Choice of a Deep Neural Networks Architecture to Monitor the Dynamics of an Object State

Andrey Puchkov[1], Maksim Dli[1], Yekaterina Lobaneva[1], Maria Vasilkova[1]

[1] National Research University «Moscow Power Engineering Institute»
(Branch) in Smolensk, Energetichesky proyezd 1, g. Smolensk, 2014013, Russia

putchkov63@mail.ru, MiDli@mail.ru, lobaneva94@mail.ru, vasilkova_mariya00@mail.ru

Abstract. The study proposes a deep neural network architecture to monitor the dynamics for a state of a complex technological object according to the data received in the form of images. The paper also contains recommendations for the architecture adaptation to a specific application. The developed architecture is based on the cascade use of convolutional neural networks for processing multi-channel video information from different technological zones of one object.

Key words: machine learning, convolutional neural networks, computer vision

Introduction

Now deep neural networks (DNN) represent the most practically significant direction in the development of artificial intelligence methods. DNN are actively used in the systems of video analytics to obtain metadata from a video stream. From the technical point of view, video analytics is a software and a hardware complex for the intellectual analysis of the events that fall into the sector of video surveillance systems and undergo deep processing by software tools. On average, only 10% of data, which a camera is able to give for processing, are used. The intellectualization of this process allows increasing the part of the useful information implementation.

Initially, in the systems of industrial safety video analytics was applied as a detection of an object movement or crossing of a control line, the objects identification (people, transport, luggage), their behavior estimation [1]. However, the opportunities of computer vision use in manufacturing are not limited by this area of application. Within the development program “Digital economy of the Russian Federation”, approved by the Russian government in 2017, the following promising directions for video information processing in automated systems for technological process control (ASTPC) with the use of artificial intelligence methods can be specified:

– reidentification of objects and processes during their transition from one production zone to the other in accordance with the accepted technological and logistic chains;

In: P. Sosnin, V. Maklaev, E. Sosnina (eds.): Proceedings of the IS-2019 Conference, Ulyanovsk, Russia, 24-27 September 2019, published at http://ceur-ws.org

– automated control of safety measures requirements;

– solution of transport problems under conditions of high dimensionality of initial data, when the methods for linear programming lead to the significant time expenditures;

– search and prediction of equipment failures to reduce the probability of incident situations;

– implementation of multidimensional machine vision based on processing information from a large number of sensors in order to monitor technological processes (TP) in real time and predict their behavior;

– business analytics based on the generation of metadata for a state of a production process by intellectual cameras, as well as the detection of hidden regularities in visualized information about the results of an enterprise commercial activity.

Most of the noted directions for the artificial intelligence methods implementation consider video data processing and include the estimation for the change rate state of the object under observation. The image changes of the object under study can be applied as a visual prompt in decision–making process as the meaning of many actions is precisely in the dynamics, it is enough to observe the movement of individual points in order to recognize the event. In this case the changes are understood as a wide set of characteristics for images, i.e. the shifting of any objects and contour in the background, changes the brightness of the elements, texture modification [2].

The problem of developing and adapting machine vision methods and algorithms to assess the rate for the change of TP state, taking into account the specificity of production, is actual due to the significant diversity of nature of the observed objects and processes.

1 Problem statement

Complex TP are characterized by the significant duration not only in time but in space as well. This fact specifies the reasonability to include into APCS not only signals from standard control and measuring equipment, but from the additionally installed systems of visual control for technological zones with high responsibilities as well [3].

Suppose there are *kz* of technological zones for which video cameras are installed. Each video camera gives a stream of shots with the resolution of $n[ip] \times m[ip]$ pixels, where $ip=1, 2, \ldots, kz$ and frequency *f*. As a result, video data from all the cameras are mapped by tensor X of the sixth rank with the form: a camera, samples, shots, height, width, color [4].

