that still showed promising performance. Li at al.[24] proposed the learning of Locally-Adaptive Decision Functions (LADF) for person verification, which can be viewed as a joint model of a distance metric and a locally adapted thresholding rule. Prosser et al.[39] formulated the person re-identification problem as a ranking problem, and applied the RankSVM to learn a subspace. In [21], local experts were considered to learn a common feature space for person re-identification across views.

Except a novel feature representation, the proposed XQDA algorithm is mostly related to Bayesian face [36], KISSME [18], Linear Discriminant Analysis (LDA) [13], local fisher discriminant analysis (LF) [38], and CFML [1]. XQDA can be seen as an extension of Bayesian face and KISSME, in that a discriminant subspace is further learned together with a metric. The LF method applies FDA together with PCA and LPP to derive a low dimensional yet discriminant subspace. The CFML algorithm aims at a different problem though learns a similar subspace to XQDA. However, both LF and CFML use the Euclidean distance on the derived subspace, while the proposed method considers a discriminant subspace as well as an integrated metric. For the traditional LDA method, though XQDA shares a similar generalized Rayleigh quotient formulation, they are essentially not equivalent, which is explained in [1].

(a) (b)

Figure 1. (a) Example pairs of images from the VIPeR database [9]. (b) Processed images in (a) by Retinex. Images in the same column represent the same person.

3. Local Maximal Occurrence Feature

3.1. Dealing with Illumination Variations

Color is an important feature for describing person images. However, the illumination conditions across cameras can be very different, and the camera settings might also be different from camera to camera. Therefore, the perceived colors of the same person may vary largely from different camera views. For example, Fig. 1 (a) shows some sample images from the VIPeR database [9]. It can be seen that images of the same person across the two camera views have a large difference in illumination and color appearance.

In this paper, we propose to apply the Retinex algorithm [20, 17, 16] to preprocess person images. Retinex considers human lightness and color perception. It aims at producing a color image that is consistent to human observation of the scene. The restored image usually contains vivid color information, especially enhanced details in shadowed regions.

We implement the multiscale Retinex algorithm according to [16], which combines the small-scale Retinex for dynamic range compression and the large-scale Retinex for tonal rendition simultaneously. As a result, the algorithm handles both the color constancy and dynamic range compression automatically, achieving a good approximation to human visual perception. Specifically, we use two scales of center/surround Retinex, with $\sigma = 5$ and $\sigma = 20$. Besides, we automatically compute the gain/offset parameters so that the resulting intensities linearly stretches in [0,255].

Fig. 1 (b) shows some examples of the processed images

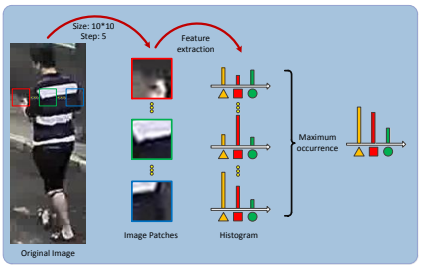


Figure 2. Illustration of the LOMO feature extraction method.

by our implementation of Retinex. Comparing to Fig. 1 (a), it can be observed that the Retinex images of the same person across cameras have a better consistency in lighting and color. This makes person re-identification easier than using the original images. With the Retinex images, we apply the HSV color histogram to extract color features.

In addition to color description, we also apply the Scale Invariant Local Ternary Pattern (SILTP) [26] descriptor for illumination invariant texture description. SILTP is an improved operator over the well-known Local Binary Pattern (LBP) [37]. In fact, LBP has a nice invariant property under monotonic gray-scale transforms, but it is not robust to image noises. SILTP improves LBP by introducing a scale invariant local comparison tolerance, achieving invariance to intensity scale changes and robustness to image noises.

3.2. Dealing with Viewpoint Changes

Pedestrians under different cameras usually appear in different viewpoint. For example, a person with frontal view in a camera may appear in back view under another camera. Therefore, matching persons in different viewpoints is also difficult. To address this, [39, 49] proposed to equally divide a person image into six horizontal stripes, and a single histogram is computed in each stripe. This feature has made a success in viewpoint invariant person representation [39, 49, 27]. However, it may also lose spatial details within a stripe, thus affecting its discriminative power.

We propose to use sliding windows to describe local details of a person image. Specifically, we use a subwindow size of 10×10, with an overlapping step of 5 pixels to locate local patches in 128×48 images. Within each subwindow, we extract two scales of SILTP histograms (SILTP$^{0.3}_{4,3}$ and SILTP$^{0.3}_{4,5}$), and an $8 \times 8 \times 8$-bin joint HSV histogram. Each histogram bin represents the occurrence probability of one pattern in a subwindow. To address viewpoint changes, we check all subwindows at the same horizontal location, and maximize the local occurrence of each pattern (i.e. the same histogram bin) among these subwindows. The resulting histogram achieves some invariance to viewpoint changes, and at the same time captures local region characteristics of a person. Fig. 2 shows the procedure of the proposed LOMO feature extraction.

