Person Re-identification based on Deep Multi-instance Learning

Domonkos Varga*†, Tamás Szirányi*‡
*MTA SZTAKI, Institute for Computer Science and Control
{varga.domonkos, sziranyi.tamas}@sztaki.mta.hu
†Budapest University of Technology and Economics, Department of Networked Systems and Services
‡Budapest University of Technology and Economics, Department of Material Handling and Logistics Systems

***Abstract*—Person re-identification is one of the widely studied research topic in the fields of computer vision and pattern recognition. In this paper, we present a deep multi-instance learning approach for person re-identification. Since most publicly available databases for pedestrian re-identification are not enough big, over-fitting problems occur in deep learning architectures. To tackle this problem, person re-identification is expressed as a deep multi-instance learning issue. Therefore, a multi-scale feature learning process is introduced which is driven by optimizing a novel cost function. We report on experiments and comparisons to other state-of-the-art algorithms using publicly available databases such as VIPeR and ETHZ.**

I. Introduction

Person re-identification has captured severe attention in the computer vision and pattern recognition community in recent decades. The goal of person re-identification is to determine whether two pedestrian images from two different camera views without common field of view possess the same identity or not. It is a challenging task due to the illumination, scale, pose, and occlusion that can change across viewpoints. In contrast with pedestrian detection, the main challenge in person re-identification is the high inter-class similarities rather than intra-class similarities.

On the other hand, person re-identification is an essential part of intelligent video surveillance systems. It is a necessary objective to help security staff to monitor the pedestrian's movement and behavior at a broad range but with low costs. In a regular real-world application, a gallery set of pedestrian images is given and the goal is to match a new probe image with one of those individuals in the gallery set.

In this paper, our goal is to introduce our person re-identification system based on deep multi-instance learning. Traditional algorithms in person re-identification are usually based on hand-crafted feature extraction and distance metric learning. Deep learning architectures have captured severe attention since 2012 when Alex Krizhevsky and his colleagues [1] used a Convolutional Neural Network to win ImageNet challenge. Since that breakthrough, CNNs have been applied to various computer vision tasks such as pedestrian detection [2], removing phantom objects from point clouds [3], grayscale image colorization [4], etc.

Because of the small person re-identification datasets, the problem of over-fitting often appears in the case of deep learning architectures. To overcome this problem, we propose a deep multi-instance approach in this paper.

We outline the main contributions of this work as follows. A deep architecture is presented which incorporates the two phases of person re-identification into a single architecture. Moreover, a novel convolutional feature representation is suggested with a novel loss function in order to facilitate the optimization of a Siamese-like CNN [5].

The rest of this paper is organized as follows. In Section II, the related and previous works are reviewed. We describe the proposed person re-identification algorithm in Section III. Section IV shows experimental results and analysis. We draw the conclusions in Section V.

II. Related Work

Broadly speaking, person re-identification algorithms can be divided into two groups. The first one primarily concentrates on hand-crafted feature extraction and distance learning. The second one applies different deep architectures in order to learn feature representation and distance metrics in a unified framework. In the rest of this section, we review first methods based on hand-crafted features then deep architectures.

Early papers mainly concentrate on the construction of effective feature representation. Numerous feature vectors are used or proposed for person re-identification, many of them are lent from pedestrian detection systems [6]. Examples of hand-crafted features incorporates color histograms, texture histograms [8], gradient histograms, symmetry-driven local features [9], and the combination of aforesaid features. The approach of Li and Wang [7] automatically partitions the image spaces of two different camera views into subregions and learns a different feature transform for a pair of configurations. Four types of features were combined together such as LBP, HSV color histogram, Gabor features, and HOG features. Bak et al. [10] introduced an appearance model based on spatial covariance regions extracted from human body parts to cope with pose variations. Similarly, Cheng et al. [11] proposed a method based on Pictorial Structures which was improved by modeling the common appearance of a given person within multiple images.

