A SPATIOTEMPORAL DEEP LEARNING SOLUTION FOR AUTOMATIC MICRO-EXPRESSIONS RECOGNITION FROM LOCAL FACIAL REGIONS

Mouath Aouayeb, Wassim Hamidouche, Kidiyo Kpalma, Amel Benazza-Benyahia

2019 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, OCT. 13–16, 2019, PITTSBURGH, PA, USA

A SPATIOTEMPORAL DEEP LEARNING SOLUTION FOR AUTOMATIC MICRO-EXPRESSIONS RECOGNITION FROM LOCAL FACIAL REGIONS

Mouath Aouayeb[†⋆], *Wassim Hamidouche*[†], *Kidiyo Kpalma*[†] *and Amel Benazza-Benyahia*[⋆]

[†]Univ Rennes, INSA Rennes, CNRS, IETR - UMR 6164, Rennes, France.
Email: wassim.hamidouche@insa-rennes.fr
[⋆]University of Carthage, SUP'COM, LR/11TIC04, COSIM Lab, Tunis 2083. Tunisia.
E-mail: mouath.aouayeb@supcom.tn

ABSTRACT

Humans always try to hide their Macro-Expressions (MaE) to conceal their real emotion, and it is hard to distinguish between true and false emotions even with artificial intelligence. Micro-Expressions (MiEs), on the contrary, are spontaneous and fast, undetectable with the naked eye and thus always inform us of true feelings. Therefore, there is plenty of studies to generate an automatic system of detecting and analyzing these MiEs.

In this paper we propose a new solution that relies on a combination of Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) applied on particular regions of the face to extract relevant spatial and temporal features, respectively, for MiEs recognition. The proposed solution achieves high recognition accuracy of 90% precision on a different databases including SMIC, CASME II and SAMM. Moreover, under the conditions of Micro-Expression Grand Challenge (MEGC) 2019, our approach performs better than the state of the art solutions including the ones proposed in the challenge.

Index Terms— Micro-Expression, CNN, LSTM, Regions of Interest.

1. INTRODUCTION

The concept of Micro-Expression (MiE) was first introduced in 1872 by Charles Darwin and Phillip Prodger [1], while Haggard and Isaacs [2] have described the conditions of MiE in 1966. Later in 1969, Paul Ekman [3] has defined the MiE based on therapeutic experiences, which is now used by most researchers as a combination of Action Unit (AU) known as the Facial Action Coding System (FACS). According to [3], there are six MiE emotions including anger, disgust, fear, happiness, sorrow, and surprise. The period of MiE is still debated, and it is estimated to be in the interval $[1/20\text{ s},\ 1/25\text{ s}]$. Since MiE occurs in a fraction of a second and has very small intensity, detecting it is very challenging task for ordinary humans. These expressions can only be spotted and identified by professionally qualified people. Even with professional training, in the literature only 47% precision of recognition was reported [3].

The handcrafted solutions [4, 5, 6] try to extract spatio-temporal features supposed to identify the MiE with Local Binary Pattern (LBP) or Optical Flow (OF). The advantage of such solutions is their independence from data quality since we are dealing with a small and imbalanced databases. However, these solutions show their limits in terms of accuracy and they are heavy computational methods preventing their adoption in a real-time context.

The most commonly used solutions in the state-of-the-art is the hybrid solutions [7, 8, 9, 10]. The primary concept is to use handcrafted solution such as Local Binary Pattern on Three Orthogonal Planes (LBP-TOP) or OF to assist Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) to extract the most significant spatio-temporal features despite the database issues. Such solutions have already reached an acceptable accuracy level but not sufficient to consider in other application domain and are also difficult to execute and strongly over moment of complexity.

The world of machine learning and deep learning has evolved in latest years to address many of its issues such as overfitting, vanishing gradient, the need for such enormous and balanced information, etc. This is why community attention has been focused on finding a pure deep learning approach, such as [11, 12, 13, 14], that can automatically detect and recognize MiE from a few imbalanced data. However, compared to hybrid techniques, it still gets a poor precision rate.