To further consider the multi-scale information, we build a three-scale pyramid representation, which downsamples the original 128×48 image by two 2×2 local average pooling operations, and repeats the above feature extraction procedure. By concatenating all the computed local maximal occurrences, our final descriptor has ($8 * 8 * 8$ color bins + $3^4 * 2$ SILTP bins) $*$ ($24 + 11 + 5$ horizontal groups) $= 26,960$ dimensions. Finally, we apply a log transform to suppress large bin values, and normalize both HSV and SILTP features to unit length. Since we only use simple HSV and SILTP features, the proposed feature extraction method is efficient to compute (see Section 5.5.4).

4. Cross-view Quadratic Discriminant Analysis

4.1. Bayesian Face and KISSME Revisit

Consider a sample difference $\Delta = \mathbf{x}_i - \mathbf{x}_j$. Δ is called the intrapersonal difference if $y_i = y_j$, while it is called the extrapersonal difference if $y_i \neq y_j$ [36]. Accordingly, two classes of variations can be defined: the intrapersonal variations Ω_I and the extrapersonal variations Ω_E. Therefore, in this way the multi-class classification problem can be solved by distinguishing the above two classes. Moghaddam et al. [36] proposed to model each of the two classes with a multivariate Gaussian distribution. This corresponds to a QDA model with the defined Ω_I and Ω_E as two classes. Furthermore, it was noticed in [36] that both Ω_I and Ω_E have zero mean. The resulting algorithm is called Bayesian face applied to face recognition. Interestingly, in [18], Köstinger et al. also derived a similar approach called KISSME via the log likelihood ratio test of the two Gaussian distributions, and applied it to person re-identification.

Formally, the Bayesian face and the KISSME algorithms are formulated as follows. Under the zero-mean Gaussian distribution, the likelihoods of observing Δ in Ω_I and Ω_E are defined as

$$P(\Delta|\Omega_I) = \frac{1}{(2\pi)^{d/2}|\Sigma_I|^{1/2}} e^{-\frac{1}{2}\Delta^T\Sigma_I^{-1}\Delta}, \quad (1)$$

$$P(\Delta|\Omega_E) = \frac{1}{(2\pi)^{d/2}|\Sigma_E|^{1/2}} e^{-\frac{1}{2}\Delta^T\Sigma_E^{-1}\Delta}, \quad (2)$$

where Σ_I and Σ_E are the covariance matrices of Ω_I and Ω_E, respectively, and n_I and n_E denotes the number of samples in the two classes. By applying the Bayesian rule and the log-likelihood ratio test, the decision function can be simplified as

$$f(\Delta) = \Delta^T(\Sigma_I^{-1} - \Sigma_E^{-1})\Delta, \quad (3)$$

and so the derived distance function between $\mathbf{x}_i$ and $\mathbf{x}_j$ is

$$d(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T(\Sigma_I^{-1} - \Sigma_E^{-1})(\mathbf{x}_i - \mathbf{x}_j). \quad (4)$$

Therefore, learning the distance function corresponds to estimating the covariant matrices Σ_I and Σ_E.

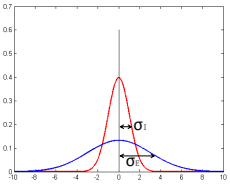


Figure 3. Distributions of Ω_I and Ω_E in one projected dimension.

4.2. XQDA

Usually, the original feature dimensions d is large, and a low dimensional space $\mathbb{R}^r$ ($r < d$) is preferred for classification. [36] suggested to decompose Σ_I and Σ_E separately to reduce the dimensions. In [18], PCA was applied, then Σ_I and Σ_E were estimated in the PCA subspace. However, both methods are not optimal because the dimension reduction does not consider the distance metric learning.

In this paper, we extend the Bayesian face and KISSME approaches to cross-view metric learning, where we consider to learn a subspace $W = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r) \in \mathbb{R}^{d\times r}$ with cross-view data, and at the same time learn a distance function in the r dimensional subspace for the cross-view similarity measure. Suppose we have a cross-view training set $\{\mathbf{X}, \mathbf{Z}\}$ of c classes, where $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d\times n}$ contains n samples in a d-dimensional space from one view, $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m) \in \mathbb{R}^{d\times m}$ contains m samples in the same d-dimensional space but from the other view. The cross-view matching problem arises from many applications, like heterogeneous face recognition [25] and viewpoint invariant person re-identification [10]. Note that $\mathbf{Z}$ is the same with $\mathbf{X}$ in the single-view matching scenario.

Considering a subspace W, the distance function Eq. (4) in the r dimensional subspace is computed as

$$d_W(\mathbf{x}, \mathbf{z}) = (\mathbf{x} - \mathbf{z})^T W(\Sigma_I'^{-1} - \Sigma_E'^{-1})W^T(\mathbf{x} - \mathbf{z}), \quad (5)$$

where $\Sigma_I' = W^T\Sigma_I W$ and $\Sigma_E' = W^T\Sigma_E W$. Therefore, we needs to learn a kernel matrix $M(W) = W(\Sigma_I'^{-1} - \Sigma_E'^{-1})W^T$. However, directly optimizing d_W is difficult because W is contained in two inverse matrices.

