Facial Action Unit Detection with Local Key Facial Sub-region Based Multi-label Classification for Micro-expression Analysis

Liangfei Zhang
University of St Andrews
St Andrews, UK
lz36@st-andrews.ac.uk

Ognjen Arandelović
University of St Andrews
St Andrews, UK
ognjen.arandjelovic@gmail.com

Xiaopeng Hong
Xi'an Jiaotong University
Xi'an, China
hongxiaopeng@ieee.org

ABSTRACT

Micro-expressions describe unconscious facial movements which reflect a person's psychological state even when there is an attempt to conceal it. Often used in psychological and forensic applications, their manual recognition requires professional training and is time consuming. Therefore, achieving automatic recognition by means of computer vision would confer enormous benefit. Facial Action Unit (AU) is a coding of facial muscular complexes which can be independently activated. Each AU represents a specific facial action. In the present paper, we propose a method for the challenging task that is the detection of activated AUs when the micro-expression occurs, which is crucial in the inference of emotion from a video capturing a micro-expression. This specific problem is made all the more difficult in the light of limited amounts of data available and the subtlety of micro-movements. We propose a segmentation method for key facial sub-regions based on the location of AUs and facial landmarks, which extracts 11 facial key regions from each sequence of micro-expression images. AUs are assigned to different local areas for multi-label classification. Considering that there is little prior work on the specific task of *detection* of AU activation in the existing literature on micro-expression analysis, for the evaluation of the proposed method we design an AU independent cross-validation method and adopt Unweighted Average Recall (UAR), Unweighted F1-score (UF1), and their average as the scoring criteria. Evaluated using the established standards in the field and compared with previous work, our approach is shown to exhibit state-of-the-art performance.

CCS CONCEPTS

• **Computing methodologies** → Activity recognition and understanding.

KEYWORDS

Micro-expression analysis, Micro-movements, Facial Action Unit detection

ACM Reference Format:
Liangfei Zhang, Ognjen Arandelović, and Xiaopeng Hong. 2021. Facial Action Unit Detection with Local Key Facial Sub-region Based Multi-label Classification for Micro-expression Analysis. In . ACM, New York, NY, USA, 8 pages.

ACM MM '21, OCT 20–24, 2021, Chengdu, China

ACM ISBN 978-1-4503-XXXX-X/18/06.

1 INTRODUCTION

Facial expressions can reflect human emotions. Due to different cultural environments, individuals use different languages to communicate, but their emotions are expressed by the same facial expressions [4]. In addition to the regular macro-expressions which take place on larger time scales, small and speedy movements that are inadvertently exhibited for short periods of time, better reveal emotions which individuals attempt to conceal. Ekman et al. [3] first reported on a case of these particular expressions. In a recording of a conversation between a psychiatrist and a depression patient, there are occasional frames with very painful expressions of a patient otherwise displaying a happy appearance in the video. Researchers call that kind of fast, unconscious, spontaneous facial movements such people produce when they experience intense emotions, *Micro-Expressions*. Micro-expressions usually happen within less than 0.5 seconds. If the occurrence of micro-expressions is detected and the emotional meaning represented by them is recognized, the real mental activities of individuals could be accurately identified.

Facial *actions* are different from facial expressions. Facial action units (AUs) refer to muscular complexes which are activated during facial movements. External factors such as a gust of wind blowing across the face also cause AU activation. Facial expressions are reflected by facial movements caused by some mental activities (such as emotion). Micro-expressions are unconscious expressions that appear when a person tries to suppress or hide emotions [6]. The correct recognition of micro-expressions helps us understand real emotions, so it can be an essential basis for identifying individuals' subjective feelings in the context of public safety or psychotherapy, for example. The most intuitive method of identifying the emotions represented in the micro-expressions is to analyse the AUs included in the micro-expressions, which is also the current method of manually recognising micro-expressions. Therefore, facial AUs could also be considered as an intermediate variable during automatic recognition between micro-expressions and emotions.