To the best of our knowledge, most of the three-category solutions consider only two MiE characteristics that are rapid reaction and low intensity and exclude that, according to [3], MiE are also local, meaning that only some parts and not the whole face, are responsible for the significant micro-movement that defines the six emotions. However, it is essential to point out that the particular character of MiE has been taken into account in latest research [15, 16, 17] since it not only improves the result in terms of accuracy, but it also decreases the time response which is essential for real time applications.

In this paper we propose a new technique that takes into account the locality of MiE. The motivation is to obtain the spatial characteristics of each frame of each particular region by a CNN and then apply a Long Short Term Memory (LSTM) to obtain the spatio-temporal characteristics and lastly classify all spatio-temporal features from the chosen areas by a Fully Connected Layer (FCL). The solution was trained and tested under the Micro-Expression Grand Challenge (MEGC) 2019 conditions.

The rest of the paper is organized as follows. Section 2 presents the state-of-the-art on MiE recognition. Section 3 introduces our proposed solution. In Section 4, more experimental details and results of the proposed solution are provided. Finally, Section 5 concludes this paper.

2. RELATED WORK

In this section, we briefly review and discuss the state-of-the-art approaches for MiE recognition. This section consists of four sub-

978-1-7281-0824-7/19/$31.00

section: handcrafted approach, hybrid approach, deep learning approach and region based approach.

2.1. Handcrafted Approach

Polikonsky *et al.* [18] have divided the face into sub-regions and introduced a 3D-gradients orientation histogram-based feature descriptor for each sub-region to capture the correlation between the frames and then they used a Support Vector Machine (SVM) to identify the MiE. Zhao *et al.* [4] have used a LBP-TOP to get the spatio-temporal information to describe the MiE. This information is then provided to an SVM classifier to perform recognition. Wang *et al.* [19] have proposed Local Binary Pattern with six intersection points to deal with the problem of redundancy of LBP-TOP and they have used also SVM as a classifier of MiE recognition. Liong *et al.* [5] have exploited the Bi-Weighted Oriented Optical Flow (Bi-WOOF) to detect the facial motion, the resulting feature's vector is provided to an SVM to identify the MiE. Recently, Davison *et al.* [6] have used a Histogram of Oriented Optical Flow (HOOF) and have proposed to re-group MiE based on AU instead of using emotion categories. The SVM has been also used as a classifier for training recognition MiE and reached better results compared to other handcrafted methods.

2.2. Hybrid Approach

The basic concept of hybrid solution is to use handcrafted methods along with deep learning algorithm, which makes it a heavy computational solution. Thanks to the recent developments in hardware and software technologies those kind of solution become possible. Khor *et al.* [10] have proposed an Enriched Long-term Recurrent Convolutional Network (ELRCN). First, several types of OF are calculated {Horizontal, Vertical, Magnitude and Strain}. Then authors have used two different blocks of CNN, one for spatial features extraction (the input was the image concatenated with the results from different OF) and another CNN block of 3 convolution blocks is used for temporal features where each OF results was provided to a different convolution block. The two CNN blocks end up with an FCL for the classification. Off-ApexNet, proposed by Liong *et al.* [7], consists in three steps identifying the Offset and Apex frame, calculate the horizontal and the vertical OF and finally provide them to a CNN. The classification step is done by an FCL. An improved solution of Off-ApexNet is STSTNet [8], authors have added to the horizontal and vertical OF the strain OF to get a better result. Xia *et al.* [9] have proposed a Spatiotemporal Recurrent Convolution Network (STRCN). The authors have developed two varieties of the network: STRCN with Appearance based Connectivity (STRCN-A) that consists of a different representation of the image as a vector and so the whole sequence as matrix is provided to a STRCN which is basically a block of recurrent CNN. The other variety is STRCN, with Geometric based Connectivity (STRCN-G), consists in applying OF that feeds the STRCN block. Many other works has been proposed, but most of the methods can be summarized into two steps: calculating OF or LBP, which feed an architecture of a CNN and RNN to extract the relevant spatio-temporal features.

