Multiregion Bilinear Convolutional Neural Networks for Person Re-Identification

Evgeniya Ustinova,[1,][*] Yaroslav Ganin,[1,][†] and Victor Lempitsky[1,][‡]

[1]*Skolkovo Institute of Science and Technology*
Skolkovo, Moscow, Russia

In this work we explore the applicability of the recently proposed CNN architecture, called Bilinear CNN, and its new modification that we call multi-region Bilinear CNN to the person re-identification problem. Originally, Bilinear CNNs were introduced for fine-grained classification and proved to be both simple and high-performing architectures. Bilinear CNN allows to build an orderless descriptor for an image using outer product of features outputted from two separate feature extractors. Based on this approach, Multiregion Bilinear CNN, apply bilinear pooling over multiple regions for extracting rich and useful descriptors that retain some spatial information.

We show than when embedded into a standard "siamese" type learning, bilinear CNNs and in particular their multi-region variants can improve re-identification performance compared to standard CNNs and achieve state-of-the-art accuracy on the largest person re-identification datasets available at the moment, namely CUHK03 and Market-1501.

I. INTRODUCTION

The task of person re-identification is to match pedestrian images coming from different and potentially non-overlapping camera views. This problem draws increasing attention of researchers and practitioners as it is one of the critical points in implementing automated surveillance systems. The ability to match people across cameras can be used among other things to track people over large spaces in safety/forensic applications or to collect statistics of typical motion trajectories in a building or an open area for planning purposes.

Despite a long history of research on re-identification [1–14], the accuracy of the existing systems is often insufficient for the full automation of such application scenarios, which stimulates further research activity. The main confounding factors stem from the notoriously high variation of the appearance of the same person (even at short time spans) due to pose variations, camera viewpoint variations, illumination variation, background clutter, etc.

The key task within person re-identification is to measure similarity for pairs of pedestrian images in such a way that a pair would get high similarity score in case of depicting the same person and low score in case of depicting different persons. This requires to construct a robust representation and an appropriate similarity measure which would facilitate assigning corresponding similarity scores. Traditionally, features/representation design and fixed distance/similarity function design have been considered separately as two steps in the pipeline.

The approach we pursue in this work is motivated by recent success of deep neural networks [16] that unify feature design and similarity learning, and have brought breakthroughs across the range of different recognition tasks. Person re-identification problem, however, is almost unique among pattern recognition tasks in that the proposed deep learning approaches [1, 12, 13], while competitive, do not clearly outperform more traditional approaches based on "hand-engineered" features [9, 17]. This is partly because of the lack of very big training sets for deep re-identification. Here, we present a deep learning approach that for the two of the biggest re-identification datasets (*CUHK03* [12] and *Market-1501* [15]) achieves state-of-the-art performance,
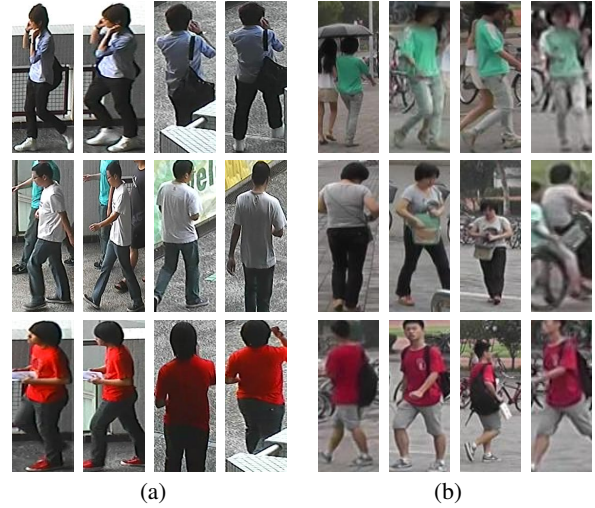

FIG. 1: (a) - sample images of the CUHK03 dataset [12]. Each person is captured by two cameras. There are six camera views in total. Each row corresponds to a person. First two images in a row are samples of first camera view for a person, 3th and 4th images are from the second camera view for a person. All the images in this paper are better viewed in color. (b) - sample images of the Market-1501 dataset [15]. Each row corresponds to a person, different camera views are shown for each identity. Each person is captured by at most six cameras.

even though the size of these datasets is still modest compared to datasets typical for other recognition tasks. We anticipate that further increase of training set sizes is likely to favour deep learning methods including ours.

Our approach is derived from the bilinear convolutional neural network [18] that was originally presented for fine-grained classification tasks and later evaluated for face recognition [19]. We note that the task of person re-identification shares considerable similarity with fine-grained categorization, as the matching process in both cases often needs to resort to the analysis of fine texture details and parts that are hard to localize. At the same time, person re-identification has some peculiarities, which are exploited in our method and bring improvement over the pure bilinear CNN approach.

Below, we first discuss the related work in person re-identification and deep learning, then gradually introduce the parts of our method, and conclude with the in-depth experimental validation.

* evgeniya.ustinova@skolkovotech.ru
† ganin@skolkovotech.ru
‡ lempitsky@skoltech.ru

II. RELATED WORK

As mentioned above, works that are not using deep learning for person re-identification pursue one of the two main directions: 1) improving feature design, or 2) metric/distance learning or pair ranking function design. Systems in the literature use various combinations of features and distance/metric/ranking learning approaches.

Among the methods aiming for better representation for pedestrian images, Ma *et al.*[2] utilise Gabor filters and Covariance descriptors to handle illumination changes and background variations, while Bazzani *et al.*[3] build a descriptor called Symmetry-Driven Accumulation of Local Features (SDALF). Exploiting person symmetry, this method helps to extract information from actual body parts on the image, avoiding background clutter. Li *et al.*[4] use SIFT and color histograms for further learning cross-view dictionaries for patch-level and image-level descriptors. Similarity scores for patches and images are then fused with chosen coefficients.

Metric learning and ranking learning methods include method introduced in [5], which solves re-identification as a ranking problem using set of RankSVMs to learn similarity parameters; Kuo *et al.*[6] use various kind of features, defining similarity measure for each of them and learns coefficients for fusing such similarities via RankBoost algorithm. In [7], the evaluation of different metric learning approaches to person re-identification is presented, in which the authors show KISSME [20] and EIML [8] to be top-performing metric learning methods. Based on a number of common visual features, Paisitkriangkrai *et al.*[9] build an ensemble of distance functions, using their weighted sum as the final distance measure. Authors formulate re-identification task as ranking problem and utilize cutting plane approach to learn those weighs. State-of-the-art results on the majority of re-identification datasets are reported in [9].

Methods combining feature learning and metric learning should also be mentioned. Ma *et al.*[10] use Fisher Vector descriptors learned for simple intensity and first- and second- order local features. Interestingly, they report that saving spatial information by constructing Fisher Vectors for patches instead of the whole image improves performance, which is the approach we also pursue here. PCCA [21] is used to learn distance in a supervised manner. Liao *et al.*[11] propose the method for building descriptor that is invariant to illumination and view point changes. The authors utilize Retinex algorithm [22] in order to achieve consistency in illumination for images from different cameras. Then, local feature occurrences for horizontal stripes are analyzed. To preserve spatial information, the occurrences are calculated withing sliding window and then are pooled for entire stripe. The authors also propose an algorithm called XQDA to learn discriminant subspace jointly with distance function in the learned subspace.

