A coarse-to-fine deep learning for person re-identification

Anonymous WACV submission

Paper ID 207

Abstract

This paper proposes a novel deep learning architecture for person re-identification. The proposed network is based on a coarse-to-fine learning (CFL) approach, attempting to acquire a generic-to-specific knowledge throughout a transfer learning process. The core of the method relies on a hybrid network composed of a convolutional neural network and a deep belief network denoising autoenconder. This hybrid network is in charge of extracting features invariant to illumination varying, certain image deformations, horizontal mirroring and image blurring, and is embedded in the CFL architecture. The proposed network achieved the best results when compared with other 12 state-of-the-arts methods, over the VIPeR and i-LIDS data sets.

1. Introduction

Person re-identification is one of the most challenging task in Computer Vision. It consists in identifying a person across a database of images, given a target image or video of that person. Person re-identification systems usually cope with inherent characteristics of the environment, such as unstructured scenes, human pose variation, lighting changing, low-resolution images, just to cite a few. A comprehensive review on person re-identification can be found in [4].

In the last decade, several methods for person re-identification have been proposed. Features with the goal of globally representing the disparities or similarities between two images can be found in [3], [11], [12] and [14]. In the same way, new distance metrics were proposed with the goal of learning similarity scores considering two person images [20], [9], [21], [7], [22]. In [16], a new discriminative model based on a ranked-SVM was introduced to solve the problem without labeling information of persons in the target domain cameras.

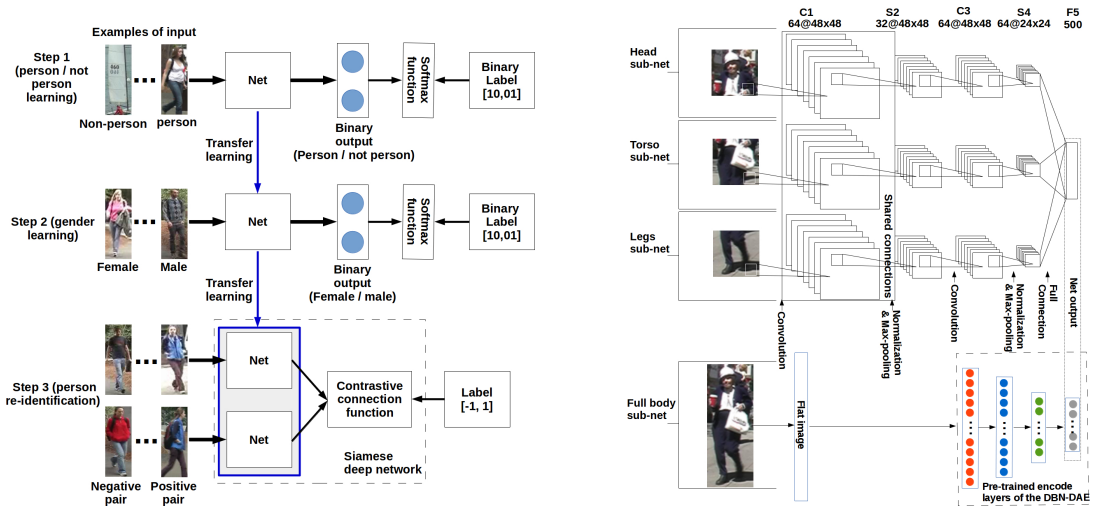
In recent years, deep learning has been adopted to solve several Computer Vision problems, such as pedestrian detection [23], face identification [24], feature generation [6], feature extraction [17] and face parsing [15]. This approach aims at learning, at the same time, some levels of abstraction in image representation, classifier parameters and/or distance metric functions. A deep network can be previously trained in an unsupervised fashion by a symmetric network, called autoencoder, in order to create a compressed representation of its input and/or to address the lack of sufficient data to learn. Some works that use an autoencoder network for different tasks can be found: In [18], a multi-modal deep belief network (DBN) was proposed to learn a sharing representation of a set of videos and their associated audio information; in [24], a normalized representation of face images, learned by a convolutional deep autoencoder, was created with the goal of generating face features invariant to pose and illumination changing; in [19], a denoising autoencoder (DAE) was introduced to learn useful image representation; in [10], a deep autoencoder was proposed to retrieve context-based image. In the context of person re-identification, some deep network architectures have been proposed [22], [2], [13], [5], achieving state-of-the-art results on almost all evaluated data sets.