2.3. Deep Learning approach

Deep Learning (CNN + LSTM) was used by Kim *et al.* [11] to encode spatial and temporal characteristics. MicroExpSTCNN has been proposed by Reddy *et al.* [16]. The solution is based on the architecture of 3D-CNN applied to the whole face. Wang *et al.* [13] proposed a CNN-based solution and added a remaining block-based attention unit to assist the network concentrate on key areas. Lateral Accretive Hybrid Network (LEARNet) is a recent CNN solution proposed by Verma *et al.* [14]. The contribution can be resumed on adding the accretion layer to refine the salient expression features. Quang *et al.* [12] have adapted the architecture of CapsuleNet to the context of MiE recognition, using only the most important frame on an MiE sequence which is the apex frame. They have less data so they used transfer learning from ImageNet and data augmentation.

2.4. Region based approach

Instead of extracting spatio-temporal features from the entire face, there is more interest in recent research to deal only with particular regions of the face because of the nature of MiE being a local micro-movements. Zhao *et al.* [15] have proposed a Necessary Morphological Patches (NMPs) which are the most interesting regions among other Active Patches (AP) which are some pre-selected regions from the entire face. The idea is to apply handcraft methods (LBP-TOP) on those NMPs instead of the entire face and then merge all the features to feed SVM for the classification. It is important to note that going with NMPs instead of the entire face have reported not only a better result in terms of accuracy rate and F1-score but also it reduces the response time which is important for real time application. Reddy *et al.* [16] have proposed in addition to the MicroExpSTCNN, MicroExpFuseNet which basically the same 3D-CNN architecture used on MicroExpSTCNN but applied in two regions {eyes, mouth} and ends up with the same FCL. Zhao *et al.* [17] have proposed another improved MiE recognition method based on the NMPs. The proposed NMPs are selected by Random Forest (RF) applied on a combined OF histogram and LBP-TOP histogram of the AP regions. This solution slightly improves the recognition performance.

3. PROPOSED SOLUTION

In this section, we present our proposed deep learning framework for MiE recognition. This framework first detects the face and extracts most important regions based on the solution proposed in [17]. We adopt CNN and LSTM to get spatio-temporal features from the selected regions and then merge and fed them to a simple FCL for the classification. Figure 1 summarizes the proposed architecture.

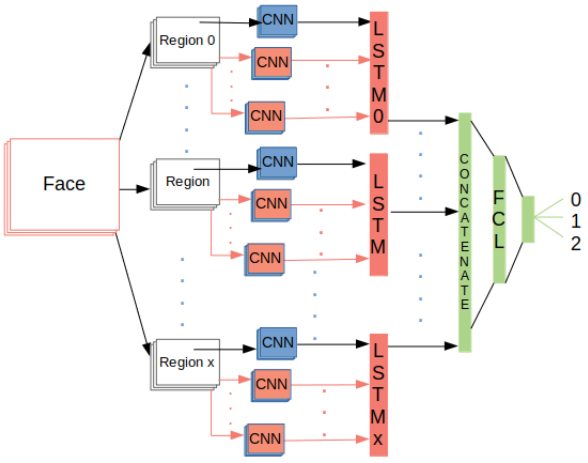

Fig. 1. Overview of the proposed architecture.

3.1. Regions Selection

Referring to the work proposed in [17], we've selected six regions: region0 and region1 for eyes + eyebrows, region2 for nose, region3 and region4 for the cheeks and region5 for the mouth. Those regions, shown in Figure 2, represent the AP. We first detect the face and we align and crop it from the entire image but since there are no interesting movements of the face in all the provided data there is no need to do the alignment step. The position of 6 boxes which should contain all the 6 regions of interest is indicated from the first frame of every sequence based on the 68 facial landmarks, and then all the frames that belong to the same sequence are cropped on those boxes in order to get the six regions.

This operation is performed to save the changes inside the box all along the sequence. For example let's assume that we have a sequence of MiE. First frame must show a neutral face and let's assume the next frame will contain micro-movements and if we detect the 68 facial landmarks for each frame and crop the regions depending on them we will get less changes inside boxes because the 68 points will change position to follow the movement and so we get very similar images for each region.

