Journal of Communication and Computer 14 (2017) 34-38
doi:10.17265/1548-7709/2017.01.004

A Quick Review of Deep Learning in Facial Expression

Mehdi Ghayoumi

Artificial Intelligence Laboratory, Computer Science Department, Kent State University, Ohio 44242, USA

Abstract: Over the last few years, deep artificial neural networks have gotten the most attention in computer science, especially in pattern recognition, machine vision and machine learning. One of its excellent applications is in the emotion recognition via facial expression area. Facial expression analysis is useful for many tasks and the application of deep learning in this area is also developing very fast. We review some recent research works in this domain, introduce some new applications and show the general steps to implementing each of them.

Key words: Deep learning, convolutional neural network, facial expression, emotion analysis.

1. Introduction

Human communication and social interaction are one of the main research areas which many scientists like psychologists, sociologists and computer scientists are interested in. As its computer science aspect, the computer should help people for better interaction with both human and computer [1]. Recently, HCI (human-computer interaction) has been an active research topic in computer science and it is about all the researches of the design and use of computer technology, which are focused on the interfaces between users and computers. Human emotion analysis plays a major role in providing proper human-computer interaction and it discusses about computer systems which attempt to analyze and recognize both facial feature changes and facial motions from all visual perceptions. Thus, facial feature extraction is one of the main parts of emotion analysis, which attempts to find the most appropriate information representation of facial images. Automated facial expression recognition has a large variety of applications, such as data-driven animation, interactive games, entertainments, sociable robotics, surveillance, crowd analytics, humanoid robots, interactive TV, and many other human-computer interaction systems [2, 3]. Although there are many researches in facial expression and recognition, with a high accuracy and performance in an online real world, it still has some difficulties due to its complexity and variability.

The way that different people express their facial emotions may vary from each other moreover, an image can be varied in brightness, background, position and many other criteria which directly or indirectly are affected in the analysis process. More accurate and higher level of knowledge and information is required for human emotion analysis [4]. Facial expressions also have some information about intention, cognitive processes, physical effort, or other intra- or interpersonal meanings and interpretation about all of these data can be complete and be more accurate by context, body gesture, voice, individual differences, and cultural factors as well as by facial configuration and timing but the automated facial expression analysis systems need to analyze the facial actions and features regardless of context, culture, gender, and so on [5].

Extracting the best facial features is one of the key factors in this area. The optimal features should minimize within-class variety of expression, and maximize between-class variance. If the extracted features were not sufficient enough, even the best classifiers might fail to achieve a very high accuracy

Corresponding author: Mehdi Ghayoumi, Ph.D. Candidate, research field: artificial intelligence.

and performance.

Automatic facial expression analysis can be done in three main steps:

(1) Face acquisition;

(2) Facial data extraction and representation, and;

(3) Facial expression recognition.

In the first step, ace acquisition can be separated in two major steps:

(a) Face detection and;

(b) Head pose estimation.

For the second step, three methods have been proposed to extract the facial expression features:

(a) Geometric feature-based method;

(b) Appearance-based method, and;

(c) Hybrid-based method.

In the first method, it is important to measure the shape and location of facial features. The Geometric measurements based on the relationships between these features are being used to construct a feature vector for training purpose. In static image the task will be done on the current image, but in dynamic images such as video frames, when we have a sequence of images, it can be done by measuring geometrical displacement of facial feature points between the current frame and initial frame or another specific frame or even other frames [6, 7].

The second method extracts the features by applying one or more filters to the facet images and here are some related approaches, such as PCA (principal component analysis), LDA (linear discriminate analysis), ICA (independent component analysis) and GW (Gabor wavelet). In appearance-based method, different facial regions contain different information, for example, eyes and mouth contain more information than the forehead and cheek [8].

And the third method is a combination of geometric and appearance based methods, which gives better results in some cases. Both the first and second methods have some issues and errors which can be covered by their fusion, and as a result, system accuracy will be increased [9].

The last step in facial expression analysis is recognition by classifying these features and many methods can be used for it, such as an ANN (artificial neural network), SVM (support vector machine), BN (Bayesian network) and many other classifiers. Deep learning is one of the most recent methods that achieves outperforming results. It provides an effective solution for approximate reasoning, and efficient greedy algorithm for applying in many applications such as facial expression recognition. Many different researches have been done in this area, but in general they have similar steps [10, 11].

In Section 2, the implementation steps in general will be reviewed. Section 3 briefly describes the CNN method, Section 4 describes the implementation steps of the CNN and the last section concludes the article.

2. Facial Expression Analysis System

As we mentioned before, in the general a FEs (facial expression systems) can be implemented in 3 major steps which we explain in detail here:

2.1 Image Acquisition and Pre-processing

The Facial image data can be picked up from a database (static) or a live video stream (dynamic), in 2D or 3D mode. Here we also have some steps as pre-processing such as de-noising (which are related to the devices that we deploy), and so on which help to have better performance.

