Semantic Video Segmentation with Using Ensemble of Particular Classifiers and a Deep Neural Network for Systems of Detecting Abnormal Situations

O. Amosov, Y. Ivanov, S. Zhiganov
Department of Industrial Electronics
Komsomolsk-on-Amur State Technical University
Komsomolsk-on-Amur, Russia

***Abstract*—A new approach based on the use of a deep neural network and an ensemble of particular classifiers is proposed. This approach is based on use of the novel block of fuzzy generalization for combines classes of objects into semantic groups, each of which corresponds to one or more particular classifiers. As result of processing, the sequence of frames is converted into the annotation of the event occurring in the video for a certain time interval.**

Keywords—semantic segmentation; automatic image annotation; deep neural network; abnormal situations

I. INTRODUCTION

Detection of abnormal situations is one of the most important tasks of a modern closed circuit television (CCTV). Traditional methods of object localization and recognition are oriented on the search of predetermined objects in the image: license plates of vehicles, people, etc. Under this approach, abnormal situations are understood as an occurence of the object from the "blacklist".

However, there are abnormal emergency situations (situations not provided by the system): unconventional behavior of people and vehicles.

The characteristic feature of such situations is a time span, and therefore it is necessary to carry out a semantic analysis of each frame and generalize the information received for a certain time interval. In this case, the application of methods based on binary classifiers, is hampered.

The problem of frame semantic analysis can be solved by using the semantic image segmentation, i.e. its splitting into separate areas and assigning each of them to a certain class.

The use of deep neural networks (DNN) is appeared as the most promising solution [1, 2].

There are various approaches for video sequences description [3-5]. The paper [3] proposes architectures of long-term recurrent convolutional networks.

Convolutional neural network (CNN) together with long short-term memory (LSTM) network for image annotation are used in the paper [6]. The operating result of the proposed

algorithm is a description of the event occurring in the frame.

However, all the approaches [3-6] are based on the following principle: convolutional network, conditional random fields (CRF) or another deep architecture processes a frame in the video sequence. The result of segmentation is generalized by the recurrent network or the LSTM network that generates the annotation for each frame or video sequence as a whole.

Thus, the completeness of the annotation will depend on the segmentation quality. To detect an abnormal situation, it is important to take into account as many classes as possible, and the annotation should be the most detailed one.

Research shows [7] that the increasing number of classes will increase the error. Besides, the errors are observed when classifying the semantically close objects: for example, "man", "woman", "child", "pedestrian", etc.

Such errors complicate the use of complex architectures [8] in real CCTV systems where the accurate determination of the visitor's sex or age, the car class, etc. is required.

In practice, the single-purpose or the particular pretrained models are used in security systems [9]. Particular classifiers detect and recognize the certain subclasses of objects: brands and models of cars, pedestrians and cyclists, etc. This approach does not allow getting a full frame annotation, and therefore there is no possibility of abnormal situation detecting.

In this paper, we propose an approach based on the use of a deep neural network and an ensemble of particular classifiers.

The application of fuzzy generalization block combining the object classes into semantic groups, eliminates the negative effects caused by the increase in the number of classes in the deep neural network. The result specification is carried out by the particular classifier.

The work was supported by the Russian Ministry of Education research project - state task in the framework of the project № 2.1898.2017 / PCH "Designing the Mathematical and Algorithmic Ware of Intelligent Information and Telecommunication System for Higher Educational Institution Security".

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

II. Statement Of a Semantic Video Segmentation Problem

Let there be a continuous video stream V, represented by the set of frames $\mathbf{I}^t$, where t – number of current frame. There is a final set of object classes $C=\{1,\ldots,k\}$. For each frame $\mathbf{I}^t$ it is necessary to create a mask $\mathbf{A}^t$, besides, the class mark will be matched with each pixel $\mathbf{I}^t(x,y)$ - $\mathbf{A}^t(x,y)\in C$ (Fig. 1).

