Measuring Customer Behavior with Deep Convolutional Neural Networks

Veaceslav Albu
Institute of Mathematics and Computer Science, 5 Academiei, Chisinau,
Republic of Moldova, MD 2028
vaalbu@googlemail.com

Abstract

In this paper we propose a neural network model for human emotion and gesture classification. We demonstrate that the proposed architecture represents an effective tool for real-time processing of customer's behavior for distributed on-land systems, such as information kiosks, automated cashiers and ATMs. The proposed approach combines most recent biometric techniques with the neural network approach for real-time emotion and behavioral analysis. In the series of experiments, emotions of human subjects were recorded, recognized, and analyzed to give statistical feedback of the overall emotions of a number of targets within a certain time frame. The result of the study allows automatic tracking of user's behavior based on a limited set of observations.

Keywords: Deep Neural Networks, Computer Vision, Emotion Classification, Gesture Classification

1. Introduction

Recognition of human behavior can be most efficiently achieved by visually detecting facial features and specific body movements, such as gestures. Using computer vision and machine learning algorithms for processing these features, recorded by infrared cameras, we can classify emotional states and behavioral patterns of multiple targets. The aim of this paper is to provide statistical observations and measurements of human behavior during the standard interaction with a user interface of a commonly used software. For academic purposes, we have chosen a very limited number of emotional states and behavioral patterns by studying only type of such standard interaction: the interaction of a user with typical ATM equipment, since it provides us with very distinctive patterns of 'typical' and 'non-typical' behavior and facial expressions. During this study, we observed the behavior of human subjects during standard interaction with the ATM versus non-standard interaction. Automated analysis of these behaviors with the machine learning techniques allowed us to train a complex convolutional neural network (CNN) to classify behavior of a user by classification both body movements and facial features. Such a feedback can provide important measures for user response to an interaction with any chosen system with a limited number of gestures involved. We use infrared cameras to automatically detect features and the movements of the limbs in order to classify user behavior into typical or untypical for the kind of task he is performing.

We restrict ourselves to only one type of interaction; however, this kind of classification task is very useful in the number of applications, where the number of gestures of the human is limited, such as:

- Customers at the various types of automated machines. For this category of users, the algorithm can be used for detection of unusual/fraudulent behavior to decrease the workload of the closed-circuit television (CCTV), or video surveillance, operators who monitor users of these machines:
    - customer at the ATM machine;
    - customer at the ticket machine in the underground;
    - customer at the automated cashier in the countries, where such payment type is widely used;
- Drivers. For this category of users, the algorithm can be used for detection of dangerous actions and preventing the unwanted consequences, such as sleeping, loss of attention etc.:
    - train driver in the train line/underground;
    - track driver;
    - automobile driver;

- Workers. Here, we could classify correct vs. incorrect actions, identify such unwanted stated as loss of attention, sickness, tiredness etc.
    - assembly line workers;
    - construction workers (e.g. on high buildings, underground, mines).

The aim of current paper is to analyze the person's actions during the interaction with a user interface and implement the algorithm, which will be able to classify the human behavior (normal vs. abnormal) in real time (Perez-Sala et al., 2014).

The processing of facial features with infrared cameras for academic and industrial purposes is rapidly developing: it is used in gaming system domain (MacCormick, 2013; Vera et al., 2011), as well as in the security systems[1]. In this study, we also use infrared cameras, imbedded in the Kinect Microsoft system, however the usage of other types of infrared cameras is also possible. The output from the infrared camera (point cloud) is used as an input to the neural network architecture, which classify the user's behavior based on his gestures and facial expressions. User's movements are classified into two categories (typical and non-typical), whereas facial expressions are classified into six basic emotions: anger, disgust, fear, happiness, sadness, and surprise (Erkman & Friesen, 1978).

To solve the problem of emotion and gesture recognition, we use convolutional neural networks (CNN), which proved to be extremely effective for classification of the large amount of data. Despite the similarity between artificial neural network and convolutional neural network, CNN is more effective because it uses alteration of convolutional and subsampling layers. The contribution of the current study is mainly in the application of the algorithms: we combine the particular type of NNs with the infrared input for recognition and classification of facial features and body movements. As far as we aware, this type of application has not been mentioned in the literature so far.