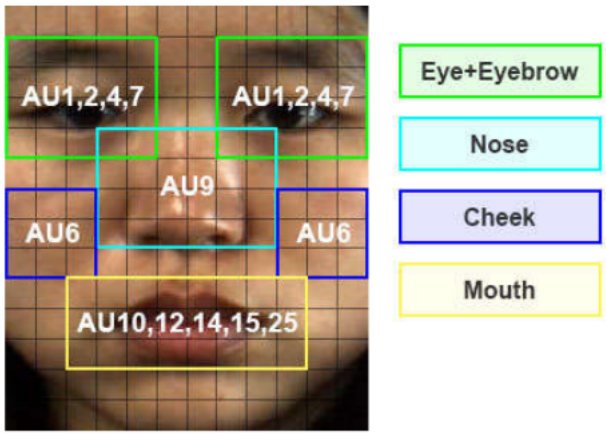

Fig. 2. Illustration of the active patches with AU annotations.

3.2. Convolution Neural Network

CNN is responsible, in the proposed architecture, for extracting spatial features. After cropping regions we train the CNN model for each region separately because each micro-expression is expressed by local micro-movement, it may appear in all the regions or just one region reaction. So it is more interesting to not use the same label for all the regions and that is why we changed the label to "not me" when the region is not responsible for that emotion. For example, if the emotion label is "happy" and the region is the nose it is obvious that nose will not give a reaction for that emotion so we change the label for this region to "not me" instead of "happy" given by the entire face label.

So we recreate six different vectors of labels from the label's vector given to the entire face to train six CNN models, based on [17] to know which regions are responsible for which micro-expressions as shown in Figure 3.

The proposed CNN architecture is inspired by the Inception CNN architecture [20]. So, the image firstly goes through two Convolution layers and two Max-pooling layers then followed by 4 different Convolution layers with filters of different sizes

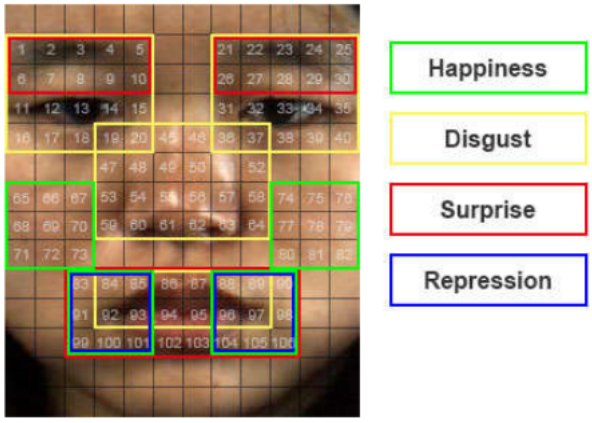

Fig. 3. Illustration of the active patches and emotional state.

{1*1,3*3,5*5,7*7} in parallel. The output of those 4 convolution layers are concatenated with the output of the second convolution layer.

After that an average-pooling is applied to reduce the problem of overfitting. Since there is no good tool to our knowledge we can use to test the effectiveness of the spatial features, we just added an FCL and try to classify the spatial features into the emotions given by the vector of labels specified for each region. Finally, we remove the last layers of the FCL and save the results as the spatial features to be used as input for the LSTM. The CNN architecture is shown in Figure 4.

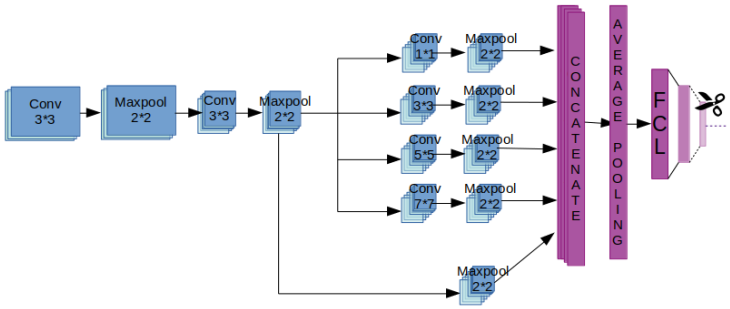

Fig. 4. CNN architecture.