Recall that Ω_I and Ω_E have zero mean, then given a basis $\mathbf{w}$, the projected samples of the two classes will still center at zero, but may have different variances, as shown in Fig. 3. In this case, the traditional Fisher criterion used to derive LDA is no longer suitable because the two classes have the same mean. However, the variances σ_I and σ_E can still be used to distinguish the two classes. Therefore, we can optimize the projection direction $\mathbf{w}$ such that $\sigma_E(\mathbf{w})/\sigma_I(\mathbf{w})$ is maximized. Notice that $\sigma_I(\mathbf{w}) = \mathbf{w}^T\Sigma_I\mathbf{w}$ and $\sigma_E(\mathbf{w}) = \mathbf{w}^T\Sigma_E\mathbf{w}$, therefore, the objective $\sigma_E(\mathbf{w})/\sigma_I(\mathbf{w})$ corresponds to the Generalized Rayleigh Quotient

$$J(\mathbf{w}) = \frac{\mathbf{w}^T\Sigma_E\mathbf{w}}{\mathbf{w}^T\Sigma_I\mathbf{w}}. \quad (6)$$

The maximization of $J(\mathbf{w})$ is equivalent to

$$\max_{\mathbf{w}} \mathbf{w}^T\Sigma_E\mathbf{w}, \ s.t. \ \mathbf{w}^T\Sigma_I\mathbf{w} = 1, \quad (7)$$

which can be solved by the generalized eigenvalue decomposition problem as similar in LDA. That is, the largest eigenvalue of $\Sigma_I^{-1}\Sigma_E$ is the maximum value of $J(\mathbf{w})$, and the corresponding eigenvector $\mathbf{w}_1$ is the solution. Furthermore, the solution orthogonal to $\mathbf{w}_1$ and corresponding to the second largest value of $J(\mathbf{w})$ is the eigenvector of the second largest eigenvalue of $\Sigma_I^{-1}\Sigma_E$, and so on. Therefore, with $W = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r)$ we learn a discriminant subspace, as well as a distance function in the learned subspace, as defined in Eq. (5). We call the derived algorithm Cross-view Quadratic Discriminant Analysis (XQDA) to reflect its connection to QDA and the output of a cross-view metric.

4.3. Practical Computation

The computation of the two covariance matrices Σ_I and Σ_E require $O(Nkd^2)$ and $O(nmd^2)$ multiplication operations, respectively, where $N = max(m, n)$, and k represents the average number of images in each class. To reduce the computation, we show that

$$n_I\Sigma_I = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T + \widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^T - \mathbf{S}\mathbf{R}^T - \mathbf{R}\mathbf{S}^T, \quad (8)$$

where $\widetilde{\mathbf{X}} = (\sqrt{m_1}\mathbf{x}_1, \sqrt{m_1}\mathbf{x}_2, \ldots, \sqrt{m_1}\mathbf{x}_{n_1}, \ldots, \sqrt{m_c}\mathbf{x}_n)$, $\widetilde{\mathbf{Z}} = (\sqrt{n_1}\mathbf{z}_1, \sqrt{n_1}\mathbf{z}_2, \ldots, \sqrt{n_1}\mathbf{z}_{m_1}, \ldots, \sqrt{n_c}\mathbf{z}_m)$, $\mathbf{S} = (\sum_{y_i=1}\mathbf{x}_i, \sum_{y_i=2}\mathbf{x}_i, \ldots, \sum_{y_i=c}\mathbf{x}_i)$, $\mathbf{R} = (\sum_{l_j=1}\mathbf{z}_j, \sum_{l_j=2}\mathbf{z}_j, \ldots, \sum_{l_j=c}\mathbf{z}_j)$, y_i and l_j are class labels, n_k is the number of samples in class k of $\mathbf{X}$, and m_k is the number of samples in class k of $\mathbf{Z}$. Besides,

$$n_E\Sigma_E = m\mathbf{X}\mathbf{X}^T + n\mathbf{Z}\mathbf{Z}^T - \mathbf{s}\mathbf{r}^T - \mathbf{r}\mathbf{s}^T - n_I\Sigma_I, \quad (9)$$

where $\mathbf{s} = \sum_{i=1}^n \mathbf{x}_i$ and $\mathbf{r} = \sum_{j=1}^m \mathbf{z}_j$. The above simplification shows that the computations of Σ_I and Σ_E are both reduced to $O(Nd^2)$. It can be observed that, Σ_I and Σ_E can be computed directly from sample mean and covariance of each class and all classes, so there is no need to actually compute the mn pairs of sample differences required in many other metric learning algorithms.

Another practical issue is that, Σ_I may be singular, resulting that Σ_I^{-1} cannot be computed. Therefore, it is useful to add a small regularizer to the diagonal elements of Σ_I, as usually done in similar problems like LDA. This will make the estimation of Σ_I more smooth and robust. Empirically we find that, when all samples are normalized to unit length, a value of 0.001 as a regularizer can be commonly applied to improve the result.