2. Background

2.1. Literature review

The type of the convolutional network, described in this study, was first proposed by Fukushima in the theoretical model, called "Neocognitron" (Fukushima, 1980). One of the earliest representatives of this class of models, Neocognitron is a hierarchical network in which feature complexity and translation invariance were alternately increased in different layers of simple and complex cells. The recognition occurred in different levels of processing hierarchy by a template match, and a pooling operation over units tuned to the same feature but at different positions. Starting with the Neocognitron for translation-invariant object recognition, several hierarchical models, employing pooling mechanisms and convolutional layers, have been proposed. The concept of pooling of units tuned to transformed versions of the same object or feature was subsequently proposed by Perrett and Oram (Perrett & Oram, 1993). They proposed a scheme, which provides 3D object recognition through 2D shape description. This scheme involves viewer-centred description of objects. Shape components were used as features for object comparison. First, these components were used to activate representations of the approximate appearance of one object type at one view, orientation and size. The invariance was achieved through a series of independent analyses with a subsequent pooling of results, which are performed at each pooling stage. Therefore, the system performed parallel processing with computations performed in a series of hierarchical steps.

In 1997 similar type of neural network was proposed by Logothetis et al (Logothetis et al., 1994). They constructed a regularization RBF network for 3D object recognition, based on radial-basis functions (RBFs). This model had a continuation, proposed by Riesenhuber and Poggio, widely known as HMAX model. The structure of HMAX model is similar to Fukushima's Neocognitron with its feature complexity-increasing simple-cell (S) layers and invariance-increasing complex cell (C) layers (Riesenhuber & Poggio, 1999; Riesenhuber & Poggio, 2000). HMAX uses another type of pooling mechanism, which is called MAX operation. This mechanism allows to

[1] Aurora face recognition system http://www.facerec.com/deep-learning/

increase invariance in the complex cell layers. It uses the following principle: the most strongly activated afferent of the C-cell determines the response of the pooling unit, providing the ability to isolate essential feature from background and thus build feature detectors invariant to scale and translation changes. More complex features in higher levels of HMAX are built from simpler features. The tolerance to deformations in their local arrangement is achieved by the invariance properties of the lower level units. From the point of view of biological plausibility, visual areas of the primate cortex in this model are considered as a number of modules (V1-V4, IT, PFC), modelled as a hierarchy of increasingly sophisticated representations, naturally extending the model of simple to complex cells of Hubel and Wiesel (Hubel & Wiesel, 1962).

An important breakthrough of CNNs came with the widespread use of the backpropagation learning algorithm for multi-layer feed-forward NNs. LeCun et al. (LeCun et al., 1998) presented the first CNN that was trained by backpropagation and applied it to the problem of handwritten digit recognition. The term Convolutional Neural Network refers to NN models that are similar to the one proposed by LeCun et al., which is actually a simpler model than the Neocognitron and its extensions, mentioned above.

Since 2012, when deep neural network first demonstrated their performance, they were used in a large number of applications for computer vision. Alex Krizhevsky et al. trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different 1000 different classes (Krizhevsky, Sutskever, & Hinton, 2012). On the test data, they achieved extremely small error rates, which were considerably better than the previous state-of-the-art. The proposed neural network had 60 million parameters and 650,000 neurons, consisted of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. The softmax function, or normalized exponential, is a generalization of the logistic function that converts a K-dimensional vector *z* of arbitrary real values to a K-dimensional vector *sigma(z)* of real values in the range (0, 1) that add up to 1. In neural network simulations, the softmax function is often implemented at the final layer of a network used for classification.

2.2. Related work

Recently, the field of applications of deep learning for sensor and infrared cameras processing is rapidly evolving. Recently, similar algorithm was applied for night vision systems in vehicles (Wang et al., 2016). In that study, vehicle candidates for classification are detected from the infrared frame, contours are generated by using a local adaptive threshold based on maximum distance. The obtained vehicle candidates are verified using a deep belief network (DBN) based classifier. Also deep neural networks were recently applied for detection of faces in airports for robust feature recognition[2]. The system recognizes faces and compares them with the database of passengers. It is implemented in Heathrow and Manchester airports.

3. Methodology

3.1. Model Architecture

A deep neural network (DNN) is an artificial neural network (NN) with multiple hidden layers of units between the input and output layers. Similar to shallow ANNs, DNNs can model complex non-linear relationships. DNN architectures, e.g., for object detection and parsing generate compositional models where the object is expressed as a layered composition of image primitives (LeCun et al., 1989). The extra layers enable composition of features from lower layers, giving the potential of modeling complex data with fewer units than a similarly performing shallow network.