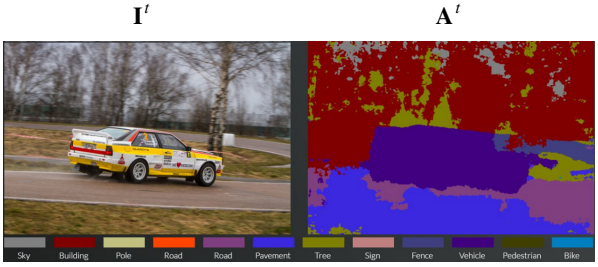


Fig. 1. Semantic Image Segmentation

Vector $\mathbf{s}^t$ contains a set of class marks of the objects that are present in the image and the polygons parameters assigned to every class, i.e. $\mathbf{s}^t=\langle \mathbf{s}_1;\mathbf{s}_2,\ldots,\mathbf{s}_r\rangle$, where r – number of objects found in the image. Each object $i\in 1..r$ is described by the vector $\mathbf{s}_i=\{C_j,x_1,y_1,\ldots,x_m,y_m\}$, where C_j – class mark, $j\in 1..k$, and $x_1,y_1,\ldots,x_m,y_m$ – polygon parameters (coordinates of the vertices). C_j can be represented as a basis vector with dimensions of k and 1 in the position of j, i.e. $\mathbf{c}^{(j)}=[0,\ldots,0,1,0\ldots,0]$. All the classes that are present in the frame t, can be written in the form of a set C^t.

Then the function of segmentation (segmentator) Seg will be a transformation of the matrix $\mathbf{I}^t$ into the vector $\mathbf{s}^t$, i.e. $Seg:\mathbf{I}^t\rightarrow\mathbf{s}^t$. Matrix $\mathbf{A}^t$ is a visualisation of the vector $\mathbf{s}^t$ and represents a set of marked areas.

Taking into account the class marks and the coordinates of the areas it is necessary to aggregate this information in the space-time coordinates of a continuous video stream, i.e. to carry out a semantic video segmentation. Then let us understand a function of aggregating (aggregator) L as a transformation of the set $\mathrm{S}_n=\{\mathbf{s}^t,\mathbf{s}^{t+1},\ldots,\mathbf{s}^{t+n}\}$, containing the segmentation results of n frames, into the $\mathbf{l}_n$ vector, having a form of the video frame annotation for n frames, i.e. $L:S_n\rightarrow\mathbf{l}_n$

Fig. 2 illustrates the given statement of the problem.

The confusion matrix method is used for assessing the quality of the segmentation algorithm [10].

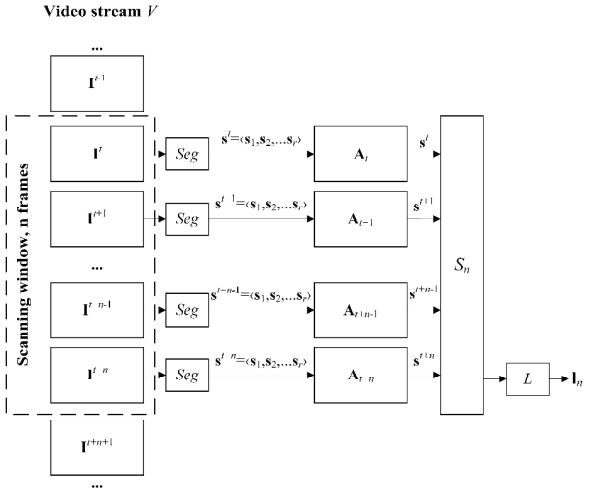


Fig. 2. Statement of a semantic video segmentation problem

III. Solving of a Semantic Video Segmentation Problem with Using an Ensemble of Particular Classifiers and a Deep Neural Network

Solving of a semantic video segmentation problem is divided into solving a number of subproblems (Fig. 3):

1) Frame capture and image preprocessing are performed.

2) Generalized segmentation and class integrating into aggregative semantic groups are performed.

3) Regions of interest tansfer to the particular classification algorithms is performed.

4) Aggregating of frame description and query of the next frame are performed.

5) Annotating of events for n frames is performed.

A. *Image Preprocessing*

Let there be a continuous video stream V, represented by the set of frames $\mathbf{I}^t$, where t – number of current frame.

It is necessary to perform an image preprocessing in order to reduce the negative impact of the following factors:

- changing of scene illumination;
- digital and analog noise;
- loss of focus;
- weather conditions.