In additon to elementary and complex features, some specialized representations have been also proposed. For instance, Implicit Shape Models [12], Spin Image [13], Panoramic Maps

ISBN 978-0-9928626-7-1

[14], or Bag-of-Words based descriptors [15] were employed to person re-identification.

Another line of works concentrate on metric learning. Weinberger et al. [16] introduced an algorithm that utilizes $k-$Nearest Neighbors classification by minimizing the distance between each training sample and its K nearest same labelled neighbors, while maximizing the distance from all the other samples. Zheng et al. [17] proposed a Probabilitic Relative Distance Comparison (PRDC) method for person-reidentification in order to maximize the likelihood of true matches having smaller distance than that of a mistaken match pair. Li et al. [18] introduced an approach that uses transferred metric optimal for different candidate sets instead of learning a general metric.

Recently, works have appeared that learns feature representation and similarity metric from raw image data utilizing deep learning techniques. Yi et al. [19] applied a Siamese CNN with two symmetrical, independent sub-networks which are connected by Cosine function. Wu et al. [20] introduced PersonNet which takes a pair RGB images as input that is passed through a pile of convolutional layers, a matching layer, and higher layers estimating relationships between them, and gives back a similarity value.

III. Our approach

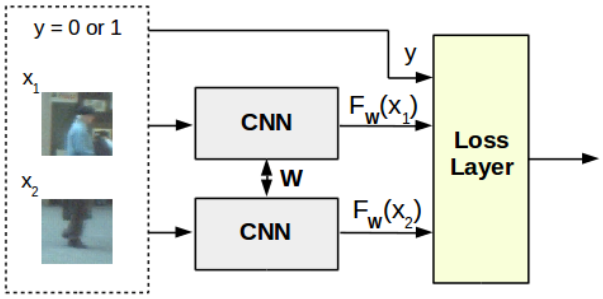
The structure of the proposed person re-identification algorithm can be seen in Figure 1. Our main goal is to construct a feature space where feature vectors extracted from the same identity lie nearby in the feature space, while feature vectors extracted from different identities lie far away from each other. To this end a novel deep architecture is proposed which is driven by a novel loss function.

Image feature learning and distance metric learning are fused together by approximating the nonlinear function $F_{\mathbf{W}}(\cdot)$. In order to approximate this function, we propose a Siamese-like structure which incorporates two Convolutional Neural Networks with multi-scale convolution. Multi-scale convolution facilitates to capture more discriminative features because persons can occur in many scales. The whole structure is driven by a novel loss function.

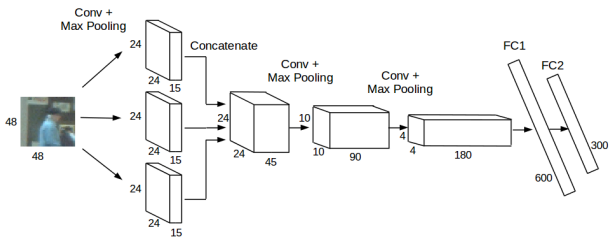
Our architecture contains two Convolutional Neural Networks (CNN) which have the same structure and share the same parameters [5]. These CNNs consist of five convolutional layers followed by max pooling operator, one concatenating layer, and two fully-connected layers. The first three convolutional layers are parallel whose filters are 3×3, 5×5, and 7×7 correspondingly. In order to get the same output size, the input data are padded with circular repetition of elements within the dimension. The sizes of the filters in the two other convolutional layers are 5×5 and 3×3, respectively. In our system, $ReLU$ [1] is utilized as an activation function because it can speed up the convergence of the training process.

A. Deep multi-instance learning

Above all, we have to create our database. Let denote I_c^p the image of person p from the point of view of camera c. The



(a) Overall architecture.



(b) The architecture of the CNN.

