A Novel Visual Word Co-occurrence Model for Person Re-identification

Ziming Zhang(✉), Yuting Chen, and Venkatesh Saligrama

Boston University, Boston, MA 02215, USA
{zzhang14,yutingch,srv}@bu.edu

Abstract. Person re-identification aims to maintain the identity of an individual in diverse locations through different non-overlapping camera views. The problem is fundamentally challenging due to appearance variations resulting from differing poses, illumination and configurations of camera views. To deal with these difficulties, we propose a novel visual word co-occurrence model. We first map each pixel of an image to a visual word using a codebook, which is learned in an unsupervised manner. The appearance transformation between camera views is encoded by a co-occurrence matrix of visual word joint distributions in probe and gallery images. Our appearance model naturally accounts for spatial similarities and variations caused by pose, illumination & configuration change across camera views. Linear SVMs are then trained as classifiers using these co-occurrence descriptors. On the VIPeR [1] and CUHK Campus [2] benchmark datasets, our method achieves 83.86% and 85.49% at rank-15 on the Cumulative Match Characteristic (CMC) curves, and beats the state-of-the-art results by 10.44% and 22.27%.

1 Introduction

In intelligent surveillance systems, *person re-identification* (*re-id*) is emerging as a key problem. *Re-id* deals with maintaining identities of individuals traversing different cameras. As in the literature we consider *re-id* for two cameras and focus on the problem of matching probe images of individuals in Camera 1 with gallery images from Camera 2. The problem is challenging for several reasons. Cameras views are non-overlapping so conventional tracking methods may fail. Illumination, view angles and configurations for different cameras are generally non-consistent, leading to significant appearance variations to the point that features seen in one camera are often distorted or missing in the other. Finer bio-metrics like face and gait thus often become unreliable [3].

The existing papers mainly focus on designing distinctive signature to represent a person under different cameras, or learning an effective matching methodology to predict if two images describe the same person. Our proposed method diverts from the literature by aiming to learn an appearance model that is based on *co-occurrence statistics* of visual patterns in different camera views. Namely, our appearance model captures the appearance "transformation" across cameras instead of some unknown invariant property among different views. Particularly,

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part III, LNCS 8927, pp. 122–133, 2015.
DOI: 10.1007/978-3-319-16199-0_9

Fig. 1. Illustration of codeword co-occurrence in positive image pairs (*i.e.* two images from different camera views per column belong to a *same* person) and negative image pairs (*i.e.* two images from different camera views per column belong to *different* persons). For positive (or negative) pairs, in each row the enclosed regions are assigned the same codeword.

our method does not assume any smooth appearance transformation across different cameras. Instead, our method learns the visual word co-occurrence pattens statistically in different camera views to predict the identities of persons.

While co-occurrence based statistics has been used in some other works [4] [5] [6], ours has a different purpose. We are largely motivated by the observation that the co-occurrence patterns of visual codewords behave similar for images from different views. In other words, the transformation of target appearances can be statistically inferred through these co-occurrence patterns. As seen in Fig. 1, we observe that some regions are distributed similarly in images from different views and robustly in the presence of large cross-view variations. These regions provide important discriminant co-occurrence patterns for matching image pairs. For instance, statistically speaking, the first column of positive image pairs shows that "white" color in Camera 1 can change to "light blue" in Camera 2. However, "light blue" in Camera 1 can hardly change to "black" in Camera 2, as shown in the first column of negative image pairs.

Thus we propose a novel visual word co-occurrence model to capture such important patterns between images. We first encode images with a sufficiently large codebook to account for different visual patterns. Pixels are then matched into codewords or visual words, and the resulting spatial distribution for each codeword is embedded to a kernel space through *kernel mean embedding* [7] with latent-variable conditional densities [8] as kernels. The fact that we incorporate the spatial distribution of codewords into appearance models provides us with locality sensitive co-occurrence measures. Our approach can be also interpreted as a means to *transfer* the information (*e.g.* pose, illumination, and appearance) in image pairs to a common latent space for meaningful comparison.

To conduct re-identification, we employ linear support vector machines (SVMs) as our classifier trained by the appearance descriptors. On the VIPeR [1] and CUHK Campus [2] benchmark datasets, our method achieves 83.86%

and 85.49% at rank-15 on the Cumulative Match Characteristic (CMC) curves, and beats the state-of-the-art results by 10.44% and 22.27%.

