UDC 004.93:159.942

4. EMOTIONIQ: EMOTION RECOGNITION BY PHOTO WITH NEURAL NETWORKS

Orsik S.P., Bachelor Degree Student, gr. 351001

Belarusian State University of Informatics and Radioelectronics
Minsk, Republic of Belarus

Ladyjenko M.V. – Senior Lecturer

Annotation. The article provides algorithms for real-time face detection and recognition in complex backgrounds. A neural network based solution combined with image processing is used in classifying the universal emotions: happiness, sadness, anger, disgust, surprise and fear. The article proposes a prototype system EmotionIQ which automatically recognises the emotion represented on a face. The main stages of emotion recognition such as face detection, feature extraction and emotion classification are considered.

Key words. EmotionIQ, online platform, emotional intelligence (EI), emotion recognition, image processing, neural networks.

Emotions are mental states brought on by neurophysiological changes that reflect a person's relationship to themselves, to other people, to the real world. Emotions perform two functions: regulatory and signaling. The regulatory function refers to basic human emotions and feelings that guide and regulate the behaviour of a human, while the signaling function includes the emotions that arise and change in accordance with changes occurring in the external and internal environment. Human emotions are accompanied by expressive movements: facial (facial and muscle movements), pantomimic (gestures, body muscle movements), as well as changes in the tone of the voice and expressiveness of speech.

Emotional intelligence (EI) is defined as the ability to recognise, understand and manage your own as well as other's emotions. The benefits of EI are better physical and mental health, improved job performance, stronger relationships, enhanced communication, better decision making skills and etc. An organisation whose team members are equipped with greater emotional intelligence can work with increased productivity. This is because EI makes team members capable of understanding the client's emotions and using them to craft an empathetic response. Further, it also makes team members be able to identify and capitalise on their relationships with staff, clients, competitors, and redirect their efforts depending on the insights.

I and my team developed an online platform "EmotionIQ" which was presented at the grand final of the republican youth innovation project "100 Ideas for Belarus". It is a specialised tool designed to assess and provide insights into an individual's emotional intelligence (EI) through a structured assessment. EmotionIQ is an online platform that allows to monitor, enhance, monetise the emotional intelligence such as happiness, satisfaction, etc. of a person, group of people, community thanks to the synergy of fundamental knowledge, artificial intelligence, gamification elements and ratings. Our platform "EmotionIQ" can help to promote the economic and social success of a person, a group of people, an organisation, a community. It includes a set of neural networks for finding the face in a photo, recognising emotions, and verifying a person. This article considers the algorithms used for real-time face detection and recognition in complex backgrounds with neural networks that are implemented in our project EmotionIQ.

First, the survey was conducted to find out how much emotions affect a person's life and productivity. 300 students of the GDU "Lyceum 1 Estate of Academician Y.M. Ostrovsky, Grodno" took part in this study. The results of this survey are presented in Figure 1.

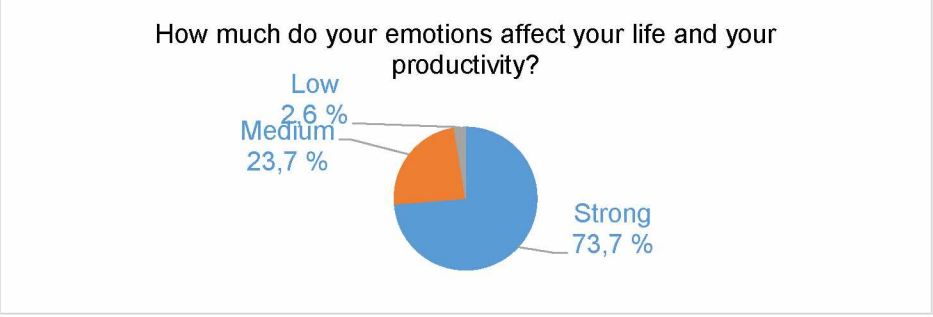

Figure 2 – Survey results

The results of the survey suggest that emotions are a crucially important aspect of people's psychological composition that affect intrapersonal, interpersonal, and social-cultural areas of life. Therefore, emotion regulation can help to improve the quality of a person's life, create a safe and supportive space where team members can share their feelings without judgment or criticism. From the employer's point of view, emotional intelligence can improve company performance by encouraging their employees to seek professional help if they are struggling with mental health issues.