Different camera models allow forming a video with the frequency of shots in a big range; usually it is from 10 to 60 *fps* (frames per second). Then, to be processed by the convolutional neural network, the discretization interval at time $\Delta t(i)$, $i=1, 2, \ldots, I$, where *I* is a number of information video channels, for *i*-th information channel, is chosen to be more than one second. This choice is based on the assumption that the inertia of TP allows to do it without violation of Kotelnikov's theory which sets the maximum value of the discretization interval, at this value the accurate restoration of initial continuous signal is possible. If the discretization interval is needed to be less than one second, then it is necessary to choose a video camera model more carefully, but the methodology of the given bellow approach to the dynamics recognition is not changed.

It is required to develop an architecture of a deep neural network to detect metadata from the forming tensor X, the metadata provide the recognition and forecast for "movement" (evolution) of TP in time. The estimation for the recognition quality is implemented on the base of confusion matrix CM [5].

2 Methods background for recognition of the processes dynamics according to video data

DNN find their application in algorithmic support of video analytics systems for various application areas, for example: to detect technological defects [6], medical diagnostics [7], sensor information processing [8], vehicles identification [9] etc.

However, in most cases these algorithms do not support simultaneous description of an object by its image and motion, that will allow recognizing the events even at a low resolution and predict the evolution of an observed object state. The exception in this case is only a direction connected with the recognition of people actions [10 – 11], the difficulty of this direction is associated with the necessity to take into account the environment. The sense interpretation of a recognizing action depends on this environment.

One of the popular methods, called "sliding window", consists of several stages: fragment selection (spatio-temporal parallelepiped); solution of the images classification problem and search for the objects for three-dimensional spatio-temporal volume. However, it does not suit the automated recognition of changes, as it requires prior indication of window borders on the image which makes it difficult to be applied when the boundaries of this window need to be shifted in the observed technological process.

In other methods for analyzing video sequences the basic tool is the concept of optical flow. This approach was first proposed by Bruce D. Lucas and Takeo Kanade in1981. Optical flow is often defined as a vector field or an image of objects apparent motion, surfaces or scene edges resulting from the motion between an observer and a scene [12]. In the process of analyzing the optical flow for each pixel of one shot, the displacement vector is calculated from the current shot to the next one. As a result a process of matching occurs: for each pixel of one shot the same point on the other shot is found. The disadvantages of this approach include the necessity to analyze the solving of optical flow problem which entails the need to control aperture problem [13]. On the basis of the mentioned methods specialized software products are created, in particular, the software for mapping and measuring particles flow rate of any environment [14].

The proposed algorithm is based on the application of DNN ensemble, the structure of which has a temporary time delay unit to enable the recognition of objects dynamics by video data.

3 Proposed solutions

Advances in the use of deep neural networks in computer vision systems for various purposes provide reasons for optimism in the case of recognition of TP processes dynamics. This approach reflects a modern paradigm of Software 2.0 which unlike Software 1.0 does not imply the explicit writing of an algorithm , it provides the creation of a neural network with a specified architecture which learns (adjust) itself to solve a specific applied task.

Deep neural network models hierarchical abstractions in data using architectures which consist of cascading ensembles of nonlinear transformations (filters). Today there are some popular DNN architectures: neocognitron, autocoder, convolutional neural networks (CNN), Boltzman machine, deep trust networks, long short-term memory networks, controlled recurrent networks, residual neural networks [15]. This study uses convolutional neural networks.

The architecture of a neural network defines the hypotheses space, i.e. the number of classes for the input data sets splitting. The proposed architecture for a deep neural network to monitor the object dynamics is in Fig.1. It uses the successful practices of neural networks ensemble application [16 – 18] but differs from them in presence of a time delay unit, the signal from which is also fed to the output cascade of the network to enable the recognition of TP state changes.

After the calculation for the intervals of discretization according to time Δt(i) for all channels the minimum $\Delta t = \min_i (\Delta t(i))$ is chosen for further synchronization and unification of video data transformation performed by the neural network.

Also, to unify further transformation the input multichannel images with the resolution of $n[ip] \times m[ip]$pixels, before being fed to the neural network input, are normalized to one dimension of $n\times m$ pixels which is smaller than the minimum from $n[ip] \times m[ip]$.