3.3. Long-Short Term Memory and Fully Connected Layer

LSTM is a well known derivative of RNN and its task is to extract temporal features from sequenced spatial features to get the spatio-temporal features. Unlike CNN, all the regions are trained together. So, there are 6 LSTM blocks as the number of regions and they are all concatenated and connected to the same FCL network to be classified into emotions given by the initial label's vector of sequences. After each LSTM block a dropout layer is used to reduce the overfitting problem.

3.4. Loss Function

The loss function used when training CNN or LSTM + FCL are the same. Since we are in a context of multi-label classification it

is obvious that Categorical Cross Entropy (CCE) loss function is a good choice to use. The Cross Entropy (CE) equation for binary classification is expressed as follows

$$CE(p,y) = \begin{cases} -\log(p) & \text{if } y = 1, \\ -\log(1-p) & \text{otherwise,} \end{cases} \quad (1)$$

where $y \in \{+1, -1\}$ specifies the ground truth class and $p \in [0, 1]$ is the model estimated probability for the class with label $y = 1$.

However, this loss function is not the best choice since we deal with a scenario of imbalanced data between classes. A modification has been added to that CCE loss function based on a research proposed by Lin *et al.* [21]. The modified CE equation is given by Equation (2)

$$CE(p,y) = \begin{cases} -\alpha(1-p)^{\gamma}\log(p) & \text{if } y = 1, \\ -(1-\alpha)p^{\gamma}\log(1-p) & \text{otherwise,} \end{cases} \quad (2)$$

where $\alpha \in [0, 1]$ is a weighting factor set by inverse class frequency to contribute the imbalance between classes but it does not differentiate between hard and easy classification task of samples. That's why another modulating factor $(1-p)^{\gamma}$ is introduced to the CCE loss function, with the $\gamma >= 0$ factor is called the focusing parameter.

4. EXPERIMENTS AND DISCUSSION

4.1. Data Preparation and Model Implementation

4.1.1. Database

Referring to the MEGC [22], we used the same Cross-DB (CDE: Cross-DB Evaluation) that includes 3 most used spontaneous micro-expression databases described in Table 1.

Table 1. Samples distribution in 3 classes of all datasets for CDE

Emotion Class	SMIC	CASME II	SAMM	3DB
Negative	70	88*	92**	250
Positive	51	32	26	109
Surprise	43	25	15	83
TOTAL	164	145	133	442

* *Negative class of CASME II includes Disgust and Repression expressions*

** *Negative class of SAMM includes Anger, Contempt, Disgust, Fear and Sadness expressions*

The following list gives details of each database: **SMIC** [23] : database contains 164 samples (Spontaneous MiE sequences) from 16 subjects. Each sequence is recorded at 100 fps and labeled as: positive, negative and surprise.

CASME II [24] : a spontaneous MiE database containing 247 sequences from 26 participants. The sequences are labeled into 5 emotions {Happiness, Disgust, Repression, Surprise, Sadness} plus the "OTHER" class and recorded at the frame rate of 200 fps.

SAMM [25] : it is a 159 spontaneous MiE samples from a demographically diverse groups of 32 subjects and an even gender male-female split. SAMM was induced based on 7 basics emotions and recorded also at a frame rate of 200 fps.

3DB-combined : the main dataset used in the MEGC 2019 is a combination of the 3 databases: SMIC, SAMM and CASME II. There are many similarities between CASME II and SAMM, meanwhile SMIC is different. To avoid clutter when combining the 3 databases, in the MEGC was introduced a common reduced set of emotion classes. The composition of the 3DB-combined data set is given in Table 1.