Due to the small range of movement and the short duration of facial movements when micro-expressions happen, individuals need professional training to recognize micro-expressions manually. The human based processes of training as well as recognition itself are time demanding, yet the recognition accuracy is still not satisfactory for most practical purposes. Many researchers have tried using and developing new computer vision techniques to recognize micro-expressions automatically. This automatically approach to the identification of micro-expressions has unique advantages,

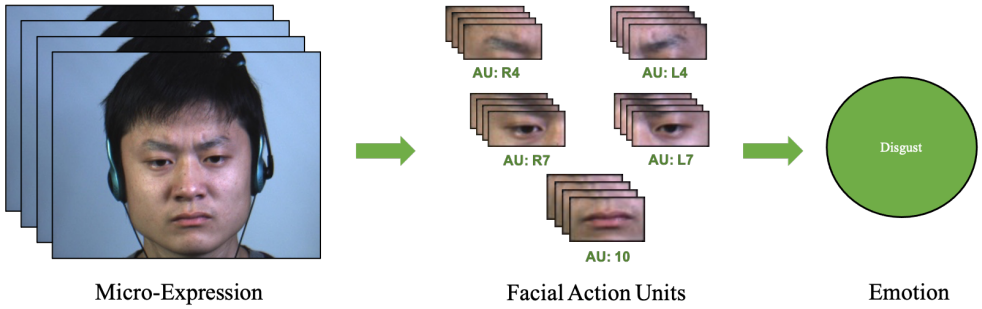

Figure 1: Example of AUs detected from a micro-expression and recognized to an emotion.

which significantly improves the feasibility of micro-expression applications. No matter how fast the facial movement is, as long as the camera records it, the computer can obtain the corresponding information and process it. In addition, once an efficient and stable model is trained, it can process large volumes of micro-expression data at low cost, which far exceeds the efficiency of manual recognition of micro-expressions by professionals. Thus, some research uses high-speed cameras for the collection of micro-expressions. Recently, following the publication of open-source micro-expression databases, the amount of work related to micro-expressions increased every year. Research thus far all but invariably uses 3DHOG [15], LBP-TOP [14], HOOF [12] and their variations or deep learning methods as features.

However, most previous work focuses on emotion recognition directly, even some considered AUs as supplementary features [9, 13, 18], there is tiny work about micro-expression AU detection task specifically. In this paper, we focus on the AU detection task for micro-expression analysis and demonstrate our proposed framework can achieve the effect of state-the-art in the task, even without deep learning methods. In addition, Since our method does not use deep learning, it does not require a lot of time for training and run. It can almost meet the requirement of real-time detection. Taking CASME II as an example, it only takes about *1s* to complete the AU detection test of all samples using our framework.

The mean contributions of this paper are as follows:

(1) We proposed a novel facial key subregion segmentation method based on the facial muscle of AU activated and a novel framework to detect multi-labelled facial micro-expression AUs by transfer a big multi-label classification to several small ones based on the segmented regions.
(2) We design an AU independent 5-fold cross-validation method for Facial AU detection in micro-expression and conduct intensive experiments on two publicly micro-expression databases with AU labels. The results represent the effectiveness of our approach.

2 RELATED WORK

The majority of research on micro-expression analysis before 2019 relies on low-level feature extraction in the form of LBP-TOP, 3DHOG, HOOF and similar extensions thereof [**?**]. In recent years, the application of deep learning has been increasing steadily, which is a trend which we expect to continue in the near to medium future.

LBP-TOP (Local Binary Patterns on Three orthogonal Planes) is a feature type which extends the traditional Local Binary Patterns (LBP) to three dimensions. It is one of the earliest methods of extracting micro-expressions features [14]. This method is representative of the majority of the work in the area and provides a reliable verification and comparison method for the subsequent micro-expression recognition work. Therefore, LBP-TOP and its extensions are most popular methods in micro-expression recognition research. LBP-SIP (Local Binary Pattern with Six Intersection Points) [17] extends LBP features for micro-expression recognition from another perspective. The main improvement of this work is to reduce the dimension of features and improve the efficiency of feature extraction. CBP (Centralized Binary Pattern) [5] is another improvement to the local binary pattern. The main contribution is that the value is the difference between the average value of the centre point and the neighbouring points, so the corresponding binary code length is half of LBP and the histogram dimension is lower.

3DHOG (3D Histograms of Oriented Gradients) [15] uses a gradient feature to describe the spatial and temporal local dynamics of the face. The feature extraction method of this work is an extension of the planar gradient histogram. The model construction process can be regarded as a k-nearest neighbour model constructed with the help of the k-means algorithm. Its key limitation stems from the fact that although the number of facial muscles involved in micro-expressions is small, the assumption that only one AU is activated is overly crude. HOOF (Histogram of Oriented Optical Flow) methods uses the optical flow field as the basic feature to describe the micro-expression movement. Liu et al. [12] proposed MDMO (Main directional mean optical flow feature) to extract the main direction in the video sequence and calculate the average optical flow feature in the face block. Xu et al. [19] proposed another feature extraction method based on HOOF, FDM (Facial Dynamics Map). This method can better reflect the movement features of micro-expressions, and the calculated features of the facial dynamic spectrum which are easy to visualize. Thus it can be used to assist in the understanding of micro-expressions. Its main drawback is that the computation of the dense optical flow field is time-consuming, which is not suitable for real-time, large-scale micro-expression recognition.