In the time delay block the shift is done by moment Δt of image receiving which makes it possible to calculate discrete analogues of derivatives when defining the values changes at neural networks outputs.

The input cascade of a deep neural network contains some CNN operating in parallel classifying images from cameras of a corresponding information channel taken at intervals Δt. This procedure consists in forming output channel vector V (j|i), j=1,2,…, cl(i), where cl(i) is a number of classes for i-th information cannel, at each moment Δt.

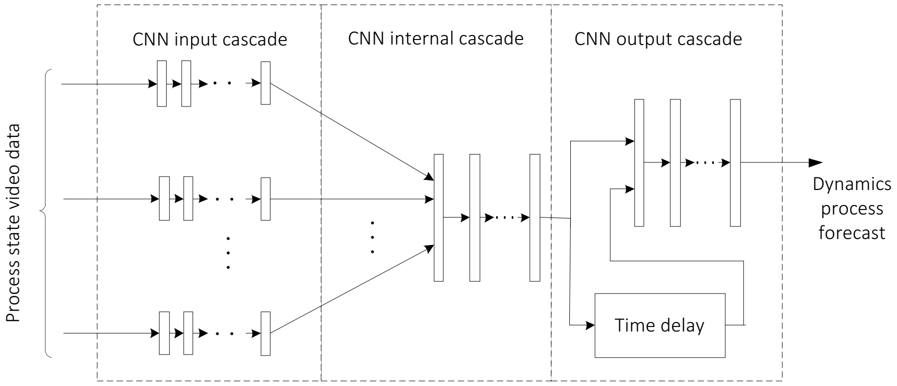


Fig. 1. Deep neural network architecture

CNN of an input cascade forms the elements of vectors V(j|i), which take the values in the range from 0 to 1. This reflects the i-th channel CNN degree of confidence in membership of the parameter controllable according to the image to a particular class at moment t(k)= kΔt, where k is a sequence number of time discrete.

The time interval during the technological process is divided into fragments with ΔT duration. Consider the fragments for all information channels are equal and defined by the requirements for the periodicity of information flow to APCS and characteristics of the most frequency-critical channel.

By moment T(ζ)= ζ ΔT, ζ =1, 2, …,ψ, where ψ is a number of fragments with ΔT duration for each i-th channel, the matrix of classification results can be formed:

$$MV(i\,|\,\xi) = \begin{pmatrix} V(1\,|\,i,1) & V(1\,|\,i,2) & \dots & V(1\,|\,i,k) \\ V(2\,|\,i,1) & V(2\,|\,i,2) & \dots & V(2\,|\,i,k) \\ \dots & \dots & \dots & \dots \\ V(cl(i)\,|\,i,1) & V(cl(i)\,|\,i,2) & \dots & V(cl(i)\,|\,i,k) \end{pmatrix}, \qquad (1)$$

where element V(j|i, k) means the j-th CNN output of the input cascade for the i-th information channel at time discrete kΔt.

Matrix (1), in fact, reflects CNN degree of confidence changes in classification results when TP passes interval ΔT under number ζ.

CNN internal cascade receives the tensor consisting from the combined particular matrixes (1) for all information channels. At the output of the internal cascade matrix MS is formed. It is an analogue of matrix (1), containing more number of elements according to the numbers of information channels and the classes of TP state for each channel.

To have an idea about the dynamics of the entire technological process tensor DV is calculated. Tensor DV contains the relations of MS matrix elements increments to Δt discretization interval. The range of this tensor is equal to three, and its ζ-th cut has a form of:

$$DV(\xi) = \begin{pmatrix} \frac{MS_{\xi}(1,1) - MS_{\xi-1}(1,1)}{\Delta t} & \cdots & \frac{MS_{\xi}(1,I) - MS_{\xi-1}(1,i)}{\Delta t} \\ \frac{MS_{\xi}(2,1) - MS_{\xi-1}(2,1)}{\Delta t} & \cdots & \frac{MS_{\xi}(2,I) - MS_{\xi-1}(2,i)}{\Delta t} \\ \cdots & \cdots & \cdots \end{pmatrix}. \tag{2}$$

Cut (2) can be interpreted as an image fed to the output of CNN cascade. The sense load for the elements of cut (2) can be matched with the analogue of derivatives for the continuous functions as they reflect the confidence change of neural network in image membership to a certain class. The rate of change can be used to forecast TP development.

