Micro Expression Detection and Recognition from High Speed Cameras using Convolutional Neural Networks

Diana Borza, Razvan Itu and Radu Danescu
Computer Science Department, Technical University of Cluj-Napoca,
28 Memorandumului Street, 400114, Cluj Napoca, Romania

Keywords: Micro Expression Recognition, Micro Expression Detection, Convolutional Neural Networks.

Abstract: In this paper, we propose a micro-expression detection and recognition framework based on convolutional neural networks. This paper presents the following contributions: the relevant features are learned by a convolutional neural network that uses as input difference images of three equally spaced frames from the video sequence, capturing important motion information. Next, a sliding time window is used to iterate through the video sequence and the output of the network in order to eliminate false positives. The method was trained using images from two publicly available micro-expression databases. The effectiveness of the proposed solution is demonstrated by the experiments we performed, from which a recognition rate of 72.22% was obtained.

1 INTRODUCTION

Micro expressions (ME) are brief facial expressions that last between 1/15 and 1/25 of a second, and they occur when people try to hide their feelings, either as a suppression (deliberate concealment) or repression (non-conscious concealment) (Ekman, 2009). Therefore, micro expressions are crucial in understanding human behavior, detecting concealed emotions, security and surveillance systems etc. For example, the Transportation Security Administration in the United States launched the SPOT program (Wikipedia, 2017), in which airport employees are trained to observe the passengers with suspicious behavior by analyzing micro expressions and conversational signs.

Although automatic recognition of facial expressions has been intensively studied over the last years, automatic analysis of micro-expressions is a relatively new subject. There are two major challenges in the automatic detection of micro-expressions field. First, they are involuntary movements and therefore training data is hard to gather. Several micro-expression databases are now publicly available (Rautio, H., 2013); (Yan et al, 2013), (Yan et al, 2014). Another challenge related to micro expression analysis is data labelling; this process is subjective, time-consuming and challenging even for trained human labelers. Therefore, not all the databases offer images labelled at the same granularity level. For example, SMIC database (Rautio, 2013) labels the images into 3 categories: positive, negative and surprise, while CASME database (Yan et al, 2013) uses emotion categories: amusement, sadness, disgust, surprise, contempt, fear, repression and tension.

The second major issue related to ME is that they are visible only a small number of frames, requiring accurate motion detection and tracking algorithms. Most solutions which tackle micro-expression analysis use high speed cameras to overcome this problem. Another challenge is the detection of micro-expressions in real world situations, where people can move their heads freely or where other movements are present.

A typical ME has three key moments in time: *onset* (the moment when the ME starts), *apex* (the moment of maximum amplitude), and *offset* (the time when the ME fades out). The symmetry and amplitude of the facial changes with respect to these three moments can be used to detect the concealed emotions and deceit.

In this paper, we propose an original micro-expression detection and recognition framework based on convolutional neural networks. The classical stages of machine learning (feature extraction and learning) are replaced by the

Borza D., Itu R. and Danescu R.
Micro Expression Detection and Recognition from High Speed Cameras using Convolutional Neural Networks.
DOI: 10.5220/0006548802010208
In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP International Conference on Computer Vision Theory and Applications)*, pages 201-208
ISBN: 978-989-758-306-3

convolutional neural network which also learns the features which are relevant in the classification problem. In order to incorporate motion information, the input volume of the network is formed by two image differences frames between three equally spaced frames. The first image difference is computed between the (potential) *onset* frame and the (potential) *apex* frame, while the second image difference is between the (potential) *apex* frame and the (potential) *offset* frame. Another original contribution highlighted by this paper is the use of a sliding time window which iterates through the video sequence and feeds the corresponding frames to the neural network. The response of the neutral network is further processed in order to eliminate micro-expression false positives and to merge together responses that belong to the same micro-expression, as described in Section 3.3.

The proposed method is scalable and modular, and it has the main advantage that by simply using more training data it can "learn" to distinguish between more micro-expression classes, or between micro-expressions and other facial movements, such as blinks and macro-expressions.

The remainder of this work is organized as follows: in section 2 we present the recent advances in the field of micro-expressions detection and recognition. In Section 3 we discuss the general outline and we detail the proposed solution. The experimental results are discussed in Section 4, and the conclusions and future directions of improvement are detailed in Section 5.