There are various ways of detection [11] and elimination of the listed clutters [12-14], including the use of fuzzy logic [15].

The result of this is a matrix $\mathbf{I}^t$, clarified from the noises and external clutters.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

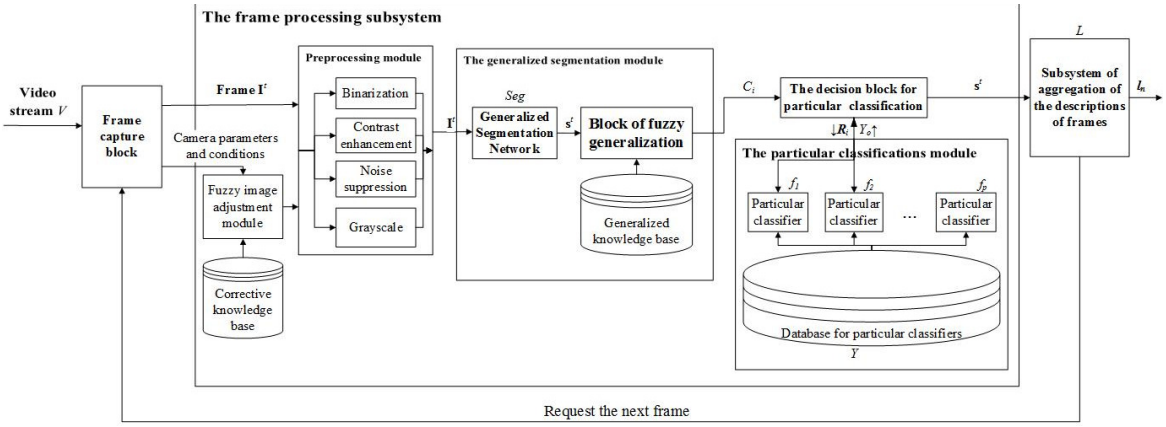


Fig. 3. Solving of a semantic video segmentation problem

B. Semantic Segmentation of a Video Frame

Let there be a video frame $\mathbf{I}^t$, where t – number of current frame. There is a final set of object classes $C = \{1,\ldots,k\}$.

It is necessary to construct an algorithm representing $Seg: \mathbf{I}^t \to \mathbf{s}^t$. According to obtained vector $\mathbf{s}^t$ it is necessary to create a mask $\mathbf{A}^t$, besides, the class mark will be matched with each pixel $\mathbf{I}^t(x, y)$ – $\mathbf{A}^t(x, y) \in C$.

It is proposed to use a deep architecture of neural networks, CNN in particular [1] as an algorithm of generalized segmentation.

Convolutional network (Fig. 4.) consists of several alternate layers that are divided into convolution layers and pooling layers.

The convolution by the kernel $\mathbf{K}$ is performed according to a formula:

$$\mathbf{x}^\tau = \sigma\left(\mathbf{x}^{\tau-1} * \mathbf{K} + b\right), \quad (1)$$

where $\mathbf{x}^\tau$ – output layer τ, σ – activation function, b – shear coefficient.

The pooling/subsampling layer reduces the image size by the formula:

$$\mathbf{x}^\tau = \sigma\left(a \cdot subsample\,(\mathbf{x}^{\tau-1}) + b\right), \quad (2)$$

where a, b – coefficients, $subsample$ – the operation of sampling local maximum values.

The last layer is a fully connected SoftMax or MLP layer:

$$\mathbf{x}^\tau = \sigma\left(\sum_\eta \mathbf{x}_\eta^{\tau-1} \cdot \omega_{\eta\kappa}^{\tau-1} + b_\kappa^{\tau-1}\right), \quad (3)$$

where ω – matrix of weight coefficients, η, κ – layer dimensions.

The learning is realized with the use of the backpropagation algorithm.

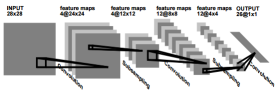


Fig. 4. Structural scheme of Lenet-5 neural network [1]

However, a high dimension of the network and a great number of parameters make the convolutional network learning a difficult task. On the one hand, a large data set is required. On the other hand, special computer equipment with GPU is needed [16].

In practice, ready-made architectures are used [17], or the re-training of the finished model is carried out using the transfer learning technology [18, 19].