1.1 Related Work

The theme of local features for matching is related to our kernel-based similarity measures. To ensure locality, [9] models the appearances of individuals using features from horizontal strips. [10] clusters pixels into similar groups and the scores are matched based on correspondences of the clustered groups. Histogram features that encode both local and global appearance are proposed in [11]. Saliency matching [2,12], one of the-state-of-the-art methods for *re-id* uses patch-level matching to serve as masks in images to localize discriminative patches. More generally low-level features such as color, texture, interest points, co-variance matrices and their combinations have also been proposed [10,13–19]. In addition high-level structured features that utilize concatenation of low-level features [18] or deformable part models (DPMs) [20] have been proposed. Metric learning methods have been proposed for *re-id* (*e.g.* [21–24]). In [25,26] distance metrics are derived through brightness transfer functions that associate color-levels in the two cameras. [27] proposes distance metrics that lend importance to features in matched images over the wrongly matched pairs without assuming presence of universally distinctive features. Low-dimensional embeddings using PCA and local FDA have also been proposed [28]. Supervised methods that select relevant features for *re-id* have been proposed by [14] using Boosting and by [15] using RankSVMs.

2 Visual Word Co-occurrence Models

We generally face two issues in visual recognition problems: (1) *visual ambiguity* [29] (*i.e.* the appearance of instances which belong to the same thing semantically can vary dramatically in different scenarios), (2) *spatial displacement* [30] of visual patterns.

While visual ambiguity can be somewhat handled through codebook construction and quantization of images into visual words, our goal of matching humans in *re-id* imposes additional challenges. Humans body parts exhibit distinctive local visual patterns and these patterns systematically change appearance locally. Our goal is to account for this inherent variability in appearance models through co-occurrence matrices that quantify spatial and visual changes in appearance.

2.1 Locally Sensitive Co-occurrence Designs

We need co-occurrence models that not only account for the locality of appearance changes but also the random spatial & visual ambiguity inherent in vision problems. Therefore, we first construct a codebook $\mathcal{Z} = \{\mathbf{z}\} \subset \mathbb{R}^D$ with M codewords. Our codebook construction is global and thus only carries information

about distinctive visual patterns. Nevertheless, for a sufficiently large codebook distinctive visual patterns are mapped to different elements of the codebook, which has the effect of preserving local visual patterns. Specifically, we map each pixel at 2D location $\boldsymbol{\pi} \in \Pi$ of image $\mathcal{I}$ into (at least one) codewords to cluster pixels.

To emphasize local appearance changes, we look at the spatial distribution of each codeword. Concretely, we let $C(\mathcal{I}, \mathbf{z}) \subseteq \Pi$ denote the set of pixel locations associated with codeword $\mathbf{z}$ in image $\mathcal{I}$ and associate a spatial probability distribution, $p(\boldsymbol{\pi}|\mathbf{z}, \mathcal{I})$, over this observed collection. In this way visual words are embedded into a family of spatial distributions. Intuitively it should now be clear that we can use the similarity (or distance) of two corresponding spatial distributions to quantify the pairwise relationship between two visual words. This makes sense because our visual words are spatially locally distributed and small distance between spatial distributions implies spatial locality. Together this leads to a model that accounts for local appearance changes.

While we can quantify the similarity between two distributions in a number of ways, the kernel mean embedding method is particularly convenient for our task. The basic idea to map the distribution, p, into a reproducing kernel Hilbert space (RKHS), $\mathcal{H}$, namely, $p \rightarrow \mu_p(\cdot) = \sum K(\cdot, \boldsymbol{\pi}) p(\boldsymbol{\pi}) \stackrel{\Delta}{=} E_p(K(\cdot, \boldsymbol{\pi}))$. For universal kernels, such as RBF kernels, this mapping is injective, *i.e.*, the mapping preserves the information about the distribution [7]. In addition we can exploit the reproducing property to express inner products in terms of expected values, namely, $\langle \mu_p, \Phi \rangle = E_p(\Phi)$, $\forall \Phi \in \mathcal{H}$ and obtain simple expressions for similarity between two distributions (and hence two visual words) because $\mu_p(\cdot) \in \mathcal{H}$.