Within the realm of traditional micro-expression feature extraction methods, LBP-TOP performs better than 3DHOG and HOOF in

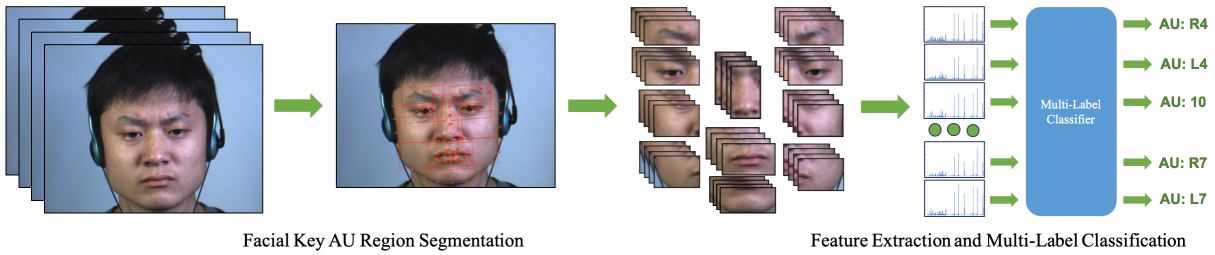

Figure 2: Framework of the proposed method.

high-resolution images. However, in low-resolution data, the performance ranking is reversed, which suggests that LBP-TOP depends more on spatial information (XY), whereas HOOF and 3DHOG are more dependent on temporal variation (XT and YT). Unlike these 'conventional' computer vision based methods which rely on hand-crafted feature descriptors, the recent emergence of deep learning methods helped further increase the accuracy of micro-expression recognition; a variety of CNN (convolutional neural network) based methods for automatic extraction micro-expression features have now been proposed.

Most of these address micro-expression recognition and a small number focus on micro-expression spotting. Only the work of Li et al. [11] focuses on AU detection in micro-expressions as we do herein. They apply a channel attention module and a spatial attention module on 3DResNet to capture regional changes and the relationship of facial regions. In addition, they treat every AU detection as a specific task instead of adopting a multi-label learning approach for all AUs, which is not a common practice in AU detection tasks.

3 PROPOSED METHOD

Recall that our main aim in the present work is the identification of AUs activated during a facial micro-expression. Thus, the method we propose can be broadly seen as comprising the following stages: facial sub-region segmentation, facial sub-region feature extraction, and multi-label classification. These are summarized in Figure 2 and explained in detail hereafter.

3.1 Local Facial Region Segmentation

Facial Action Coding System [4] is currently recognized as the universal norm for encoding facial actions, which associates each facial action with an AU. It also forms the basis for the standard labelling of facial movements in micro-expression datasets. Hence, the overarching aim of the present work is to detect from a sequence of images all AUs activated when a micro-expression is displayed (n.b. the majority of micro-expressions using in emotion inference involve the activation of multiple AUs). Since AUs are by design canonical, elementary primitives used to describe facial movement, it is a corollary of the anatomical structure of the face that a specific AU is spatially localized i.e. it corresponds to a specified sub-region of the face. For example, the common units AU1, AU2, and AU4 describe movement only in the eyebrow area. Therefore, segmenting the face into multiple sub-regions and identifying the AUs that appear in each specific sub-region is less complicated than concurrently identifying all AUs that activate across the entire face area. Mirroring the spatial layout of AUs, we propose a segmentation method which segments the facial region into 11 sub-regions, specifically: Left and Right Brow, Left and Right Eye, Left and Right Cheek, Left and Right Nasolabial Area, Nose, Mouth, and Chin.

Our local facial region segmentation method is based on facial landmark detection – a crucial step in many face recognition and analysis algorithms. The task involves the localization of salient areas of a face, such as the eyebrows, eyes, nose, mouth, or the face contour, from a given image of a face.