Fig. 1: The proposed architecture.

multi-instance set is created by sampling I_c^p with the help of a scanning rectangle $(0.4h, w)$ where h and w stand for the height and the width of the person image. The scanning rectangle is moved downwards until it reaches the bottom of the image in such a way that five instances can be sampled (see Figure 2). Formally, the instance set of person p from the point of view of camera c can be defined as $S_c^p = \{R_c^{p,i} | i = 1, 2, ..., 5\}$ where $R_c^{p,i}$ stands for ith sample from I_c^p.

These samples are divided further into six sub-samples which are in different instance spaces. We just take the left, middle, and right chunks and their mirrored versions (six chunks on the whole) to produce a so-called bag. In Figure 2, the three chunks are in blue squares. These three chunks and their mirrored versions form the first bag. In total, five bags are extracted from a person's image. Formally, $B_c^{p,i,j} = \{x_c^{p,i,j,k} | j = 1, 2, ..., 5, k = 1, 2, ..., 6\}$ where $B_c^{p,i,j}$ is the jth bag of $R_c^{p,i}$ and $x_c^{p,i,j,k}$ is the kth chunk in bag $B_c^{p,i,j}$. Hence, five instance spaces are characterized as following:

$$\mathscr{L}_i = \bigcup_{j,p,c} B_c^{p,i,j}, \text{ where } \begin{cases} \{i, j\} = 1, ..., 5 \\ p = 1, ..., P \\ c = 1, ..., C. \end{cases} \quad (1)$$

In the above equation, P stands for the number of person identity and C denotes the number of different cameras in the system.

The next step is the approximation of a nonlinear function for each instance space in order to jointly model feature representation and distance metric. Let $\mathscr{P} = \{(x_1, x_2, y)\}_{i=1}^N$ be a set of sample pairs, where x_1 and x_2 are image chunks from the same instance space but from different camera views, y is 1 if x_1 and x_2 belongs to the same identity and 0 if not, and N is the number of training pairs. As we mentioned $\mathbf{W}$

ISBN 978-0-9928626-7-1 © EURASIP 2017

Fig. 2: The sampling process. The scanning rectangle is moved downwards along the dashed red lines. Each instance is sampled further into bags.

stands for the parameters of the nonlinear function and $F_{\mathbf{W}}(\cdot)$ is the nonlinear function.

The initial point of the loss function's deduction is the following constraint that we would like to satisfy:

$$||F_{\mathbf{W}}(x_1) - F_{\mathbf{W}}(x_2)||_2|_{y=1} < ||F_{\mathbf{W}}(x_1) - F_{\mathbf{W}}(x_2)||_2|_{y=0}. \quad (2)$$

This means that the loss produced by a true pair ($y = 1$) must be always less than the loss produced by a mistaken pair ($y = 0$). To satisfy this constraint, the loss produced by a true pair is construed as:

$$l|_{y=1} = \frac{1}{2}||F_{\mathbf{W}}(x_1) - F_{\mathbf{W}}(x_2)||_2^2. \quad (3)$$

On the other hand, the loss produced by a mistaken pair is determined as:

$$l|_{y=0} = \frac{1}{2}max(0, m - ||F_{\mathbf{W}}(x_1) - F_{\mathbf{W}}(x_2)||_2)^2, \quad (4)$$

where m is a margin and equated to 1. Consequently, the overall loss for one pair is determined as:

$$\begin{aligned} l(x_1, x_2, y) &= y \cdot l|_{y=1} + (1-y) \cdot l|_{y=0} = \\ &= y \cdot \frac{1}{2}||F_{\mathbf{W}}(x_1) - F_{\mathbf{W}}(x_2)||_2^2 + \\ &+ (1-y) \cdot \frac{1}{2}max(0, m - ||F_{\mathbf{W}}(x_1) - F_{\mathbf{W}}(x_2)||_2)^2. \end{aligned} \quad (5)$$