Several CNN-based methods for person re-identification have been proposed recently. One of the main advantages of using CNNs for re-identification problem is the ability to combine separate steps of traditional re-identification pipeline into end-to-end learning procedure, enabling automatic interactions between feature extraction, feature transforms and similarity estimation parts of the model [12]. Yi *et al.*[1] introduce "siamese" architecture consisting of two subnetworks, connected by the similarity layer (cosine similarity is used). Two images are processed independently in the network, and only their final representations are connected in the similarity calculation step. The authors consider the performance of their method on comparatively small datasets such as VIPER and PRID-2011. We use the same architecture as a starting point and investigate its effectiveness on newer and larger datasets CUHK03 and Market-1501.

Differently to [1], [12] and [13] learn classification networks that can categorize a pair of images as either depicting the same subjects or different subjects. In both approaches, the two images are fed into the network, and the difference in their middle representation is processed within additional special layers (Patch matching in [12] and cross-input neighbourhood difference in [13]). After that there are some processing steps on already combined middle representation of the pair of images and classification for two classes (same subject or different subjects) as the last layer.

Chen *et al.*[14] introduce deep ranking framework consisting of a CNN that learns similarity from pairs of stitched images and a ranking part that enhances the re-identification quality, notably on small datasets. It is worth mentioning, that [13] and [14] demonstrate high results on datasets of different sizes, even moderate and very small ones (CUHK01 [23] and VIPeR [24]).

When searching for matches in a dataset, the methods proposed in [12], [13] and [14] need to process pairs that include the query and every image in the dataset, and hence cannot directly utilize fast retrieval methods based on Euclidean and other simple distances. Here we aim at the approach that can learn per-image descriptors and then compare them with simple distances. This justifies starting with the architecture proposed in [1] and then modifying it by inserting new layers.

III. THE APPROACH

The difficulty of person re-identification problem lies in several factors of person appearance variation, including lighting changes, different poses, different view points. Many challenges of person re-identification are shared with the fine-grained recognition task: one person can look very differently when captured by different cameras, and different persons can look similar if they wear similar clothes and can only be distinguishable based on small outfit element. Likewise, the appearance of a bird species can vary greatly from shot to shot and different related bird species from the same family can have similar appearance and only differ in an appearance of a small body part. Therefore our solution combines the state-of-the-art method for person re-identification (deep metric learning) and the state-of-the-art fine-grained recognition method (bilinear CNN).

We now give an in-depth review of both components, and then discuss how our architecture combines them. Finally, we motivate and discuss a further improvement over the straightforward combination that helps us to boost the re-identification performance.

A. Deep Metric Learning

Deep Metric learning (DML) proposed by [1] incorporates "siamese" neural net, initially introduced for signature verification [25] and later for face recognition in [26]. Siamese architecture consists of two similar sub-networks for separate pro-

cessing of two input images. This sub-networks are responsible for calculating descriptors for a pair of images which are then connected via distance function.

In [1] special architecture specific to pedestrian images is proposed. Each of the two "siamese halves" includes three independent sub-networks, in which three overlapping parts of person images are processed separately, namely, the top part (head, upper torso), the middle part (torso), and the lower part (legs, lower torso). This is done in order to take into account different statistics of textures, shapes, and pose changing in different parts of the image, while still utilizing convolutional layers with spatially-invariant kernels. Each of the two sub-networks produces 500-dimensional descriptor vector as an output, and for each training pair of images the cosine similarity of these descriptors is calculated. At training time, the similarity value with the pair label (+1 for the same identity, -1 - for different identities) is then fed into the *Binomial Deviance* loss function [1].

As in [1], we consider the view-invariant case with weights shared for the two "siamese" sub-networks. This case is more general as the trained networks can be used for new out-of-sample cameras, unlike the view-specific case that lacks weights sharing (which makes networks to capture view-specific appearance peculiarities).

As in other deep learning re-identification methods, DML has an advantage of jointly learning representation and similarity function, which seems to be prospective approach, considering difficulty of the problem. Unlike other approaches incorporating CNNs ([12], [13] and [14]), DML uses a simple cosine similarity (equivalent to Euclidean distance between normalized descriptors) for the obtained image descriptors. Thus, photometric and geometric transforms specific for a pair of images are not modeled or infered. On the other hand, as discussed above, descriptors produced with DML are suitable for large-scale systems that need to use simple metrics.

B. Bilinear CNNs

Bilinear convolutional networks (B-CNN), introduced in [18] achieved state-of-the-art results for a number of fine-grained recognition tasks, and has also showed potential for face verification [19]. B-CNN consists of two CNNs (where the input of these two CNNs is the same image) without fully-connected layers, and the outputs of the two CNNs are combined in a special way via bilinear pooling[1]. In more detail, the outer product of deep features are calculated for each spatial location, resulting in the quadratic number of feature maps, to which sum pooling over all locations is then performed. The resulting orderless image descriptor is then used in subsequent processing steps, for example, in [18] and [19] it is normalized and fed into the softmax layer for classification.

An intuition, given in [18], is that the two CNNs combined by bilinear operation may correspond to part and texture detectors and therefore facilitate localization when significant pose variation is present. The great advantage of Bilinear CNN is that it

---

[1] The two sub-networks within the B-CNN are not to be confused with the two sub-networks within the Siamese network. Below, we mostly refer to the B-CNN sub-networks when talking about "halves" of the network or "subnetwork".

does not need any part labeling for learning. As the Bilinear architecture represents a directed acyclic computation graph, both feature extractors can be learned by back-propagating gradients of loss function specific to the task.

As mentioned above, bilinear CNNs were successfully applied to different fine-grained classification tasks (bird species, car model, aircraft variant, faces), outperforming most of existing methods, including more complicated ones, which additionally use explicit part annotations for training. Particularly large performance gap between ordinary CNN and B-CNN were demonstrated for the bird species dataset without tight bounding boxes, while the gap for similar architectures, trained on the dataset with bounding was smaller. This indirectly confirms the part localization abilities of the B-CNN architecture.

In [19] face identification is viewed as fine-grained classification and is also approached by B-CNN method. Significant performance boost of B-CNN architecture was demonstrated in comparison to standard CNN for very challenging face verification dataset IJB-A [27], capturing faces in different positions and lighting conditions. Results shown is this paper [19] are especially interesting for us as the task of face verification is also related to retrieval and has to deal with human appearance (although at different scale than pedestrian re-identification).

As to person re-identification in the context of bilinear models, Fisher Vector has been previously used for it (along with metric learning) in [10]. It is shown in [18], that Fisher Vector and can also be represented in a form of outer product (between features and weights of soft assignments to cluster centers of GMM), followed by pooling across all the locations.