We pursued the standard 68 key-point positioning strategy. The specific landmarks recognition results is summarized in Figure 3. We used the ERT (Ensemble of Regression Trees) [8] algorithm, a regression tree method based on gradient improvement learning, to localize these key points of the face. The ERT uses cascading regression factors and several GBDTs (Gradient Boosted Decision Trees) whose leaf nodes store the residual. When the input falls onto a node, the residual is added to the input for the purpose of regression correction. Finally, all the residuals are superimposed to obtain the final face point position.

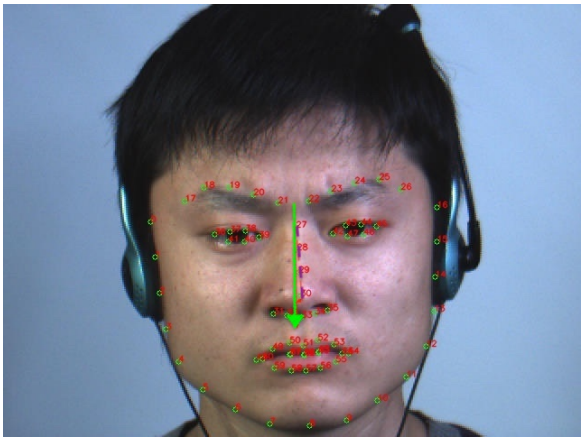

Figure 3: An example of detected landmarks in a micro-expression image from CASME II database.

As each micro-expression is exhibited over a short period, and the range of the corresponding muscular movement is small, the person's pose does not change significantly. Therefore, it is not

necessary to detect all the facial key points of each frame in a micro-expression image sequence. We select the central frame of the image sequence as the reference image to identify facial landmarks and extend the result to each frame of this micro-expression image sequence. After that, we propose a segmentation method of facial sub-regions based on 68 facial landmarks in each micro-expression sequence.

As shown in Figure 3, points 27–30 can be used to determine the centre line of the nose that can be established within the facial range in the image. This line is used as the vertical line of the entire face area during our segmentation process. During the decision of each sub-region of face, the sub-region image will rotate according to this vertical line to ensure that images of all facial sub-regions are in vertical state. The specific distribution of choice 11 face sub-regions is shown in Figure 4. Eyebrows, eyes, nose, and mouth area are all determined based on the landmarks that mark them. The cheek and nasolabial area are determined by the upper and lower contour points of the face and the landmarks of the upper part of the lips (for example, points 0, 4 and 50). The chin is determined by the lower lip point 57 and the lowest point 8 of the facial contour.

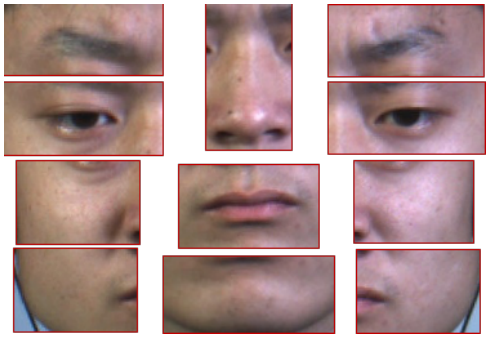

Figure 4: An example of segmentation areas of micro-expression image from CASME II database.

3.2 Sub-regional Feature Extraction and Multi-label Classification

The LBP-TOP feature extraction method is representative of the bulk of micro-expression recognition research and is often used as a baseline model in new micro-expression research. LBP-TOP features capture the relationship between the appearance of a pixel and its neighbourhood. In order to encode the spatio-temporal co-occurrence model, feature extraction is performed on the XY plane, XT plane and YT plane in the image sequence. We used the method provided in [7] to set the radius RX, RY, RT on three space-time axes (X, Y, T). After setting the number of samples on three spaces PXY, PXT, PYT, on the space-time plane, an ellipse determined by the corresponding axes in each space-time plane is used to sample points uniformly to calculate the local binary mode on each plane. After that, the histogram of data in each facial sub-region is used to find the unified features of each facial AU. In the proposed method, LBP-TOP feature extraction is performed on the image sequence of each facial key sub-region, instead of the entire face. In this way, features can be focused on the key parts of the face that are meaningful in micro-expressions. We can also ignore the facial information unrelated to emotion and AUs, which can make the feature used in learning more specific.

After extracting the features of each facial sub-region, we perform multi-label classification based on multiple AU labelled micro-expressions. Traditional supervised learning by and large focuses on single-label learning. However, the target samples in real life are often more complex, with multiple semantics, and containing multiple labels. In particular, in our micro-expression AU detection work, most of the expressions are composed of more than one active facial AUs, making multi-label learning the natural choice.