It should be mentioned, that the number of classes for different information channels and different technological zones can be different. Thus, to provide the proportionality when feeding tensor DV for neural network processing some cuts contain zeros on the places of redundant classes.

Tensor DV is formed on the base of the data about the CNN confidence changes in class membership of information channels parameters reflected in matrix (1). It allows assuming that hypotheses space formation can provide good forecasting results with the help of the proposed deep neural network architecture.

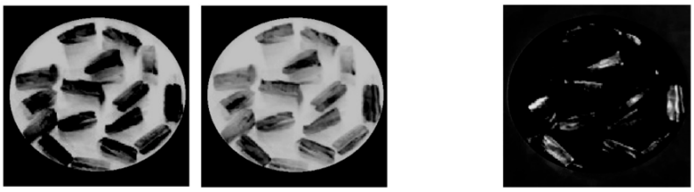
4 Results

The recommended choice for the deep neural network architecture means to insure the correspondence for the number of channels of the video information flow with the number of convolutional neural networks in the input cascade. In this case the authors give the results of simulation experiment when processing ingots images of aluminum alloy with the aim to forecast the time from its complete melting. The melting process takes approximately 300 seconds; the aggregate state is estimated by the surface image observing through a viewing window fitted on the furnace door [19].

The model program was performed in IDE Spyder from Anaconda (version for Linux) in Python 3.6. CNN were created with the special neural network library Keras which is the add-on the framework of Tensor Flow calculations [20]. To visualize TensorFlow process framework TensorBoard is used.

In the considered example only one informational channel from the technological melting zone is used, thus the structure of the applied network is significantly simplified. The network contains seven alternate convolution layers and subsamples and one output fully connected layer with four outputs according to the number of defying classes for the substance aggregate state: class «solid» (0 – 269 sec.), «initial transit » (270 – 279 sec.), «final transit» (280 – 289 sec.), «liquid» (290 – 300 sec.). Thus, the time interval is connected with the class, therefore, when making the classification the time of the aggregate state occurrence is forecasted.

Fig.2 shows the images of a melting zone taken in different moments, Fig.2b shows tensor DV image reflecting the dynamics of images presented in Fig.2a. The sequence of 2b images forms the dynamics trend for the melting process which is recognized by CNN.



a) b)

Fig. 2. Processed images

The initial learning sampling for each class had 1600 examples; the testing sampling had 400 examples. In addition to the standard Dropout method, used in CNN, augmentation was implemented to reduce the possibility of the network relearning. Shifts, scale changes, rotation, mirror reflection, affine transformations were realized for the initial images of the ingots working surface. As a result of these procedures, the total size of the training sampling was 32,000 examples; the total size of testing sampling was 8000 examples. The network was trained during 110 epochs. Categorical entropy was set as a loss function for CNN.

The splitting of video data into shots with conversion into jpg format is performed with software utility Free Video to JPG Converter with a given discrecity of one second. This approach allows obtaining sufficient learning sample volume for CNN as a great number of images can be detected even from the video recording of one melting process.

The learning is conducted on video card GeForce GTX 1060 installed on Asus FX502VM notebook with CPU IntelCore i7-7700HQ, which provides more than twenty fold time gain comparing with the learning on a regular notebook processor. The quality of CNN learning is reflected in accuracy metric which is 77 % on testing samples. The graphics for loss function and accuracy function are shown in Fig. 3.

Confusion matrix (CM), used to estimate the classification quality, has four rows and four columns in this case. Its columns reflect the factual data; its rows show the results of the classifier work.