C. Combined architecture

We now discuss the combination of DML and B-CNN. As in DML [1], we use the siamese architecture, where two pedestrian images are fed into identical neural networks. The two output vectors can then be regarded as the descriptors of the two input images. Similarity and cost function value are then calculated for obtained descriptors. Similar to [1], we use cosine similarity function and the Binomial Deviance loss function:

$$J_{dev} = \sum_{i,j} W \circ \ln(\exp^{-\alpha(S-\beta)\circ M} + 1) \qquad (3.1)$$

where $\circ$ is elementwise multiplication, i and j are the numbers of the training images, and $S = [S_{i,j}]_{n \times n}$ is the similarity matrix for image pairs (i.e. $S_{i,j} = cosine(x_i, x_j)$).

Furthermore, $M = [M_{i,j}]_{n \times n}$ is the learning supervision defined as: $M_{i,j} = \begin{cases} 1, \text{positive pair} \\ -1, \text{negative pair} \end{cases}$

Finally, the weighting matrix $W = [W_{i,j}]_{n \times n}$ is defined as: $W_{i,j} = \begin{cases} \frac{1}{n_1}, \text{positive pair} \\ \frac{1}{n_2}, \text{negative pair} \end{cases}$ where α and β are hyperparameters.

Following the original architecture [1], we use the two convolution layers with 64 output feature maps (filter size of the first one is 7×7, and of the second one is 5×5), followed by the rectified linear (ReLU) non-linearity and max pooling with the kernel size of two pixels and the stride of two pixels. The overall structure of the siamese neural net is shown in figure 2a. Also following [1], we use separate sub-networks for three parts of pedestrian images (first convolution layers have weights shared

for three sub-networks) to take into account different semantic nature of the three parts.

We investigate the following modification of the original architecture: instead of using one sub-network for image parts, we use two feature extractors with the same architecture described above. The outputs are combined by the bilinear operation [18] after the second convolution. Three bilinear outputs for each of the image parts are then concatenated and turned into 500-dimensional image descriptor by an extra fully connected layer. The overall scheme of Bilinear DML net for each of the two siamese sub-networks used is this work is shown in figure 2b.

D. Multiregion Bilinear Model for person re-identification

Bilinear CNNs are motivated by the specialized pooling operation that aggregates the correlations across maps coming from different halves of the network. The aggregation however discards all spatial information that remains in the network prior to the application of the operation. This is justified when the images lack even loose alignment (as e.g. in the case of the birds dataset), however is sub-optimal in our case, where person detector ensures that there is some loose geometric alignment between images. Therefore we modify bilinear layer and replace it with *multiregion bilinear layer*, which allows us to retain some of the geometric information. Our modification is, of course, similar to many other approaches in computer vision, notably to spatial pyramids of [28].

Similar to [18], we introduce bilinear model for image similarity as follows: $\mathcal{B} = (f_A, f_B, \mathcal{P}, \mathcal{S})$, where f_A and f_B are feature extractor functions, $\mathcal{P}$ - pooling function, $\mathcal{S}$ - similarity function. This feature function takes an image $\mathcal{I}$ at location $\mathcal{L}$ and outputs the feature of determined size $\mathcal{D}$ (unlike [18], we use vector notation for features for simplicity): $f : \mathcal{I} \times \mathcal{L} \to \mathcal{R}^{1 \times \mathcal{D}}$.

For each of the two images in the pair, the outputs of two feature extractors f_A and f_B are combined using the bilinear operation introduced in [18]: $bilinear(l, \mathcal{I}, f_A, f_B) = f_A(l, \mathcal{I})^T f_B(l, \mathcal{I})$. In this way, we compute the bilinear feature vector for each spatial location of the image. If feature extractor f_A outputs local feature vectors of size M and f_B outputs feature vectors of size N, their bilinear combination will have size $M \times N$, and for notational simplicity we assume that the output of bilinear operation is resized to $MN \times 1$.

We then suggest to aggregate obtained bilinear features across locations situated in predefined set of image regions: $r_1, ..., r_R$, where R is number of chosen regions. After such pooling, we get the pooled feature vector for each image region i (as opposed to the feature vector that is obtained in [18] for the whole image):

$\phi_{r_i}(\mathcal{I}) = \phi(\mathcal{I}_{r_i}) = \sum_{l \in r_i} bilinear(l, \mathcal{I}, f_A, f_B)$.

In order to get a descriptor for image $\mathcal{I}$, we combine all region descriptors into a matrix of size $R \times MN$:

$$\phi(\mathcal{I}) = [\phi_{r_1}(\mathcal{I})^T; \phi_{r_2}(\mathcal{I})^T; ...; \phi_{r_R}(\mathcal{I})^T]. \quad (3.2)$$

This matrix can be turned into vector descriptor by resizing, or some transforms can be applied beforehands. Formally, such additional transforms can be incorporated into similarity function $\mathcal{S}$. For example, fully connected layer performs additional transform in our architecture. Note, that for some types of transforms

(for example, for convolution) spatial order of pooling regions should be preserved.

In our experiments, we simply used the grid of equally-sized non-overlapping patches as a set of pooling regions, and concatenate the matrix columns in (3.2) into a single global descriptor. The resulting Multiregion Bilinear architecture is shown in figure 3.

IV. DATASETS AND EVALUATION PROTOCOLS

A. Evaluation metrics

At test time, re-identification methods are usually evaluated by matching images from one set (the *probe set*) to images from another set (the *gallery set*), often corresponding to disjoint camera views. Given a query image, a method is supposed to sort gallery images by similarity with the query and give k most similar images (*the top-k list*) as an output. Commonly, performance of person re-identification methods is evaluated with CMC (Cumulative Matching Characteristic curve) metric which represents methods quality as number of recall values for different k.

There are two common settings for evaluating person re-identification algorithms: *single-shot* and *multi-shot*. Single-shot setting implies that there is only one image for each person in each camera. In multi-shot scenario, there may be multiple images foe each person in each camera.

Although most of person re-identification papers use CMC metric for evaluating performance, mean average precision (mAP) is more applicable in the case when multiple images of the same person are present in the gallery set [15] as it takes into account all the occurences of right answers for a query, not only the first one (as the CMC metric). Therefore, we report single-shot results in a form of CMC and multi-shot results in both forms of CMC and mAP.

B. Datasets

We investigate the performance of the DML method and its Bilinear variant for two re-identification datasets: CUHK03 [12] and Market-1501 [15]. They are the largest among person re-identification datasets, and therefore are most appropriate for deep learning methods. At the same time, it is reasonable to expect that the datasets will continue to grow and even bigger datasets will emerge.

The CUHK03 dataset includes 13,164 images of 1,360 pedestrians captured from 6 surveillance cameras (3 pairs of cameras). Each identity is observed by two cameras and has 4.8 images in each camera on average. The two versions of the dataset are provided: *CUHK03-labeled* (pedestrian bounding boxes are manually labeled), and *CUHK03-detected* (bounding boxes are obtained automatically by the pedestrian detector [29]). We provide results for both versions.