Our strategy of multi-label classification is to transform the problem structurally, to make the extracted features more readily usable by existing single-label learning algorithms. Firstly, we apply the Label Powerset (LP) algorithm to transform a multi-label learning problem into a multi-class (single-label) classification task. This is achieved by learning one single-label classifier $h : X \rightarrow P(L)$, where L is a set of disjoint labels, $P(L)$ is the powerset of L, containing all possible label subsets. The label set predicted by LP is already in the training set, and it cannot be generalized to the unseen label set. In order to overcome this limitation of LP, the LP classifier used by Random k-labelsets (RAkELd) [16] only trains a subset of length k in Y output dataset and then integrates a large number of LP classifiers to predict. In general, this type of method considers the relationship between the class labels, but for datasets with many class labels and a large amount of data, the computational complexity of problem transformation is an obvious limitation. However, micro-expression datasets are not big enough and our facial sub-region segmentation work reduced the number of labels of each sample. So, this limitation of these methods has little effect in the present context.

The main purpose of employing the LP algorithm is to convert the multi-label classification problem into a single-label one. Every combination of different labels is henceforth considered as a class in itself. This algorithm will generate more classes when there are more labels. Therefore, if it is applied to the entire face image since the number of all micro-expression AUs appearing on the entire face is large, the result of learning when they are all used as labels in one multi-label classification is very poor. However, after we segment the face according to the range of AUs, the number of AUs that may appear in each salient area is much lower across than the whole facial region. Hence, segmentation is crucial in preventing an excessive increase of the computational cost of the LP algorithm, resulting in far better performance. RAkELd algorithm is a variant of the LP algorithm. It converts an LP from multiple labels into multiple LPs of length k to predict the results jointly. This method can effectively reduce complexity when there are too many types of labels in the LP algorithm. When we apply these two multi-label classifiers in actual experiments, the results of the LP are better than RAkELd. Therefore, LP is used as the final multi-label classification strategy in our approach. Finally, we apply adopt the Gaussian Naïve Bayes algorithm on the extracted sub-regional features to learn a model of multi-labelled AUs activated during micro-expressions.

4 EVALUATION

The four datasets most widely used in the existing literature on micro-expression recognition and analysis are SMIC [10] (without AU labels), CASME [21], CASME II [20] and SAMM [2]. It is important to emphasise that all of the aforementioned corpora were acquired in relatively controlled conditions for the specific purpose of micro-expression analysis. In particular, the data acquisition process involved the participants watching emotional videos while attempting to hide the facial expression of the aroused emotions. Thus, they are more standardized and easier to process than the datasets of faces in the wild. Considering that the duration of a micro-expression is usually under 0.5s, in order to capture more frames of images during the occurrence of micro-expressions, data is usually acquired using high-speed cameras. SAMM and CASME II contain data with the highest frame rate of 200 frames per second. In order to test the effectiveness of our method, we apply it first on CASME II and then the combination of CASME II and SAMM.

4.1 Data Preparation

For our experiments we chose to adopt the use of the CASME II database [20], which is widely used in the field owing to its size and high video frame rate. Its collection method is spontaneously induced, which is representative of the real-world conditions. Professional psychologists have marked all AUs in each micro-expression image sequence. A total of 19 AUs were included in CASME II, namely AU 1, 2, 4–7, 9, 10, 12, 14–18, 20, 24–26 and 38. In our experiment, the 11 facial sub-regions are the smallest modules. Therefore, these 19 AUs are separated into each sub-region according to the area where they appear. However, because some AUs may appear in both left and right half of the face, such as AU1, Inner Brow Raiser, they are included in both left and right facial areas. In addition, the original labels in CASME II also include some single-side AUs, such as L1, L2 and R4. Therefore, we divide AUs which are activated in both sides into two parts. For example, the initial both-side label *AU1* is relabelled as *L1&R1*, the initial single-side label *L1* remains. Finally, in our experiment, a total of 26 AUs were included. The specific AUs included in each facial sub-region are shown in Table 2.

The types of AUs labelled in the SAMM micro-expression dataset are more abundant than those in CASME II. However, there are several rare AUs only activated in one or two samples of micro-expression. After analysis the AU labels in SAMM and CASME II, we find the relabelled AUs we previously described for CASME II are the most common ones in both datasets. In order to unify the evaluation criteria of the experiment, we only used the 26 AUs as we described and relabeled AUs of samples in SAMM. The other rare AU labels were deleted, and only the AUs in Table 2 were used for the experiment. Due to there are no AU16 and AU38 labelled in SAMM, the final AU number applied in SAMM is 24.