2 STATE OF THE ART

Automatic micro expression recognition implies three main steps: (1) selecting the facial regions of interest, (2) defining and extracting the classification features and (3) the actual recognition of the micro expression using the selected features and state of the art machine learning algorithms.

The first step involves establishing the facial area which will be analyzed; for this step, the face is either divided into several rectangular segments around the most prominent facial features (eyes, lips, nose) (Polikovsky et al, 2009) (Polikovsky et al, 2013) (Godavarthy et al, 2011) or a more complex deformable model is used to split the face into more accurate regions (Liu et al, 2015), (Pfister et al, 2011). Other methods (Liong et al, 2016), (Li et al, 2017) divide the face geometrically into *n* equal cells and analyze the movement within these cells.

In the second step, several spatiotemporal descriptors are extracted from the defined regions of interest in order to describe the facial transformations that occur over time. Several types of descriptors can be used: dense optical flow (Liu et al, 2015), optical strain (Liong et al, 2016), texture based descriptors – Local Binary Patterns in Three Orthogonal Planes (LBP-TOP) (Pfister et al, 2011) or 3D histogram of oriented gradients (Polikovsky et al, 2009), (Polikovsky et al, 2013). Finally, in the last step, a machine learning algorithm (which can be supervised (Pfister et al, 2011) or non-supervised (Polikovsky et al, 2009), (Polikovsky et al, 2013)) is used for the actual recognition of micro expressions.

Recently, several works tackle both the problem of micro expression detection and the problem of micro expression recognition. The method presented in (Liong et al, 2016) uses optical strain features in two different ways: first the classification is performed based solely on optical strain information, and second, the optical strain information is used for weighting the LBP-TOP features. The best results are obtained by the second method. In (Li et al, 2017), the authors propose a general micro expression analysis framework that performs both micro expression detection and recognition. The detection phase does not require any training and exploits frame difference contrast to determine the frames where movement occurred. For the recognition phase, several descriptors are extracted (LBP-TOP, Histogram of Oriented Gradients (HOG) and Histogram of Image Gradient Orientation (HIGO)) and a support vector machine (SVM) classifier is used to recognize the micro expression.

The majority of the works reported in the literature follow "classical" stages of machine learning: region of interest selection, the extraction of motion descriptions and a classifier (usually a SVM) to recognize the exact class of the micro-expression. In this paper, we tackle the problem of micro-expression detection and recognition using a deep convolutional neural network, which is able to automatically learn the motion features and classify micro-expression from high speed video sequences. As opposed to other method, the proposed solution does not require complex face alignment or normalization (Li et al, 2017) or the extraction of motion features (as they are automatically learned by the network). The network takes as input two frame differences in order to capture the motion variation across video frames.

3 PROPOSED SOLUTION

3.1 Solution Outline

We propose an original micro-expression detection and recognition framework based on convolutional neural networks; the neural network automatically selects the relevant features from the input image and performs the classification. The outline of the proposed solution is depicted in Figure 1.

We will first define the concepts involved in the proposed solution. A sliding time window will be used to iterate over the input video. Based on the frame rate of the capturing device we compute the average number of frames Δt a micro expression is visible. This parameter was chosen based on the physiological duration of a micro-expression: 1/15th of a second (Ekman, 2009). At each step, we inspect the current frame *Ft* in order to determine if a micro expression occurred at this frame and to recognize it if the case. We also use the onset *Ft* - *k* and offset frames *Ft* + *k* in order to extract the movement features, where $k = (\Delta t + 1)/2$. The first and last *k* frames from the video are excluded, because in this case the onset and offset frames will exceed the video boundaries.

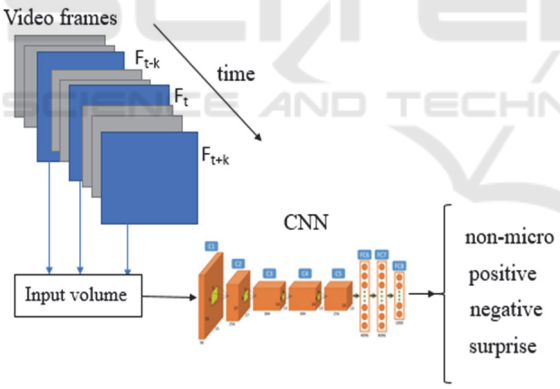

Figure 1: Solution outline.