Following [12], we use Cumulative Matching Characteristic curve (CMC) metric to report our results. The evaluation protocol accepted for the CUHK03 is the following: 1,360 identities are split into 1,160 identities for training, 100 for validation and 100 for testing. We use the validation set to choose the best Bilinear architecture variant and the number of training iterations.
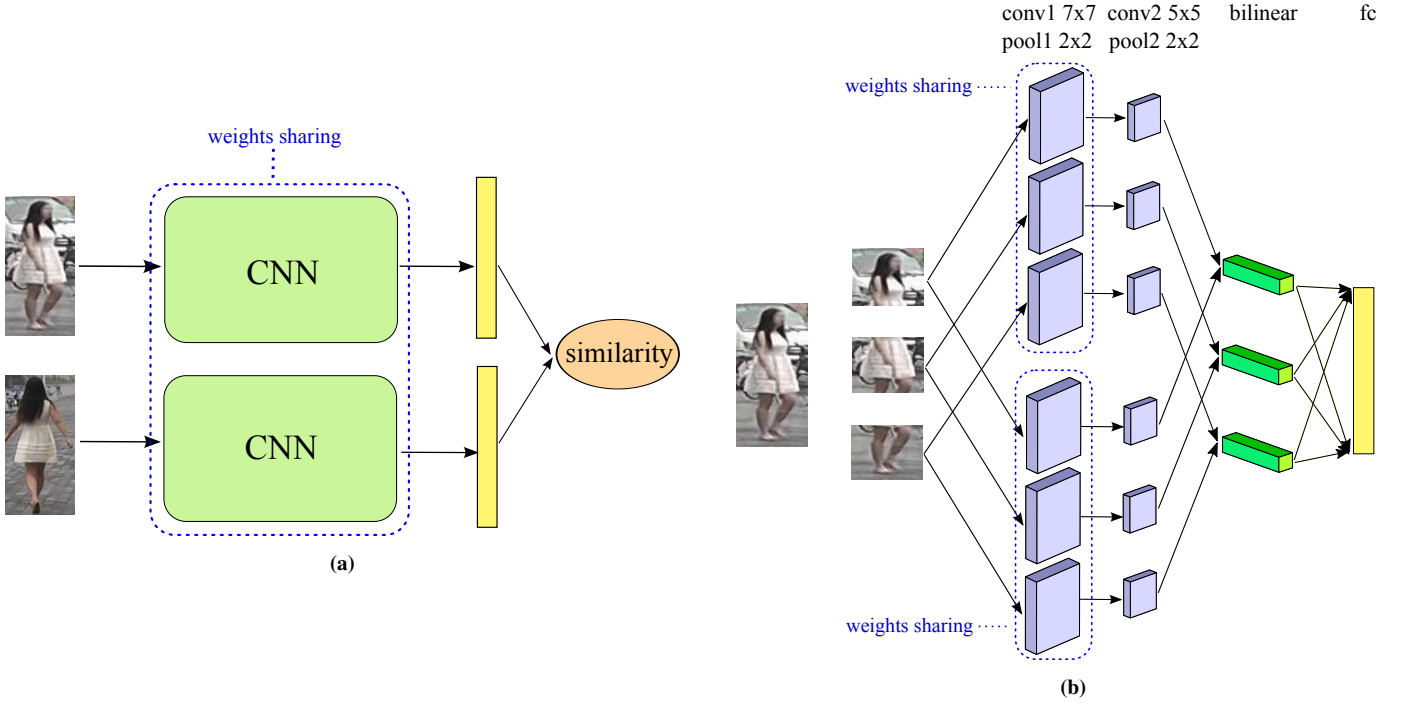
conv1 7x7 conv2 5x5   bilinear   fc
pool1 2x2 pool2 2x2

weights sharing

weights sharing

(b)

weights sharing

(a)

FIG. 2: Siamese network architecture used (a) - overall structure of the neural net introduced in [1], (b) - bilinear modification for each of the siamese "halves" used in this work. There are two separate feature extractors for each of three image parts, which are then combined by bilinear operation.
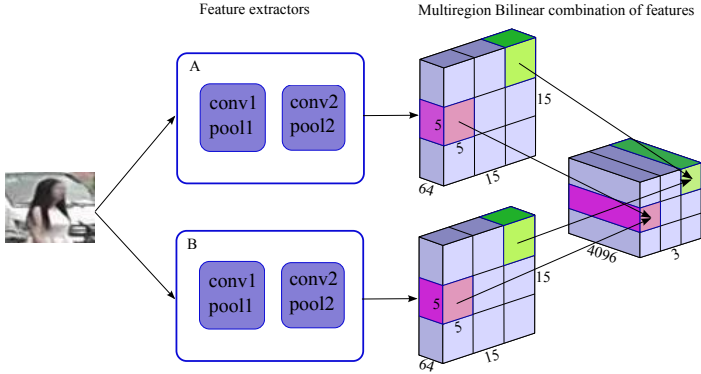

Feature extractors

Multiregion Bilinear combination of features

FIG. 3: Multiregion Bilinear architecture used in this work. There are two feature extractors for three image parts. Threir outputs are combined by bilinear operation in each location, then obtain features are pooled within equal patches of size 5×5. See section III D for the formal model description.

At test time single-shot CMC curves are calculated. This means that one image for each identity from test set is chosen randomly in each of its two camera views. CMC curves for 100 of such tries are averaged to compute the final CMC for a given split. We use one split for choosing architecture and parameters (figures 4, 5) and five random splits to calculate the resulting average CMC (figures 8, 6). Some sample images of CUHK03 dataset are shown in figure 1a.

We also report our results on the Market-1501 dataset, introduced in [15] (See figure 1b for sample images). This dataset contains 32,643 images of 1,501 identities, each identity is captured by at most six cameras and by at least two cameras. The dataset is randomly divided into the test set of 751 identities and the train set of 750 identities. For each identity in the test set one image in each camera is selected and used as a query. Manually

drawn bounding boxes are used for query images, and automatically detected ones for training images as well as for the gallery images in the test set. Matching is done across different cameras, and in the sitauation when the gallery images of the same person and form the same camera as the query are ignored and not counted as true positives during the evaluation.

The additional difficulty of this dataset is the presence of "distractor" images in the test gallery set, which are actually "bad" bounding boxes obtained by the DPM detector [29]. These images serve as "negative answers" for all queries.

As in [15], we also evaluate the re-identification performance using multi-shot protocol (multiple images from the same camera are present for each identity in the gallery set), using the CMC and the mAP measures (which is more appropriate in the multi-shot scenario).

Following [15], we demonstrate results with Multi Query approach, when all images of a person from one camera are used for a query, instead of using only one randomly chosen image as in standard protocol. As in [15] we investigate pooling methods for descriptors of query images for CUHK03 and Market-1501. Although descriptors are merged only for images of a person from one camera, they still capture different poses, thus using pooled query descriptor might be better for re-identification performance.