4.2 Metrics

The frequency of activation of AUs is different across facial expressions. Some AUs are more commonly activated than others, such as AU4, Brow Lowerer, which is the most frequently engaged AU. AU26, Jaw Drop, is activated less in micro-expressions than others, especially when participants are asked to suppress their expressions. Therefore, in the model training process, in order to make sure that all the AUs' features could be learned, we randomly separate data that each AU appears in each facial sub-region into 5 folds and each time 4 of them as a training set. Thus, we ensure that samples of all AUs are in our training corpus. The remaining subset of the micro-expression data is used as a test dataset to evaluate the final algorithm results. In this way, an AU independent 5-fold cross-validation strategy is applied to evaluate the proposed method.

As emphasised already, there is virtually no AU detection work in the context of micro-expressions and no standard metrics which could be used for evaluation in this realm. Therefore, we adopt the evaluation approaches from other AU detection work, as well as the metrics used in as related as possible micro-expression analysis problems. Accuracy and F1-score are widely used criteria in both AU detection and micro-expression recognition. The distribution of the number of each AU in the micro-expression database is unbalanced, so we choose *Unweighted F1-score (UF1)*, and *Unweighted Average Recall (UAR)* to show the performance of our approach and equalize the influence of each AU.

$$Precision_c = \frac{\sum_{i=1}^{F} TP_{i,c}}{\sum_{i=1}^{S} TP_{i,c} + \sum_{i=1}^{F} FP_{i,c}}, \tag{1}$$

$$Recall_c = \frac{\sum_{i=1}^{F} TP_{i,c}}{\sum_{i=1}^{S} TP_{i,c} + \sum_{i=1}^{F} FN_{i,c}}, \tag{2}$$

$$F1\text{-}score_c = 2 \times \frac{Precision_c \times Recall_c}{Precision_c + Recall_c}, \tag{3}$$

$$UF1 = \frac{\sum_{c=1}^{C} F1\text{-}score_c}{C}, \tag{4}$$

$$UAR = \frac{\sum_{c=1}^{C} \frac{\sum_{i=1}^{S} TP_{i,c}}{N_c}}{C}, \tag{5}$$

where $TP_{i,c}$, $FP_{i,c}$ and $FN_{i,c}$ are true positive, false positive and false negative for each class c (of C AUs, 26 in our experiments), when samples of fold i as test set. $Precision_c$ and $Recall_c$ represent the fraction of AU_c is correctly identified and the number of correct detections of AU_c over the actual number of samples with AU_c active. F is the number of fold (5), and N is the total number of samples. After obtaining the average of *UAR* and *UF1*, this quantity is used as the final evaluation score, which is also the comparison criterion used in the EmotionNet Challenge [1] (a popular AU detection challenge "in the wild").

In addition, in order to make a clearer comparison with the work of Li et al. [11], we also adopted their subject independent 4-fold cross validation on CASME II and SAMM. Nevertheless, we still use our multi-label learning for 26 AUs, as all of the AUs chosen by Li et al. are amongst these.

4.3 Summary of Results

The test results of the models trained on CASME II, SAMM and CASME II & SAMM by our method are shown in Table 1. Firstly, observe that the proposed approach achieves excellent results across the different micro-expression databases, testifying to the value of our multi-label AU detection based approach. It is also important to note the model performed equally well across the entire set of AUs. This finding demonstrates that our method can effectively