MTCNN (multi-task cascaded convolutional neural networks) is commonly used to understand how to find a face in an image [1]. The model consists of three networks: P-Net, R-Net and O-Net (Figure 2). Each subsequent neural network increases the accuracy of the prediction.

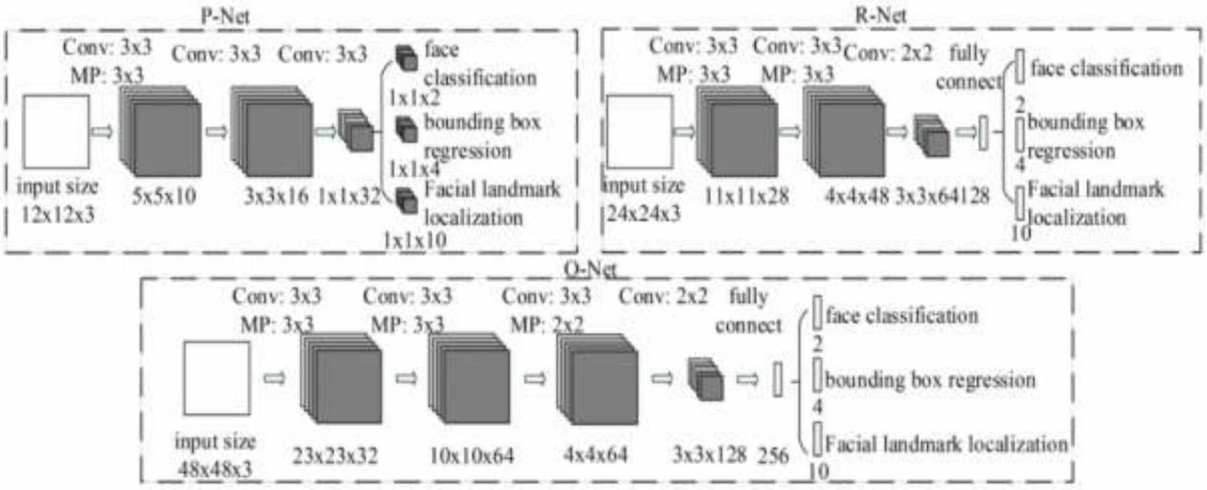

Figure 3 – Structure of MTCNN

The P-Net network at the output returns the coordinates of the bounding rectangles of the intended persons. Next, R-net trims areas where individuals are likely to be absent and adds a level of confidence to those areas that remain. In the O-Net network we again delete the areas with a low level of confidence and add the coordinates of five facial landmarks.

To find a face, Viola-Jones method [2, 3] is used including Haar features (Figure 3), which is a division of a given rectangular region into sets of different types of rectangular sub-fields [4]:

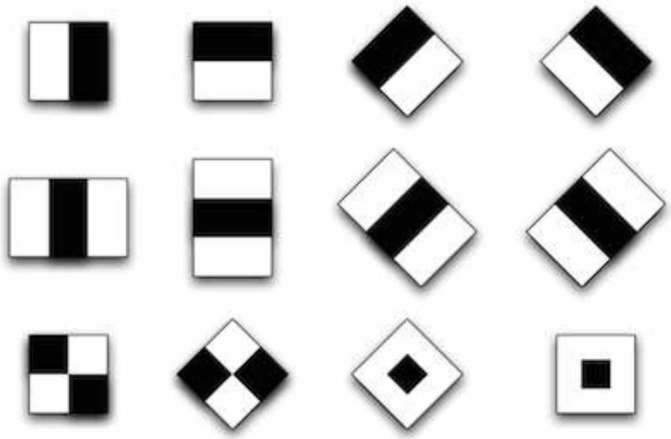

Figure 4 – Haar features

The original version of the algorithm used features without rotation, and to calculate the feature value, the sum of pixels of one sub-field was subtracted from the sum of the brightness of the other sub-field. During the development of the method, features with a 45-degree inclination and asymmetric configurations were proposed. In addition, instead of calculating the normal margin, it was suggested that weighted weights (1) and attributes should be assigned to each subgroup, which should be calculated as the weighted sum of pixels of the different types of domains:

$$feature = \sum_{i \in I=1...N} w_i RectSum(r_i)$$

(1).

The method is based on Haar features. The main reason was to try not to use the pixel representation while maintaining the feature calculation speed. It is difficult to extract any meaningful information from the values of pixel pairs for classification, while from two Haar features (Figure 4) the first cascade of the face recognition system is constructed, for example, which has a fully meaningful interpretation:

Figure 5 – Haar features in practice

From each sub-field, you can calculate by combining 4 values of the integral representation SAT (Summed Area Table), which in turn can be constructed one time in advance for the whole image in O(n), where n is the number of pixels in the image using the formula (2):

$$SAT(x,y) = SAT(x,y-1) + SAT(x-1,y) + I(x,y) - SAT(x-1,y-1) \quad (2)$$
$$SAT(-1,y) = SAT(x,-1) = SAT(-1,-1) = 0.$$

This allows the creation of a fast object search algorithm that has been successful for more than a decade. The sum of the values of the weak classifiers of this cascade is to be found in each cascade. Each weak classifier produces two values, depending on the feature belonging to that classifier greater than or less than the specified threshold. At the end, the sum of the values of the weak classifiers is compared with the threshold of the cascade and solutions are given to the found object or not by this cascade.

Once the face is found in the image, it is necessary to recognise emotions. One common method of training neural networks is “supervised learning”. To look at this in greater depth, it is necessary to understand how to measure recognition in order to start training our network. This article addresses the most common mean square error (MSE) and squared deviations from the mean (SDM) function in neural network theory [5]:

$$E^p = \frac{1}{2}(D^p - O(I^p, W))^2, \quad (3)$$

in this formula, E^p is a recognition error for the p-learning pair, D^p is the desired network output, $O(I^p, W)$ is the network output, depending on p-input and W weights, which include the cores of the package, offset, and S- and F-layers weights.

The task of training is to adjust the weights of W so that they for any learning pair (I^p, D^p) give minimal error E^p. To calculate the error for the whole learning sample, simply take the arithmetic mean for all learning couples. Such an average error is denoted as E.

Gradient methods are the most effective methods to minimise E^p errors. Consider the essence of gradient methods on the example of the simplest one-dimensional case (i.e., when we have only one weight). If we decompose the Taylor error function of the E^p, we get the following expression (4):

$$E(W) = E(W_C) + E(W + W_C) \cdot \frac{dE(W_C)}{dW} + \frac{1}{2}(W - W_C)^2 \cdot \frac{d^2E(W_C)}{dW^2} + \cdots, \quad (4)$$

here E is the same error function, W_C is some initial weight. It is crucial to remember that to find an extremum of a function we need to take its derivative and equal zero. Take the derivative of the error function by weights (5), discarding the terms above order 2:

$$\frac{dE(W)}{dW} = \frac{dE(W_C)}{dW} + (W - W_C) \cdot \frac{d^2E(W_C)}{dW^2}, \quad (5)$$

it follows from this statement that the weight at which the value of the error function will be minimal can be calculated from the following expression (6):

$$W_{min} = W_c - \left(\frac{d^2E(W_C)}{dW^2}\right)^{-1} \frac{dE(W_C)}{dW}, \quad (6)$$

i.e. the optimal weight is calculated as the difference between the current weight and the derivative of the weight error function divided by the second derivative of the error function.

