https://doi.org/10.2352/ISSN.2470-1173.2017.13.IPAS-211

Feature representation learning by sparse neural network for multi-camera person re-identification

S. Makov., A. Minaev., A. Nikitin, V. Voronin, E. Semenishchev, V. Marchuk; Don State Technical university, Dept. of Radio-Electronics Systems, Gagarina 1, Rostov on Don, Russian Federation

Abstract

In this paper we study ability to sparse the neural networks which are used in the task of person re-identification in multi-camera CCTV systems. Sparse neural network allows significant reducing of the computation complexity. It means increasing of processing speed for huge volume of data. The main idea of our research is decreasing of computation complexity with simultaneous preserving of neural network efficiency in the task of person re-identification. The paper consists of 4 parts and conclusion. The first part – introduction, describes sphere of person re-identification applying and the key problems in this field. The second part – related work, describes main state-of-the-art approaches related to person re-identification techniques. In the third part called proposed approach, we formulate technique of sparse neural network learning. The fourth part – experiment results, describes experiment conditions, constrains, training datasets, and results. In the conclusion we make our proposal on new technique usage.

Introduction

Nowadays many objects have multi-camera CCTV systems. The main goal of cameras quantity growing is providing more view zones in public places to control. At same time there are other goals:

- counting of authentic or non-recurring byers in big shops and molls. It allows us to make detailed marketing researches, measure level of interest to particular goods according with target groups:
- searching for criminals on archived video-data obtained from cameras around the place of crime scene or accident place. It allows to take less time for identification of offender;
- controlling of unauthorized access to some territories, zones or objects.

All these applications require large number of security system operators. Otherwise all this data in the best case are just stored in some video-archive for not very long time. That's why we should use automatic systems to help operators – attract their attention to some concrete person on any camera. Also we should have ability of fast search in video-archive by sample of person image and re-identify him on records from other cameras. Thus the person re-identification problem is very important.

Growing of number of cameras in CCTVs leads to increasing of load on computation systems. Thus, we need efficient and easy to compute algorithms of person re-identification.

Direct comparison of images does not allow us to decide that at images are showed same or different person. The image features are used for their comparison. Feature is a vector of values. These values to be invariant from many factors such as scale, rotation, brightness, contrast, color saturation and balance, etc. To re-identify people from different cameras we also have to provide invariant or weak dependence of feature value from person pose or camera shooting angle.

There are many techniques of features obtaining. Most of them are hand-crafted features like GIST, CIFT, HoG, RIFT, Textons etc. They allow us to search similar images [1] with very good efficiency but in case of changing form of object or changing camera position their efficiency is very low.

In the present time, most popular technique for features obtaining is neural networks. The neural networks allow finding very complex transformation based on real data to obtain image features with required properties. For example, we can apply neural network in similar task – image retrieval [2].

The main criterion of feature efficiency is discriminative generalizing and ability. In one hand, we should have near values for same persons and in other hand for not same persons the values should be fare from each other. For computing distance between pair of vector many metrics are used. It depends on concrete problems and features.

Very often good features have very big computation complexity. In practice we should have ability to get result of re-identification algorithm on-the-fly. Thus other important criteria of feature efficiency are time to compute feature and time to compare it with feature of search sample.

The goal of the research is to minimize the computation complexity with simultaneous preserving of its discriminative generalizing and ability.

Related Work

At the present time re-identification task is very relevant and that there is a large amount of research related to this topic.

In paper of De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, Nanning Zheng «Person Re-Identification by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function» [3], is described method, which calculates cross-input differences to determine the correlation between the two input images. Higher level features are calculated using mid-level features. The main idea of this paper is illustrated on the figure 1.

The paper proposes to use four methods for re-identification: «Pairwise Constrained» - regularization component analysis, «Local Fisher» - core of discriminant analysis, Marginal Fisher Analysis (MFA) and Locally Adaptive Decision Functions(LADF), which is used in combination with different sized sets of histograms. This method has following disadvantages: there are a large number of neurons, and thus the calculation complexity, the efficiency of the method is 2-3% better in comparison with analogues.