	CASME2			SAMM			CASME2&SAMM		
AU	**Accuracy**	**F1**	**Score**	**Accuracy**	**F1**	**Score**	**Accuracy**	**F1**	**Score**
L1	0.7729	0.6563	0.7146	0.9466	0.7525	0.8495	0.6361	0.5314	0.5838
R1	0.7649	0.5970	0.6810	0.9084	0.6006	0.7545	0.6516	0.4380	0.5448
L2	0.7656	0.6020	0.6838	0.7863	0.5492	0.6677	0.6411	0.5166	0.5789
R2	0.8433	0.6170	0.7301	0.7634	0.4889	0.6261	0.7018	0.4356	0.5687
L4	0.6813	0.6720	0.6767	0.8626	0.6796	0.7711	0.7030	0.6555	0.6792
R4	0.5037	0.4866	0.4951	0.8168	0.4876	0.6522	0.5664	0.5374	0.5519
L5	0.9728	0.7597	0.8662	0.7109	0.4155	0.5632	0.7818	0.4815	0.6317
R5	0.9572	0.4891	0.7231	0.7344	0.4505	0.5924	0.7714	0.4764	0.6239
L6	0.5804	0.3758	0.4781	0.9844	0.4961	0.7402	0.4569	0.3177	0.3873
R6	0.6902	0.4084	0.5493	0.9531	0.4880	0.7206	0.6527	0.3949	0.5238
L7	0.6070	0.4974	0.5522	0.5156	0.4507	0.4832	0.8234	0.6891	0.7563
R7	0.5720	0.4448	0.5084	0.5156	0.4572	0.4864	0.8234	0.6688	0.7461
9	0.7490	0.4431	0.5960	0.7734	0.4678	0.6206	0.7363	0.4241	0.5802
10	0.7985	0.5580	0.6782	0.9044	0.4749	0.6897	0.8947	0.5967	0.7457
12	0.6844	0.5645	0.6245	0.6471	0.5810	0.6140	0.6291	0.5158	0.5724
L14	0.7255	0.4205	0.5730	0.7344	0.4234	0.5789	0.4465	0.3312	0.3888
R14	0.7569	0.4987	0.6278	0.7422	0.4539	0.5980	0.3760	0.2919	0.3339
15	0.8821	0.6634	0.7728	0.9926	0.9425	0.9676	0.5564	0.4227	0.4896
16	0.9848	0.8294	0.9071	-	-	-	0.9900	0.8308	0.9104
17	0.7137	0.5118	0.6128	0.7891	0.4410	0.6151	0.4909	0.4015	0.4462
18	0.9962	0.4990	0.7476	0.9779	0.7801	0.8790	0.9424	0.4852	0.7138
20	0.9924	0.4981	0.7452	0.9265	0.4809	0.7037	0.8997	0.4736	0.6867
24	0.9620	0.6331	0.7975	0.9706	0.8256	0.8981	0.9599	0.7039	0.8319
25	0.9924	0.4981	0.7452	0.9338	0.7176	0.8257	0.9599	0.5897	0.7748
26	0.9961	0.4990	0.7475	0.9141	0.5543	0.7342	0.9086	0.4761	0.6923
38	0.9922	0.4980	0.7451	-	-	-	0.9948	0.4987	0.7467
	UAR	**UF1**	**Score**	**UAR**	**UF1**	**Score**	**UAR**	**UF1**	**Score**
Final	**0.7309**	**0.5347**	**0.6328**	**0.7823**	**0.5128**	**0.6476**	**0.6642**	**0.4884**	**0.5763**

Table 1: Experimental scores on CASME II, SAMM and CASME II & SAMM with AU independent 5-fold cross-validation

Facial Sub-region	AUs
Left Brow	L1, L2, L4
Right Brow	R1, R2, R4
Left Eye	L5, L7
Right Eye	R5, R7
Left Cheek	L6
Right Cheek	R6
Nose	9, 38
Mouth	10, 12, 15, 16, 18, 20, 24, 25
Chin	17, 26
Left Nasolabial Area	L14
Right Nasolabial Area	R14

Table 2: AUs in each local key facial sub-regions

address the challenge posed by highly unbalanced multi-labelled data, which is crucial for real-world applicability.

The performance of our method evaluated by AU independent 5-fold cross-validation of three experiments is summarized in Table 1. The results of the experiments conducted on CASME II & SAMM show little deterioration s compared with those obtained by using only CASME II or SAMM data. A possible cause of the slight performance drop may lie in the fact that the SAMM database is more ethnically diverse – CASME II contains data from only one ethnic group, whereas SAMM includes 13 different ethnicities. It is also worth noting that the data acquisition protocols utilized for the collection of the two datasets are different, making the AU detection task on their composite is harder than when no such confounding is present.

As for the subject independent 4-fold cross-validation in Table 3 and Table 4, only F1-score is applied for a fair comparison. The advantages of our approach in addressing the problem of unbalanced data are clearly demonstrated by this comparison. The results show that the proposed method's F1-score corresponding to each individual AUs lies between 0.4 and 0.6. This is in contrast with other methods, which exhibit dependency on the frequency of AU activations. For example, AU4 is the most commonly activated AU in CASME II, so our competitors' detection of other AU is much worse than that of AU4. In summary, our method comprehensively exhibits state-of-the-art performance, out performing the otherwise leading methods in the literature.