Each frame of the input video, *Ft*, together with its corresponding onset and offset frames, are fed to a convolutional neural network (Li et al, 2017), which will classify the state at time *t*, into one of the following classes: micro (positive, negative, surprise) or non-micro. The raw responses of the convolutional neural network are further processed in order to establish the exact time frame when the micro expression occurred and to filter out false positives.

It is generally acceptable that the micro-expressions correspond to the seven universal emotions: surprise, anger, fear, sadness, disgust, contempt, happiness and fear. However, as convolutional neural networks require large amounts of training data we chose a taxonomy with only three classes: positive, negative and surprise for the micro-expression recognition part.

The distribution of micro-expressions classes in the CASME II database (Yan et al, 2014) is the following: happiness – 32 sequences, surprise – 25 sequences, fear – 2 sequences, disgust - 63 sequences, sadness -7 sequences, repression – 27 sequences and others – 99 sequences. Therefore, we choose a three-class taxonomy: positive (happiness), surprise and negative (disgust, fear, sadness). This taxonomy is often used in the literature (Rautio, H., 2013).

3.2 Classification using Convolutional Neural Networks

Convolutional neural networks (CNN) have recently received particular interest from the scientific community since they achieved near human performance in a variety of computer vision tasks. These networks replace the traditional stages of classification: feature pre-processing, feature extraction. Classification is done by "learning" the features that are relevant (to the classification problem). The structure of these networks mimics the mechanisms found in the visual cortex of the biological brain. The neurons in a CNN are arranged into 3 dimensions: (width, height and depth) and the neurons within a layer are only connected to a small region of the previous layer, to the receptive field of the neuron. Numerous topologies of convolutional neural networks have been proposed, but they all share three types of basic layers: the *convolutional layer* (*CONV*) – flowed by a non-linearity, e.g. Rectified Linear Unit (ReLU) –, the *pooling layer* and the *fully connected* (*FC*) *layer*.

The CONV layer is the main building block of a CNN and it consists of a set of learnable kernels. During the forward pass, several feature maps are computed by convolving each kernel across the width and height of the input volume. Pooling layers are usually placed immediately after the CONV layers and their main purpose is to simplify the information of the output of the convolutional layers by reducing the spatial size of the representation. Pooling layers control over-fitting by reducing the number of parameters and computations. The fully connected layer is composed of neurons that have full connections to all the activations from the previous layer, as seen in classical neural networks.

The FC layers are used to perform high level reasoning. In addition to these three, general type of layers, some works also add normalization layers in order to mimic the inhibition schemes observed in the biological brain.

The proposed solution uses a variant of AlexNet (Krizhevsky et al, 2012) network, originally proposed for the ILSVRC (ImageNet Large Scale Visual Recognition Challenge) contest, which aimed at classifying different objects into 1000 different classes. The network consists of three convolutional layers and three fully connected layers. In order to avoid overfitting, after each FC layer a dropout layer is used (i.e. randomly setting the output value of network neurons to zero). We modified the initial topology of the network, such that the input has a depth equal to 2 and, of course, the last fully-connected layer of the network contains only three neurons which are mapped to the probability of each micro-expression class.

All the images from the training and test datasets are cropped so that they only contain the regions where micro expressions cause facial movements: eyes, eyebrows, nose, mouth and chin area. We first apply a facial landmark localization algorithm (Cox et al, 2013), and based on these landmarks and anthropometric constraints, the image is cropped so that it contains only the relevant features. Figure 2 shows an example of this preprocessing step.

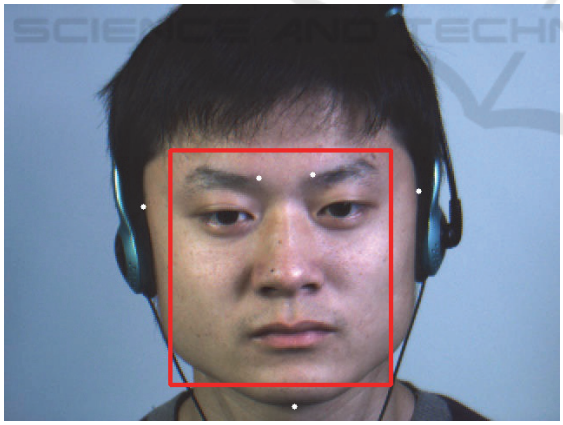

Figure 2: Region of interest (in red) selection based on facial landmarks (in white) and anthropometric constraints.