[2] Aurora face recognition system http://www.facerec.com/deep-learning/

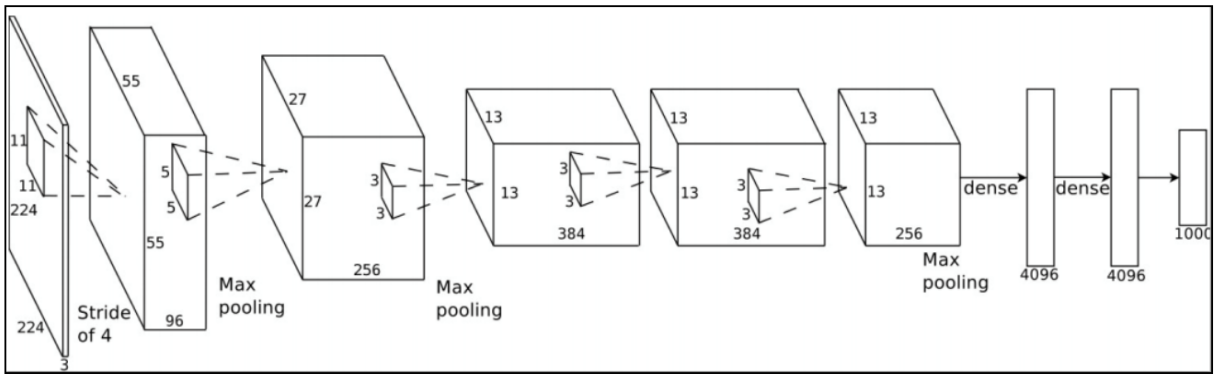

Figure 1. CNN architecture (adopted from Krizhevsky et al. '12)

The architecture of a CNN can be described as following. A small pixel region goes to input neurons and then connects to a first convolution hidden layer (Figure1). There we can see a set of learnable filters, which are activated during the presentation some particular type of feature in pixel region in the input. On this phase, CNN does shift invariance, which is carried by feature map. Subsampling layer goes next. There we have two processes: local averaging and sampling. As a result, we get declining resolution of feature map. To correspond this task CNN needs supervised learning. Before starting the experiment, we gave a set of labeled videos with different emotional experience. The system analyses images and finds similar features. Then the system creates a map, where it arranges videos in accordance with similar features. Thereby, images with similar emotions form certain class. To test the system, we add other videos and correct the system when it refers them improperly.

The proposed model consists of four convolutional layers, followed by max-pooling layers, and three fully-connected layers with a final classificatory presented with MLP (with six basic outputs, corresponding to basic emotions for emotion classification and two outputs for motion classification for typical and non-typical behavior). The input data was presented as infrared camera output.

3.2. Model implementation and training

The computations were performed on Python, using Theano CNN library. Theano is a Python library that allows defining, optimizing, and evaluating mathematical expressions involving multi-dimensional arrays efficiently (Bastien et al., 2012). The model was trained with the trained data and model evaluation was performed on the test data with the the *k*-fold cross-validation (for details, see next subsection). The computations were performed on the Amazon EC2 machine[3].

3.3. Model evaluation

The validation of the neural network model was performed with the leave one out cross validation (LOOCV) technique. The use of LOOCV was essential for appropriate estimation of optimal level of regularization and parameters (connection weights) of neural network obtained (1). Cross-validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. LOOCV is a particular case of leave-*p*-out cross-validation. Leave-*p*-out cross-validation (LpOCV) involves using *p* observations as the validation set and the remaining observations as the training set. This is repeated on all ways to cut the original sample on a validation set of *p* observations and a training set. LpO cross-validation requires to learn and validate C^n_p times (where *n* is the number of observations in the original sample). In Leave-*one*-out cross-validation we assume $p = 1$. However, for our purpose LOOCV appeared to be relatively slow. Therefore, the validation of the CNN network results was performed with the K-fold cross-validation technique (Golub & Van loan, 1996). In *k*-fold cross-validation, the original sample

[3] https://portal.aws.amazon.com

is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $k - 1$ subsamples are used as training data. The cross-validation process is then repeated k times (the *folds*), with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method over repeated random sub-sampling (see below) is that all observations are used for both training and validation, and each observation is used for validation exactly once. 10-fold cross-validation is commonly used, but in general k remains an unfixed parameter. When $k=n$ (the number of observations), the k-fold cross-validation is exactly the leave-one-out cross-validation.

4. Results and discussion

4.1. Infrared input processing

In this study, we used one of the approaches to recognition of gestures is body tracking: classification of the body movements. One of the classification techniques for this method is pattern recognition: i.e. special video/infrared camera recognizes human actions: waving, jumping, hand gestures etc. Among the first successful representatives of this technology are Kinect from Microsoft (MacCormick, 2013). The Kinect uses structured light and machine learning as follows:

- The depth map is constructed by analyzing a speckle pattern of infrared laser light.
- Body parts are inferred using a randomized decision forest, learned from over 1 million training examples.
- Starts with 100,000 depth images with known skeletons (from a motion capture system).
- Transforms depth image to body part image.
- Transforms the body part image into a skeleton.