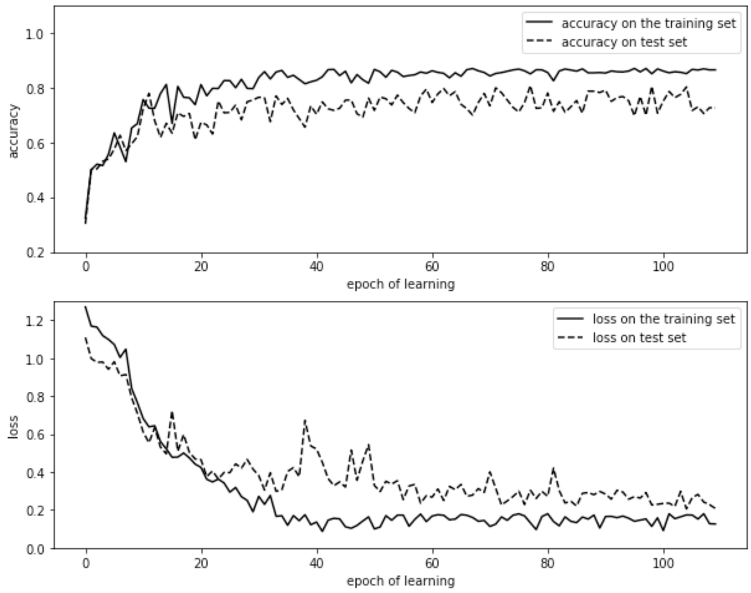


Fig 3. Graphics for behavior of loss function and accuracy function.

When filling the matrix the number on the cross of the class row, returned by the classifier, and the class column, to which the object really belongs to, increases by one. In this example every time interval corresponding to different melting stages is divided into 10 subintervals ΔT. After the conducted experiments matrix CM is filled:

$$\mathrm{CM} = \begin{pmatrix} 7 & 2 & 1 & 0 \\ 1 & 6 & 2 & 1 \\ 2 & 1 & 6 & 1 \\ 0 & 2 & 1 & 7 \end{pmatrix}.$$

The analysis of CM matrix elements values shows that the majority of classes are recognized correctly as matrix diagonal elements are clearly expressed.

5 Conclusion

In the process of the conducted study for the possibility of deep convolutional neural networks application to monitor the dynamics of technological objects state the following results were obtained:

1. The architecture of a deep neural network to obtain information about the dynamics of the technological object state based on the video data fed from different technological zones is proposed. Its basis contains convolutional neural networks with cascades differentiation at input, internal and output which allows integrating

different information flows into the entire set and increase the level of the presented data abstraction. In the output cascade of a neural network the time delay in images processing is used to enable the object state recognition dynamics and its forecast.

2. Recommendations on the choice and adaptation of the neural network architecture depending on the number of technological zones and the number of information channels are given.

3. The study presents the results of simulating experiment which show the efficiency of the proposed neural network architecture. The results can indicate the reasonability of the architecture use in various application areas where it is necessary to control the dynamics of the processes by the available video information.

6 Acknowledgment

The reported study was funded by RFBR according to the research project № 19-01-00425.

References

1. Rubio D. Videoanalytics: possibilities and solutions. (2016) Modern automation technologies, 4, pp. 86-92.
2. Li H., Qian X., Li W. (2017) Image Semantic Segmentation Based on Fully Convolutional Neural Network and CRF. In: Yuan H., Geng J., Bian F. (eds) Geo-Spatial Knowledge and Intelligence. GRMSE 2016. Communications in Computer and Information Science, vol 698. Springer, Singapore.
3. Pokhabov Y.P. Problems of dependability and possible solutions in the context of unique highly vital systems design. Dependability. 2019; 19(1): S. 10 – 17. (In Russ) https://doi.org/10.21683/1729-2646-2019-19-1-10-17.
4. Chollet F. Deep learning with Python (2018) Peter, SPb.
5. Shunina Yu. S., Alekseeva V.A., Klyatchkin V.N. Performance criteria for classifiers (2015) UlGTU bulletin, 2(70) URL: https://cyberleninka.ru/article/n/kriterii-kachestva-raboty-klassifikatorov
6. Cha YJ., Choi W. (2017) Vision-Based Concrete Crack Detection Using a Convolutional Neural Network. In: Caicedo J., Pakzad S. (eds) Dynamics of Civil Structures, Volume 2. Conference Proceedings of the Society for Experimental Mechanics Series. Springer, Cham
7. Kori A., Soni M., Pranjal B., Khened M., Alex V., Krishnamurthi G. (2019) Ensemble of Fully Convolutional Neural Network for Brain Tumor Segmentation from Magnetic Resonance Images. In: Crimi A., Bakas S., Kuijf H., Keyvan F., Reyes M., van Walsum T. (eds) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2018. Lecture Notes in Computer Science, vol 11384. Springer, Cham
8. Kasnesis P., Patrikakis C.Z., Venieris I.S. (2019) PerceptionNet: A Deep Convolutional Neural Network for Late Sensor Fusion. In: Arai K., Kapoor S., Bhatia R. (eds) Intelligent Systems and Applications. IntelliSys 2018. Advances in Intelligent Systems and Computing, vol 868. Springer, Cham