1.1. Contributions

Our work brings two main contributions: (i) A machine transfer learning approach motivated by the human skill of obtaining coarse-to-fine knowledge; and (ii) a novel hybrid deep network topology. The basic idea of the first contribution is to train a deep network to learn specific concepts, having previously learned a more generic knowledge. The goal of the second contribution relies on merging a set of convolutional neural networks (CNN) and a pre-trained DBN-DAE into a network able to learn image local features (CNN) and noise-invariant global features (DBN-DAE). Our proposed approach was compared with other 12 state-of-the-art methods [22], [11], [3], [16], [2], [12], [9], [14], [21], [5], [20], [7] over VIPeR [8] and i-LIDS [1] data sets, presenting the best performance.

2. A coarse-to-fine deep network architecture for person re-identification

Usually, if a human being desires to identify a person, first that one should know what a person is (usually learned



(a) (b)

Figure 1: (a) Outline of our proposed network. The "Net" box contains a network composed of three CNNs (each one for each human body part - head, torso and legs) and a pre-trained DBN-DAE, as shown in (b).

Figure 2: Noise filters on samples of VIPeR data set. From left to right: Original image, randomly changed brightness, horizontal mirroring, blurring and image distortion.

in infancy), discriminate gender (male/female), and then compare each one of the parts of a person with his/her own mental data base of person characteristics. The goal of our proposed network is to follow this rationale to have a coarse-to-fine knowledge acquisition with respect to the structure of a person to be identified.

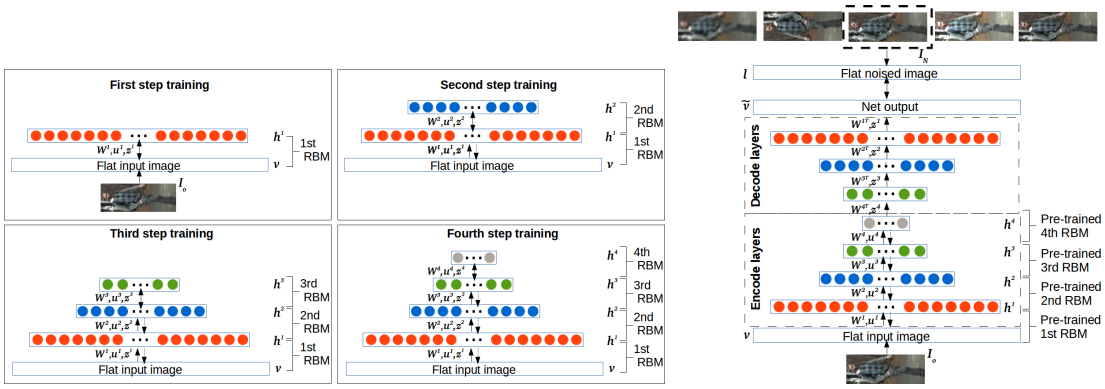
Figure 1a depicts the outline of our proposed architecture. The "Net" box denotes a network, which is a hybrid of a CNN and a DBN-DAE (see Fig. 1b). In the hybrid architecture (Sec. 2.1), while the CNN extracts local features from the person images, a pre-trained DBN-DAE (Sec. 2.2) extracts global features, invariant to certain types of noises, such as randomly changed brightness, horizontal mirroring, blurring and small image distortions, according to the examples of Fig. 2. Although it would be possible to have all CNNs or all DBN-DAEs in the hybrid of Fig. 1b, experiments show that a DBN-DAE along with a CNN form the best configuration (Sec. 3). After having the DBN-DAE trained, the overall network is now able to be trained, including the CNN of the hybrid network. The learning is transferred (Sec. 2.3) from the person (step 1) to the gender network (step 2), and then lately to the Siamese network (step 3) in order to accomplish the final person identification by means of a two identical networks (see Fig. 1a). In turn, the Siamese (Sec. 2.4) learns which pairs of images belong to the same or different persons.