4.2. Psychological experiments

We conducted a series of experiments in order to evaluate how effectively the proposed system can detect normal vs. abnormal behavior of customer during interaction with ATM. For the purposes of experiment, we developed an ATM simulation software that was used in the stand-alone terminal. During the interaction session, body movements of users and facial expressions were recorded by a camera, mounted on the top of the terminal. These records were later evaluated by human observers; and behavior, displayed on these records, was classified as typical or non-typical. In order to preserve the uniformity of the data, we showed the videos on the same equipment which were used during the ATM experiment session. Thirty healthy subjects, age 21-37, with normal or corrected-to-normal vision, participated in the experiment. Simultaneously, the data from two series of experiments was processed with an infrared camera and used as an input to the CNN algorithm. Each subject performed 10 sessions with the ATM-simulation software and 5 video session. During each session, the recognition of the upper-body movements (in the range of the camera, mounted on the top of the typical ATM machine) was performed together with facial features classification and recognition. Among thirty subjects, we used 22 as examples of 'normal' behavior and 8 as examples of 'abnormal' behavior.

4.3. Comparison to related work

The field of application of DNN to sensor input recognition is relatively new, but rapidly developing. A large number of studies exist, but to our knowledge, this particular application has not been studies so far. Among similar works, we can mention our previous works, in which we have applies RBFn-SOM model to the same problem (Veaceslav & Cojocaru, 2015; Veaceslav, 2016). In comparison to this type of architecture, we have managed to achieve a better accuracy (1,5 – 2%), but the computational costs of application of DNNs is much higher.

4.3. Results

In this study, we developed a NN model for recognition of body movements on two types (typical and non-typical) and facial expression accuracy (of 18% and 38% error rate, respectively). These results are achieved independently and combined afterwards with a simple classification

algorithm. To improve system's results, the proposed model requires a large amount of training data, which we cannot be easily obtained. Therefore, the natural continuation of current research would be conducting further field tests to obtain more training data and improve performance.

References

Perez-Sala, X., Escalera, S., Angulo, C., & Gonzalez, J. (2014). Survey on Model Based Approaches for 2D and 3D Visual Human Pose Recovery. *Sensors*, vol. 14, pp. 4189-4210.

MacCormick, J. (2013). How does Kinect work. John MacCormick, "How does the kinect work?" Retrieved from *http://users.dickinson.edu/~jmac/selected-talks/kinect.pdf.*

Vera, L., Gimeno, J., Coma, I., & Ferńandez, M. (2011). Augmented mirror: interactive augmented reality system based on Kinect," *Human-Computer Interaction–INTERACT 2011*, pp. 483–486, 2011.

Aurora face recognition system http://www.facerec.com/deep-learning/

Erkman, P. & Friesen, W. (1978). Facial action coding system: A technique for the measurement of facial movement. *Consulting Psychologists Press*, Palo Alto, 1978.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for mechanism pattern recognition unaffected by shift in position, *Biological Cybernetics*, 36(4):193-202.

Perrett, D. I., & Oram, M. W. (1993). Neurophysiology of shape processing. Image and Vison Computing, 11(6), 317-333.

Riesenhuber, M. & Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nature Neuroscience 2:1019–1025.

Riesenhuber, M. & Poggio, T. (2000). Models of Object Recognition. Nature Neuroscience 3(supp.): 1199-1204.

Hubel, D. & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, Journal of Physiology, 160(1): 106–154.2.

LeCun, Y.A., Bottou, L., Orr, G.B. & Müller, K.R. (1998). Efficient BackProp, *Neural networks: Tricks of the trade*, pp. 9-48.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012.) ImageNet Classification with Deep Convolutional Neural Networks. *Proc. Neural Information and Processing Systems*. Available at http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., & Jackel, L.D. (1989) Backpropagation Applied to Handwritten Zip Code Recognition, *Neural Computation,* vol. 1, pp. ,541-551.

Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I., Bergeron, A., Bouchard, N., Warde-Farley, D., & Bengio, Y. (2012). "Theano: new features and speed improvements". *NIPS 2012 deep learning workshop*.

Golub, G.H. & Van loan, C.G. (1996). Matrix Computations (3 ed.) Ithaka, NY, p. 784.

Wang, H., Cai, Y., Chen, X., & Chen, L. (2016). Night-Time Vehicle Sensing in Far Infrared Image with Deep Learning, Journal of Sensors, v.2016, p. 8.

Veaceslav, A. & Cojocaru, S. (2015). Measuring human emotions with modular neural networks and computer vision based applications. Computer Science Journal of Moldova, vol.23, no.1(67), pp. 40-61.

Veaceslav, A. (2016). Measuring human emotions with modular neural networks. The proceedings of the 7th International Multi-Conference on Complexity, Informatics and Cybernetics: IMCIC 2016, March 8 - 11, 2016, Orlando, Florida, USA.

Logothetis, N., Bricolo, E., Poggio, T. (1994). 3D Object Recognition: A Model of View-Tuned Neurons. Retrieved from http://papers.nips.cc/paper/1296-3d-object-recognition-a-model-of-view-tuned-neurons.pdf