To this end, consider the codeword $\mathbf{z}_m$ in image $\mathcal{I}_i^{(1)}$ and codeword $\mathbf{z}_n$ in image $\mathcal{I}_j^{(2)}$. The co-occurrence matrix (and hence the appearance model) is the inner product of visual words in the RKHS space, namely,

$$\begin{aligned} \phi(\mathbf{x}_{ij})_{mn} &= \left\langle \mu_{p(\cdot|\mathbf{z}_m, \mathcal{I}_i^{(1)})}, \mu_{p(\cdot|\mathbf{z}_n, \mathcal{I}_j^{(2)})} \right\rangle \\ &= \sum_{\boldsymbol{\pi}_u} \sum_{\boldsymbol{\pi}_v} K(\boldsymbol{\pi}_u, \boldsymbol{\pi}_v) p(\boldsymbol{\pi}_u|\mathbf{z}_m, \mathcal{I}_i^{(1)}) p(\boldsymbol{\pi}_v|\mathbf{z}_n, \mathcal{I}_j^{(2)}), \end{aligned} \tag{1}$$

where we have used the reproducing property in the last equality. We now have several choices for the kernel $K(\boldsymbol{\pi}_u, \boldsymbol{\pi}_v)$ above. We list some of them here:

Identity: $K(\cdot, \boldsymbol{\pi}) = \mathbf{e}_{\boldsymbol{\pi}}$, where $\mathbf{e}_{\boldsymbol{\pi}}$ is the usual unit vector at location $\boldsymbol{\pi}$. We get the following appearance model:

$$\phi(\mathbf{x}_{ij})_{mn} \propto \left| C(\mathcal{I}_i^{(1)}, \mathbf{z}_m) \bigcap C(\mathcal{I}_j^{(2)}, \mathbf{z}_n) \right|, \tag{2}$$

where $|\cdot|$ denotes set cardinality. This choice often leads to poor performance in *re-id* because it is not robust to spatial displacements of visual words, which we commonly encounter in *re-id*.

Radial Appearance Model (RBF): This leads to the following appearance model:

$$\phi(\mathbf{x}_{ij})_{mn} = \sum_{\boldsymbol{\pi}_u}\sum_{\boldsymbol{\pi}_v} \exp\left(\frac{\|\boldsymbol{\pi}_u - \boldsymbol{\pi}_v\|_2^2}{2\sigma^2}\right) p(\boldsymbol{\pi}_u|\mathbf{z}_m, \mathcal{I}_i^{(1)}) p(\boldsymbol{\pi}_v|\mathbf{z}_n, \mathcal{I}_j^{(2)}) \tag{3}$$
$$\leq \sum_{\boldsymbol{\pi}_u} \max_{\boldsymbol{\pi}_v} \left\{ \exp\left(\frac{\|\boldsymbol{\pi}_u - \boldsymbol{\pi}_v\|_2^2}{2\sigma^2}\right) p(\boldsymbol{\pi}_v|\mathbf{z}_n, \mathcal{I}_j^{(2)}) \right\} p(\boldsymbol{\pi}_u|\mathbf{z}_m, \mathcal{I}_i^{(1)}).$$

The upper bound above is used for efficiently computing our appearance model by removing the summation over $\boldsymbol{\pi}_v$. This appearance model is often a better choice than the previous one because RBF accounts for some spatial displacements of visual words for appropriate choice of σ.