9. Xiang, L., et al.: Automatic vehicle identification in coating production line based on computer vision. In: International Conference on Computer Science and Engineering Technology, pp. 260 – 267. World Scientific Publication Co. Pvt. Ltd. (2016)
10. Ahlawat S., Batra V., Banerjee S., Saha J., Garg A.K. (2019) Hand Gesture Recognition Using Convolutional Neural Network. In: Bhattacharyya S., Hassanien A., Gupta D., Khanna A., Pan I. (eds) International Conference on Innovative Computing and Communications. Lecture Notes in Networks and Systems, vol 56. Springer, Singapore.
11. Fan Y., Lam J.C.K., Li V.O.K. (2018) Multi-region Ensemble Convolutional Neural Network for Facial Expression Recognition. In: Kůrková V., Manolopoulos Y., Hammer B., Iliadis L., Maglogiannis I. (eds) Artificial Neural Networks and Machine Learning – ICANN 2018. ICANN 2018. Lecture Notes in Computer Science, vol 11139. Springer, Cham
12. Solovich I.O., Belov Yu. S. Lucas-Kanade method application to calculate optical flow. (2014) Engineering journal: science and innovations, 7 URL: http://engjournal.ru/catalog/pribor/optica/1275.html
13. Nagiev A.G., Sasyhkov V.V. The problem of aperture delay in digital measurement systems and its analytical solution using the matrix exponential method. (2017) Measuring engineering, 9, p.p..16 – 20.
14. Thielicke W. (2019). PIVlab - particle image velocimetry (PIV) tool(https://www.mathworks.com/matlabcentral/fileexchange/27659-pivlab-particle-image-velocimetry-piv-tool), MATLAB Central File Exchange. Retrieved April 24, 2019.
15. Sozykina A. V. Overview of deep neural network learning methods. (2017) YuUrGU bulletin: computational mathematics informatics,6(3), p.p. 28-59.
16. Frazão X., Alexandre L.A. (2014) Weighted Convolutional Neural Network Ensemble. In: Bayro-Corrochano E., Hancock E. (eds) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2014. Lecture Notes in Computer Science, vol 8827. Springer, Cham
17. Koitka S., Friedrich C.M. (2017) Optimized Convolutional Neural Network Ensembles for Medical Subfigure Classification. In: Jones G. et al. (eds) Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2017. Lecture Notes in Computer Science, vol 10456. Springer, Cham
18. Puchkov A., Dli M., Kireyenkova M. (2020) Fuzzy Classification on the Base of Convolutional Neural Networks. In: Hu Z., Petoukhov S., He M. (eds) Advances in Artificial Systems for Medicine and Education II. AIMEE2018 2018. Advances in Intelligent Systems and Computing, vol 902. Springer, Cham.
19. Shkundin S.Z., Kolistratov M.V., Belobokova Yu. Testing the performance of algorithms for determining changes in the aggregate state of a metal. (2018)System administrator, 10 (191), p.p. 90-93.
20. Geron Au.. Applied machine learning using Scikit-Learn and TensorFlow: concepts, tools and the technique for intellectual systems creation. (2018) Dialectic, Moscow.