In the long run, the minimizing of the loss compels the model to map the members of a coherent pair near to each other in the feature space. Hence, our model is able to identify the common features that are present in distinct samples of the same identity. The loss function for the whole training sample pairs can be written as:

$$L = \sum_{i=1}^{N} l((x_1, x_2, y)_i). \quad (6)$$

To minimize the above loss fucntion, we optimize it using stochastic gradient descent (SGD) with batches. The partial derivatives of the loss are:

$$\frac{\partial L}{\partial \mathbf{W}}\bigg|_{y=1} = ||F_{\mathbf{W}}(x_1) - F_{\mathbf{W}}(x_2)|| \frac{\partial ||F_{\mathbf{W}}(x_1) - F_{\mathbf{W}}(x_2)||}{\partial \mathbf{W}}, \quad (7)$$

$$\frac{\partial L}{\partial \mathbf{W}}\bigg|_{y=0} = \begin{cases} 0, \\ \text{if } ||F_{\mathbf{W}}(x_1) - F_{\mathbf{W}}(x_2)|| \geqq m \\ (||F_{\mathbf{W}}(x_1) - F_{\mathbf{W}}(x_2)|| - m) \frac{\partial ||F_{\mathbf{W}}(x_1) - F_{\mathbf{W}}(x_2)||}{\partial \mathbf{W}}, \\ \text{otherwise.} \end{cases} \quad (8)$$

B. Training

First, we represent each image as a multi-instance set containing 6 bags which are in distinct instance spaces. For each instance space, a set of sample pairs for training is constructed. If the ratio between true pairs and mistaken pairs are strongly imbalanced, reducing the loss would conduct to the breakdown of the training process by equalizing all parameters to zero. That is why, we set this ratio to 0.2. This means that the number of mistaken pairs is five times greater than the number of true pairs. After generating the sets of pairs, 5 models are trained separately for each instance space.

We have adopted VIPeR [23] database (consists of 632 image pairs) as our training data because it is the most challenging collection of pedestrians in outdoor environment. We have used 380 identities for training and 252 identities for testing. The system was trained on VIPeR and was tested on VIPeR and ETHZ [24]. We report on experiments in Section IV.

C. Metrics

As we described images in the test set are present in the form of multi-instance sets. As we mentioned a model is trained for each instance space and every instance is projected to one point in the multi-instance feature space. Accordingly, there are five points in every single feature space which belongs to one identity in one camera view. We determine the distance between a gallery (G) instance and a probe (P) instance as:

$$d(x_G^{p,i,j,k}, x_P^{q,i,j,k}) = ||F_{\mathbf{W}}(x_G^{p,i,j,k}) - F_{\mathbf{W}}(x_P^{q,i,j,k})||_2. \quad (9)$$

Then Chamfer distance [21] is applied to figure out the distance between a bag in the probe set and another one in the gallery set. The formula of the distance is derived as followings:

$$d_1(x_G^{p,i,j,k}, B_P^{q,i,j}) = min\{d(x_G^{p,i,j,k}, x_P^{q,i,j,k})|k = 1, ..., 6\}, \quad (10)$$

$$d_2(x_P^{q,i,j,k}, B_G^{p,i,j}) = min\{d(x_P^{q,i,j,k}, x_G^{p,i,j,k})|k = 1, ..., 6\}, \quad (11)$$

ISBN 978-0-9928626-7-1 © EURASIP 2017

$$D_{ch}(B_G^{p,i,j}, B_P^{q,i,j}) = \frac{1}{2 \times 5}(\sum_{j=1}^{6} d_1(x_G^{p,i,j,k}, B_P^{q,i,j}) + \\ + \sum_{j=1}^{6} d_2(x_P^{q,i,j,k}, B_G^{p,i,j})). \quad (12)$$

Note that Eq. 12 consists of Eq. 10 as first term and Eq. 11 as second term. Eq. 10 is the sum of the minimal distances between an instance in the gallery set and every instance in the probe image. Eq. 11 is just the opposite. These results are integrated. The distance between a probe image and gallery images is computed as:

$$D(S_G^p, S_P^q) = \sum_{i=1}^{5} D_{ch}(B_G^{p,i,j}, B_P^{q,i,j}). \quad (13)$$

Finally, the gallery set is ordered against the probe image with respect to Eq. 13.

