# An Improved Deep Learning Architecture for Person Re-Identification

Ejaz Ahmed[1], Michael Jones[2], Tim K. Marks[2]
[1]Department of Computer Science, University of Maryland. [2]Mitsubishi Electric Research Labs (MERL), Cambridge, MA.

Person re-identification is the problem of identifying people across images that have been taken using different cameras, or across time using a single camera. The problem of re-identification is usually formulated in a similar way to face recognition. A typical re-identification system takes as input two images, each of which usually contains a person's full body, and outputs either a similarity score between the two images or a classification of the pair of images as *same* (if the two images depict the same person) or *different* (if the images are of different people). In this paper, we follow this approach and use a novel deep learning network to assign similarity scores to pairs of images of human bodies.

Our deep network, shown in Figure 1, begins with two layers of convolution and max pooling that act separately on each of the two input images to learn a set of features for comparing the two images. The weights in the convolutional layers are tied (i.e., they are the same for each input path), which ensures that the same filters are applied to each input image. We next use a novel layer to compare the features computed from the two images. It computes cross-input neighborhood difference features, which compare the features from one input image with the features computed in neighboring locations of the other image. The motivation behind taking differences in a neighborhood is to add robustness to positional differences in corresponding features of the two input images. This is followed by a subsequent novel layer that distills these local differences into a smaller patch summary feature, by computing a linear combination of differences for each feature's neighborhood. Next, we use another convolutional layer with max pooling, followed by two fully connected layers and softmax output.

We implement our deep learning architecture in Caffe [2]. We augment our training sets by randomly shifting input images by small amounts. We also use hard negative mining and retraining to further improve our networks.

We tested our method on the CUHK03, CUHK01, and VIPeR data sets and compared against many other algorithms that were also tested on these data sets. The CUHK03 data set [5] is the largest existing data set for re-identification, containing 13,164 images of 1,360 pedestrians. There are two versions of the images: one with manually labeled pedestrian bounding boxes, and one with automatically detected boxes that are less accurate. We show results on both versions. The CUHK01 data set [4] has 971 identities with 2 images per person in each of 2 views. Previous work has tested on this data set using either 100 identities for testing (which leaves 871 for training) or 486 identities (leaving 485 for training). Again, we test and compare on both variations. Finally, the VIPeR data set [1] is one of the most popular for evaluating re-identification algorithms, but is also very small with only 632 identities and 1 image per identity in each of 2 different views. Half of the data is used for training and half for testing. This is a very small amount of data for deep learning approaches, but we still achieve good results by first training on CUHK03 data and then fine tuning on VIPeR data. Table 1 summarizes our results in terms of rank-1 recognition rates for our method and the best-performing methods from the literature.

Our results show a very large improvement over previous state-of-the-art methods when the amount of training data is relatively large (CUHK03, and CUHK01 with 100 test identities). These results demonstrate the importance of the novel layers of our deep network architecture: the cross-input neighborhood differences layer, and the subsequent layer that summarizes these differences. We also show that models learned by our method on a large data set can be adapted to new, smaller data sets.

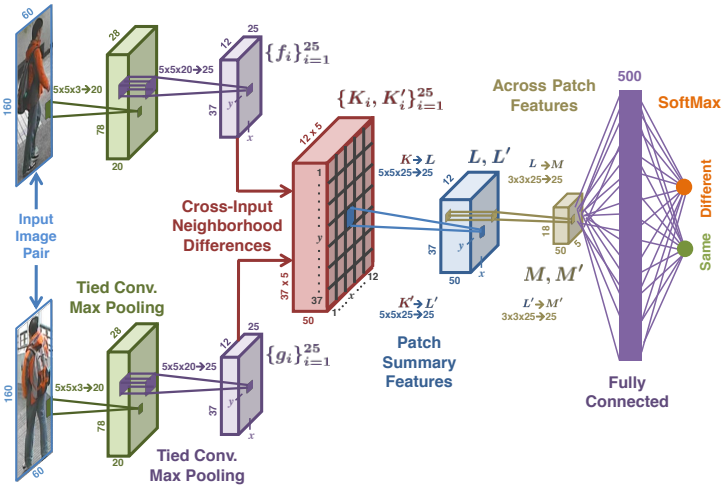


Figure 1: Proposed Architecture. Paired images are passed through the network. While initial layers extract features from the two views individually, higher layers compute relationships between the views. The number and size of convolutional filters that must be learned are shown. For example, in the first tied convolution layer, $5 \times 5 \times 3 \rightarrow 20$ indicates that there are 20 convolutional features in the layer, each with a kernel size of $5 \times 5 \times 3$. There are 2,308,147 learnable parameters in the entire network.

| | CUHK03 labeled | CUHK03 detected | CUHK01 100 test IDs | CUHK01 486 test IDs | VIPeR |
|---|---|---|---|---|---|
| Our Method | **54.74** | **44.96** | **65.00** | **47.53** | 34.81 |
| FPNN [5] | 20.65 | 19.89 | 27.87 | N/A | N/A |
| KISSME [3] | 14.17 | 11.70 | 29.40 | N/A | 19.60 |
| visWord [7] | N/A | N/A | N/A | 44.03 | N/A |
| mFilter+LADF [6, 8] | N/A | N/A | N/A | N/A | **43.49** |

Table 1: Rank-1 recognition rates (%) for our method and other top performing methods on various data sets.

[1] Doug Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *In IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, Rio de Janeiro*, 2007.

[2] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

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

[4] W. Li, R. Zhao, and X. Wang. Human re-identification with transferred metric learning. In *ACCV*, 2012.

[5] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[6] Z. Li, S. Chang, F. Liang, T. Huang, L. Cao, and J. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, 2013.

[7] Z. Zhang, Y. Chen, and V. Saligrama. A novel visual word co-occurrence model for person re-identification. In *ECCV Workshop on Visual Surveillance and Re-identification*, 2014.

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

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.