Latent Spatial Kernel: This is a type of probability product kernel that has been previously proposed [8] to encode generative structures into discriminative learning methods. In our context we can view the presence of a codeword $\mathbf{z}_m$ at location $\boldsymbol{\pi}_u$ as a noisy displacement of a true latent location $\mathbf{h} \in \Pi$. The key insight here is that the spatial activation of the two codewords $\mathbf{z}_m$ and $\mathbf{z}_n$ in the two image views $\mathcal{I}_i^{(1)}$ and $\mathcal{I}_j^{(2)}$ are conditionally independent when conditioned on the true latent location $\mathbf{h}$, namely, the joint probability factorizes into $Pr\{\boldsymbol{\pi}_u, \boldsymbol{\pi}_v \mid \mathbf{h}, \mathcal{I}_i^{(1)}, \mathcal{I}_j^{(2)}\} = Pr\{\boldsymbol{\pi}_u \mid \mathbf{h}, \mathcal{I}_i^{(1)}\} Pr\{\boldsymbol{\pi}_v \mid \mathbf{h}, \mathcal{I}_j^{(2)}\}$. We denote the noisy displacement likelihoods, $Pr\{\boldsymbol{\pi}_u \mid \mathbf{h}, \mathcal{I}_i^{(1)}\} = \kappa_1(\boldsymbol{\pi}_u, \mathbf{h})$ and $Pr\{\boldsymbol{\pi}_v \mid \mathbf{h}, \mathcal{I}_j^{(2)}\} = \kappa_2(\boldsymbol{\pi}_v, \mathbf{h})$ for simplicity. This leads us to $K(\boldsymbol{\pi}_u, \boldsymbol{\pi}_v) = \sum_{\mathbf{h}} \kappa_1(\boldsymbol{\pi}_u, \mathbf{h}) \kappa_2(\boldsymbol{\pi}_v, \mathbf{h}) p(\mathbf{h})$, where $p(\mathbf{h})$ denotes the spatial probability at $\mathbf{h}$, which we assume here to be uniform. By plugging this new K into Eq. 1, we have

$$\phi(\mathbf{x}_{ij})_{mn} = \sum_{\boldsymbol{\pi}_u}\sum_{\boldsymbol{\pi}_v}\sum_{\mathbf{h}} \kappa_1(\boldsymbol{\pi}_u, \mathbf{h}) \kappa_2(\boldsymbol{\pi}_v, \mathbf{h}) p(\mathbf{h}) p(\boldsymbol{\pi}_u|\mathbf{z}_m, \mathcal{I}_i^{(1)}) p(\boldsymbol{\pi}_v|\mathbf{z}_n, \mathcal{I}_j^{(2)})$$
$$\leq \sum_{\mathbf{h}} \max_{\boldsymbol{\pi}_u} \left\{ \kappa_1(\boldsymbol{\pi}_u, \mathbf{h}) p(\boldsymbol{\pi}_u|\mathbf{z}_m, \mathcal{I}_i^{(1)}) \right\} \max_{\boldsymbol{\pi}_v} \left\{ \kappa_2(\boldsymbol{\pi}_v, \mathbf{h}) p(\boldsymbol{\pi}_v|\mathbf{z}_n, \mathcal{I}_j^{(2)}) \right\} p(\mathbf{h}), \tag{4}$$

where the inequality follows by rearranging the summations and standard upper bounding techniques. Again we use an upper bound for computational efficiency, and assume that $\mathcal{P}_{\mathcal{H}}$ is a uniform distribution for simplicity without further learning. The main idea here is that by introducing the latent displacement variables, we have a handle on view-specific distortions observed in the two cameras. We only show the performance using the latent kernel in our experimental section, since it produces much better performance than the other two in our preliminary results.

2.2 Implementation of Latent Spatial Kernels

Fig. 2 illustrates the whole process of generating the latent spatial kernel based appearance model given the codeword images, each of which is represented as

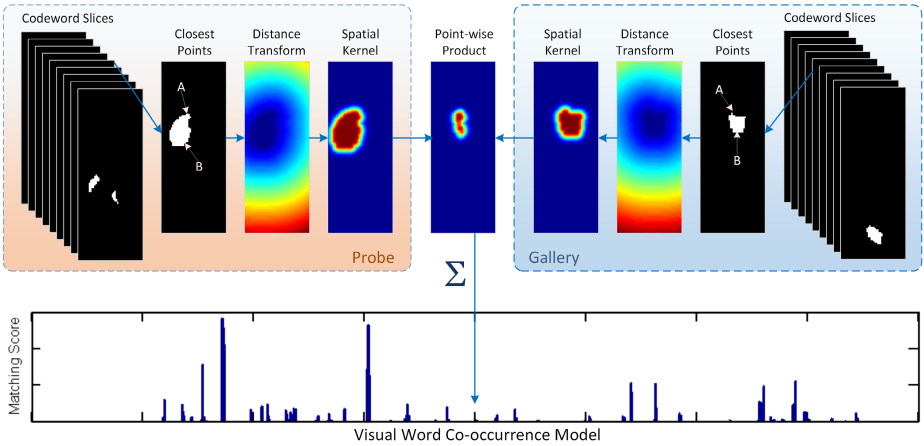


Fig. 2. Illustration of our visual word co-occurrence model generation process. Here, the white regions in the codeword slices indicate the pixel locations with the same codeword. "A" and "B" denote two arbitrary pixel locations in the image domain. And "Σ" denotes a sum operation which sums up all the values in the point-wise product matrix into a single value $\phi(\mathbf{x}_{ij})_{mn}$ in our model.