IV. Experimental results

Our system was implemented with the help of Keras[1] deep learning library [22]. In the training process, we used SGD with initial learning rate of 0.01. If the training loss stops reducing, we divide the learning rate by ten. We evaluate the performance of the proposed system and other state-of-the-art algorithms on VIPeR[2] [23] and ETHZ[3] [24] datasets.

We compared our method to several state-of-the-art methods such as ELF [8], salience [26], RPML [27], LMNN [28], SCR [10], PRDC [17], Transfer [18], and DML [19]. DML is based on deep metric learning, while the other algorithms apply hand-crafted feature representations and metric learning.

Cumulative Match Characteristics (CMC) [25] was used to evaluate the performance of a person re-identification algorithm. CMC gives us how well an identification framework ranks the identitites in a database with respect to a new image.

Table I demonstrates that our system outperforms most of the state-of-the-art methods. Although our top 10 and 20 are somewhat lower than DML [19]. On the other hand, our method outperforms DML [19] by 3.6% in top 1 and by 0.1% in top 5, respectively. We attribute these improvements to the observation that our method avoids better the overfitting problem. Furthermore, with the help of the multi-scale convolution our method is able to capture more discriminative features.

As we mentioned, our algorithm was trained on VIPeR. In order to study our system's generalization ability, we carried out comparison on ETHZ person re-identification database. The results are summarized in Table II. As one can see, our method outperforms all the other state-of-the-art algorithms in this case. Consequently, our algorithm has a good ability for generalization.

To evaluate the robustness of our system, we test it on artificially modified VIPeR test images. Namely, we add salt & pepper noise with different noise density (see Figure 4). Then we measured the Cumulative Matching Rate with respect to the noise's density. The results are summarized in Table III. We can see that our system is rather robust to salt & pepper noise and the significant disimprovement of the performance begins approximately at 0.2 noise density.

[1]https://keras.io/

[2]Available: https://vision.soe.ucsc.edu/?q=node/178

[3]Available: http://www.ssig.dcc.ufmg.br/ethz-dataset-for-appearance-based-modeling/

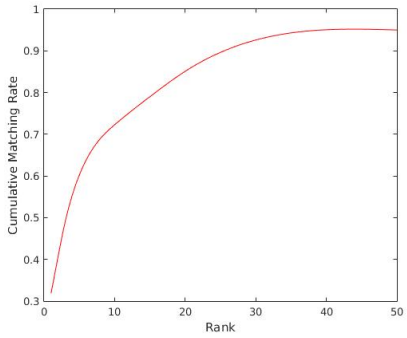


Fig. 3: Performance of the proposed system on VIPeR. The vertical axis represents the Cumulative Matching Rate and the horizontal axis shows rank between 1 and 50.

TABLE I: Comparison to other state-of-the-art methods on VIPeR. The table demonstrates the top ranked cumulative matching rate (%) with respect to the rank. If the data cannot be purchased, we indicate it by '-'. The best result is typed by **bold**, the second best result is typed by *italic*.

Method	$rank = 1$	$rank = 5$	$rank = 10$	$rank = 20$
ELF [8]	12.0	31.0	41.0	58.0
salience [26]	26.7	50.7	62.4	76.4
RPML [27]	27.0	-	69.0	-
LMNN [28]	6.3	19.7	32.6	52.3
SCR [10]	7.3	18.3	39.4	50.5
PRDC [17]	15.7	39.3	53.7	71
Transfer [18]	15.9	38.1	52.4	68.0
DML [19]	*28.2*	*60.0*	**73.4**	**86.4**
Ours	**31.8**	**60.1**	*72.2*	*85.1*

TABLE II: Comparison to other state-of-the-art methods on ETHZ. The table demonstrates the top ranked cumulative matching rate (%) with respect to the rank. The number of identities in the test set is 70. If the data cannot be purchased, we indicate it by '-'. The best result is typed by **bold**, the second best result is typed by *italic*.