4.1.2. Experimental setup

First we extract the six regions: left eye + eyebrow, right eye + eyebrow, nose, left cheek , right cheek and mouth, respectively, at size {[100,80], [100,80], [120,80], [60,60], [60,60], [160,60]} using 68 facial landmarks based on detector from the dlib library[1]. Each database has a different size of images for each cropped region and we only resized them to the same size by adding a black background. The first Convolution layer has 4 filters with a size of $5 * 5$, the next one has 8 filters with a size of $3 * 3$ and the next 4 parallel convolution layers have the same number of filters 16 and respectively with the size of [$1 * 1$, $3 * 3$, $5 * 5$ and $7 * 7$] then a dense layer with 1024 neurons followed by Dropout with 0.2 as a parameter, then the dense layer that will give us the spatial features with 20 neurons and finally another dense layer with 4 neurons for the classification that will be cut after we get all the spatial features. We used Rectified Linear Unit as an activation function for all previous layers while for the last one we used Softmax as an activation function. For LSTM, we have resized all input sequences to the same size 500 by adding zeros at the end to be suitable for LSTM architecture and we used the same length of inputs 120 for all LSTMs and after each LSTM we added a dropout layer with a value of 0.2. All the LSTM's outputs are concatenated and fed to a dense layer of 64 neurons with Leaky Rectified Linear Unit as an activation function then another dense layer with 3 neurons (for the three classes 0: Negative, 1: positive, 2: surprise) with Softmax as an activation function. The CNN was trained with 64 batch size and 100 epochs, and the LSTM was trained with 244 batch size and 60 epochs. All the experiments are performed on Ubuntu 18.04.2 LTS, python3.6 with keras-gpu 2.2.4, tensorflow-gpu 1.12.0 and with GeForce GTX 1080 Ti GPU (32 GB memory) and Intel Xeon Processor.

4.1.3. Evaluation Protocol and Metrics

The LOSO-CV is used to determine the training-testing splits (i.e each subject group is held out as the testing set while all remaining samples are used for training). This protocol mimics a realistic scenario where different subjects are enrolled separately in different environment and settings into a single recognition system. The LOSO-CV also ensures subject-independent evaluation. The performance is assessed based on two balanced metrics, the Unweighted F1-score (UF1) known also as the macro-averaged F1-score

$$F1_c = \frac{2TP_c}{2TP_c + FP_c + FN_c}, \quad (3)$$

$$UF1 = \frac{F1_c}{C}, \quad (4)$$

where TP_c , FP_c and FN_c are respectively true positive, false positive and false negative for the class c and $C = 3$ is the number of classes. The other metric is the Unweighted Average Recall (UAR) also known as balanced accuracy of the system

$$UAR = \frac{1}{C}\sum_{c=1}^{C} ACC_c$$
$$ACC_c = \frac{TP_c}{N_c} \quad (5)$$

[1] $http://dlib.net/face_landmark_detection.py.html$

Table 2. The LOSO-CV performance of our proposed method, baselines and the recent methods (* references from the MEGC 2019 challenge)

Models	FULL		SMIC		CASAME II		SAMM	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
LBP-TOP [22]$^{\diamond}$	0.5882	0.5785	0.2000	0.5280	0.7026	0.7429	0.3954	0.4102
Bi-WOOF [5]$^{\diamond}$	0.6296	0.6227	0.5727	0.5829	0.7805	0.8026	0.5211	0.5139
OFF-ApexNet [7]†	0.7196	0.7096	0.6817	0.6695	0.8764	0.8681	0.5409	0.5392
Micro-Attention [13]$^{\oplus}$	0.5080	0.4930	0.4730	0.4660	0.5390	0.5170	0.4030	0.3400
ATNet *(Fusion)* [26]$^{\oplus}$	0.6310	0.6130	0.5530	0.5430	0.7980	0.7750	0.4960	0.4820
Quang *et al.* [12]$^{*\oplus}$	0.6520	0.6506	0.5820	0.5877	0.7068	0.7018	0.5882	0.5989
Zhou *et al.* [27]*†	0.7322	0.7278	0.6645	0.6726	0.8621	0.8560	0.5868	0.5663
Liong *et al.* [8]*†	0.7353	0.7605	0.6801	0.7013	0.8382	0.8686	0.6588	0.6810
Liu *et al.* [28]*†	0.7885	0.7824	0.7461	0.7530	0.8293	0.8209	0.7754	0.7152
Our proposed method $^{\oplus}$	**0.9022**	**0.9018**	**0.8886**	**0.8828**	**0.9857**	**0.9857**	**0.7855**	**0.8103**

$^{\diamond}$ handcrafted approach, † hybrid approach, $^{\oplus}$ deep learning approach.

where ACC_c is the accuracy rate of the class c and N_c is the number of samples of the same class c.