We compute the image differences between the onset and current frame, and the onset and offset frame, respectively. These two difference images are stacked together to form a volume with depth 2 and are used as input of the CNN. Figure 3 shows the input used for the convolutional neural network.

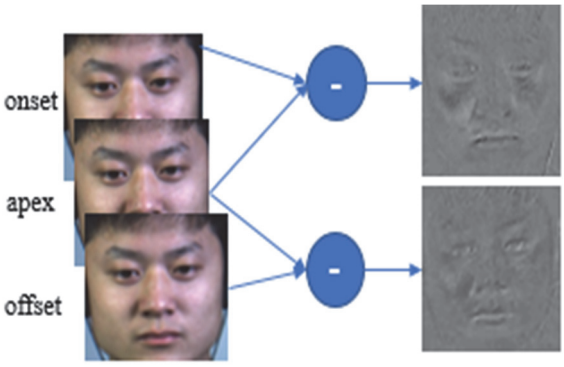

Figure 3: CNN input.

These difference images are scaled to 227×227 resolution and processed directly by the network. The layers of the network are defined as follows:

- In the first CONV layer (C1), 96 filters of size 2×7×7 are applied to the input volume, followed by a non-linearity, a max pooling layer and a local response normalization layer;
- The second CONV layer (C2) contains 256 filters of size 96 × 5 × 5 pixels. C2 is followed by ReLU, a max pooling layer and a local response normalization layer;
- The last CONV layer (C3) operates on the 256 × 14 × 14 result from C2 by applying a set of 384 filters of size 256 × 3 × 3 pixels, followed by ReLU and a max pooling layer;
- The first FC layer (FC1) contains 512 neurons and is followed by a non-linearity and a dropout layer;
- A second FC layer (FC2) receives the 512-dimensional output of FC1 and contains 512 neurons, followed by a ReLU and a dropout layer;
- The last FC layer (FC3) maps to the micro expressions classes.

Finally, the output of the last fully connected layer is fed to a soft-max layer that assigns a probability for each class. For the minimization of the loss function, we used the recently proposed Adam optimizer (Kingma and Ba, 2014). The prediction itself is made by taking the class with the maximal probability for the given test image. The learning rate is initially set to 0.01, $\beta_1 = 0.9$ and $\beta_2 = 0.9$. The learning rate decays after each epoch.

3.3 ME Detection and Recognition

In the remainder of this section we describe the micro expression spotting and recognition algorithms. Micro expression spotting or detection refers to the process of determining if a micro

expression occurred at frame *t*. The Micro expression recognition algorithm is used to determine the actual type of the micro expression (positive, negative or surprise) that occurred.

For the micro expression detection part, a sliding time window is used. Each frame *t*, together with its corresponding onset and offset frames are fed to the neural network and its response is saved to a preliminary result feature vector *R* at position *t*. As the first and last *k* are ignored, their corresponding positions are set to neutral class, where *k* is half the average micro-expression duration.

This feature vector is further analyzed in order to determine the exact time frame when a micro expression occurred. First, the feature vector is binarized as follows:

$$RB_t = \begin{cases} 0, & R_t = neutral \\ 1, & R_t \in \{positive, negative, surprise\} \end{cases}$$

, where R_t is the CNN output class for frame *t* and *RB* is the corresponding binarized feature vector. In other words, we transform the feature vector so that it contains only two classes: non-micro (neutral, blink etc.) and micro. One would expect that the feature vector contains agglomerations / clusters of 1s around the micro expression time and 0 values in rest.

We select all the intervals that contain only values of '1' as micro expression interval candidates. The intervals that are too close to each other (the distance between the start of the first cluster and the end of the second one is less than $\Delta t/4$) are merged together. Finally, all the intervals with a length lower than $\Delta t/10$ are considered false positives as their duration is too short. Each interval represents a micro-expression and we select the apex frame as the centroid of the micro-expression.

Figures 4 and 5 show the raw response of the CNN on an input video sequence and the filtered response of the CNN using the proposed algorithm. The ground truth onset, apex and offset frames of the video sequence are also marked on the plot.

In Figure 5, it can be noticed that the false positive micro-expression interval (starting from frame 113 to frame 129) is filtered out because its length is too short. Also, some intervals are merged together as they are too close to each other.

