Operator's State Estimation Based on the Face's Video Images Analysis Using Deep Convolutional Neural Networks

O N Korsun\textsuperscript{1}, V N Yurko\textsuperscript{1} and E I Mikhaylov\textsuperscript{1}

\textsuperscript{1}State Research Institute of Aviation Systems, Moscow, Russia

E-mail: marmotto@rambler.ru

Received xxxxxx
Accepted for publication xxxxxx
Published xxxxxx

Abstract

The paper deals with the problem of operator’s state estimating. For this purpose various approaches based on using deep convolutional neural networks are proposed. The approach using automatic emotion recognition methods is considered in the most detail. During the experiment video records of the operator's face registered during operator performing the flight task on the flight simulator were processed. To determine the type of operator's activity the studies based on using the emotional background of the face are also carried out. The experimental results of this approach confirmed the efficiency of the selected methods, especially for monitoring the operator's state when falling asleep.

Keywords: convolutional neural network, emotion, operator's state

1. Introduction

During the analysis of the operator’s face video images using deep convolutional neural networks two approaches were applied:

- neural network was trained according to the traditional recognition way in order to distinguish normal operator’s state and fatigue. For this training experimental data received during operators perform flight task was used;
- possibility of using the emotions automatic recognition by face video images was investigated. Now there are many databases from the open access that can be used for neural network training.

Nowadays deep convolutional neural networks are considered the most perspective approach to pattern recognition tasks. The peculiarity of the deep learning methods is that it gradually, layer by layer constructs more and more complex representations and takes into account their interaction, so each layer is updated in accordance with the needs of both the previous and subsequent layers. There are very functional software tools for research in the field of deep learning that effectively implement the back propagation method and operations with tensors, which greatly simplifies the realization of new structures of neural networks.

Emotion is a special kind of medium duration psychic processes, representing a subjective evaluative attitude to existing or possible situations and the objective world. Despite the individual and cultural differences between people, there are general, genetically determined paradigms that determined exactly how our emotions are expressed as a sort of contractions of perfectly defined groups of facial muscles: forehead, eyebrows, eyelids, cheeks, lips, chin.

2. Description of the research and results of the experiments

The following are the stages of automatic emotions recognition:
1. Real-time image registration by capturing the video stream of the camera, or previously recorded video files analysis;
2. Recognition of facial contours in the image;
3. Facial images transmit to a trained convolutional neural network and the emotions classification results output.

For the image facial contours recognition, the Haar cascade was chosen. This is one of the methods for recognizing objects’ classes with a high speed of operation. In this article an approach based on the sliding window technology was used: the image was scanned by the sliding window and the Haar attribute was calculated for each image area scanned by the window. The object presence or absence in the window was determined by the difference between the value of the attribute and the threshold value, that was being trained. This system is completely automated and does not require human intervention, so this approach works quickly.

For deep convolutional neural network training the FER-2013 data set [1] was chosen. This database consists of 35887 monochrome images, 48 × 48 pixels in size, with 7 types of emotions (anger, disgust, fear, joy, sadness, astonishment and neutral state). Here with there were 28709 images used for training, 3589 images — for checking and 3589 for testing. Data from the database was converted to tensors with real numbers, pixel values were scaled from the range [0 – 255] to the range [0 – 1], cause it is unsafe to transmit to the neural network data that can take very large values or dissimilar data. Failure to these rules can result in significant gradient changes that will prevent network convergence.

The architecture of the convolutional neural network constructed in this work with the use of the accepted terminology is presented in the figure 1.

![Figure 1. The architecture of the convolutional neural network.](image)

Convolution layers use the ReLU (rectified linear unit) activation function. First convolutional layer contains weights regularization to reduce the network complexity by limiting its weight coefficients. In this article L2-regularization was used whose main idea is that the penalty added to the loss function is proportional to the squares of the weight coefficients. Also in the work “dropout” was used as one of the most effective and common regularization methods for neural networks. Dropout consists in the randomly selected attributes removal at the training stage (the percentage of attributes is chosen in accordance with the dropout coefficient). The last completely connected layer is the loss layer, which returns an array with seven probability estimates.

Examination of the proposed structure on a test set containing 3589 images showed 62.4% of correct recognitions. The received results analysis shows that it is possible to improve the neural network architecture in order to reach the level of 92% correct recognitions.