Also, in research of McLaughlin N., Martinez del Rincon J. & Miller P. «Recurrent Convolutional Network for Video-based Person Re-Identification» [4], is described system built based on a few low-level hand-crafted and high level features obtained by CNN. Then, the two optimization algorithms are formed to

optimize the measures evaluation which commonly used in re-ID, method Cumulative Matching Characteristic (CMC). This method has the disadvantage that consists of low efficiency of re-ID in case of occlusions present on the pictures.

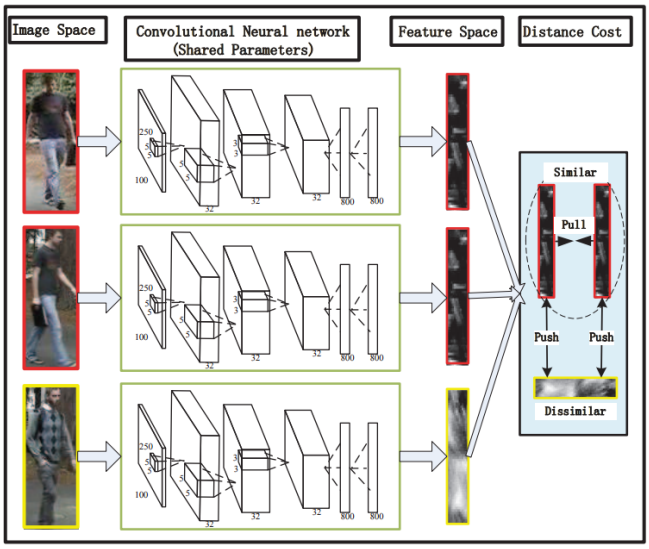

Figure 1. Training framework for «Person Re-Identification by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function» method.

In method, Yeong-Jun Cho and Kuk-Jin Yoon describes in paper «Improving Person Re-identification via Pose-aware Multi-Shot Matching» [5] Identification is done by analyzing the field of view of the camera and the human posture, called Pose-aware Multi-Shot Matching (PaMM), which gives an estimate posture roughly and effectively coordinates the two channels of video stream. Then we generate multi-pose model containing four functions extracted from the four clusters of images, grouped by human postures (i.e., front, right, back, left). The main idea of this paper is illustrated on the figure 2.

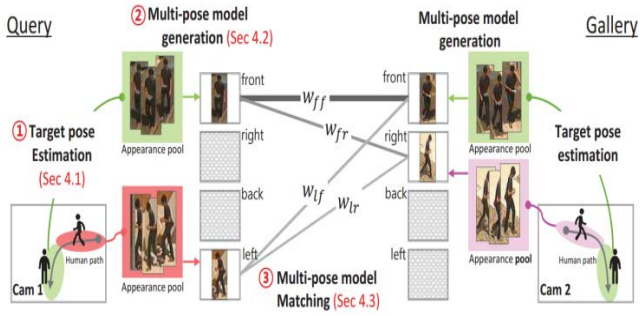

Figure 2. Processing multi-pose model for person re-identification

After creating the multi-pose model, they calculate the relation between the model of multi-pose weighted sum based on a matching weight. Weak sides of the method are using of the Siamese neural network to learning and analysis of two synchronized video streams, which increases the amount of calculation and consequently slows down the processing speed.

In paper of Tong Xiao Hongsheng, Li Wanli, Ouyang Xiaogang Wang «Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification» [6] is described a method of training the neural network on features of several domains, which are effective every trained domain. The method is generalized to the problems with data sets of several domains. The main idea of this paper is illustrated on the figure 3.

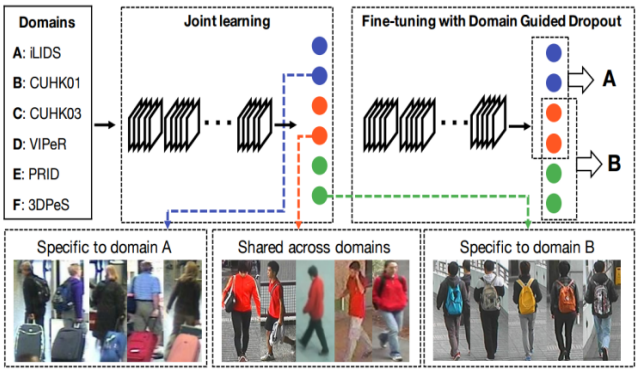

Figure 3. Analyze the effectiveness of each neuron on each domain.