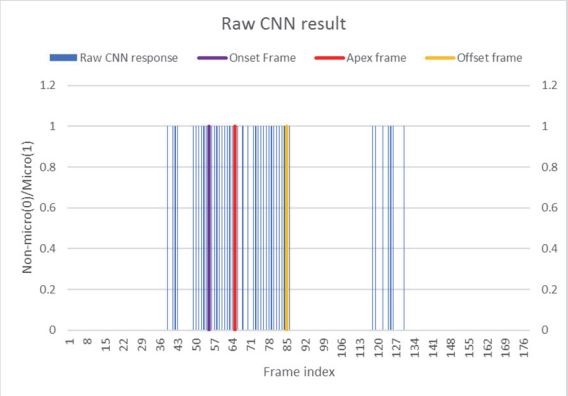

Figure 4: Raw CNN result on a video sequence.

Next, all the intervals that were selected as micro expression candidates are examined in order to recognize the actual micro expression that occurred at that time. Given a candidate micro expression interval [t_{start}, t_{end}], we determine the actual micro-expression that occurred at that time by a simple voting algorithm: the class that is predicted the most by the CNN.

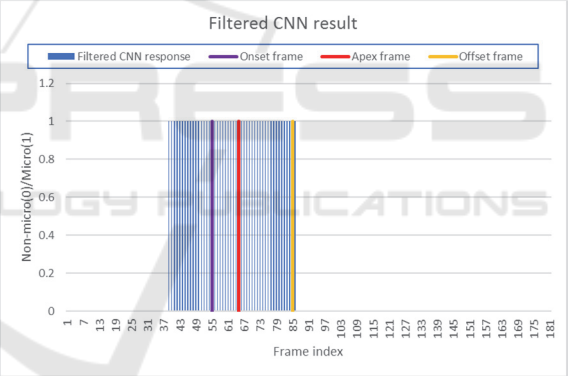

Figure 5: Filtered CNN result (false positives are eliminated and intervals that are "too" close are merged together).

4 EXPERIMENTAL RESULTS

4.1 Dataset Gathering

An important issue that needs to be addressed in the context of machine learning is the gathering of the training data, a time-consuming process, which also requires domain specific knowledge. As stated before, there are few available datasets for genuine micro-expression recognition. For the training data, we combined video sequences from CASME I (Yan

et al, 2013), CASME II (Yan et al, 2014) and SMIC (Rautio, H., 2013) databases.

CASME I (Yan et al, 2013) database contains 195 micro-expression clips captured with a 60-fps camera at 1280 × 720 (CASME A) and 640×480 (CASME B) spatial resolutions. 35 participants (having a mean age of 22.03) were involved in the gathering of the data. An extended version of (Yan et al, 2013), CASME II (Yan et al, 2014) database contains 247 micro-expression sequences, captured from 26 participants; the mean age of the subjects is 22.03 years, with 1.6 standard deviation. Each video sequence was captured by a high-speed camera (200 fps) with a resolution of 640 × 480 pixels. The video sequences are labelled with the following classes: happiness, disgust, surprise, repression and tense. Finally, the SMIC (Rautio, 2013) database contains 164 micro-expression videos, of 16 participants with mean age 28.1 years. The video sequences were captured using a high-speed camera (100 fps) at 640 × 480 pixels resolution, and the sequences were labelled using only three classes: positive (happiness), negative (sadness, fear and disgust) and surprise. Also, SMIC database also contains some examples with neutral sequences annotation.

One of the main drawbacks of CNNs is that they require a large amount of training data in order to give accurate results and to avoid overfitting. A common approach to enlarge the training dataset is to use augmentation.

We used the following augmentation techniques: (1) random horizontal flips, (2) brightness enhancement and (3) "reversing" the video sequences. As the micro-expressions are symmetrical: the subject starts from a neutral expression, the micro-expression briefly occurs and then the subject returns to a neutral expression, we also used the data in reversing order: offset – apex – onset.

In order to gather the training data, we use a sliding time window to navigate through the video sequence. If Δt is the average micro-expression duration for the current dataset, and apex is the ME apex frame, we use the following labelling criteria for the current frame t:

- If $t \in [0, t_{apex} - \delta \cdot \Delta t]$ or $t \in [t_{apex} + \delta \cdot \Delta t]$, then the frame t is labelled as non-micro;
- If $t \in (t_{apex} - \delta \cdot \Delta t, t_{apex} + \delta \cdot \Delta t)$, then the frame t is labelled with the corresponding ME label (negative, positive or surprise).

, where $\delta = 0.5$.

The total training (after data augmentation) set consists of approximately 15,000 images.