Further on the trained neural network there was conducted an experiment to estimate the operator’s state. During the experiment 7 video records of the operator’s face received during operator performing the flight task on the flight simulator were processed. The flight task durations were 14, 18, 19, 20, 12, 19, 12 minutes. The frame rate in the video records was 24 Hz.

In the normal operator’s state during the piloting there were 2 emotions observed – sadness and a neutral state (figure 2). It should be noted that the operator’s yawn and closed eyes were classified as astonishing or fear (figure 3).

![Figure 2. Neutral state (a) and sadness (b).](image)

![Figure 3. Yawn classification as astonishment (a) or fear (b).](image)
The obtained results have an obvious physical interpretation. Operators perform monotonous work during a long time, which corresponds to a normal or sad facial expression. The tendency to fall asleep, expressed in yawning, is classified as astonishment or fear, because yawning as a separate class was not introduced during the neural network training.

According to psychologists [2] fatigue like a fear causes the desire to stop doing something and, if activities do not change, the operator’s body will go into a state of sleep.

The following experiments were carried out to study operator's fatigue. A deep convolutional neural network with a trained convolutional basis of the VGG16 [3] model was constructed (it was trained for 1.4 million images belonging to 1000 classes); the last convolutional layer and a fully connected classifier were retrained. The output layer that up to 9.5 minutes the neural network classified the images to the normal state, and then to the “fatigue” state, that confirms the efficiency of the method. Figure 5 shows the image areas by which neural network classifies the video data as a one or another class [4, 5].

| Table 1. The results of emotions classification into the categories after processing seven flight records using a convolutional neural network (the recognized emotions percentage of the total emotions’ number during the performing a flight task is presented). |
|-----------------|------|------|-----|------|------|------|
| Pilot1          | 2.4  | 0    | 2.1 | 0    | 36   | 4    | 56   |
| Pilot2          | 0    | 0    | 0   | 0    | 37.7 | 0    | 62.3 |
| Pilot3          | 21   | 0    | 0.2 | 0    | 78.3 | 0    | 0    |
| Pilot4          | 0    | 0    | 0.06| 0    | 1.8  | 1.45 | 96.7 |
| Pilot5          | 0.1  | 0    | 1.37| 0    | 6.7  | 3.88 | 87.9 |
| Pilot6          | 0.02 | 0    | 0.78| 0    | 1.8  | 1.62 | 95.75|
| Pilot7          | 0    | 0    | 1.53| 0    | 2.21 | 8.05 | 88.2 |

Analysis of the record with duration of 14 minutes showed that up to 9.5 minutes the neural network classified the images to the normal state, and then to the “fatigue” state, that confirms the efficiency of the method. Figure 5 shows the image areas by which neural network classifies the video data as a one or another class [4, 5].

| Table 2. The experimental results of distinguishing between two classes (“normal state” and “fatigue”). |
|-----------------|-----|-----|-----|
| t, min          | Normal | Fatigue |
| 2.3             | 100   | 0    |
| 3.4             | 100   | 0    |
| 4.5             | 100   | 0    |
| 5.6             | 100   | 0    |
| 6.7             | 100   | 0    |
| 7.8             | 95    | 5    |
| 8.9             | 93.6  | 6.4  |
| 9.10            | 76    | 24   |
| 10.11           | 0     | 100  |
| 11.12           | 5     | 95   |
| 12.13           | 3     | 97   |

Acknowledgements

This work was supported by Russian Foundation for Basic Research (RFBR), project 18-08-01142-a.

References

[1] Challenges in representation learning: Facial expression recognition challenge http://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge

[2] Ekman P and Friesen W V 2003 Unmasking the Face: A Guide to Recognizing Emotions From Facial Expressions (Malor Books, 8/16/03 edition)

[3] Simonyan K and Zisserman A 2015 Very Deep Convolutional Networks for Large-Scale Image Recognition. InICLR

[4] Simonyan K, Vedaldi A and Zisserman A 2015 Deep inside convolutional networks: Visualizing image classification models and saliency maps CoRR abs/1312.6034

[5] Selvaraju R R, Das A, Vedantam R, Cogswell M, Parikh D and Batra D 2016 Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization CoRR abs/1610.0239