2.1. Hybrid Network

Our hybrid network has four sub-nets: one for each body part (head, torso and legs) and a full-body sub-net (see Fig. 1b). Each one of the three body-part sub-nets is a CNN with two convolutional layers (C1 and C3), two max-pooling layers (S2 and S4), and one full-connected layer (F5). This latter one has 500 units shared with the three body-part sub-nets. The full-body sub-net is composed of the pre-trained DBN-DAE, which provides an 500-dimensional feature vector in its output. At the end, the hybrid network output is given by F5 layer concatenated with the output of the pre-trained DBN-DAE, forming a 1000-dimensional feature vector.

2.2. DBN-DAE

An autoencoder is a symmetric network whose output is equal to the input. The main goal is to learn a compact representation of the input. There are two types of layers in that network: Encode and decode. A trained autoencoder com-



(a) Steps of the DBN-DAE training

(b) DBN-DAE fine-tuning with the flat noisy image layer

Figure 3: DBN-DAE topology and training steps. In (a), a cascade of layer-wise Restricted Boltzmann Machines (RBM) training is performed in four steps. In the first step, the input layer of the first RBM is the flat version of an original image; when the first RBM is trained, a second RBM is stacked on the top of the first one; the output of the first RBM becomes the input layer of the second one. In the second step, while the new RBM is trained, the weights of the first one is fine-tuning. In the same way, the third and fourth steps follow the first and second ones. In (b), a fully symmetric DBN-DAE with all pre-trained RBM is trained to minimize a cross-entropy error between the output of the network and the flat noisy image.

putes a compact representation from its input by the encode layers, and recover a version of its input by the decode ones. A DAE attempts to create a representation robust to certain types of noises (embedded in the input images), placing a noisy input image on the output of the network.

The goal of the proposed DAE was to create a compressed representation of an image person, invariant to some conditions of illumination, blurring, horizontal mirroring and small image distortions (see Fig. 2). The training set of the DBN-DAE was composed of a set of input and output image pairs. After applying noise filters for each image of VIPeR and i-LIDS data sets, the total number of image pairs to feed the DBN-DAE was 91008 and 32256, respectively. This is so, once the training set was formed by matching one image against all the others, for each image person (also considering pairs formed by the same images).

The training process was the same as in [15], but in a single modality. The topology of our DBN-DAE is structured by four pre-trained Restricted Boltzmann Machines (RBM) layers. According to Fig. 3, $\mathbf{v}$ and $\mathbf{h}^i$, with $i = 1$ to 4, are the visible and hidden units, respectively. $\mathbf{v}$ is the flat version of the original image $\mathbf{I}_o$. $\widetilde{\mathbf{v}}$ is the output of the network, while $\mathbf{l}$ is the flat version of the noisy image $\mathbf{I}_N$. The weights between the layers are represented by $\mathbf{W}^i$ vectors. $\mathbf{z}^i$ and $\mathbf{u}^i$ are the offset vectors for input and hidden units, respectively. There are two steps to reach a fully trained DBN: (i) A stacked layer-wise training for each one of the four RBM, as shown in Fig. 3a, and (ii) a DBN fine-tuning to minimize the cross-entropy error between $\widetilde{\mathbf{v}}$ and $\mathbf{l}$, as shown in Fig. 3b. The weights of the encode layers is initialized by the weights of the pre-trained RBMs. The weights of the decode layers is the transpose of the weights of the encode ones. The number of DBN-DAE units for the input and output layers are 6912 (the output layer is the flat version of the 48×48 re-sized image with the 3 RGB channels). The number of hidden units are 4000 in $\mathbf{h}^1$, 2000 in $\mathbf{h}^2$, 1000 in $\mathbf{h}^3$ and 500 in $\mathbf{h}^4$. Each RBM is trained to maximize the product of the probability of a given training set T, given by