a collection of codeword slices. For each codeword slice, the max operation is performed at every pixel location to search for the spatially closest codeword in the slice. This procedure forms a distance transform image, which is further mapped to a spatial kernel image. It allows each peak at the presence of a codeword to be propagated smoothly and uniformly. To calculate the matching score for a codeword co-occurrence, the spatial kernel from a probe image and another from a gallery image are multiplied element-wise and then summed over all latent locations. This step guarantees that our descriptor is insensitive to the noise data in the codeword images. This value is a single entry at the bin indexing the codeword co-occurrence in our descriptor for matching the probe and gallery images. As a result, we have generated a high dimensional sparse appearance descriptor.

3 Experiments

We test our method on two benchmark datasets, VIPeR [1] and CUHK Campus [2]. For each dataset, images from separate camera views are split into a gallery set and a probe set. Images from the probe set are treated as queries and compared with every person in the gallery set. For each query, our method produces a ranking of matching individuals in the gallery set. Performance can be evaluated with these resultant rankings, since the identity label of each image is known. The rankings for every possible query is combined into a Cumulative Match Characteristic (CMC) curve, which is a standard metric for re-identification performance. The CMC curve displays an algorithm's recognition rate as a function of rank. For instance, a recognition rate at rank-r on the CMC curve denotes

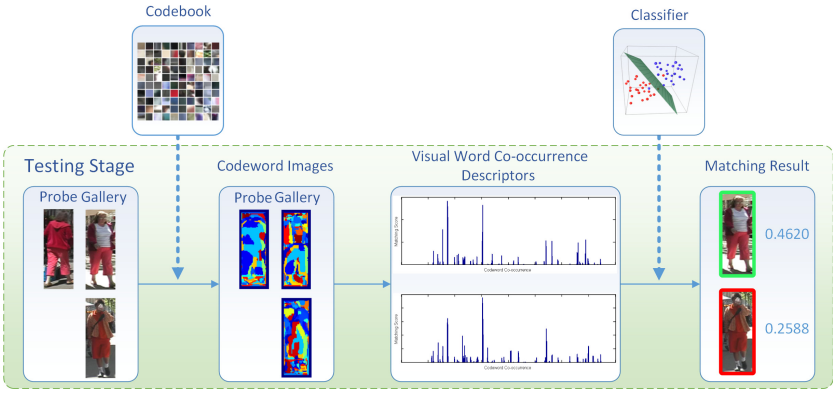


Fig. 3. The pipeline of our method, where "codebook" and "classifier" are learned using training data, and each color in the codeword images denotes a codeword. This figure is best viewed in color.

what proportion of queries were correctly matched to a corresponding gallery individual at rank-r or better. Experimental results are reported as the average CMC curve over 3 trials.

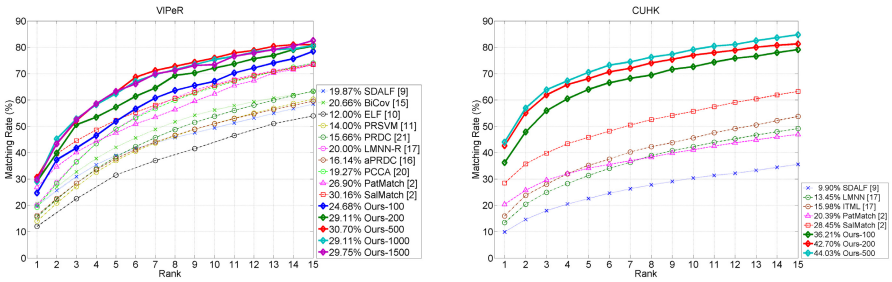
3.1 Implementation

We illustrate the schematics of our method in Fig. 3. At training stage, we extract low-level feature vectors from randomly sampled patches in training images, and then cluster them into codewords to form a codebook, which is used to encode every image into a codeword image. Each pixel in a codeword image represents the centroid of a patch that has been mapped to a codeword. Further, a visual word co-occurrence model (descriptor) is calculated for every pair of gallery and probe images, and the descriptors from training data are utilized to train our classifier, performing re-identification on the test data.