2.2 Feature Extraction

Extracting the best features is one of the most important steps of any successful facial expression recognition system. The efficiency and effectiveness of the facial image representation could influence on the robustness during the recognition process.

2.3 Classification and Facial Expression Recognition

Many classifiers such as KNN, LDA, ANNs, HMMs, SVMs and CNN can be applied to the

automatic expression recognition problem.

3. Convolutional Neural Network

For the first time, CNN was introduced by Lecun et al.[12]. Having different types of information representation, is the key point in CNN functionality. Each layer can react to the different information, and when the layers stacked together, they can create a complex representation. Recently it showed good results in emotion analysis, and facial expression, especially in HRI and RRI. [13-15]. Fig. 1 shows the general architecture for CNN.

4. Facial Expression with CNN

Here we describe four main steps of the facial expression recognition process by CNN:

4.1 Normalization

The images in the database vary in many parameters which can affect directly on recognition accuracy and performance. These are some difficulties such as rotation, brightness and illumination changes even for the same person's images. To address this problem, a normalization of the face image such as detecting, de-noising and some other preprocessing such as correcting the rotation is performed. The image brightness and contrast variations increase the complexity of the problem.

4.2 Image Cropping

The original face images have background information that is not important and could make the output to be less accurate. The cropping region also tries to remove facial parts that do not contribute to the expression.

4.3 Downsampling

It is performing to ensure the same location of eyes, mouth, eyebrows, and other face components every face image. Down sampling helps the CNN to learn which regions are related to each specific expression and also enables the CNN to be performed on the GPU more efficiently.

4.4 Convolutional Network

The network receives an $n \times m$ image (can be specified in the down sampling step) as an input and then returns the confidence of each expression as an output. The first layer of the CNN is a convolution layer that applies a convolution kernel of $m \times m$ and outputs an image of mxn2 pixels. This layer is followed by a subsampling layer that uses Max-pooling with kernel size $k \times k$ to reduce the size of the image by half. Subsequently, a new convolution is applied to the feature vector and is followed by another subsampling. The output is given

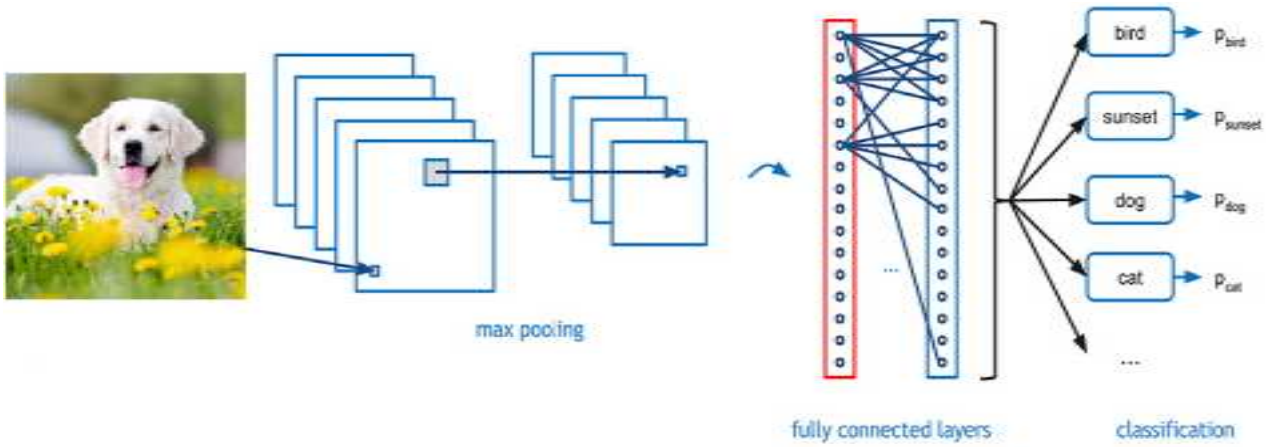

Fig. 1 The CNN architecture.

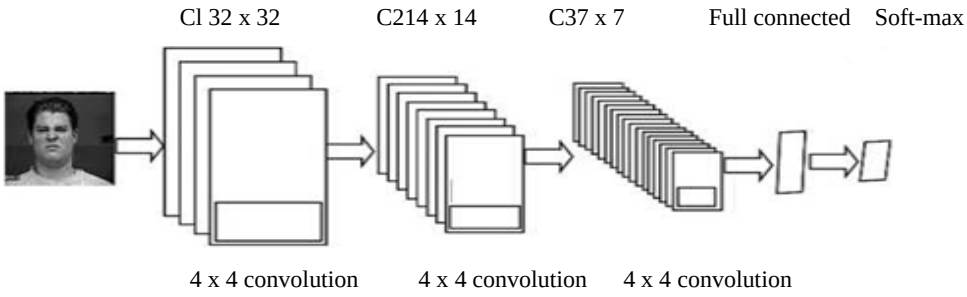

Fig. 2 The ER architecture.

to a fully connected layer that has **L** neurons. The network has six outputs.