$$\underset{\mathbf{W},\mathbf{z},\mathbf{u}}{\operatorname{argmax}} \prod_{\mathbf{v}\in T} P(\mathbf{v})\,, \tag{1}$$

where

$$P(\mathbf{v}) = \frac{1}{Z}\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}\,, \tag{2}$$

where Z is a normalizing constant to ensure the probability distribution sums to 1. The energy function E of the equation 2 is given by

$$E(\mathbf{v},\mathbf{h}) = -\mathbf{u}^T\mathbf{v} - \mathbf{z}^T\mathbf{h} - \mathbf{v}^T\mathbf{W}\mathbf{h}\,. \tag{3}$$

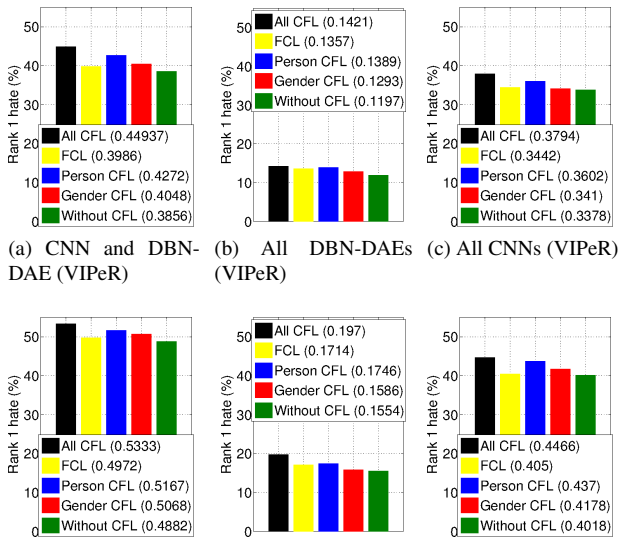
The conditional probabilities of $P(\mathbf{h}|\mathbf{v})$ and $P(\mathbf{v}|\mathbf{h})$ are modeled by a product of Bernoulli distributions, given by

$$P(\mathbf{h}_i = 1|\mathbf{v}) = \sigma(\mathbf{u}_i + \mathbf{W}_i\mathbf{v}) \tag{4}$$

and

$$P(\mathbf{v}_j = 1|\mathbf{h}) = \sigma(\mathbf{z}_j + \mathbf{W}_j^T\mathbf{h})\,, \tag{5}$$

where $\sigma(.)$ is a *sigmoid* function.



(a) CNN and DBN-DAE (VIPeR) (b) All DBN-DAEs (VIPeR) (c) All CNNs (VIPeR)

(d) CNN and DBN-DAE (i-LIDS) (e) All DBN-DAEs (i-LIDS) (f) All CNNs (i-LIDS)

Figure 4: Comparative evaluation of different configurations of our hybrid deep network. Black, yellow, blue, red and green bars depict the network performance: With full CFL, FCL, with only knowledge about person, with only knowledge about gender and without CFL, respectively.