As the output of the convolutional network is a vector containing a set of classes that are present in the image, the classification problem is solved. Fully convolutional networks are used for image segmentation [20]. The peculiarity of the architecture is as follows: after the image is compressed to the vector containing the classes, a return to the original image size is performed by alternating the operations of convolution and upsample.

Nowadays there are many algorithms [21] solving the problem of semantic image segmentation with the use of convolutional architecture. The best results [22] among convolutional networks are shown by the Inception-BN model [23]. This model is learnt on the ImageNet dataset [24] and contains 21K classes

However, increasing the number of classes increases the degree of uncertainty which reduces the classification accuracy. The accuracy of segmentation depends also on the quality of the learning sample.

When analyzing the results of tests of deep models [25], it can be seen that the highest errors are observed when

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

classifying the semantically close objects: “car”, “bus”, “truck” or ‘man”, “woman”, “child”, etc.

As a rule, such errors are neglected in the image annotation, but such errors can become crucial in security systems during the abnormal situations detecting.

In order to improve the accuracy of classification the papers

[26, 27] proposed to introduce the notion of semantic similarity of the detectable objects. The paper [28] proposed the “general-to-specific” approach, i.e. the semantic attributes were derived for each object (for example, “paws”, “hair”, “tail” for the object “dog”, etc.).

We propose the contrary approach – “specific-to-general”. The key part of the approach is the following:

1) it is proposed to generalize the classes by the thematic categories using their semantic similarity;

2) it is necessary to transfer the area containing the semantic group to the entry of the respective particular classifier for class specification.

C. *Class Integrating into Generalized Semantic Groups and Regions of Interest Transfer to the Particular Classification Algorithms*

Let there be a vector $\mathbf{s}^t$, containing a set of class marks of the objects that are present in the image and the polygons parameters assigned to every class. The set $D = \{d_1, d_2 \dots d_h\}$ has the marks of semantic groups, at that $h < k$. $F = \{f_1, f_2 \dots f_p\}$ – set with dimensions of p contains classifiers corresponding to a certain semantic group. Let us call such classifiers as particular classifiers.

It is necessary to create a decision rule g corresponding each of the found classes C_j to a certain semantic group. Algorithm g can be written as follows:

$$g : C^t \to D^t, \tag{4}$$

where the set D^t contains the marks of semantic groups found in the image $\mathbf{I}^t$.

As an algorithm g it is proposed to use a semantic graph presented in the Fig. 5. The belonging of a class to a semantic group can be determined using fuzzy logic methods [29].

Let us call a rectangle described around the polygon of the i object as a region of interest (ROI) $\mathbf{R}_i$. Then the task of the particular classification can be written as follows:

$$f_{\mathrm{p}} : \mathbf{R}_i \to Y_p, \tag{5}$$

where Y_p – final set of classes for the classifier f_{p}.

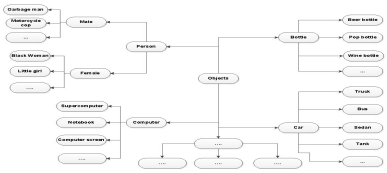


Fig. 5. Fragment of the semantic network

The result of the work is a specified vector $\mathbf{s}^t$, where every element C_j is corrected according to the following rule (Fig. 6):

$$C_j = \begin{cases} f_{\mathrm{o}} : \mathbf{R}_i \to Y_o, if \ \ for \ d_u \exists f_{\mathrm{o}} \\ C_j, otherwise \end{cases}, \tag{6}$$

where $u \in 1..h$ and $o \in 1..p$.

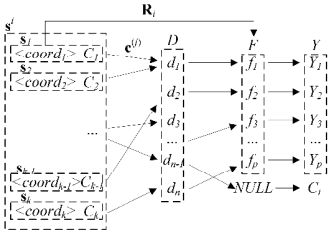


Fig. 6. Application of the particular classifiers

In order to detect the abnormal situations in CCTV systems, it is possible to use the following particular classifiers:

- recognition of vehicle type, model and brand [30];
- recognition of license plate symbols of vehicles [31];
- UAV detection [32];
- recognition of pedestrians, cyclists, sex, race, age of people [33];
- person's identity [34], etc.