Method	$r = 1$	$r = 5$	$r = 10$	$r = 20$
ELF [8]	64.3	82.1	90.1	95.2
salience [26]	60.4	67.9	86.0	91.9
RPML [27]	53.6	-	76.6	-
LMNN [28]	57.5	78.2	86.3	92.8
SCR [10]	48.8	62.7	78.3	79.1
PRDC [17]	58.7	77.6	85.2	91.8
Transfer [18]	56.2	70.4	83.7	86.8
DML [19]	80.5	89.0	92.4	97.3
Ours	**80.9**	**89.8**	**93.9**	**97.8**

ISBN 978-0-9928626-7-1 © EURASIP 2017

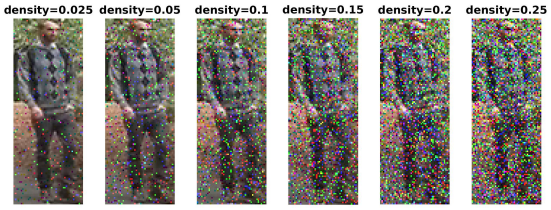


Fig. 4: Some images with different noise density. Noise density denotes the average ratio of pixels which are replaced by salt & pepper noise.

TABLE III: Performance evaluation in the presence of salt & pepper noise. Noise density is abbreviated by $'dens.'$.

	$r = 1$	$r = 5$	$r = 10$	$r = 20$
$dens. = 0$	31.8	60.1	72.2	85.1
$dens. = 0.025$	30.9	59.1	72.0	84.1
$dens. = 0.05$	30.5	58.7	71.9	83.9
$dens. = 0.1$	28.4	58.2	70.4	83.1
$dens. = 0.15$	25.3	53.2	65.5	78.1
$dens. = 0.2$	10.8	44.7	49.5	61.6
$dens. = 0.25$	6.7	32.5	44.7	60.3

V. Conclusion

We have introduced a novel person re-identification system based on multi-instance learning. In order to predict an identity efficiently, we have applied multi-instance deep learning and a novel architecture incorporating multi-scale convolutional layers. We have reported on experiments and comparisons to other state-of-the-art algorithms using publicly available databases such as VIPeR and ETHZ. In our future work, we want to do further measurements to test the robustness of the learned features. This may include robustness test against occlusion, background variation, and lighting variation.

Acknowledgment

The research was supported by the Hungarian Scientific Research Fund (No. OTKA 120499). We are very thankful to Levente Kovács for helping us with professional advices in high-performance computing.

References

[1] A. Krizhevsky, I. Sutskever, and G.E. Hinton. Imagenet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 1097–1105, 2012.

[2] E. Bochinski, V. Eiselein, and T. Sikora. Training a convolutional neural network for multi-class object detection using solely virtual world data. *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 278–285, 2016.

[3] B. Nagy and C. Benedek. 3D CNN Based Phantom Object Removing from Mobile Laser Scanning Data. *International Joint Conference on Neural Networks*, 4429–4435, 2017.

[4] X. Liang, Z. Su, Y. Xiao, J. Guo, and X. Luo. Deep patch-wise colorization model for grayscale images. *SIGGRAPH ASIA 2016 Technical Briefs*, 13, 2016.

[5] R.R. Varior, M. Haloi, and G. Wang. Gated Siamese Convolutional Neural Network Architecture for Human Re-identification. *European Conference on Computer Vision*, 791–808, 2016.

[6] A. Guo, M. Xu, F. Ran, and Q. Wang. A Real-time Pedestrian Detection System in Street Scene. *International Journal on Smart Sensing and Intelligent Systems*, **9**(3):1592–1613, 2016.