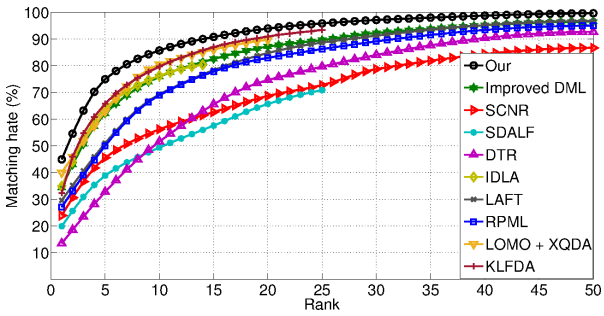
The encode layers of the pre-trained DBN-DAE is coupled to CNN to form the hybrid network (see Fig. 1b). The last encode layer corresponds to the global features of the image person that will be tuning in the training phase of the Siamese network.

2.3. Coarse-to-fine transfer learning

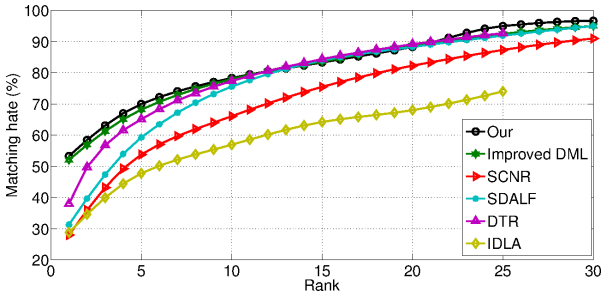
Before the Siamese network training, the coarse-to-fine transfer learning (CFL) approach takes place by means of a cascade of transfer learning (person $\rightarrow$ gender $\rightarrow$ identification). The goal of the transfer learning is to initialize the parameters of a network by using those pre-trained parameters of another network; this latter one trained in a different problem domain. Particularly, as the problem domains in step 1 and 2 of Fig. 1a reside in a binary classification, a binary layer in the output of the networks was added, and the networks were trained by a backpropagation algorithm with a softmax loss function. The learning rate in the training process in a step is decreased by 10 times regarding to the previous step. This is so, since the network of a higher step is tuning the parameters already learned in the previous one.

2.4. Siamese Network

Usually, the outputs of the two networks inside a Siamese topology are connected by one connection function and a cost function. The connection function evaluates



(a) Results on VIPeR dataset



(b) Results on i-LIDS dataset

Figure 5: Cumulative curves of the methods over VIPeR and i-LIDS datasets

the relationship between the two network outputs, while the cost function converts this relationship into a cost. A sample in the supervised training phase of the Siamese is composed of a pair of images and a label, y. In our Siamese network, the two networks were connected by a contrastive connection function, which ultimately measures the similarity between the two network outputs and the cost, at the same time. A contrastive function L is defined as

$$L(X1(\phi), X2(\phi), y) = (1-y)\frac{1}{2}D^2 + y\frac{1}{2}(max(0, m-D))^2 , \quad (6)$$

where $D = \|X1(\phi) - X2(\phi)\|_2$, and $X1(\phi)$ and $X2(\phi)$ denote the output of the Nets (step 3 of Fig. 1a), m represents a constant (in our case, equal to 1), and ϕ represents the network parameters. The Siamese is trained to find the values of ϕ that minimize L.

Siamese training was performed using the stochastic gradient descend with mini-batch size equals to 100 and 30000 iterations. The learning rate of the pre-trained DBN-DAE was set to 0.001, while the CNN learning was set to 0.01.

The contrastive function is not used in the prediction phase and the output of the two networks are evaluated by an Euclidean distance. The smaller the distance, the higher the similarity between the two persons in the input of the

Table 1: Comparative analysis on VIPeR data set. Each value corresponds to a hit rate score of a method in a specific rank (rank 14 instead of 15, for IDLA).