Nodes that are fully connected to the previous layer. Each output node is corresponding to one of the expressions that outputs its confidence level [16, 17]. Fig. 2 shows the expression recognition system architecture. Some operations such as Max pooling and convolution will be applied to the original image as filters to extract a different representation of images in each layer. The maximum pooling operation is used to reduce the dimensions of the extracted hidden features for training. Finally, in the last step, the Soft-max classifier is used to classify the facial expressions of the test samples from extracted features. Some issues in applying CNN are such as computation time for the many layers and data which we have here and the type of the filter which determine the new presentation of data as the core of CNN. There are some solutions for these issues such as using the GPU (graphics processing unit). The experimental results of researches, show the performance and the generalization ability of the CNN for creating a very accurate facial expression recognition.

5. Conclusion

In this paper, a brief description of facial expression recognition with CNN has presented. We conclude by saying that the technology of facial expression recognition has enormous market potential and, in the near future, it will enhance most human computer interfaces. The new learning technologies (especially CNN) help to achieve better accuracy and performance.

References

[1] Li, S. Z., and Jain, A. K. 2011. “Handbook of Face Recognition.” *Springer Science & Business Media*.

[2] Michel, V., and Maja, P. 2010. “Induced Disgust, Happiness and Surprise: An Addition to the Facial Expression Database.” In *Proc. 3rd Intern. Workshop on EMOTION*.

[3] Ghayoumi, M. 2016. “Follower Robot with an Optimized Gesture Recognition System.” Presented at the Robotics: Science and Systems (RSS), USA.

[4] Susskind, J. M., Anderson, A. K., and Geoffrey, E. H. 2011. *The Toronto Face Database*. University of Toronto, Toronto, ON, Canada, Tech. Rep.

[5] Tian, Y., Kanade, T., and Cohn, J. F. 2005. “Facial Expression Analysis.” In Handbook of Face Recognition: 247-75.

[6] Ghayoumi, M., and Bansal, A. 2015. “Unifying Geometric Features and Facial Action Units for Improved Performance of Facial Expression Analysis.” *New Developments in Circuits, Systems, Signal Processing, Communications and Computers* (CSSCC): 259-66.

[7] Zee, T., and Ghayoumi, M. 2016. “Comparative Graph Model for Facial Recognition.” Presented at the 2016 International Conference on Computational Science and Computational Intelligence (CSCI’16).

[8] Ghayoumi, M., and Bansal, A. 2014. “An Integrated Approach for Efficient Analysis of Facial Expressions.” Presented at the International Conference on Signal Processing and Multimedia Applications (SIGMAP).

[9] Ghayoumi, M. 2015. “A Review of Multimodal Biometric Systems Fusion Methods and Its Applications.” Presented at the 14th International Conference on Computer and Information Science (ICIS), USA.

[10] Ghayoumi, M., Tafar, M., and Arvind, K. B. 2016. “Towards Formal Multimodal Analysis of Emotions for Affective Computing.” Presented at the 22nd

International Conference on Distributed Multimedia Systems (DMS).

[11] Ghayoumi, M., Tafar, M., and Bansal, A. K. 2016. “A Formal Approach for Multimodal Integration to Derive Emotions.” *Journal of Visual Languages and Sentient Systems*: 48-54.

[12] Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. 1998. “Gradient-Based Learning Applied to Document Recognition.” In *Proc. IEEE*.

[13] Ghayoumi, M., and Bansal, A. 2016. “Emotion in Robots Using Convolutional Neural Networks.” Presented at the ICSR.

[14] Ghayoumi, M., and Bansal, A. 2016. “Architecture of Emotion in Robots Using Convolutional Neural Networks.” Presented at the RSS, USA.

[15] Ghayoumi, M., and Bansal, A. 2016. “The Multimodal Architecture of Emotion in Robots Using Deep Learning.” Presented at the Future Technologies Conference, San Francisco, United States.

[16] Liu, P., Han, S., Meng, Z., and Tong, Y. 2014. “Facial Expression Recognition via a Boosted Deep Belief Network.” Presented at the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[17] Ghayoumi, M., and Bansal, A. K. 2017. “Improved Human Emotion Recognition Using Symmetry of Facial Key Points with Dihedral Group.” *International Journal of Advanced Studies in Computer Science and Engineering (IJASCSE)*.