[7] W. Li and X. Wang. Locally Aligned Feature Transforms across Views. *IEEE Conference on Computer Vision and Pattern Recognition*, 3594–3601, 2013.

[8] D. Gray and H. Tao. Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. *European Conference on Computer Vision*, 262–275, 2008.

[9] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person Re-identification by Symmetry-driven Accumulation of Local Features. *IEEE Conference on Computer Vision and Pattern Recognition*, 2360–2367, 2010.

[10] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Person Re-identification Using Spatial Covariance Regions of Human Body Parts. *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 435–440, 2010.

[11] D.S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom Pictorial Structures for Re-identification. *British Machine Vision Conference*, 2010.

[12] J. Kai, C. Bodensteiner, and M. Arens. Person Re-identification in Multi-camera Networks. *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 55–61, 2011.

[13] K. Aziz, D. Merad, and B. Fertil. Person Re-identification using Appearance Classification. *International Conference Image Analysis and Recognition*, 170–179, 2011.

[14] T. Gandhi and M. Trivedi. Person Tracking and Re-identification: Introducing Panoramic Appearance Map (PAM) for Feature Representation. *Machine Vision and Applications*, **18**(3):207–220, 2007.

[15] W. Zheng, S. Gong, and T. Xiang. Associating Groups of People. *British Machine Vision Conference*, 2009.

[16] K.Q. Weinberger and L.K. Saul. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *Journal of Machine Learning Research*, **10**:207–244, 2009.

[17] W.S. Zheng, S. Gong, and T. Xiang. Reidentification by Relative Distance Comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**(3):653–668, 2013.

[18] W. Li, R. Zhao, and X. Wang. Human Reidentification with Transferred Metric Learning. *Asian Conference on Computer Vision*, 31–44, 2012.

[19] D. Yi, Z. Lei, S. Liao, and S.Z. Li. Deep Metric Learning for Person Re-identification. *International Conference on Pattern Recognition*, 34–39, 2014.

[20] L. Wu, C. Shen, and A. Hengel. PersonNet: Person Re-identification with Deep Convolutional Neural Networks. *arXiv preprint arXiv:1601.07255*, 2016.

[21] H.G. Barrow, J.M. Tenenbaum, R.C. Bolles, and H.C. Wolf. Parametric Correspondence and Chamfer Matching: Two New Techniques for Image Matching. *Proc. International Joint Conference on Artificial Intelligence*, 1977.

[22] F. Chollet. Keras. *https://github.com/fchollet/keras*, 2015.

[23] D. Gray, S. Brennan, and H. Tao. Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. *IEEE International Workshop on Performance Evaluation for Tracking and Surveillance*, 2007.

[24] W.R. Schwartz and L.S. Davis Learning Discriminative Appearance-based Models using Partial Least Squares. *XXII Brazilian Symposium on Computer Graphics and Image Processing*, 322–329, 2009.

[25] A. Kale, A. Sundaresan, A.N. Rajagopalan, N.P. Cuntoor, A.K. Roy-Chowdhury, V. Krüger, and R. Chellappa. Identification of Humans Using Gait. *IEEE Transactions on Image Processing*, **13**(9):1163–1173, 2004.

[26] R. Zhao, W. Ouyang, and X. Wang. Unsupervised Salience Learning for Person Re-identification. *IEEE Conference on Computer Vision and Pattern Recognition*, 3586–3593, 2013.

[27] M. Hirzer, P.M. Roth, M. Köstinger, and H. Bischof. Relaxed Pairwise Learned Metric for Person Re-identification. *European Conference on Computer Vision*, 780–793, 2012.

[28] W.S. Zheng, S. Gong, and T. Xiang. Person Re-identification by Probabilistic Relative Distance Comparison. *IEEE Conference on Computer Vision and Pattern Recognition*, 649–656, 2011.

ISBN 978-0-9928626-7-1