Method \ Rank	1	5	10	15*	20	25	30	50
Our	**0.4494**	**0.7500**	**0.8576**	**0.9082**	**0.9399**	**0.9589**	**0.9715**	**0.9968**
improved DML [22]	0.3440	0.6215	0.7589	0.8256	0.8722	0.8965	0.9228	0.9652
SCNR [11]	0.2392	0.4557	0.5623	0.6266	0.6873	0.7278	0.7880	0.8671
SDALF [3]	0.1987	0.3889	0.4937	0.5759	0.6573	0.7089	-	-
DTR [16]	0.1345	-	0.5158	-	0.7468	-	-	0.9272
IDLA [2]	0.3481	0.6424	0.7627	0.8038	-	-	-	-
LAFT [12]	0.2960	-	0.6931	-	-	0.8870	-	0.9680
RPML [9]	0.2700	-	0.6900	-	0.8300	-	-	0.9500
LOMO+XQDA [14]	0.4000	-	0.8051	-	-	0.9108	-	-
KLFDA [21]	0.3233	0.6578	0.7972	0.8699	0.9095	0.9346	-	-

Table 2: Comparative analysis over i-LIDS data set. Each value corresponds to a hit rate score of a method in a specific rank.

Method \ Rank	1	5	10	15	20	25	30
our	**0.5333**	**0.7000**	**0.7833**	0.8333	0.8832	**0.9333**	**0.9500**
DFLRDC [5]	0.5210	0.6820	0.7800	0.8360	0.8880	-	0.9500
LMNN [20]	0.2800	0.5380	0.6610	0.7550	0.8230	-	0.9100
MCC [7]	0.3130	0.5930	0.7560	0.8400	0.8830	-	0.9500
KLFDA [21]	0.3802	0.6512	0.7738	**0.8440**	**0.8919**	0.9267	-
SDALF [3]	0.2880	0.4778	0.5696	0.6424	0.6804	0.7405	-

Siamese. A combination of 4 images (original plus noisy produced ones), 2 by 2, has generated 16 distance scores from each pair of persons evaluated by the Siamese. The final score was computed by the maximum value of the 16 ones.

3. Experimental analysis

The performance of the proposed network was evaluated over VIPeR and i-LIDS data sets. VIPeR is composed of 632 pedestrian image pairs taken from two non-overlapping cameras. i-LIDS data set contains 476 images of 119 pedestrians taken from two non-overlapping cameras. Here the experiments were repeated 10 times, considering a random selection of 316 persons on VIPeR, for training and testing, and 60 persons for training and 59 for testing on i-LIDS. Training and testing sets were disjunct with regard to the person image.

A first step in the performance assessment of the proposed method was to define the best hybrid architecture, which is lately used inside the coarse-to-fine deep network. Three types of hybrid network was experimentally evaluated, considering a general structure depicted in Fig. 1b, but varying the type of network inside: (i) all CNNs, (ii) all DBN-DAEs and (iii) CNN and DBN-DAE. The second step was to assess the performance of the overall network depicted in Fig. 1a by varying its architecture, according to: (i) person transfer learning (Person CFL); (ii) fine-to-coarse learning (FCL) – training gender before person; (iii) only person transfer learning (Person CFL); (iv) only gender transfer learning (Gender CFL); (iii) person and gender transfer learning (all CFL); and, (iv) with only the Siamese deep network (without CFL). Figure 5 shows that the use of the **DBN-DAE with a full CFL** increases the hit rate, in the top rank, of our model by 11%, over VIPeR, and 13% over i-LIDS, in comparison with the single CNN Siamese deep network without CFL. The use of the hybrid topology, instead of the network with only CNNs, increases the top rank performance of our model by at least 5% over both data sets. The network pre-trained by CFL increases the hit rate of our model, in the top rank performance, by at least 4%, in comparison with the network without CFL.