Specific of training is a large-scale classification dataset, are mixed together and all domains of learning convolution Neural Network (CNN). CNN model consists of several BN-Inception [7] modules and their capacities are well suited work mixed dataset. CNN model provides a strong enough base level, but a simple joint training scheme does not give the full variations in multiple domains. Neurons, which are effective for one domain may be useless for another domain because of the presence of domain characteristics such as brightness, color saturation, baggage on people, occlusions etc. Thus, the method uses complex algorithms for processing, but with the change in conditions of efficiency dramatic decreased.

The article of Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, Yoichi Sato «Hierarchical Gaussian Descriptor for Person Re-Identification» [8] describes the usage of new descriptors for image domain based on hierarchical Gaussian distribution functions. Retrieves the local areas within the region and considers the region as a set of local spots. The main idea of this paper is illustrated on the figure 4.

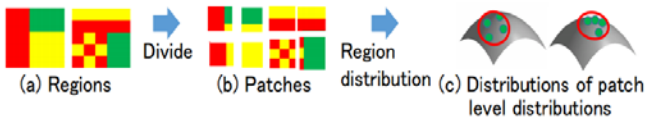

Figure 4. Hierarchal distribution in regions by method of Hierarchical Gaussian Descriptor for Person Re-Identification

The model of the region is described as a set of multiple Gaussian distributions, each of them exist in a local patch. Characteristics set of Gaussians patches again described by a Gaussian distribution to others. Parameters fields with Gaussian distribution are used as a feature vector representation of the regions. But the method has disadvantages. The mean value is not included in every hierarchy. The loss of the mean values is the key problem. This is due to the fact that the wear is typically composed

of a small set of colors in each local area. If we have a similar average color on the different objects it can cause of errors and efficiency decreasing.

For all the above methods the main problem lies in the large computational complexity and thus slow the system, not big advantage in the effectiveness from 2 to 5%. In our research, we decided to focus on reducing the size of the neural network by sparsing that is to increase the speed of work, to ensure ease of calculation, under the same values of the efficiency of both in similar articles. It was based on the article of Suraj Srinivas, Akshayvarun Subramanya, R. Venkatesh Babu «Training Sparse Neural Networks» [9] that describes a way to decrease the number of working neurons using additional coefficients.

Proposed approach

Proposed feature representation learning method is based on receiving feature descriptor of pedestrian using neural network with sparsed weights. Sparsed neural network was researched in [9]. Method, that was proposed there, showed good results, and allowed to reach more than 90% sparsity on dense layers and 80% on convolutional layers and keep high accuracy rates of classification. We decide to fine tune pre-trained network. Researches in creating advanced image recognition algorithms are widely conducted by corporations. In result, there are many datasets appear in open access on the internet. Also they share their results in form of neural network such as ImageNet, GoogLeNet. Such pre-trained image recognition networks already have learned main features on real natural pictures. In this case our task is fine tune existing network in new condition and get rid of redundant weights that make no impact in our case. It is presumed that images of one person from different angles will have similar descriptors. So we basically look for the closest match of descriptor.

To compare descriptor we used several metrics. The simplest one is Euclidian distance between

We used VGG-16, presented in [10] as pre-trained model with convolutional layers initialized with ImageNet [11] weights. To perform re-identification task, network was fine-tuned on datasets designed special for this task CUHK03 [12], CUHK01 [13], Shinpuhkan [14]. During learning network performed classification task, matching pedestrians in dataset to their IDs. After the end of fine-tuning convolutional layers of network learned descriptors, that describe pedestrian from different view angles.

To keep neural network sparsed, we used method, described in work “Training Sparse Neural Networks” [14]. Method involves adding additional variables, called “Gate Variables” in each sparse layer of network. Although these variables can have values between 0 and 1, during the calculation of layer’s output gate variables are sampled to have values 1 or 0 that are multiplied on layer’s weights (Fig.5). By the sampling we mean performing maximum likelihood (ML) draw, that thresholds variables at value 0.5.

Thus weights, which are multiplied on value 0, are no longer involved into further process of calculation, don’t make any influence on network’s result and can be removed.

Changing of gate variables values is provided by adding new expression to the objective function. We define, that complexity of layer of gate variables equals to the sum of all its values.