Specifically, for each image a 672-dim ColorSIFT [2][1] feature vector is extracted for a 10×10 pixel patch centered at every possible pixel. Further, we decorrelate each feature using the statistics learned from training data, as suggested in [31].

For codebook construction, we randomly sample 1000 patch features per image in the training set, and cluster these features into a codebook using K-Means. Then we encode each patch feature in images from the probe and gallery sets into a codeword whose Euclidean distance to the patch feature is the minimum among all the codewords. As a result, each image is mapped into a codeword image whose pixels are represented by the indices of the corresponding encoded codewords. We also normalize our appearance descriptors using min-max normalization. The min value is for our descriptors is always 0, and the max value is the maximum among all the codeword co-occurrence bins over every training descriptor. This max value is saved during training and utilized for normalization during testing.

[1] The authors' code can be downloaded at http://www.ee.cuhk.edu.hk/~rzhao/.



(a) VIPeR (b) CUHK Campus

Fig. 4. Matching rate comparison between different methods on (a) VIPeR and (b) CUHK Campus datasets. Numbers following "Ours-" in the legends denote the size of the codebook used in each experiment. Expect for our results, the other CMC curves are cited from [2]. This figure is best viewed in color.

In the end for classifiers, we employ LIBLINEAR [32], an efficient linear SVMs solver, with the ℓ_2 norm regularizer. The trade-off parameter c in LIBLINEAR is set using cross-validation.

3.2 VIPeR

Since introduced in [33], the VIPeR dataset has been utilized by most person re-identification approaches as a benchmark. VIPeR is comprised of 632 different pedestrians captured in two different camera views, denoted by CAM-A and CAM-B, respectively. Many cross-camera image pairs in the dataset have significant variations in illumination, pose, and viewpoint, and each image is normalized to 128×48 pixels.

In order to compare with other person re-identification methods, we followed the experimental set up described in [2]. The dataset is split in half randomly, one partition for training and the other for testing. In addition, samples from CAM-A form the probe set, and samples from CAM-B form the gallery set. The parameter σ in the spatial kernel function is set to 3 for this dataset.

Fig. 4(a) shows our matching rate comparison with other methods on this dataset. When the codebook size is 100, which is pretty small, our performance is close to that of SalMatch [2]. With increase of the codebook size, our performance is improved significantly, and has outperformed that of SalMatch by large margins. For instance, at rank-15, our best matching rate is 10.44% higher. Using larger sizes of codebooks, the codeword representation of each image is finer by reducing the quantization error in the feature space. However, it seems that when the codebook size is beyond 500, our performance is saturated. Therefore, in the following experiments, we only test our method using 100/200/500 codewords.

Fig. 5. Examples of codeword co-occurrence with relatively high positive/negative weights in the learned weighting matrix. Same as Fig. 1, in each row the regions enclosed by red (or cyan) color indicate that the codeword per pixel location in these regions is the same. This figure is best viewed in color.

Fig. 5 illustrates some codeword co-occurrence examples with relatively high positive/negative weights in the learned weighting matrix. These examples strongly support our intuition of learning codeword co-occurrence based features in Section 1.

3.3 CUHK Campus

The CUHK Campus dataset is a relatively new person re-identification dataset explored by two state-of-the-art approaches outlined in [2] and [34]. This dataset consists of 1816 people captured from five different camera pairs, labeled P1 to P5. Each image contains 160×60 pixels. Following the experimental settings from [2] and [34], we use only images captured from P1 as our dataset. This subset contains 971 people in two camera views, with two images per view per person. One camera view, which we call CAM-1, captures people either facing towards or away from the camera. The other view, CAM-2, captures the side view of each person.

For our experiments, we adopt the settings described in [2] for comparison[2]. We randomly select 485 individuals from the dataset and use their 4 images for training, and the rest are used for testing. The gallery and probe sets are formed by CAM-1 and CAM-2, respectively. To re-identify a person, we compare the probe image with every gallery image, leading to 486×2=972 decision scores. Then per person in the gallery set, we average the 2 decision scores belonging to this person as the final score for ranking later. The parameter σ in the spatial kernel function is set to 6 for this dataset, since the image size is larger.