After choosing the best overall architecture (as in Fig. 1a), the performance of our proposed method was compared with 12 state-of-the-art methods: Improved Deep Metric Learning (DML) [22], Semantic Color Names and Rankboost (SCNR) [11], Symmetric-driven accumulation of local features (SDALF) [3], Domain Transfer support vector Ranking (DTR) [16], Improved Deep Learning Architec-

ture (IDLA) [2], Locally Aligned Feature Transformation (LAFT) [12], Relaxed Pairwise Learned Metric (RPLM) [9], Local Maximal Occurrence Representation and Metric Learning (LOMO+XQDA) [14], Kernel-based Metric Learning (KFLDA) [21], Deep Feature Learning with Relative Distance Comparison (DFLRDC) [5], Large Margin Nearest Neighbor (LMNN) [20] and Metric Learning by Collapsing Classes (MCC) [7]. Figure 5 shows that, in general, our proposed network achieved the best performance on both data sets. In Tables 1 and 2, after sampling some ranks, it is noteworthy that our method shows a slightly lower performance than MCC [7] and DFLRDC [5], at top rank 15, and than DFLRDC [5] and KFLDA [21], at top rank 20, over i-LIDS.

4. Conclusion

A novel coarse-to-fine deep network architecture was proposed here. The proposed network relies on acquiring the necessary knowledge to identify a person, from a generic-to-specific information, by transferring the learning achieved in each step of the training. The proposed architecture presented the best performance against 12 other state-of-the-art methods. For future work, we are investigating different ways of transfer learning.

References

[1] Uk home office, i-lids multiple camera tracking scenario definition, 2007.

[2] E. Ahmed, M. Jones, and T. Marks. An improved deep learning architecture for person re-identification. *IEEE Conference on Computer Vision and Pattern Recognition*, June 2015.

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

[4] A. Bedagkar-Gala and S. K. Shah. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270–286, 2014.

[5] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015. Discriminative Feature Learning from Big Data for Visual Recognition.

[6] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *CoRR*, 1310.1531, 2013.

[7] A. Globerson and S. Roweis. Metric learning by collapsing classes. In *NIPS*, 2005.

[8] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, 2007.

[9] M. Hirzer, P. Roth, M. Kstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision ECCV 2012*, volume 7577 of *Lecture Notes in Computer Science*, pages 780–793. Springer, 2012.

[10] A. Krizhevsky and G. Hinton. Using very deep autoencoders for content-based image retrieval. In *ESANN*, 2011.

[11] C.-H. Kuo, S. Khamis, and V. Shet. Person re-identification using semantic color names and rankboost. In *IEEE Winter Conference on Applications of Computer Vision*, pages 281–287, 2013.

[12] W. Li and X. Wang. Locally aligned feature transforms across views. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3594–3601, June 2013.

[13] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, June 2014.

[14] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[15] P. Luo, X. Wang, and X. Tang. Hierarchical face parsing via deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2480–2487. IEEE Computer Society, 2012.

[16] A. Ma, P. Yuen, and L. Jiawei. Domain transfer support vector ranking for person re-identification without target camera label information. In *IEEE International Conference on Computer Vision*, pages 3567–3574, 2013.

[17] J. Masci, U. Meier, D. Cirean, and J. Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In T. Honkela, W. Duch, M. Girolami, and S. Kaski, editors, *International Conference on Artificial Neural Networks and Machine Learning*, volume 6791 of *Lecture Notes in Computer Science*, pages 52–59. Springer, 2011.

[18] J. Ngiam, A. Khosla, M. Kim, J. N. amd Honglak Lee, and A. Y. Ng. Multimodal deep learning. In *International Conference on Machine Learning*, 2011.

[19] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408, 2010.

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

[21] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *Computer Vision - ECCV 2014*, pages 1–16. Springer, 2014.

[22] D. Yi, Z. Lei, and S. Z. Li. Deep metric learning for practical person re-identification. *CoRR*, 1407.4979, 2014.

[23] X. Zeng, W. Ouyang, and X. Wang. Multi-stage contextual deep learning for pedestrian detection. In *IEEE International Conference on Computer Vision*, pages 121–128, 2013.

[24] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity-preserving face space. In *IEEE International Conference on Computer Vision*, pages 113–120, 2013.