Denote gate variable as g, than complexity of layer $\| \Phi \|$ will be:

$$\| \Phi \| = \sum_{i=1}^{m} g_i ,$$

where m – number of gate variables in layer.

With complexity given, we can say, that objective function states as follows:

$$\hat{\theta}, \hat{\Phi} = \arg\min_{\theta, \Phi} l(\hat{y}(\theta, \Phi), y) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} g_{i,j}$$

The first part of expression is standard objective function that is used to train weights of neural network.

However, in this case all of the gate variables will be reaching toward decreasing of its values that eventually will set all gate variables below 0.5.

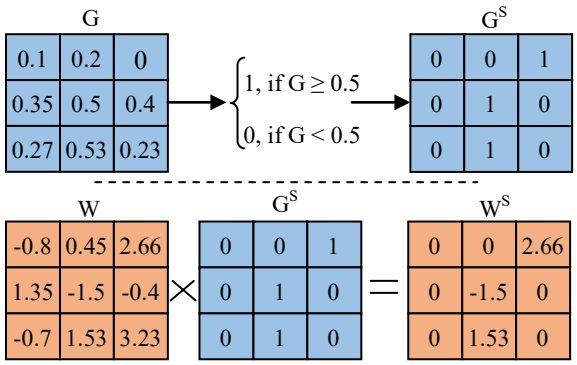

Figure 5. Process of deleting redundant weights. First, we sample gate variable, than multiply them with weights.

Article proposes to use bi-modal regulizer that is given as $w \times (1\text{-}w)$ [15], instead of L_1 and L_2, that do not ensure existing variable above value 0.5. (Fig. 6).

Now we can state objective function as:

$$\hat{\theta}, \hat{\Phi} = \arg\min_{\theta, \Phi} l(\hat{y}(\theta, \Phi), y) + \lambda_1 \sum_{i=1}^{m} \sum_{j=1}^{n_i} g_{i.j}(1 - g_{i,j}) + \lambda_2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} g_{i,j}$$

Using of proposed regulizer allows decreasing number of active neurons, insuring existing of:

- parameters λ_1 and λ_2 are used to regulate the sparsity of network.

- λ_2 should be increased to increase sparsity, while λ_1 stabilizes it.

To sample gate variables during direct flow of data in neural network we round it to the nearest integer number.

At the same time, we use restriction to prevent variables having negative values, as it will influence objective function.

Because sampling function is not differentiable, in proposed method custom gradient function was used. In original paper for this purpose was used identity function, i.e.

$$\frac{\partial g^S}{\partial g} = 1 .$$

We presume that the speed of variable decreasing suppose to depend on the value of the weights *w*. This assumption is based on idea, that weights with small values have weaker impact on final result of convolution network. In this work we propose to use gradient function, which depends on neuron weight. It helps to faster decrease gate variables that have small values. At the same time it keeps learning efficiency of neural network.

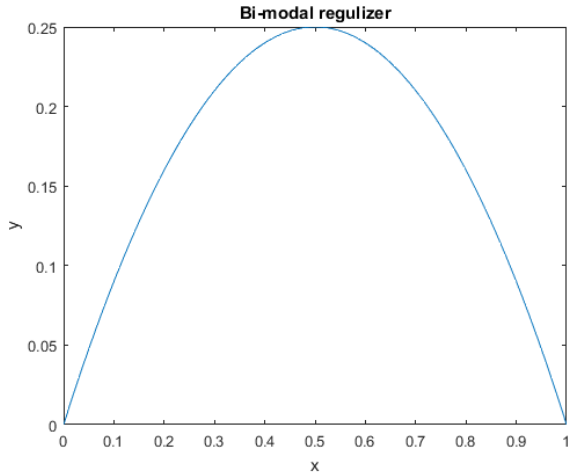

Figure 6. Bi-modal regulizer that was used during training

So we propose to calculate gradient as follows:

$$\frac{\partial g^S}{\partial g} = \frac{w_{\max}}{w},$$

where $w_{\max}$ – maximum value of neuron, w – value of current weight.

We expect, that using of this function will allow us to increase speed of training of gate variables that will increase speed of sparsing.

During experiments of studying re-identification accuracy we used network without top (Dense) trained layers.