4.2. Results and Discussion

We report the UF1 and UAR scores of our proposed model with the baseline and the participant's results of the challenge MEGC 2019 as well as some results of other proposed solutions. Table 2 compares the UF1 and UAR scores on FULL cross-database, and on separate datasets including SMIC, CASME II and SAMM . Our proposed model obtains the UF1 score of 0.9022, and the UAR of 0.9018. Apparently, our model performance is the best among all state of the art solutions. The micro-expression performance of the proposed method outperforms the state-of-the-art {handcrafted, hybrid, Deep Learning} solutions with large margins, approximately between 40% and 12%. We have the highest results ever reported on CASME II database with 0.9857 UF1 and also 0.9857 UAR scores.

Figure 5 shows the Leave-One-Subject-Out (LOSO) confusion matrix of our proposed solution and the negative class has the highest Recall rate 0.924 because it's the dominant class in the cross database.

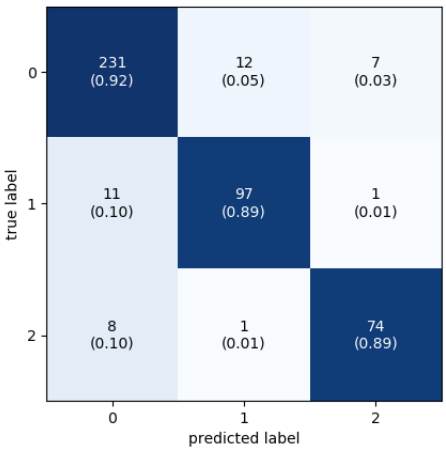

Fig. 5. Confusion matrix on cross-database with the LOSO cross-evaluation method.

5. CONCLUSION

In this paper, we have proposed two contributions to enhance the performance of MiE recognition. The proposed solution applies on regions of interest instead of the whole face, and uses a combination of CNN and LSTM to extract the most relevant spatio-temporal features. We have obtained the highest accuracy score in MiE recognition with LOSO-CV evaluation on a combined 3 databases {SMIC, CASME II, SAMM} with 0.9018 UAR and 0.9022 UF1. Moreover, it is essential to point out that we have not used any Transfer Learning methods or Data Augmentation that could lead to an even better performance.

Experimental results demonstrate the efficiency of the proposed technique that exceeds the state-of-the-art solution. With only about 1M (1.093.531) parameters to train the networks, our proposed solution can be used in real time applications. However, spotting MiE from a long video sequence or Multi-subject are still challenging tasks that can be investigated in our future work.

6. REFERENCES

[1] C. Darwin and P. Prodger, "The expression of the emotions in man and animals," *Oxford University Press, USA*, 1998.

[2] E. A. Haggard and K. S Isaacs, "Micromomentary facial expres-sions as indicators of ego mechanisms in psychotherapy," *Methods of research in psychotherapy,*, pp. 154 – 165, Springer 1966.

[3] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, pp. 88 – 106, 1969.

[4] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal.*, , no. 6, pp. 915 – 928, Mach 2007.

[5] S. T. Liong, J. See, R. C. W. Phan, and K. Wong, "Less is more: Micro-expression recognition from video using apex frame," *arXiv preprint arXiv:1606.01721.*, 2016b.

[6] A. K. Davison, W. Merghani, and M. H. Yap, "Objective classes for micro-facial expression recognition.," *arXiv preprint arXiv:1708.07549.*, 2017.