[2] We thank the authors for the response to their experimental settings.

Fig. 4(b) summarizes our matching rate comparison with some other methods. Clearly, using only 100 codewords, our method has already outperformed others dramatically, and it works better when using larger sizes of codebooks, similar to the behavior in Fig. 4(a). At rank-15, our best performance is 22.27% better than that of SalMatch.

4 Conclusion

In this paper, we propose a novel visual word co-occurrence model for person re-identification. The intuition behind our model is that the codeword co-occurrence patterns behave similarly and consistently in pairs of gallery/probe images and robustly to the changes in images. To generate our descriptor, each image is mapped to a codeword image, and the spatial distribution for each codeword is embedded to a kernel space through *kernel mean embedding* with latent spatial kernels. To conduct re-identification, we employ linear SVMs as our classifier trained by the descriptors. We test our method on two benchmark datasets, VIPeR and CUHK Campus. On both datasets, our method consistently outperforms other methods. At rank-15, our method achieves matching rates of 83.86% and 85.49%, respectively, which are significantly better than the state-of-the-art results by 10.44% and 22.27%.

Several questions will be considered as our future work. It would be useful to reduce the computational complexity of calculating our pair-wise latent spatial kernels. One possibility is to modify the learning algorithm by decomposing the weight matrix into two separable parameters, because our appearance model can be decomposed into two parts, one from the probe image and the other from the gallery image. Such decomposition will accelerate the computation. Second, in our preliminary experiments, latent spatial kernel yields significantly better results over the other two choices. It would be interesting to explore other selection of kernels (or even learn the optimal kernels) and how they affect the behavior of our visual word co-occurrence model. Building a *re-id* system for natural images using object proposal algorithms (*e.g.* [35,36]) and our model with different classifiers (*e.g.* [37–39]) would be interesting as well.

Acknowledgments. This work is supported by the U.S. DHS Grant 2013-ST-061-ED0001 and ONR award N00014-13-C-0288 respectively. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the agencies.

References

1. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: 10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS) (September 2007)
2. Zhao, R., Ouyang, W., Wang, X.: Person re-identification by salience matching. In: ICCV (2013)

3. Vezzani, R., Baltieri, D., Cucchiara, R.: People reidentification in surveillance and forensics: A survey. ACM Comput. Surv. 46(2), 29:1–29:37 (2013)
4. Banerjee, P., Nevatia, R.: Learning neighborhood cooccurrence statistics of sparse features for human activity recognition. In: AVSS, pp. 212–217 (2011)
5. Galleguillos, C., Rabinovich, A., Belongie, S.: Object categorization using co-occurrence, location and appearance. In: CVPR (June 2008)
6. Ladicky, L., Russell, C., Kohli, P., Torr, P.H.S.: Graph Cut Based Inference with Co-occurrence Statistics. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 239–253. Springer, Heidelberg (2010)
7. Smola, A.J., Gretton, A., Song, L., Schölkopf, B.: A Hilbert Space Embedding for Distributions. In: Hutter, M., Servedio, R.A., Takimoto, E. (eds.) ALT 2007. LNCS (LNAI), vol. 4754, pp. 13–31. Springer, Heidelberg (2007)
8. Jebara, T., Kondor, R., Howard, A.: Probability product kernels. JMLR **5**, 819–844 (2004)
9. Bird, N.D., Masoud, O., Papanikolopoulos, N.P., Isaacs, A.: Detection of loitering individuals in public transportation areas. Trans. Intell. Transport. Sys. **6**(2), 167–177 (2005)
10. Gheissari, N., Sebastian, T.B., Hartley, R.: Person reidentification using spatiotemporal appearance. CVPR **2**, 1528–1535 (2006)
11. Bazzani, L., Cristani, M., Perina, A., Murino, V.: Multiple-shot person re-identification by chromatic and epitomic analyses. Pattern Recogn. Lett. **33**(7), 898–903 (2012)
12. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: CVPR, pp. 3586–3593 (2013)
13. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: CVPR, pp. 2360–2367 (2010)
14. Gray, D., Tao, H.: Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 262–275. Springer, Heidelberg (2008)
15. Prosser, B., Zheng, W.S., Gong, S., Xiang, T., Mary, Q.: Person re-identification by support vector ranking. In: BMVC, vol. 1, p. 5 (2010)
16. Bauml, M., Stiefelhagen, R.: Evaluation of local features for person re-identification in image sequences. In: AVSS, pp. 291–296 (2011)
17. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Multiple-shot human re-identification by mean riemannian covariance grid. In: AVSS, pp. 179–184 (2011)
18. Ma, B., Su, Y., Jurie, F.: Bicov: a novel image representation for person re-identification and face verification. In: BMVC (2012)
19. Liu, C., Gong, S., Loy, C.C., Lin, X.: Person Re-identification: What Features Are Important? In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012 Ws/Demos, Part I. LNCS, vol. 7583, pp. 391–401. Springer, Heidelberg (2012)
20. Nguyen, V.-H., Nguyen, K., Le, D.-D., Duong, D.A., Satoh, S.: Person Re-identification Using Deformable Part Models. In: Lee, M., Hirose, A., Hou, Z.-G., Kil, R.M. (eds.) ICONIP 2013, Part III. LNCS, vol. 8228, pp. 616–623. Springer, Heidelberg (2013)
21. Dikmen, M., Akbas, E., Huang, T.S., Ahuja, N.: Pedestrian Recognition with a Learned Metric. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part IV. LNCS, vol. 6495, pp. 501–512. Springer, Heidelberg (2011)
22. Li, W., Zhao, R., Wang, X.: Human Reidentification with Transferred Metric Learning. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part I. LNCS, vol. 7724, pp. 31–44. Springer, Heidelberg (2013)