Experiments

At this moment the most popular datasets that used to train and test neural networks for re-identification tasks are: CUHK03 (13,164 images of 1,360 pedestrians), CUHK01 (3,884 images of 972 pedestrians), PRID [16] (475 person trajectories from one view and 856 from the other one, with 245 persons appearing in both views), VIPeR [17] (632 pedestrian image pairs, varying illumination conditions), 3DPeS [18] (1012 snapshot of 200 persons). Also should be motioned Shinpuhkan (24 pedestrians) dataset, that despite having less number of pedestrians includes more images of each person filmed from different angles and in different light conditions. Dataset examples are shown on figures 9, 10, 11.

VGG-16: Neural network structure

Name	Input	Output
InputLayer	128, 48, 3	128, 48, 3
SparseConvolution2D	128, 48, 3	128, 48, 64
SparseConvolution2D	128, 48, 64	128, 48, 64
MaxPooling2D	128, 48, 64	64, 24, 64
SparseConvolution2D	64, 24, 64	64, 24, 128
SparseConvolution2D	64, 24, 128	64, 24, 128
MaxPooling2D	64, 24, 128	32, 12, 128
SparseConvolution2D	32, 12, 128	32, 12, 256
SparseConvolution2D	32, 12, 128	32, 12, 128
SparseConvolution2D	32, 12, 128	32, 12, 128
MaxPooling2D	32, 12, 128	16, 6, 256
SparseConvolution2D	16, 6, 256	16, 6, 512
SparseConvolution2D	16, 6, 512	16, 6, 512
SparseConvolution2D	16, 6, 512	16, 6, 512
MaxPooling2D	16, 6, 512	8, 3, 512
SparseConvolution2D	8, 3, 512	8, 3, 512
SparseConvolution2D	8, 3, 512	8, 3, 512
SparseConvolution2D	8, 3, 512	8, 3, 512
MaxPooling2D	8, 3, 512	4, 1, 512
Flatten	4, 1, 512	2048

Implementation Details:

To fine-tune convolution network we used Keras framework with TensorFlow backend. Training was performed on batches of 32 images. During training the task of neural network was to classify pedestrians and match each one with corresponding ID. As result, network learned features that best fit to describe pedestrians, photographed from different angles.

We initialize neural network with weights of pre-trained model for image classification with weights trained on ImageNet. That allowed us to use smaller training dataset and faster reach high accuracy rates.

To faster reach acceptable levels of sparsity we initialize gate variables with constant value higher, but close to 0.5. This way in the beginning of training all weights were involved into calculating process.

During experiments a conclusion was made, that the descending speed of gate variables depends on chosen learning rate and values of global variables λ_1 and λ_2. As learning rate also influences on weights of neural network during training, it is preferably to change values λ_1 and λ_2.

As we can see from the plot, the training is performed much faster, compared to increasing of sparsity (Fig 7.).

However if we use proposed gradient function, the process of sparsifying goes much faster (Fig. 8).

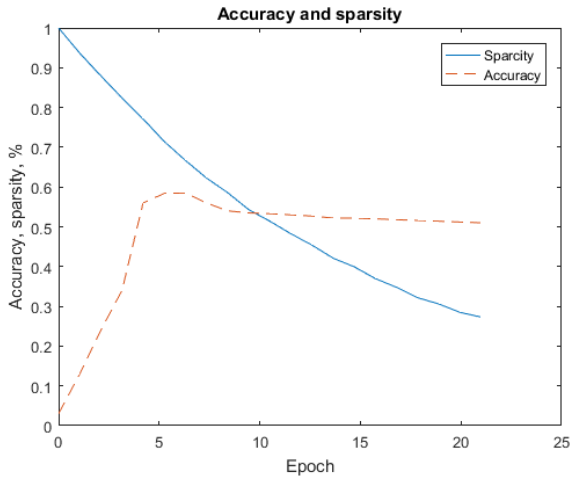

Figure 7. Accuracy rate grows much faster than sparsity of neural network

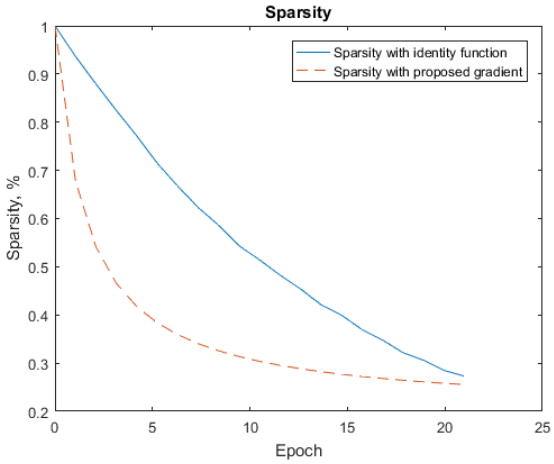

Figure 8. Proposed method increases sparsifying rate

After training fully connected layers of neural network are no longer used, and output of the last convolution layer (after Max Pooling) is used as pedestrian descriptor.