For the multidimensional case (i.e., the matrix of weights), only the first derivative is transformed into a gradient (partial derivative vector), and the second derivative becomes Hessian (second partial derivative matrix). And there are two possibilities. If the second derivative is omitted, the gradient descent algorithm is obtained as soon as possible. If we are going to use the second derivative, we are going to need enough productive resources to count the full Hessian and then reverse it. To avoid this, Hessian is replaced by something simpler. One well-known and successful method is the Levenberg-Marquardt method, which replaces the Hessian, its Jacobian square approximation.

However, consider the fact that the Levenberg-Marquardt algorithm requires the processing of the whole learning sample, whereas the gradient descent algorithm can work with each individual learning sample. In the latter case, the algorithm is called a stochastic gradient. Given that, in most cases, training samples contain tens of thousands of training samples, a stochastic gradient is more appropriate. Another advantage of the stochastic gradient is that it is less prone to local minimum than the Levenberg-Marquardt algorithm.

At this point, neural network that can recognise a person's emotions from a photograph already can be trained. The next stage in development will be the transition from a cascade of super exact neural networks to a flexible comparison method on graphs (Elastic graph matching) [6]. The essence of the method is an elastic comparison of graphs that describe images of a person's faces. Faces are represented as graphs with weighted vertices and edges. In the recognition phase, one graph, the reference graph, remains unchanged, while the other graph is deformed to best fit the former. In such recognition systems, graphs can be both a rectangular lattice and a structure formed by characteristic (anthropometric) points of the face. Such a system has several types: rectangular grid, structure formed by anthropometric facial points (Figure 5).

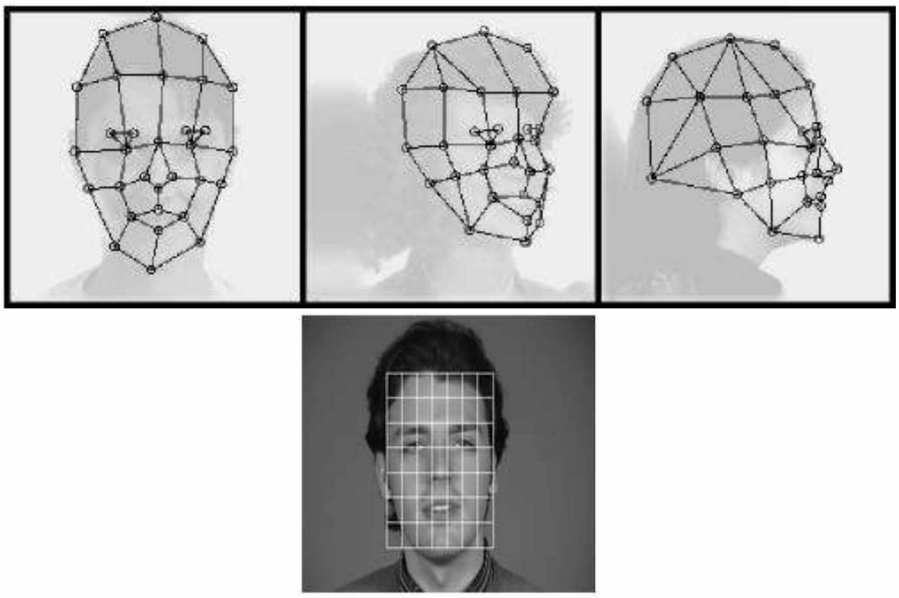

Figure 5 – Structure formed by rectangle and anthropometric facial points

At the vertices of the graph, feature values are computed, most often using the complex values of Gabor filters or their ordered sets are Gabor wavelets, which are locally (by convolution) in some regions of the graph vertex are calculated from the brightness values of pixels with Gabor filters (Figure 6).