Finally, there is a remaining issue of selecting the dimensionality of the derived XQDA subspace. In real applications, there should be a consideration to have a low dimensional subspace to ensure the processing speed. Beyond this

consideration, we find that having the selected eigenvalues of $\Sigma_I^{-1}\Sigma_E$ larger than 1 is a useful signature to determine the dimensions. This is because the eigenvalue of $\Sigma_I^{-1}\Sigma_E$ corresponds to σ_E/σ_I in Fig. 3, and $\sigma_E < \sigma_I$ may not provide useful discriminant information.

5. Experiments

5.1. Experiments on VIPeR

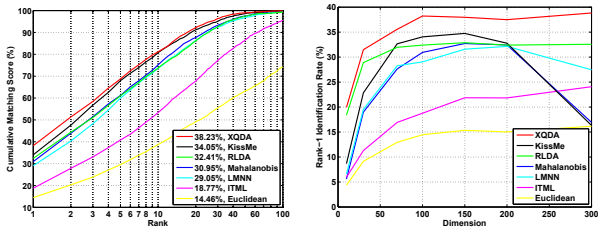
VIPeR [9] is a challenging person re-identification database that has been widely used for benchmark evaluation. It contains 632 pairs of person images, captured by a pair of cameras in an outdoor environment. Images in VIPeR contains large variations in background, illumination, and viewpoint. Fig. 1 (a) shows some example pairs of images from the VIPeR database. All images are scaled to 128×48 pixels. The widely adopted experimental protocol on this database is to randomly divide the 632 pairs of images into half for training and the other half for testing, and repeat the procedure 10 times to get an average performance. We followed this procedure in our experiments.

5.1.1 Comparison of Metric Learning Algorithms

We evaluated the proposed XQDA algorithm and several metric learning algorithms, including Euclidean distance, Mahalanobis distance trained with genuine pairs[18], LMNN v2.5[43], ITML [4], KISSME [18], and RLDA [45], with the same LOMO feature. For the compared algorithms, PCA was first applied to reduce the feature dimensionality to 100.The proposed XQDA algorithm and RLDA also learned a 100-dimensional subspace. The resulting Cumulative Matching Characteristic (CMC) curves are shown in Fig. 4 (a). It can be seen that the proposed method is better than the compared metric learning algorithms. This indicates that XQDA successfully learns a discriminant subspace as well as an effective metric. Besides, we also investigate how the performance varies with different subspace dimensions, as shown in Fig. 4 (b). It can be observed that XQDA consistently performs the best with all dimensions.

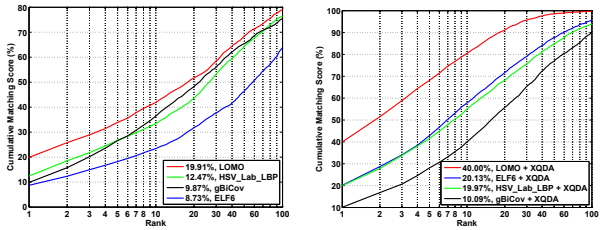
5.1.2 Comparison of Features

Next, we compared the proposed LOMO feature with other three available person re-identification features. The first feature is called Ensemble of Local Features (ELF), proposed in [10], and later modified by [39, 49]. We used the implementation in [50], denoted by ELF6, which is computed from histograms in six equally divided horizontal stripes. Eight color channels (RGB, HSV, and YCbCr) and 21 texture filters (8 Gabor filters and 13 Schmid filters) are used for the histogram representation. The other feature is proposed in [18], which applied the HSV, and Lab color feature, as well as a texture feature extracted by LBP. The third



(a) (b)

Figure 4. Comparison of metric learning algorithms with the same LOMO feature on the VIPeR database [9] (P=316). (a) CMC curves with feature reduced to 100 dimensions. (b) Rank-1 identification rates with varying subspace dimensions.



(a) Cosine (b) XQDA

Figure 5. CMC curves and rank-1 identification rates on the VIPeR database [9] (P=316), by comparing the proposed LOMO feature to three available features, ELF6 [50], HSV+Lab+LBP [18], and gBiCov [33].

feature called gBiCov[1] [33] is a combination of Biologically Inspired Features (BIF) and Covariance descriptors. We applied both the direct Cosine similarity measure and the XQDA algorithm to compare the four different kinds of features, resulting in the CMC curves shown in Fig. 5. For consistency, in the following experiments we determined the subspace dimensions of XQDA automatically by accepting all eigenvalues of $\Sigma_I^{-1}\Sigma_E$ that are larger than 1, as discussed earlier. From Fig. 5 (a) it can be seen that the raw LOMO feature outperforms the other existing features. What's more, Fig. 5 (b) shows that the performance improvement is more significant with the help of XQDA. Since these kinds of features are similar in fusing color and texture information, the improvement made by the proposed LOMO feature is mainly due to the specific consideration of handling illumination and viewpoint changes.

5.1.3 Comparison to the State of the Art

Finally, we compare the performance of the proposed approach to the state-of-the-art results reported on the VIPeR database, which are summarized in Fig. 6 and Table 1. Four methods, the SCNCD [44], kBiCov [33], LADF [24], and SalMatch [46] report the best performances on the VIPeR

[1]We used the author's implementation (available in http://vipl.ict.ac.cn/members/bpma) and the default parameters, which may not reflect the best status.