Figure 9. CHUK03 dataset examples

Figure 10. Shinpuhkan2014 dataset examples

Figure 11. CHUK01 dataset examples

Testing conditions and results

To test the efficiency of the proposed model we used the following approach:

Using the neural network we receive descriptor of a person filmed from one angle and with Euclidian distance metric we look for the closest descriptor calculated for images, taken from another angle. Proposed method showed results on the state of art level, while using neural network with lower number of parameters.

Top-1 accuracy rate for CUHK03 and CUHK01 datasets [19-22] :

Method\Dataset	CUHK03	CUHK1
IDLA	54.74%	65%
DNS	62.55%	69.09%
EDM	61.32%	86.59%
CAN	65.65%	81.04%
Our method	48.52%	52.32%

Accuracy rates for Ranks 1,5,10,15 and 20 showed on figure 12 for 40% sparsity.

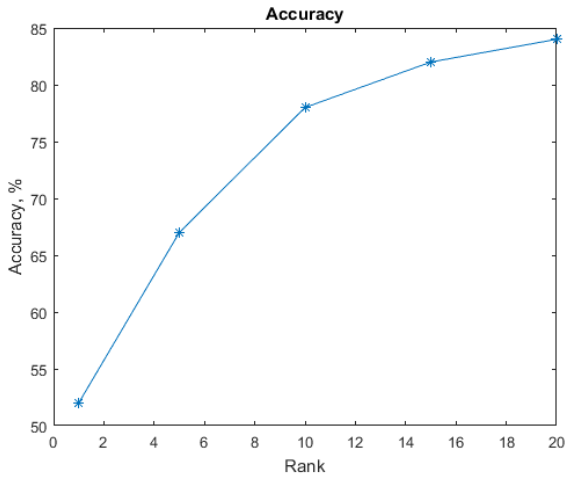

Figure 12. Accuracy levels for Ranks 1, 5, 10, 15, 20

Conclusions

The re-identification task requires real-time processing, that why the main goal of our research was develop and study sparsed convolutional neural network with state-of-art efficiency level in person re-identification field. In the paper was researched modification of new method of neural network sparsing, that satisfies these requirements. Thus we have neural network with less computational complexity but with similar efficiency. This method allows fine-tuning pre-trained neural networks, removing redundant weights in the same time.

Further Work

Future works will be linked to person feature propagation. For pair of cameras using autoencoder we get some feature transformation law. When person disappears from the one camera field of view we are trying to find propagated feature for all persons appeared in the other cameras fields of views. We use several discriminative features to compare results of re-identification efficiency with and without applying of our approach.

Acknowledgment

The reported study was supported by the Russian Foundation for Basic research (RFBR), research projects №15-07-99685, №15-01-09092, №16-37-00386.

References

[1] Voronin V.V., Marchuk V.I., Semenishchev E.A., Creutzburg R., Makov S..V. Digital inpainting with applications to forensic image and video processing. IS&T International Symposium on Electronic Imaging, Mobile Devices and Multimedia: Enabling Technologies, Algorithms, and Applications 2016, pp. 1-7(7).

[2] Frantc V.A., Makov S.V., Voronin V.V., Marchuk V.I., Semenishchev E.A., Egiazarian K.O. and Agaian S. Simultenious binary hash and features learning for image retrieval. Proc. SPIE 9869, Mobile Multimedia/Image Processing, Security, and Applications 2016, 986902 (May 26, 2016).

[3] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, Nanning Zheng. Person Re-Identification by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function. CVPR 2016 IEEE Conference.

[4] McLaughlin N., Martinez del Rincon J. & Miller P.. Recurrent Convolutional Network for Video-based Person Re-Identification. CVPR 2016 IEEE Conference.