V. EXPERIMENTS

A. Training the network

As in [1], we form training pairs inside each batch consisting of 128 randomly chosen training images (from all cameras). The label for each training pair is assigned accordingly to identity numbers (+1 for the same identity, -1 for different identities).

The training set is shuffled after each epoch, so the network can see many different image pairs while training. All images are resized to height 160 and width 60 pixels.

Binomial Deviance, also defined in [1] is used as loss function, and its parameters, as in [1], are set to $\alpha = 2$, $\beta = 0.5$. Cosine similarity is used to compute the distance between a pair of image descriptors.

As images for each mini-batch are chosen randomly, there may be very small number of positive image pairs (only few tens from 8,128 pairs in mini-batch, depending on dataset). It does not impede the learning process, because Binomial Deviance loss is weighted for numbers of negative and positive pairs. Moreover, we observed, that increasing the number of positive pairs at the expense of the number of negative pairs causes performance drop.

We train networks (baseline DML and the variants of Bilinear DML) with the weight decay rate of 0.0005. The learning rate is changing according to "step" policy, the initial learning rate is set to 0.001 and it is multiplied by 0.1 when the performance on the validation set stops improving (50,000 iterations for CUHK03, and 40,000 for Market-1501). The dropout layer with probability of 0.5 is inserted before the fully connected layer.

Baseline DML and Bilinear DML with pooling across all locations (baseline Bilinear DML) are learned for 90,000, and Multiregion Bilinear DML nets are learned for 55,000 iterations on CUHK03 and for 50,000 iterations on Market-1501.

Bilinear DML networks are initialized by the weights of the baseline DML net trained on the same dataset for 60,000 iterations [2].

B. Experiments with Bilinear DML.

We have investigated various ways to initialize the bilinear DML and multi-region bilinear DML networks. Following the intuition behind B-CNN that suggests that the two halves of the network should play different roles (e.g. parts detector and texture descriptor), we tried different initializations for B-CNN feature extractors to find the optimal configuration[3]. We have always initialized one of the halves of the B-CNN with the baseline DML network, and for the other half we have considered the following options:

- re-identification – the other half is initialized to the same state as the first one using the baseline DML network.

- segmentation – the other half is initialized by the weights of a pretrained semantic segmentation network that used a pixel-wise softmax loss and was trained on Pedestrian Parsing in Surveillance Scenes Dataset (PPSS) [30] (which has 8 class labels corresponding to body parts, and 1,908 images that we downsampled to match the size of the second convolutional layer output in our architecture).

- random – the other half is initialized with random weights.

- detection – the other half is initialized using weights of a trained binary classifier which predicts the presence of pedestrian in the image (binary classification: pedestrian/background). We used NICTA Pedestrian Dataset [31] dataset for training and added CUHK03 training images to enlarge number of images of positive class in the training set. We thus used 200,000 negative images, 37,344 positive images from NICTA and also 11,235 images from CUHK03. After pretraining we took the convolutional layers of the resulting network.

The CMC comparisons of various initialization choices are shown in figure 4. The experiments showed that initializing both feature extractors with weights of pretrained baseline DML is more effective than other initialization schemes both for the labeled and the detected variants of the CUHK03 dataset. This is the initialization strategy that we use in the subsequent experiments.

Interestingly, for the CUHK03-labeled dataset, the baseline DML network outperformed all plain bilinear variants, while for the CUHK03-detected version the situation was reversed. This confirms that the latter scenario is more beneficial for the bilinear networks due to their ability to handle pose variations and/or lack of registration.

C. Further Variations of the Bilinear CNN

Before adding multi-region pooling, we tried to improve the performance of B-CNN in some simpler ways. We observed that in our case pooling across all locations in Bilinear DML network causes dramatic decrease of number of parameters compared to the well-tuned DML architecture (21,916,608 parameters in the baseline DML network vs 6,777,216 parameters in the Bilinear DML network). Hence, we decided to check if the increase in the number of parameters of Bilinear DML network will improve its performance, and tried two ways to make the number of parameters of the baseline Bilinear DML network to match the number of parameters of the original baseline DML model.

First, we tried to insert additional fully connected layer with 1,689 output channels followed by ReLU non-linearity after the bilinear operation in the baseline Bilinear DML network. Second, we tried to increase the number of feature maps in the second convolution layer in one of the two feature extractors from 64 to 214. These experiments were carried out on the CUHK03-labeled dataset.

As shown in figure 5a, Bilinear DML net with additional fully connected layer (Bilinear DML, plus fc) outperforms baseline DML and baseline Bilinear DML networks for a number of rank values between 3 and 15, but it shows much worse identification rate for rank 1, which is arguably the most important point on the curve. Besides, it again becomes worse than the baseline DML and the Bilinear DML after the rank 15.

On the other hand, the Bilinear DML network with the increased number of feature maps for one of the two feature extractors (Bilinear DML, 64x214) performed better than the baseline Bilinear DML network only for ranks 1-4. And it was still worse than the baseline DML for all rank values.

We also showed that two separate feature extractors are important for performance of bilinear architecture: we tried to

---

[2] We tried learning baseline for more iterations (because the total number of learning iterations for Bilinear nets, including pretraining, is more than 90,000) and also decreasing the learning rate later. However this did not improve the baseline performance.

[3] In a sequel, that the same kind two feature extractors are taken for each of three parallel streams.
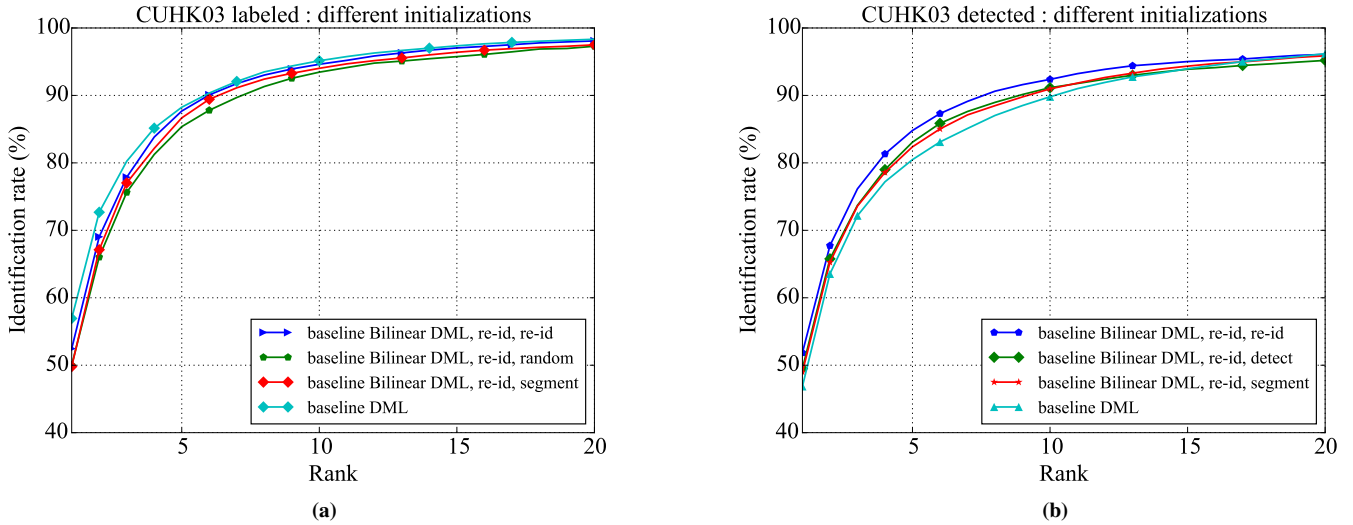
FIG. 4: Single-shot results (CMC) for different initializations of "plain" (i.e. single region) bilinear architecture achieved on (a) - CUHK03-labeled, (b) - CUHK03-detected. After the name of the method, the initialization is indicated. "Re-id" means that one of two feature extractors for bilinear combination is initialized by weights of baseline DML net learned on the same data; "random" means that one of the two feature extractors is initialized randomly; "segment" means that one of the two feature extractors is initialized by the weights of semantic segmentation CNN, "detect" means that one of the two feature extractors is initialized by the weights of the network trained to classify images into "pedestrians" and "non-pedestrians". Symmetric initialization with the same initialization for both B-CNN halves by the baseline re-identification CNN shows better results than other variants of B-CNN.