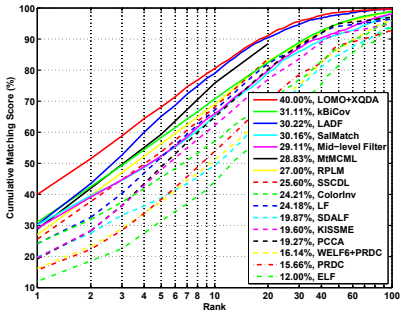


Figure 6. CMC curves and rank-1 identification rates on the VIPeR database [9] (P=316) by comparing the proposed LOMO+XQDA method to other state of the art algorithms.

Table 1. Comparison of state-of-the-art results reported with the VIPeR database (P=316). The cumulative matching scores (%) at rank 1, 10, and 20 are listed.

Method	rank=1	rank=10	rank=20	Reference
LOMO+XQDA	**40.00**	80.51	**91.08**	Proposed
SCNCD	37.80	**81.20**	90.40	2014 ECCV [44]
kBiCov	31.11	70.71	82.45	2014 IVC [33]
LADF	30.22	78.92	90.44	2013 CVPR [24]
SalMatch	30.16	65.54	79.15	2013 ICCV [46]
Mid-level Filter*	29.11	65.95	79.87	2014 CVPR [48]
MtMCML	28.83	75.82	88.51	2014 TIP [34]
RPLM	27.00	69.00	83.00	2012 ECCV [14]
LDFV	26.53	70.88	84.63	2012 ECCVW [32]
SSCDL	25.60	68.10	83.60	2014 CVPR [28]
ColorInv	24.21	57.09	69.65	2013 TPAMI [19]
LF	24.18	67.12	82.00	2013 CVPR [38]
SDALF	19.87	49.37	65.73	2013 CVIU [2]
KISSME	19.60	62.20	77.00	2012 CVPR [18]
PCCA	19.27	64.91	80.28	2012 CVPR [35]
WELF6+PRDC	16.14	50.98	65.95	2012 ECCVW [27]
PRDC	15.66	53.86	70.09	2013 TPAMI [50]
ELF	12.00	44.00	61.00	2008 ECCV [10]

*Note that [48] reports a 43.39% rank-1 accuracy by fusing their method with LADF [24]. Fusing different methods generally improves the performance. In fact, we also tried to fuse our method with LADF, and got a 50.32% rank-1 identification rate.

dataset to date, which exceed 30% at rank 1. From Table 1 it can be observed that the proposed algorithm achieves the new state of the art, 40% at rank 1, outperforming the second best one SCNCD by 2.2%.

5.2. Experiments on QMUL GRID

The QMUL underGround Re-IDentification (GRID) dataset [31] is another challenging person re-identification test bed but have not been largely noticed. The GRID dataset was captured from 8 disjoint camera views in a underground station. It contains 250 pedestrian image pairs, with each pair contains two images of the same person from different camera views. Besides, there are 775 additional

Figure 7. Example pairs of images from the GRID database [31]. Images in the same column represent the same person.

Table 2. Comparison of state-of-the-art results on the GRID database (P=900) without camera network information. Red and blue numbers are the best and second best results, respectively.

Method	rank=1	rank=10	rank=20
ELF6 + L1-norm [30]	4.40	16.24	24.80
ELF6 + RankSVM [39]	10.24	33.28	43.68
ELF6 + PRDC [50]	9.68	32.96	44.32
ELF6 + MRank-RankSVM [30]	12.24	36.32	46.56
ELF6 + MRank-PRDC [30]	11.12	35.76	46.56
ELF6 + XQDA	10.48	38.64	**52.56**
LOMO + XQDA	**16.56**	**41.84**	52.40

Table 3. Comparison of state-of-the-art results on the GRID database (P=900) with camera network information. Red and blue numbers are the best and second best results, respectively.

Method	rank=1	rank=10	rank=20
ELF6 + MtMCML [34]	14.08	45.84	59.84
ELF6 + XQDA	16.32	40.72	51.76
LOMO + XQDA	**18.96**	**52.56**	**62.24**

images that do not belong to the 250 persons which can be used to enlarge the gallery set. Sample images from GRID can be found in Fig. 7. It can be seen that these images have poor image quality and low resolutions, and contain large variations of illumination and viewpoint.

An experimental setting of 10 random trials is provided for the GRID dataset. For each trial, 125 image pairs are used for training, and the remaining 125 image pairs, as well as the 775 background images are used for test. The ELF6 feature set described in [27] is provided for developing machine learning algorithms.

We first applied the proposed method on the provided feature set of GRID. This leads to results of "ELF6+XQDA" listed in Table 2. We compared available results from [30] where the same feature set was used. Results shown in Table 2 indicates that the proposed joint dimension reduction and metric learning approach outperforms other distance learning algorithms such as RankSVM [39], PRDC [49], and MRank [30], except that the rank-1 accuracy of XQDA is slightly worse than MRank-RankSVM.