23. Mignon, A., Jurie, F.: PCCA: a new approach for distance learning from sparse pairwise constraints. In: CVPR, pp. 2666–2672 (2012)
24. Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by probabilistic relative distance comparison. In: CVPR, pp. 649–656 (2011)
25. Porikli, F.: Inter-camera color calibration by correlation model function. In: ICIP, vol 2. pp. II-133 (2003)
26. Javed, O., Shafique, K., Rasheed, Z., Shah, M.: Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. Comput. Vis. Image Underst. **109**(2), 146–162 (2008)
27. Zheng, W.S., Gong, S., Xiang, T.: Re-identification by relative distance comparison. IEEE TPAMI **35**(3), 653–668 (2013)
28. Pedagadi, S., Orwell, J., Velastin, S., Boghossian, B.: Local fisher discriminant analysis for pedestrian re-identification. In: CVPR, pp. 3318–3325 (2013)
29. van Gemert, J., Veenman, C.J., Smeulders, A.W.M., Geusebroek, J.M.: Visual word ambiguity. IEEE Trans. Pattern Anal. Mach. Intell. **32**(7), 1271–1283 (2010)
30. Felzenszwalb, P.F., Girshick, R.B., McAllester, D.A., Ramanan, D.: Object detection with discriminatively trained part-based models. TPAMI **32**(9), 1627–1645 (2010)
31. Hariharan, B., Malik, J., Ramanan, D.: Discriminative Decorrelation for Clustering and Classification. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 459–472. Springer, Heidelberg (2012)
32. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. JMLR **9**, 1871–1874 (2008)
33. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: PETS, pp. 47–47 (2007)
34. Li, W., Wang, X.: Locally aligned feature transforms across views. In: CVPR, pp. 3594–3601 (June 2013)
35. Zhang, Z., Warrell, J., Torr, P.H.S.: Proposal generation for object detection using cascaded ranking svms. In: IEEE CVPR, pp. 1497–1504 (2011)
36. Cheng, M.M., Zhang, Z., Lin, W.Y., Torr, P.H.S.: Bing: Binarized normed gradients for objectness estimation at 300fps. In: IEEE CVPR (2014)
37. Zhang, Z., Li, Z.N., Drew, M.S.: Adamkl: A novel biconvex multiple kernel learning approach. In: IEEE 2010 20th International Conference on Pattern Recognition (ICPR), pp. 2126–2129 (2010)
38. Zhang, Z., Sturgess, P., Sengupta, S., Crook, N., Torr, P.H.: Efficient discriminative learning of parametric nearest neighbor classifiers. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2232–2239. IEEE (2012)
39. Zhang, Z., Ladicky, L., Torr, P., Saffari, A.: Learning anchor planes for classification. In: Advances in Neural Information Processing Systems, pp. 1611–1619 (2011)