share the first convolution layer for the two feature extractors, which resulted in lower identification rate in comparison with the baseline Bilinear DML net for all rank values. The same situation takes place for re-id quality on the training set.

Finally, we note that we checked that increasing the number of parameters in the baseline DML by adding more hidden units in the fully-connected layer did not improve performance on the test set either.

D. Adding multi-region pooling

Finally, we have added the multi-region pooling and this gave considerable performance boost for all rank values (figure 5). The improvement is consistent for both variants of CUHK-03. The experiments were conducted with baseline DML initialization. Without pretraining, performance of Multiregion Bilinear DML drops approximately by 2% at rank 1.

Same figure also shows that the performance for Multiregion Bilinear DML net with pooling within patches of size 5×5 (Multiregion Bilinear DML, patches 5×5) noticeably outperforms Multiregion Bilinear DML with pooling within patches of size 8x8 for the first 7 ranks, and for the other ranks values it is only slightly worse.

E. Multi-shot results

Following [15], we also evaluated our method using multi-shot re-identification protocol. It turned out that multi-shot results are much higher than single-shot for both versions of CUHK03. And when using several images for a query ("Multiple Queries" protocol), the results improve further for both CUHK03 and Market-1501.

As in [15], we performed multi-shot evaluation with multiple queries, using two pooling approaches: 1) average pooling, when descriptors of several queries are pooled into one by averaged sum, 2) max pooling, when resulting descriptor consists of maximum values of query descriptors in each dimension.

Our experiments showed that pooling strategies for queries improve performance significantly, which can be explained by the fact, that pooled descriptors incorporate information about different poses present in multiple query images. Average pooling strategy turned out to be more effective for our descriptors than max pooling. It is only slightly better for CUHK03 (see figure 6) and noticeably better for Market-1501 (see figure 7).

F. Comparison with state-of-the-art methods

Finally, we compare our results with two other deep learning methods that were evaluated on CUHK03 dataset: IDLA [13] and FPNN [12]. We also include the results for baseline DML [1] evaluated on CUHK03 by us. Other methods do not use deep learning but show very good results : Metric Ensembles [9], LOMO+XQDA [11]. KISSME [20] and SDALF [3] are older methods, and we reproduced their results for CUHK03 from [15].

Overall, Multiregion Bilinear DML networks outperforms all the mentioned methods by a noticeable margin, except for Metric Ensembles on CUHK03 labeled (it is no more than 2% worse for first 20 rank values). It should be mentioned, however, that results of Metric Ensembles method, reported in [9], are achieved after learning on only part of CUHK03 images.

Market-1501 dataset is relatively new and there is only one paper [15] reporting re-identification results for it. It utilises good unsupervised method for re-identification. We include its best results for multi-query approach with max pooling strategy into the final table I.
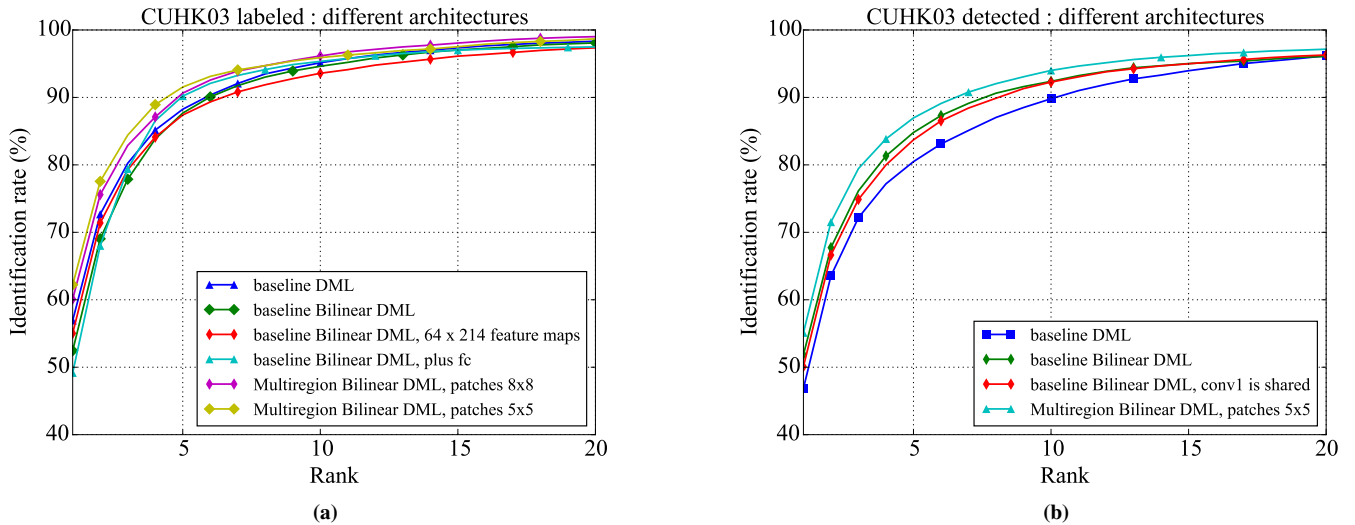
FIG. 5: Single-shot results (CMC) of different architectures for CUHK03 (a) – CUHK03-labeled, (b) – CUHK03-detected (single-shot CMC). Baseline DML is compared to baseline Bilinear DML (pooling of bilinear features is done across all locations) and the Multiregion Bilinear DML proposed in this work (pooling of bilinear features is done across locations within certain patches). "Baseline Bilinear DML, plus fc" is the variant of the baseline Bilinear architecture with an additional fully connected layer, "Baseline Bilinear DML, 64x214 features" is the variant of the baseline Bilinear architecture with 64 feature maps output from one feature extractor and 214 feature maps output from another one. Overall, multiregion Bilinear DML networks demonstrate the best performance for both variants of the CUHK-03 dataset.
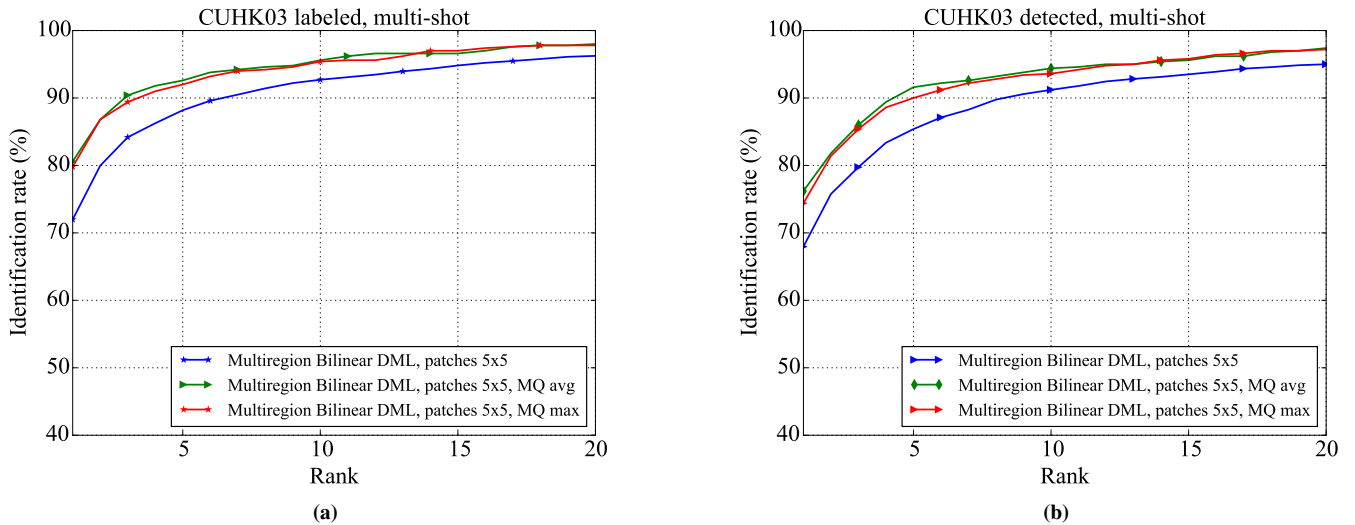