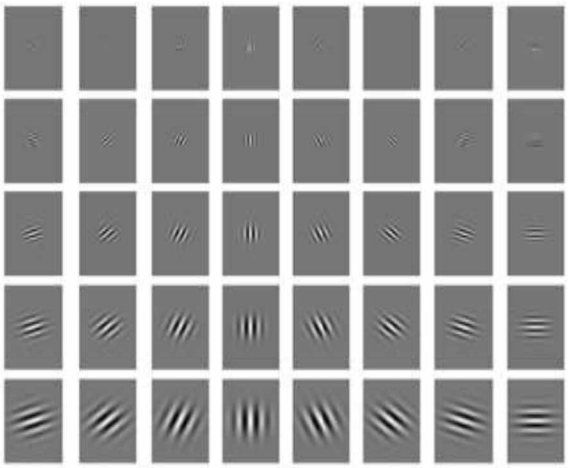

Figure 6 – Gabor filters

The edges of the graph are weighted by the distances between adjacent vertices. The difference (distance, discriminative characteristic) between the two graphs is computed by some price deformation function, taking into account both the difference between the feature values computed at the vertices and the degree of deformation of the edges of the graph. By moving each of its vertices for a certain distance in certain directions relative to its original location and choosing its position such that the difference between the values of the features (Gabor filter responses) at the vertex of the deformable graph and the corresponding vertex of the reference graph, there is minimal deformation of the graph.

The operation is performed alternately on all vertices of the graph until the smallest total distinction between the features of a deformable and a reference graph is reached. The value of the price deformation function at this position of the deformable graph will be the measure of the difference between the input image and the reference graph. This “relaxation” deformation procedure should be performed for all reference persons stored in the system database. The result of system recognition is the standard with the best value of the price deformation function. Some publications indicate 95–97 % efficiency of recognition even in the presence of various emotional expressions and the change of the facial angle to 15 degrees.

In conclusion, with such a surge in demand and innovation, choosing the right face recognition software is very crucial. It is also important to underline that recognition of the emotional state of a person using neural networks can have application in various spheres: government, healthcare, education, business, banking, security, etc. The relevance of this technology lies in the fact that modern society is striving for a solution that aligns with specific needs and integrates seamlessly into existing systems.

However, with a plethora of tools available in the market, selecting the right face recognition software is a critical decision for businesses and organisations. One implementation of this technology is presented in our project EmotionIQ [7]. Our algorithms can detect and recognise faces with high accuracy in real-time. It has a faster detection speed compared to other detection methods.

References:

1. Обнаружение и распознавание лиц MTCNN и FaceNet [Electronic resource]. – Mode of access: https://russianblogs.com/article/5106791164/. – Date of access: 27.02.2024.

2. Rapid object detection using a boosted cascade of simple features [Electronic resource]. – Mode of access: https://www.researchgate.net/publication/3940582. – Date of access: 10.03.2024.

3. Лайенхарт, Р. Эмпирический анализ каскадов обнаружения усиленных классификаторов для быстрого обнаружения объектов / Р. Лайенхарт, Е. Куранов, В. Писаревский. – В: PRS 2003. – 297-304 с.

4. Модель OpenCV для поиска лиц [Electronic resource]. – Mode of access: https://github.com/opencv/opencv/tree/4.x/data/haarcascades/haarcascade_frontalface_default.xml. – Date of access: 27.02.2024.

5. Müller, B. Neural networks: an introduction / B. Müller, J. Reinhardt, M.T. Strickland. – Berlin; Heidelberg; New York, 1995. – 88-153 p.

6. Face recognition by face bunch and graph method Electronic resource]. – Mode of access: https://www.researchgate.net/publication/255599615. – Date of access: 20.02.2024.

7. EmotionIQ [Electronic resource]. – Mode of access: https://emotioniq.by/. – Date of access: 1.03.2024.