D. *Annotating of Events for n Frames*

Let there be a vector $\mathbf{s}^t$, represented as class marks that are present in the image $\mathbf{I}^t$. A scanning window with dimensions of n frames moves in the video stream V, then the set $S_n = \{\mathbf{s}^t, \mathbf{s}^{t+1}, \dots, \mathbf{s}^{t+n}\}$ contains the segmentation results of n frames.

It is necessary to create an algorithm L transforming the set S_n into the vector $\mathbf{l}$, represented as a video frame annotation for n frames, i.e. $L : S_n \to \mathbf{l}_n$.

It is proposed to use a neural network with LSTM, shown in the Fig. 7 as such a function. A key feature of such networks is the presence of a “memory cell”, that is composed of four main elements: an input gate, a neuron with a self-recurrent connection (a connection to itself), a forget gate and an output gate.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

To calculate the cell memory update, we take the following notation:

- $\mathbf{s}^t$ – is the input to the memory cell layer at time t, γ_t – output vector;
- $\mathbf{W}_\iota$, $\mathbf{W}_\varsigma$, $\mathbf{W}_\phi$, $\mathbf{U}_\iota$, $\mathbf{U}_\varsigma$, $\mathbf{U}_\phi$, $\mathbf{V}_o$ – weight matrices;
- β_ι, β_ς, β_ϕ – are bias vectors.

The calculation algorithm is the following:

- compute the values for ι_t, the input gate, and $\tilde{\varsigma}_t$ the candidate value for the states of the memory cells at time :

$$\begin{aligned} \iota_t &= \sigma_g^L\left(\mathbf{W}_\iota \mathbf{s}^t + \mathbf{U}_\iota \gamma_{t-1} + \beta_\iota\right) \\ \tilde{\varsigma}_t &= tanh\left(\mathbf{W}_\varsigma \mathbf{s}^t + \mathbf{U}_\varsigma \gamma_{t-1} + \beta_\varsigma\right), \end{aligned} \quad (7)$$

- compute the value for ϕ_t, the activation of the memory cells "forget gates" at time t :

$$\phi_t = \sigma_g^L\left(\mathbf{W}_\phi \mathbf{s}^t + \mathbf{U}_\phi \gamma_{t-1} + \beta_\phi\right), \quad (8)$$

- update the previous state of the cell ς_{t-1} up to the current state ς_t :

$$\varsigma_t = \iota_t * \tilde{\varsigma}_t + \phi_t * \varsigma_{t-1}, \quad (9)$$

- compute the value of output gates and output vector:

$$\begin{aligned} \mathrm{o}_t &= \sigma_g^L\left(\mathbf{W}_\mathrm{o} \mathbf{s}^t + \mathbf{U}_\mathrm{o} \gamma_{t-1} + \mathbf{V}_\mathrm{o} \varsigma_t + \beta_\mathrm{o}\right) \\ \gamma_t &= \mathrm{o}_t * \tanh\left(\varsigma_t\right) \end{aligned}. \quad (10)$$

The result of work of such network will be a vector $\mathbf{l}$, each element of which has a form of an output γ_t at the current time.

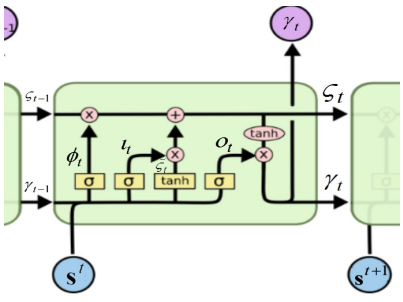


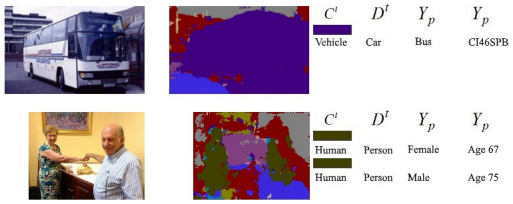
Fig. 7. LSTM neural network

IV. Experiment

The proposed approach was implemented in the Python language using the Caffe library. The following computer configuration was used for testing: Intel Core i5, 8 Gb RAM, Nvidia Geforce 1080 Ti.