The evaluation of the network was conducted using the *leave one subject out* cross validation (LOSOCV) method; i.e. we selected video sequences from 3 subjects that were not used in the training process, and the network was evaluated on these sequences. We evaluated the proposed solutions only on images from CASME (Yan et al, 2013), (Yan et al, 2014) databases, as the sequences from SMIC database are too short in order to apply the proposed sliding time window micro-expression detection algorithm.

The raw performance of the convolutional neural network and the confusion matrix is shown in Tables 1 and 2 respectively:

Table 1: Raw performance of the CNN.

	Precision	Recall	F1-score	Support
neutral	0.96	0.76	0.85	2565
positive	0.29	0.92	0.44	62
negative	0.38	0.80	0.52	422
surprise	0.87	0.58	0.70	112
Average	0.87	0.76	0.79	3161
Accuracy	75.89%			

From the confusion matrix, it can be noticed that the majority of confusions are between the neutral-negative and surprise-neutral classes. We observed that the majority of the false positives are eye blinks. This is an expected behavior, as blinking and surprise and negative micro expressions cause facial movements in the eyebrow area.

Table 2: Confusion matrix for the raw CNN classification.

		Predicted				Accuracy
		Neutral	*Positive*	*Negative*	*Surprise*	
Actual	*Neutral*	1941	81	533	10	75.67%
	Positive	5	57	0	0	91.93%
	Negative	25	61	336	0	79.62%
	Surprise	41	0	6	65	58.02%

After the analysis of the CNN output, all the detected ME intervals are compared with the ground truth labels, to decide whether they correspond to the annotated location of the micro expression. In Table 3 we show the detection rate performance of the proposed solution after the raw response from the CNN was processed. If the centroid of a detected micro-expression interval is within the frame range of [onset, offset] of a labeled micro expression sequence it is considered a true positive, otherwise it will be counted as a false positive. All the false negatives correspond to the "negative" micro

expression class. As stated before, the video sequences belonging to 3 subjects (22 micro expression clips) were not included in the training data, and these videos were used only for the validation of the proposed solution.

Table 3: Detection performance.

True positives	17
False negatives	5
False positives	8

Finally, in Table 4 we show the confusion matrix for the proposed micro expression recognition algorithm.

Table 4: Recognition performance.

	Precision	*Recall*	*F1 Score*
Positive	0.67	0.67	0.67
Negative	0.71	0.91	0.80
Surprise	1.00	0.25	0.40
Average	0.77	0.72	0.69
Accuracy	72.22%		

Our method is at least comparable, when not better with recent state of the art works. In Table 5 we present the comparison of the proposed solution with other state of the art works. ACC stands for accuracy, FPR for false positive rate and TPR for true positive rate.

Table 5: Detection performance.

Method	Detection	Recognition accuracy
(Rautio, 2013) *	ACC: 65.49 %	49.30 %
(Yan et al, 2014)	N/A	63.41%
(Liong et al, 2016) *	ACC: 74.16%	63.16%
(Li et al, 2017)	TPR: 70%	57.49%
Proposed solution	**TPR: 77.27%**	**72.22%**

Methods marked with an asterisk * were evaluated on SMIC database. To detect the micro expressions, most of the works were only evaluated on SMIC database. Therefore, the numerical comparison with these methods might not be relevant.

A video example of the detection performance using convolutional neural networks can be found at: https://drive.google.com/open?id=0ByAKFSXshk1AdkdlelpCV29nWFk.

The network was trained in 50 epochs, using a batch size of 128 samples on an NVidia Tesla K80 GPU.

In Figure 6 we plot the minimization of the loss function over the training process and the accuracy evolution.

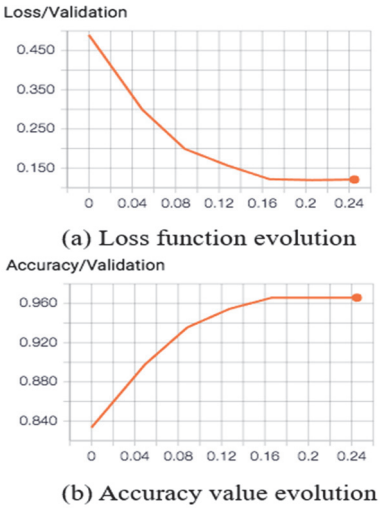

Figure 6: CNN learning process: (a) loss function evolution, (b) accuracy evolution.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we have presented an original method of detecting and recognizing micro expressions from high speed camera using convolutional neural networks. The CNN processes difference images between the onset frame and the apex frame and the apex frame and the offset frames, respectively, and outputs one of the following classes: non-micro, positive, negative and surprise. The network was trained using images from three publicly available micro-expression databases.