[7] Y. Gan, S.-T. Liong, W. C. Yau, Y. C. Huang, and L. K. Tan, "OFF-ApexNet on Micro-expression Recognition System," *Signal Processing: Image Communication*, 2019.

[8] S.-T. Liong, Y. Gan, J. See, and H.-Q. Khor, "A Shallow Triple Stream Three-dimensional CNN (STSTNet) for Micro-expression Recognition System," *arXiv preprint arXiv:1902.03634*, 2019.

[9] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, "Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions," *arXiv preprint arXiv:1901.04656*, 2019.

[10] H.-Q.Khor, J. See, R. C. W. Phan, and W. Lin, "Enriched long-term recurrent convolutional network for facial micro-expression recognition," *13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 667 – 674, 2018.

[11] D. H. Kim, W. J. Baddar, and Y. M. Ro, "Micro-expression recognition with expression-state constrained spatio-temporal feature representations," *Proceedings of the 2016 ACM on Multimedia Conference (Amsterdam)*, pp. 382 – 386, 2016.

[12] N. V. Quang, J. Chun, and T. Tokuyama, "CapsuleNet for Micro-Expression Recognition," *14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 2019.

[13] C. Wang, M. Peng, T. Bi, and T. Chen, "Micro-Attention for Micro-Expression recognition," *arXiv preprint arXiv:1811.02360*, 2018.

[14] M. Verma, S. K. Vipparthi, G. Singh, and S. Murala, "LEARNet Dynamic Imaging Network for Micro Expression Recognition," *arXiv preprint arXiv:1904.09410*, 2019.

[15] Y. Zhao and J. Xu, "Necessary Morphological Patches Extraction for Automatic Micro-Expression Recognition," *applied sciences, mdpi (2018)*, 2018.

[16] S. P. T. Reddy, S. T. Karri, S. R. Dubey, and S. Mukherjee, "Spontaneous Facial Micro-Expression Recognition using 3D Spatiotemporal Convolutional Neural Networks," *arXiv preprint arXiv:1904.01390*, 2019.

[17] Y. Zhao and J. Xu, "An Improved Micro-Expression Recognition Method Based on Necessary Morphological Patches," *Symmetry, mdpi (2019)*, 2019.

[18] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor," *3rd International Conference on Crime Detection and Prevention (ICDP 2009) (London, UK)*.

[19] Y. Wang, J. See, R. C. W. Phan, and Y. H. Oh, "LBP with Six Intersection Points: reducing redundant information in LBP-TOP for micro-expression recognition," *Computer Vision-ACCV 2014 (Singapore)*, pp. 525 – 537, Springer 2014.

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Er-han, V. Vanhoucke, and A. Rabinovich, " Going deeper with convolutions," *CVPR*, 2015.

[21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dolla, "Focal loss for dense object detection," *arXiv preprint-arXiv:1708.02002*, 2017.

[22] J. See, M. H. Yap, J. Li, X. Hong, and S.-J. Wang, "MEGC 2019 The Second Facial Micro-Expressions Grand Challenge," *14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 2019.

[23] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietika, "A Spontaneous Micro-expression Database: Inducement, Collection and Base-line," *10th Proc Int Conf Autom Face Gesture Recognit (FG2013) Shanghai, China. DOI: 10.1109/FG.2013.6553717*, 2013.

[24] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "CASME II: An Improved Spontaneous Micro-Expression Database and the Baseline Evaluation," *PloS one*, vol. 9, no. 1, 2014.

[25] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "SAMM: A Spontaneous Micro-Facial Movement Dataset," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 116 – 129, Jan 2018.

[26] M. Peng, C. Wang, T. Bi, T. Chen, X. Zhou, and Y. Shi, "A Novel Apex-Time Network for Cross-Dataset Micro-Expression Recognition ," *arXiv preprintarXiv:1904.03699*, 2019.

[27] L. Zhou, Q. Mao, and L. Xue, "Dual-inception network for cross- database micro-expression recognition," *14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 2019.

[28] Y. Liu, H. Du, L. Zheng, and T. Gedeon, "A neural micro-expression recognizer," *14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 2019.