The testing of the segmentator *Seg* was carried out in Coco dataset [36] containing 3K marked images. The Fig. 8 shows the results of work of the original Inception-BN and after the application of generalization algorithm g and the particular classifiers f_p .

Fig. 8. Segmentation result during the Coco dataset



The Fig. 9 demonstrates the work of the segmentation algorithm in the operational surveillance system of University.

The following classifiers are used as particular ones [30-35].

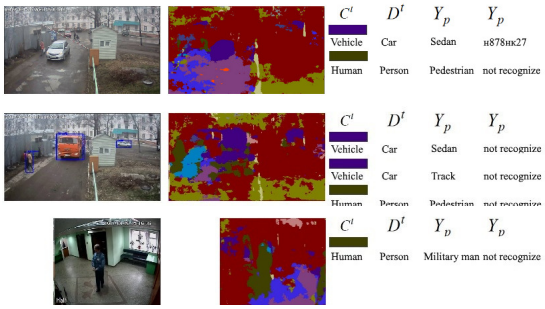


Fig. 9. Segmentation result in CCTV of University

Coco dataset (annotation) was used for algorithm L learning.

For preliminary testing of annotation algorithm DAVIS dataset [37] containing marked and annotated video frames, as well as video frames received from CCTV of University, was used (Fig. 10).

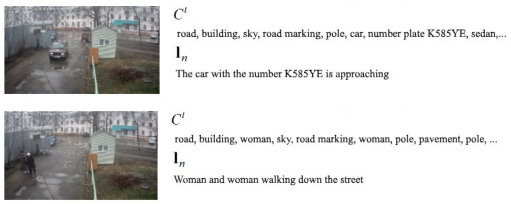


Fig. 10. Annotation result

V. Conclusion

The problem of semantic segmentation of the video stream is solved using an ensemble of particular classifiers and a convolutional neural network is solved.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

The possibility of applying the proposed approach in CCTV of University is demonstrated.

The novelty is the use of the block of fuzzy generalization combining the object classes into semantic groups, each of which corresponds to one or more particular classifiers.

As result of processing, the sequence of frames is converted into the annotation of the event occurring in the video for a certain time interval.

Hereafter, it is planned to analyze the obtained description of the video fragment for detecting the abnormal situation and elaborating the system response.

REFERENCES

[1] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," *in Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, Nov 1998.

[2] E. Shelhamer, J. Long, T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.* vol. 39, no. 4, pp. 640-651, 2017.

[3] J. Donahue, L. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, T. Darrell. "Long-term Recurrent Convolutional Networks for Visual Recognition and Description," *IEEE Trans Pattern Anal Mach Intell*, vol. 39, no. 4, pp. 677-691, 2017.

[4] Y Pan, T Mei, T Yao, H Li, Y Rui, "Jointly Modeling Embedding and Translation to Bridge Video and Language," *The IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4594-4602, 2016.

[5] Rohrbach, M. Rohrbach, B. Schiele, "The Long-Short Story of Movie Description," *Proceedings of the German Confeence on Pattern Recognition*, vol. 9358, pp. 209-221, 2015.

[6] Karpathy, L. Fei-Fei. "Deep visual-semantic alignments for generating image descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 664-676. April 2017.

[7] (2016) Full ImageNet Network. [Online]. Available: https://github.com/dmlc/mxnet-model-gallery/blob/master/imagenet-21k-inception.md.

[8] (2016) Large Scale Visual Recognition Challenge 2016 (ILSVRC2016). [Online]. Available: http://image-net.org/challenges/LSVRC/2016/results.

[9] (2016) Deep Learning GPU Training System. [Online]. Available: https://github.com/NVIDIA/DIGITS.

[10] C. Sammut, G. Webb "Encyclopedia of machine learning" Springer US, p. 1031, 2010

[11] P. Barnum, S. Narasimhan, T. Kanade "Analysis of Rain and Snow in Frequency Space," International Journal of Computer Vision, vol. 86, p. 256, 2010.

[12] A. Yu, H. Bai, Q. Jiang, Z. Zhu, C. Huang, B. Hou, "Blurred license plate recognition via sparse representations," *IEEE Conference on Industrial Electronics and Applications*, pp. 1657-1661, 2014.