We also tried the proposed feature extraction method, and applied the same XQDA algorithm for metric learning.

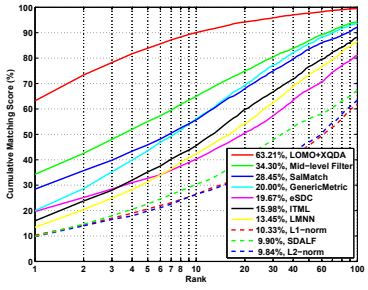


Figure 8. Multi-shot CMC curves and rank-1 identification rates on the CUHK Campus database [22] (P=486, M=2). The compared results are from [48].

Table 4. Comparison of state-of-the-art rank-1 identification rates (%) on the CUHK03 database [23] with both labeled and detected setting (P=100). The compared results are from [23].

	Labeled	**Detected**
LOMO+XQDA	**52.20**	**46.25**
DeepReID [23]	20.65	19.89
KISSME [18]	14.17	11.70
LDML [11]	13.51	10.92
eSDC [47]	8.76	7.68
LMNN [43]	7.29	6.25
ITML [4]	5.53	5.14
SDALF [2]	5.60	4.87

This corresponds to the results of the last row in Table 2. The comparison shows that the new feature improves the performance at rank 1-10. Especially, a 4.32% performance gain can be obtained for the rank-1 accuracy. This indicates that the new feature helps to reduce intra-class variations, so that the same person can be recognized at a higher rank.

Note that the above methods all trained a general model independent of camera views. A research in [34] show that the performance can be improved by utilizing the camera network information. Namely, their method MtMCML trained various metrics, each for a given camera view pair. We also followed this approach and trained several metrics depending on known camera pairs. Results listed in Table 3 show that, while with the ELF6 feature the proposed method only improves the rank-1 accuracy over MtMCML, with the new LOMO feature the proposed method is clearly better than MtMCML. However, in practice we do not suggest this way of training because the camera views under evaluation are usually unseen, and it is not easy to label data for new camera views to retrain the algorithm.

5.3. Experiments on CUHK Campus

The CUHK Campus dataset was captured with two camera views in a campus environment. Different from the above datasets, images in this dataset are of higher resolution. The CUHK Campus dataset contains 971 persons, and each person has two images in each camera view. Camera A captures the frontal view or back view of pedestrians, while camera B captures the side views. All images were scaled to 160×60 pixels. The persons were split to 485 for training and 486 for test (multi-shot). The results are shown in Fig. 8. Our method largely outperforms existing state of the art methods. The best rank-1 identification rate reported to date is 34.30% [48], while we has achieved 63.21%, with an improvement of 28.91%.

5.4. Experiments on CUHK03

The CUHK03 dataset [23] includes 13,164 images of 1,360 pedestrians. It is currently the largest publicly available person re-identification dataset. The CUHK03 dataset was captured with six surveillance cameras over months, with each person observed by two disjoint camera views and having an average of 4.8 images in each view. In addition to manually cropped pedestrian images, samples detected with a state-of-the-art pedestrian detector is also provided. This is a more realistic setting considering misalignment, occlusions and body part missing.

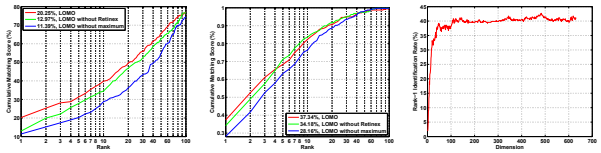
We run our algorithm with the same setting of [23]. That is, the dataset was partitioned into a training set of 1,160 persons and a test set of 100 persons. The experiments were conducted with 20 random splits and all the CMC curves were computed with the single-shot setting. The rank-1 identification rates of various algorithms in both labeled and detected setting are shown in Table 4. The proposed method achieved 52.20% and 46.25% rank-1 identification rates with the labeled bounding boxes and the automatically detected bounding boxes, respectively, which clearly outperform the state-of-the-art method DeepReID [23], with an improvement of 31.55% for the labelled setting, and 26.36% for the detected setting.

5.5. Analysis of the Proposed Method

To better understand the proposed method, we analyze it in several aspects: role of Retinex, role of the local maximal occurrence operation, influence of subspace dimensions, and the running time. The analysis was performed on the VIPeR database, by randomly sampling a training set of 316 persons, and a test set of the remaining persons.

5.5.1 Role of Retinex

We compared the proposed LOMO feature with and without the Retinex preprocessing, with results shown in Fig. 9 (a) and (b). This comparison was done by using the direct Cosine similarity measure and the XQDA algorithm, respectively. From Fig. 9 (a) we can see that, for direct matching, the performance can be obviously improved by applying the Retinex transform, with rank-1 accuracy being 12.97% without Retinex, and 20.25% with Retinex. This result indicates that Retinex helps to derive a consistent col-



(a) (b) (c)

Figure 9. CMC curves comparing the proposed feature with and without Retinex and the local maximal occurrence operation ((a) Cosine and (b) XQDA). (c) Rank-1 accuracy with varying subspace dimensions for the XQDA algorithm with LOMO feature.

or representation across different camera views, as can also be observed from Fig. 1 (b). However, From Fig. 9 (b) it can be seen that the two features are boosted by XQDA to a similar performance. This may indicate that XQDA is able to learn a robust metric against illumination variations.