FIG. 6: Results (CMC) of multi-shot experiments for CUHK03 (a) - labeled, (b) - detected. Comparison for query pooling techniques for Multiregion Bilinear DML network. Using multiple queries shows much better results than using only one image for query. Average pooling is slightly better than max pooling for CUHK03.

Rank-1 accuracies of the above-mentioned methods are listed in table I along with our results. Following [15], we also include mean average precision for experiments on Market-1501.

VI. DISCUSSION

In this paper we demonstrated an application of new multi-region Bilinear CNN architecture to the problem of person re-identification. Having tried different variants of bilinear architecture, we showed that such architectures are competitive for this task and that keeping some spatial information is important for building better descriptors. Multiregion Bilinear DML gives a way to keep such spatial information, which allowed us

to extract more complex features and increase the number of parameters over the baseline CNN without overfitting. The resulting method improved state-of-the-art results for two challenging datasets: CUHK03 and Market-1501. We also demonstrated notable gap between performance of Multiregion Bilinear DML and the performance of standard CNN (we used DML [1] as baseline) on these two dataset (where CUHK03 was considered in two variants).

There are some negative results that we have encountered. First, despite our efforts, we did not manage to get any improvement by trying to diversify the initialization of the two halves of the bilinear CNN. Also, we have done some experiments with smaller datasets (such as VIPeR [24]) and there we found the baseline DML method to outperform all the bilinear variants.

Methods	CUHK03 labeled	CUHK03 detected	Market-1501	
	r = 1	r = 1	r = 1	mAP
Single-shot results				
FPNN [12]	20.65	19.89		
SDALF [3]	5.60	4.87		
KISSME [20]	14.17	11.70		
IDLA [13]	54.74	44.96		
LOMO+XQDA [11]	52.2	46.25		
Metric Ensembles [9]	62.1			
Baseline DML [1]	55.2	49.84		
Multiregion Bilinear DML (ours)	**63.87**	**59.17**		
Multi-shot results				
Zheng et al.[15], MQ, max		24.33	47.25	21.88
Multiregion Bilinear DML	71.98	67.96	45.58	26.11
Multiregion Bilinear DML, MQ, max (ours)	79.80	74.4	53.62	30.76
Multiregion Bilinear DML, MQ, avg (ours)	**80.60**	**76.2**	**56.59**	**32.26**

TABLE I: Rank-1 identification rates for different re-identification methods for single-shot and multi-shot experiments. Mean average precision is reported for the Market-1501 dataset. Multi-shot results for different pooling strategies for multiple queries are also presented.
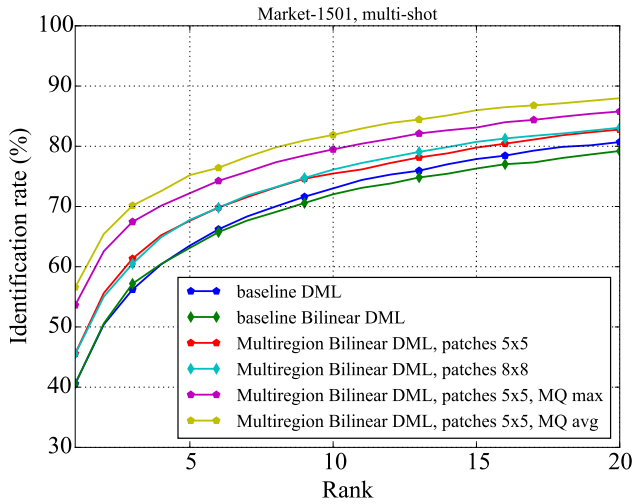

FIG. 7: Results (CMC) of multi-shot experiments for Market-1501. Baseline DML method is compared with the Bilinear DML and Multi-region Bilinear DML, which is substantially better than other methods. Multiple queries pooling strategies are also presented. As on CUHK03, pooling query descriptors significantly improves results. Average pooling strategy is also more effective than max pooling.

One interesting direction for future work would be using meaningful regions to pool bilinear features, which could help, for example, to eliminate background. We believe that these pooling regions could also be learned or be derived from some explicit body part detector/segmenter. It would be also interesting to investigate an architecture without separate processing streams for the three different parts of the image, and figure out whether bilinear approach can help in learning such neural network.