[13] K. K. Pal and K. S. Sudeep, "Preprocessing for image classification by convolutional neural networks," *IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology*, pp. 1778-1781, 2016.

[14] O.S. Amosov, Y.S. Ivanov and S.V. Zhiganov, "Human localiztion in video frames using a growing neural gas algorithm and fuzzy inference," Computer Optics, 2017, vol. 41(1), pp. 46-58. DOI: 10.18287/2412-6179-2017-41-1-46-58.

[15] O.S. Amosov, S.G. Baena, Y.S. Ivanov and Soe Htike, "Roadway Gate Automatic Control System with the Use of Fuzzy Inference and Computer Vision Technologies," *The 12th IEEE Conference on Industrial Electronics and Applications, ICIEA*, 2017, p. 6.

[16] D. Cireşan, U. Meier, J. Masci, L. Gambardella, J. Schmidhuber. "Flexible, high performance convolutional neural networks for image classification," *AAAI Press. In Proceedings of the Twenty-Second international joint conference on Artificial Intelligence*, vol. 2, pp. 1237-1242, 2011.

[17] (2017) CNN Models by CVGJ. [Online]. Available: https://github.com/cvjena/cnn-models.

[18] (2017) Transfer Learning - Machine Learning's Next Frontier. [Online]. Available: http://sebastianruder.com/transfer-learning/.

[19] S. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Trans. on Knowl. and Data Eng*, vol. 22, no. 10, pp. 1345-1359, October 2010.

[20] J. Long, E. Shelhamer, T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," CoRR abs/1411.4038, 2014.

[21] (2017) Model Zoo. [Online]. Available: https://github.com/BVLC/caffe/wiki/Model-Zoo.

[22] (2016) Large Scale Visual Recognition Challenge. [Online]. Available: http://image-net.org/challenges/LSVRC/.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," CoRR abs/abs/1512.00567, 2015.

[24] (2017) ImageNet. [Online]. Available: http://image-net.org/.

[25] (2016) CNN-benchmarks. [Online]. Available: https://github.com/jcjohnson/cnn-benchmarks

[26] T. Deselaers and V. Ferrari, "Visual and semantic similarity in ImageNet," CVPR 2011, pp. 1777-1784, 2011.

[27] A. Farhadi, I. Endres, D. Hoiem and D. Forsyth, "Describing objects by their attributes," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1778-1785, 2009.

[28] Y. Su, F. Jurie, "Improving Image Classification Using Semantic Attributes," *International Journal of Computer Vision*, vol. 100, no.1, pp 59–77, October 2012.

[29] M. Omri, N. Chouigui, "Measure of Similarity between Fuzzy Concepts for Optimization of Fuzzy Semantic Nets", CoRR abs/1206.1624, 2012.

[30] L. Yang, P. Luo, C. C. Loy, X. Tang, "A Large-Scale Car Dataset for Fine-Grained Categorization and Verification" CoRR abs/1506.08959, 2015.

[31] (2017) Deep ANPR [Online]. Available: https://github.com/matthewearl/deep-anpr

[32] (2016) Computer Vision Assisted UAV detection and tracking [Online]. Available: https://github.com/mukeshmodi/Computer-Vision-Assisted-UAV-detection-and-tracking

[33] G. Levi and T. Hassncer, "Age and gender classification using convolutional neural networks," *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 34-42, 2015.

[34] I. Masi, A. Tran, J. Leksut, T. Hassner, G. Medioni, "Do We Really Need to Collect Millions of Faces for Effective Face Recognition?" CoRR abs/1603.07057, 2016.

[35] O.S. Amosov, Y.S. Ivanov, and S.V. Zhiganov "Human Localization in the Video Stream Using the Algorithm Based on Growing Neural Gas and Fuzzy Inference," *XII Intelligent Systems Symposium, INTELS'16, Procedia Computer Science*, 2017, no 103, pp. 403–409. – DOI: 10.1016/j.procs.2017.01.128.

[36] (2016) COCO dataset. [Online]. Available: http://mscoco.org/.

[37] (2017) DAVIS: Densely Annotated Video Segmentation. [Online]. Available: http://davischallenge.org/.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731