5.5.2 Role of Local Maximal Occurrence

The person re-identification performance is largely affected by viewpoint changes, which should be addressed in feature design or classifier learning. The proposed local maximal occurrence feature extraction is a strategy towards pose or viewpoint robust feature representation. By comparing the proposed feature with and without the local maximal occurrence operation, we find that this operation does largely improve the performance of cross-view person re-identification, as shown in Fig. 9 (a) and (b). Without the local maximal occurrence operation, the rank-1 accuracy by applying the Cosine similarity measure (Fig. 9 (a)) is 11.39%, while applying this strategy, the rank-1 accuracy is improved to 20.25%. When further applying XQDA (Fig. 9 (b)), the performance gap is reduced, but the proposed feature still performs quite better with the local maximal occurrence operation than without it.

5.5.3 Subspace Dimensions

For the proposed XQDA algorithm, the dimension of the learned subspace has an influence in performance. This influence is shown in Fig. 9 (c), obtained by applying XQDA with different subspace dimensions on the VIPeR dataset. Roughly, the performance is increasing with increasing dimensions, but it becomes stable after 100 dimensions. Therefore, it is not too difficult to determine a proper number of subspace dimensionality. We use an automatic way as specified by accepting all eigenvalues of $\Sigma_I^{-1}\Sigma_E$ that are larger than 1, which works quite well in all the experiments. However, one can also select a small value considering the computational complexity. As can be observed from Fig. 9 (c), the rank-1 accuracy is consistently larger than 30% when the subspace dimensions are larger than 16.

Table 5. Training time (seconds) of metric learning algorithms.

	XQDA	KISSME	RLDA	ITML	LMNN
Time	1.86	1.34	1.53	36.78	265.28

5.5.4 Running Time

The training time comparison of metric learning algorithms is shown in Table 5 (including subspace learning time). The training time was averaged over 10 random trials on the VIPeR dataset. All algorithms are implemented in MATLAB. The LMNN algorithm has MEX functions implemented in C or C++ to accelerate the computation. The training was performed on a desktop PC with an Intel i5-2400 @3.10GHz CPU. Table 5 shows that the KISSME, RLDA, and XQDA algorithms, which have closed-form solutions, are very efficient, while ITML and LMNN, which require iterative optimizations, are time consuming.

Besides, we also evaluated the running time of the proposed feature extractor. In processing 128×48 person images, the LOMO feature extractor requires 0.012 seconds per image on average, which is very efficient. This code is also implemented in MATLAB, with a MEX function implemented for Retinex. Considering the effectiveness and efficiency of both the proposed LOMO feature and XQDA algorithm, we release both codes[2] for future research and benchmark on person re-identification.

6. Summary and Future Work

In this paper, we have presented an efficient and effective method for person re-identification. We have proposed an efficient descriptor called LOMO, which is shown to be robust against viewpoint changes and illumination variations. We have also proposed a subspace and metric learning approach called XQDA, which is formulated as a Generalized Rayleigh Quotient, and a closed-form solution can be obtained by the generalized eigenvalue decomposition. Practical computation issues for XQDA have been discussed, including the simplified computation, the regularization, and the dimension selection. Experiments on four challenging person re-identification databases, VIPeR, QMUL GRID, CUHK Campus, and CUHK03, show that the proposed method improves the state-of-the-art rank-1 identification rates by 2.2%, 4.88%, 28.91%, and 31.55% on the four databases, respectively. Due to the promising performance of the LOMO feature, it would be interesting to study other local features (e.g. Gabor, other color descriptors, etc.) or feature coding approaches with the same LOMO idea for person re-identification. It is also interesting to see the application of XQDA to other cross-view matching problems, such as the heterogeneous face recognition.

[2] http://www.cbsr.ia.ac.cn/users/scliao/projects/lomo_xqda/

Acknowledgments

This work was supported by the Chinese National Natural Science Foundation Projects #61203267, #61375037, #61473291, National Science and Technology Support Program #2013BAK02B01, Chinese Academy of Sciences Project No. KGZD-EW-102-2, and AuthenMetric R&D Funds.

References

[1] B. Alipanahi, M. Biggs, A. Ghodsi, et al. Distance metric learning vs. fisher discriminant analysis. In *International conference on Artificial intelligence*, 2008. 2

[2] L. Bazzani, M. Cristani, and V. Murino. Symmetry-driven accumulation of local features for human characterization and re-identification. *Computer Vision and Image Understanding*, 117(2):130–144, 2013. 1, 6, 7

[3] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, volume 2, page 6, 2011. 2

[4] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM, 2007. 1, 2, 5, 7

[5] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *Computer Vision–ACCV 2010*, pages 501–512. Springer, 2011. 1, 2

[6] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367, 2010. 2

[7] N. Gheissari, T. B. Sebastian, and R. Hartley. Person reidentification using spatiotemporal appearance. In *CVPR (2)*, pages 1528–1535, 2006. 2