As a future work, we plan to gather more data for the training process so that more data variation is present and to have enough information to classify into the six canonical facial expressions – disgust, sadness, happiness, fear, anger and surprise.

In order to increase the magnitude of the micro-expression movement, we plan to apply Eulerian motion magnification (Li et al., 2017).

Several other features for the input volume of the network are envisioned, dense optical flow or dense trajectories.

ACKNOWLEDGEMENTS

This work was supported by the MULTIFACE grant

(Multifocal System for Real Time Tracking of Dynamic Facial and Body Features) of the Romanian National Authority for Scientific Research, CNDI–UEFISCDI, Project code: PN-II-RU-TE-2014-4-1746.

REFERENCES

Ekman, P., 2009. *Telling Lies: "Clues to Deceit in the Marketplace, Politics, and Marriage"*, W. W. Norton & Company.

Wikipedia, 2017 (28.04.2017), *SPOT (TSA program)*, [Online]. Available: https://en.wikipedia.org/wiki/SPOT_(TSA_program).

Rautio, H., 2013. *SMIC - Spontaneous Micro-expression Database*. University of Oulu, [Online], Available: http://www.cse.oulu.fi/SMICDatabase.

Yan, W.-J., Wu, Q., Liu, Y.-J., Wang, S.-J. and Fu, X., 2013. *CASME database: a dataset of spontaneous micro-expressions collected from neutralized faces*, In FG. IEEE, pp. 1–7.

Yan, W.-J., Li, X., Wang, S.-J., Zhao, G., Liu, Y.-J., Chen, Y.-H. and Fu, X., 2014. *CASME II: An improved spontaneous micro-expression database and the baseline evaluation*, PloS one, vol. 9, no. 1.

Pfister, T., Li, X., Zhao, G. and Pietikainen, M., 2011. *Recognising Spontaneous Facial Micro-expressions*, In 2011 IEEE International Conference on Computer Vision (ICCV), Barcelona.

Polikovsky, S., Kameda, Y. and Ohta, Y., 2009. *Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor*, In 3rd International Conference of Crime Detection and Prevention, ICDP, London.

Polikovsky, S., Kameda, Y. and Ohta, Y., 2013. *Facial Micro-Expression Detection in Hi-Speed Video Based on Facial Action Coding System (FACS)*, In IEICE TRANSACTIONS on Information and Systems, vol. E96, no. 1.

Godavarthy, S., Goldgof, D., Sarkar, S., and Shreve, M., 2011. *Macro- and micro-expression spotting in long videos using spatio-temporal strain*, In Automatic Face & Gesture Recognition and Workshops (FG 2011).

Liu, Y.-J., Zhang, J.-K., Yan, W.-J., Wang, S.-J., and Zhao, G., 2015. *A Main Directional Mean Optical Flow Feature for Spontaneous Micro-Expression Recognition*, IEEE Transactions On Affective Computing, vol. 99.

Liong, S.T., See, J., Phan, R. C-W., Oh, Y.H., Le Ngo, A.C., Wong, K.S., and Tan, S.W., 2016. *Spontaneous subtle expression detection and recognition based on facial strain*. Signal Processing: Image Communication 47, pp: 170-182.

Li, X., Xiaopeng, H.O.N.G., Moilanen, A., Huang, X., Pfister, T., Zhao, G., and Pietikainen, M., 2017. *Towards Reading Hidden Emotions: A Comparative Study of Spontaneous Micro-expression Spotting and Recognition Methods*. IEEE Transactions on Affective Computing, Volume: PP Issue: 99.

Krizhevsky, A., Sutskever, I., and Hinton, G. E., 2012. *Imagenet classification with deep convolutional neural networks*. Advances in neural information processing systems.

Cox, M., Nuevo, J., Saragih, J., and Lucey, S., 2013. *CSIRO Face Analysis SDK*, [Online]. Available: http://face.ci2cv.net/doc/.

Kingma, D. and Ba, J., 2014. *Adam: A method for stochastic optimization*, [Online]. Available: https://arxiv.org/abs/1412.6980.