[5] Yeong-Jun Cho and Kuk-Jin Yoon describes in paper. Improving Person Re-identification via Pose-aware Multi-shot Matching. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1354-1362.

[6] Tong Xiao Hongsheng, Li Wanli, Ouyang Xiaogang Wang. Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification. CVPR 2016 IEEE Conference.

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In CVPR, 2015.

[8] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, Yoichi Sato. Hierarchical Gaussian Descriptor for Person Re-Identification. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1363-1372.

[9] Suraj Srinivas, Akshayvarun Subramanya, R. Venkatesh Babu. Training Sparse Neural Networks. CVPR 2016 IEEE Conference.

[10] K. Simonyan, A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv:1409.1556

[11] L. Fei-Fei, ImageNet: crowdsourcing, benchmarking & other cool things, CMU VASC Seminar, March, 2010.

[12] W.Li,R.Zhao,T.Xiao,andX.Wang. Deepreid: Deepfilter pairingneuralnetworkforpersonre-identification. InCVPR, 2014.

[13] W. Li and X. Wang. Locally aligned feature transforms across views. In CVPR, 2013.

[14] Y. Kawanishi, Y. Wu, M. Mukunoki, and M. Minoh. Shinpuhkan2014: A multi-camera pedestrian dataset for tracking people across multiple cameras. In 20th Korea-Japan Joint Workshop on Frontiers of Computer Vision, 2014.

[15] Murray,W.,andNg,K.-M. 2010. An algorithm for nonlinear optimization problems with binary variables. Computational Optimization and Applications 47(2):257–288.

[16] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In SCIA, 2011.

[17] D. Gray, S. Brennan, and H. Tao. Evaluating appearance modelsforrecognition,reacquisition,andtracking. InPETS, 2007.

[18] D. Baltieri, R. Vezzani, and R. Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In Proc. of the ACM workshop on Human gesture and behavior understanding, 2011.

[19] Qiqi XiaoKelei CaoHaonan ChenFangyue Peng Chi Zhang Cross Domain Knowledge Transfer for Person Re-identification.

[20] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang and Shuicheng Yan,Senior Member, IEEE End-to-End Comparative Attention Networks for Person Re-identification.

[21] Li Zhang Tao Xiang Shaogang Gong Queen Mary University of LondonLearning a Discriminative Null Space for Person Re-identification.

[22] Hailin Shi, Yang Yang, Xiangyu Zhu, Shengcai Liao, Zhen Lei, Weishi Zheng, Stan Z. Li Embedding Deep Metric for Person Re-identi cation: A Study Against Large Variations.

Author Biography

Sergey Makov was born in 1978. In 2001 he has graduated Don State University of Service as radio engineer. Since then he is working in the engineering company as designer of telecommunication hardware. He has got PhD in 2011. Since then he has worked as a professor assistant in the department of electronic systems of Don State Technical University

Alexander Minaev received his bachelor degree in radioengineering from Donetsk National Technical University in 2014. Enrolled into Don State Technical University and started to work in «Digital Signal Processing and Computer Vision» laboratory in 2016.

Anton Nikitin student fourth year of Don State Technical University. Enrolled in university in 2013. He is involved in the research in laboratory «Digital Signal Processing and Computer Vision» since 2015.

Viacheslav Voronin was born in Rostov (Russian Federation) in 1985. He received his BS in radio engineering from the South-Russian State University of Economics and Service (2006), his MS in radio engineering from the South-Russian State University of Economics and Service (2008) and his PhD in technics from Southern Federal University (2009). Voronin V. is member of Program Committee of conference SPIE. His research interests include image processing, inpainting and computer vision.

Semenishchev Evgeny Alexandrovich received her BS in «Radioelectronics system» from South-Russian State University of Economics and Service, Shakhty, Russia (2005), received her MS in «Radioelectronics system» from South-Russian State University of Economics and Service, Shakhty, Russia (2007), PhD in applied radiotechnics from Southern University fereralny (2009). Associate Professor of Dept. "Radio-Electronics Systems" Shakhty, Russia.

Vladimir Marchuk was born in 1951. He received the D.Tech. degree in technics from Southern Federal University (Russian Federation) in 2006. Since 2006, he has been a Professor. His research interests are in the areas of applied statistical mathematics, signal and image processing.