[8] S. Gong, M. Cristani, S. Yan, and C. C. Loy. *Person Re-Identification*. Springer, 2014. 1

[9] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE International workshop on performance evaluation of tracking and surveillance*, 2007. 2, 5, 6

[10] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European Conference on Computer Vision*, 2008. 1, 2, 4, 5, 6

[11] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *International Conference on Computer Vision*, 2009. 2, 7

[12] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *ICDSC*, pages 1–6, 2008. 2

[13] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer, 2009. 2

[14] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *European Conference on Computer Vision*. 2012. 1, 2, 6

[15] Y. Hu, S. Liao, Z. Lei, D. Yi, and S. Z. Li. Exploring structural information and fusing multiple features for person re-identification. 2

[16] D. J. Jobson, Z.-U. Rahman, and G. A. Woodell. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *Image Processing, IEEE Transactions on*, 6(7):965–976, 1997. 2

[17] D. J. Jobson, Z.-U. Rahman, and G. A. Woodell. Properties and performance of a center/surround retinex. *Image Processing, IEEE Transactions on*, 6(3):451–462, 1997. 2

[18] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 1, 2, 3, 4, 5, 6, 7

[19] I. Kviatkovsky, A. Adam, and E. Rivlin. Color invariants for person reidentification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(7):1622–1634, 2013. 6

[20] E. H. Land and J. McCann. Lightness and retinex theory. *JOSA*, 61(1):1–11, 1971. 2

[21] W. Li and X. Wang. Locally aligned feature transforms across views. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 2

[22] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *Assian Conference on Computer Vision*, 2012. 7

[23] W. Li, R. Zhao, T. Xiao, and X. Wang. DeepReID: Deep filter pairing neural network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 7

[24] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 1, 2, 5, 6

[25] S. Liao, D. Yi, Z. Lei, R. Qin, and S. Z. Li. Heterogeneous face recognition from local structures of normalized appearance. In *International Conference on Biometrics*, 2009. 4

[26] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikäinen, and S. Z. Li. Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, June 2010. 3

[27] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: what features are important? In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 391–401. Springer, 2012. 3, 6

[28] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu. Semi-supervised coupled dictionary learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 6

[29] Y. Liu, Y. Shao, and F. Sun. Person re-identification based on visual saliency. In *Intelligent Systems Design and Applications (ISDA), 2012 12th International Conference on*, pages 884–889. IEEE, 2012. 2

[30] C. C. Loy, C. Liu, and S. Gong. Person re-identification by manifold ranking. In *IEEE International Conference on Image Processing*, volume 20, 2013. 6

[31] C. C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1988–1995. IEEE, 2009. 6

[32] B. Ma, Y. Su, and F. Jurie. Local descriptors encoded by fisher vectors for person re-identification. In *European Conference on Computer Vision Workshops*, 2012. 1, 2, 6

[33] B. Ma, Y. Su, and F. Jurie. Covariance descriptor based on bio-inspired features for person re-identification and face verification. *Image and Vision Computing*, 32(6):379–390, 2014. 1, 5, 6

[34] L. Ma, X. Yang, and D. Tao. Person re-identification over camera networks using multi-task distance metric learning. *IEEE Transactions on Image Processing*, 2014. 6, 7

[35] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 6

[36] B. Moghaddam, T. Jebara, and A. Pentland. Bayesian face recognition. *Pattern Recognition*, 33(11):1771–1782, 2000. 2, 3, 4

[37] T. Ojala, M. Pietikäinen, and D. Harwood. "A comparative study of texture measures with classification based on feature distributions". *Pattern Recognition*, 29(1):51–59, January 1996. 3

[38] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. Local fisher discriminant analysis for pedestrian re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 1, 2, 6

[39] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *BMVC*, 2010. 2, 3, 5, 6

[40] R. Vezzani, D. Baltieri, and R. Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys (CSUR)*, 46(2):29, 2013. 1

[41] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. H. Tu. Shape and appearance context modeling. In *ICCV*, pages 1–8, 2007. 2

[42] X. Wang, M. Yang, S. Zhu, and Y. Lin. Regionlets for generic object detection. In *IEEE International Conference on Computer Vision*, 2013. 2

[43] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. *Advances in neural information processing systems*, 2006. 1, 2, 5, 7

[44] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li. Salient color names for person re-identification. In *Proceedings of the European Conference on Computer Vision*, 2014. 5, 6

[45] J. Ye, T. Xiong, Q. Li, R. Janardan, J. Bi, V. Cherkassky, and C. Kambhamettu. "Efficient model selection for regularized linear discriminant analysis". In *Proceedings of the ACM Conference on Information and Knowledge Management*, pages 532–539, 2006. 5

[46] R. Zhao, W. Ouyang, and X. Wang. Person re-identification by salience matching. In *International Conference on Computer Vision*, 2013. 1, 2, 5, 6

[47] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3586–3593. IEEE, 2013. 2, 7

[48] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 1, 6, 7

[49] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 649–656. IEEE, 2011. 1, 2, 3, 5, 6

[50] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(3):653–668, 2013. 5, 6