REFERENCES

[1] D. Yi, Z. Lei, S. Z. Li, Deep metric learning for practical person re-identification, arXiv prepzrint arXiv:1407.4979.
[2] B. Ma, Y. Su, F. Jurie, Bicov: a novel image representation for person re-identification and face verification, in: British Machive Vision Conference, 2012, pp. 11–pages.
[3] L. Bazzani, M. Cristani, V. Murino, Sdalf: modeling human appearance with symmetry-driven accumulation of local features, in: Person Re-Identification, Springer, 2014, pp. 43–69.
[4] S. Li, M. Shao, Y. Fu, Cross-view projective dictionary learning for person re-identification, in: Proceedings of the 24th International Conference on Artificial Intelligence, AAAI Press, 2015, pp. 2155–2161.
[5] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, Q. Mary, Person re-identification by support vector ranking., in: BMVC, Vol. 2, 2010,
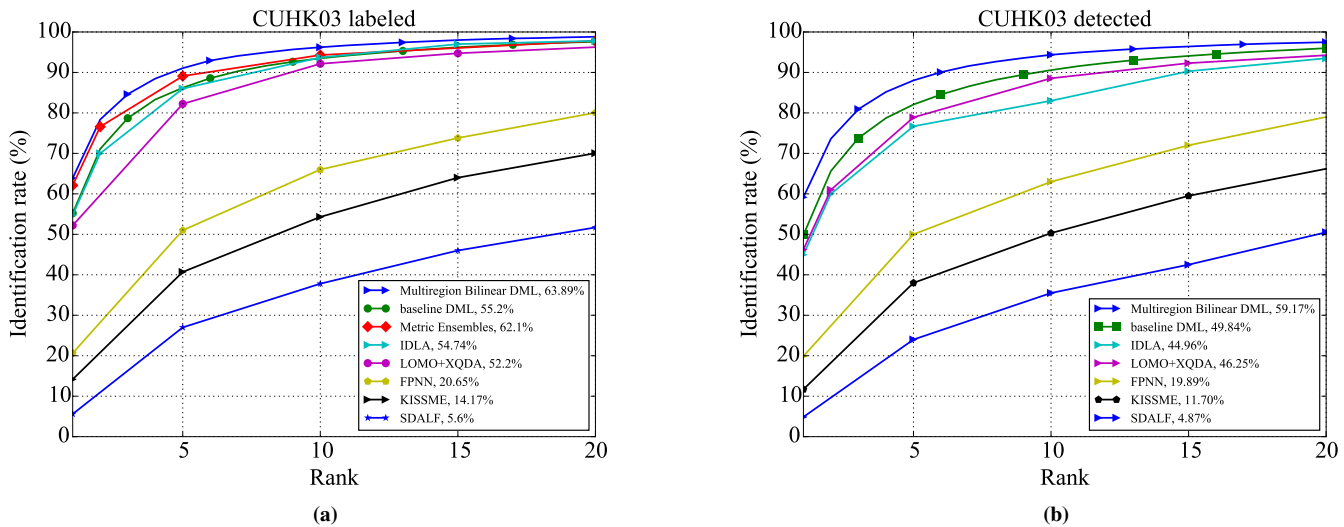
FIG. 8: Single-shot results (CMC) of different methods for CUHK03 (a) - labeled, (b) - detected. Multi-region Bilinear DML outperforms the majority of state-of-the-art methods on both versions of CUHK03. The only exception is the Metric Emsembles method [9] that shows similar performance to our method for CUHK03-labeled dataset.

p. 6.

[6] C.-H. Kuo, S. Khamis, V. Shet, Person re-identification using semantic color names and rankboost, in: Applications of Computer Vision (WACV), 2013 IEEE Workshop on, IEEE, 2013, pp. 281–287.

[7] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, H. Bischof, Mahalanobis distance learning for person re-identification, in: Person Re-Identification, Springer, 2014, pp. 247–267.

[8] M. Hirzer, P. M. Roth, H. Bischof, Person re-identification by efficient impostor-based metric learning, in: Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on, IEEE, 2012, pp. 203–208.

[9] S. Paisitkriangkrai, C. Shen, A. van den Hengel, Learning to rank in person re-identification with metric ensembles, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15), 2015.

[10] B. Ma, Y. Su, F. Jurie, Local descriptors encoded by fisher vectors for person re-identification, in: Computer Vision–ECCV 2012. Workshops and Demonstrations, Springer, 2012, pp. 413–422.

[11] S. Liao, Y. Hu, X. Zhu, S. Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2197–2206.

[12] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: Deep filter pairing neural network for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, USA, 2014.

[13] E. Ahmed, M. Jones, T. K. Marks, An improved deep learning architecture for person re-identification, 2015.

[14] S.-Z. Chen, C.-C. Guo, J.-H. Lai, Deep ranking for person re-identification via joint representation learning, arXiv preprint arXiv:1505.06821.

[15] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: Computer Vision, IEEE International Conference on, 2015.

[16] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, L. D. Jackel, Handwritten digit recognition with a back-propagation network, in: Advances in Neural Information Processing Systems 2, [NIPS Conference, Denver, Colorado, USA, November 27-30, 1989], 1989, pp. 396–404.

[17] R. Zhao, W. Ouyang, X. Wang, Person re-identification by saliency learning, arXiv preprint arXiv:1412.1908.

[18] A. R. Tsung-Yu Lin, S. Maji, Bilinear cnn models for fine-grained visual recognition, in: International Conference on Computer Vision (ICCV), 2015.

[19] A. RoyChowdhury, T.-Y. Lin, S. Maji, E. Learned-Miller, Face identification with bilinear cnns, arXiv preprint arXiv:1506.01342.

[20] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 2288–2295.

[21] A. Mignon, F. Jurie, Pcca: A new approach for distance learning from sparse pairwise constraints, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 2666–2672.

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

[23] W. Li, R. Zhao, X. Wang, Human reidentification with transferred metric learning., in: ACCV (1), 2012, pp. 31–44.

[24] D. Gray, S. Brennan, H. Tao, Evaluating appearance models for recognition, reacquisition, and tracking, in: Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), Vol. 3, Citeseer, 2007.

[25] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, R. Shah, Signature verification using a siamese time delay neural network, International Journal of Pattern Recognition and Artificial Intelligence 7 (04) (1993) 669–688.

[26] S. Chopra, R. Hadsell, Y. LeCun, Learning a similarity metric discriminatively, with application to face verification, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, Vol. 1, IEEE, 2005, pp. 539–546.

[27] B. F. Klare, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, A. K. Jain, Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a, algorithms 13 (2015) 4.

[28] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA, 2006, pp. 2169–2178.

[29] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan,

Object detection with discriminatively trained part-based models, Pattern Analysis and Machine Intelligence, IEEE Transactions on 32 (9) (2010) 1627–1645.

[30] P. Luo, X. Wang, X. Tang, Pedestrian parsing via deep decompo-sitional network, in: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE, 2013, pp. 2648–2655.

[31] G. Overett, L. Petersson, N. Brewer, L. Andersson, N. Pettersson, A new pedestrian dataset for supervised learning, in: Intelligent Vehicles Symposium, 2008 IEEE, IEEE, 2008, pp. 373–378.