AU	LBP-TOP [14]	LBP-SIP [17]	3DHOG [15]	SCA[11]	Ours
1	0.1057	0.2308	0.2771	0.2857	**0.4678**
2	**0.4985**	0.3892	0.2769	0.4532	0.4786
4	0.7324	0.7354	0.7012	**0.8877**	0.5706
7	0.0635	0.0888	0.0000	0.2473	**0.5160**
12	0.2386	0.2143	0.0526	0.4792	**0.5528**
14	0.2185	0.2979	0.0000	0.3327	**0.5070**
15	0.0000	0.4318	0.0000	0.3954	**0.4754**
17	0.1667	0.4287	0.1212	**0.5159**	0.4776
UF1	0.2530	0.3521	0.1786	0.4496	**0.5057**

Table 3: F1-scores on CASME II dataset, with subject independent 4-fold cross-validation

AU	LBP-TOP [14]	LBP-SIP [17]	3DHOG [15]	SCA[11]	Ours
2	0.2652	0.2144	0.0000	0.3289	**0.4873**
4	0.1538	0.0556	0.1667	0.1297	**0.4692**
7	0.4603	0.0400	0.2330	**0.4876**	0.4072
12	0.2376	0.0000	0.0833	0.4218	**0.4541**
UF1	0.2792	0.0775	0.1208	0.3420	**0.4545**

Table 4: F1-scores on SAMM dataset, with subject independent 4-fold cross-validation

5 CONCLUSION

In this paper, we proposed a method for facial Action Unit (AU) detection in micro-expressions achieved by facial sub-region segmentation and multi-label classification in this paper. The approach was empirically evaluated on two popular and publicly available micro-expression databases, namely CASME II and SAMM, on which it is shown to achieve state of the art results. The proposed facial sub-region segmentation method is based on facial landmarks and facial AU distribution positions. The features of key facial areas are extracted, and the micro-expression AUs separated into 11 key facial sub-regions to perform the multi-label classification. The essence of the novelty is the division of labels in a large number of multi-labels into a set of multiple small-number multi-label classifications to determine the final result of each label jointly. We focus on achieving AU multi-label classification by refining facial sub-regions by where the AUs are located.

By means of AU independent 5-fold cross-validation and a comprehensive comparison with the leading methods in the literature using subject independent 4-fold cross-validation, it is shown that the proposed method is successful in addressing the difficulty of micro-expression AU detection caused by unbalanced data – our approach was shown to achieve state-of-the-art results, outperforming its competitors. The proposed methods also opens a range of avenues for future research and further improvement. Amongst these, one of the most obvious ones is the optimization of feature extraction and multi-label classification algorithms. In addition, in future we will attempt to use the proposed method to the learn the mapping from micro-expressions to actual human emotions.

ACKNOWLEDGMENTS

REFERENCES

[1] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, Qianli Feng, Yan Wang, and Aleix M Martinez. 2017. Emotionet challenge: Recognition of facial expressions of emotion in the wild. *arXiv preprint* (2017), 1703.01210.

[2] Adrian K. Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap. 2018. SAMM: A Spontaneous Micro-Facial Movement Dataset. *IEEE Transactions on Affective Computing* (2018).

[3] Paul Ekman and Wallace V Friesen. 1969. Nonverbal leakage and clues to deception. *Psychiatry* (1969).

[4] Paul Ekman and Wallace V. Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology* (1971).

[5] Yingchun Guo, Cuihong Xue, Yingzi Wang, and Ming Yu. 2015. Micro-expression recognition based on CBP-TOP feature with ELM. *Optik* (2015).

[6] Ernest A Haggard and Kenneth S Isaacs. 1966. Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. In *Methods of research in psychotherapy*. Springer.

[7] Xiaopeng Hong, Yingyue Xu, and Guoying Zhao. 2016. LBP-TOP: a Tensor Unfolding Revisit. *ACCV Workshop on Spontaneous Facial Behavior Analysis* (2016).

[8] Vahid Kazemi and Josephine Sullivan. 2014. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

[9] Ling Lei, Tong Chen, Shigang Li, and Jianfeng Li. 2021. Micro-Expression Recognition Based on Facial Graph Representation Learning and Facial Action Unit Fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1571–1580.

[10] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikainen. 2013. A Spontaneous Micro-expression Database: Inducement, collection and baseline. In *IEEE International Conference on Automatic Face and Gesture Recognition*.

[11] Yante Li, Xiaohua Huang, and Guoying Zhao. 2021. Micro-expression action unit detection with spatial and channel attention. *Neurocomputing* 436 (2021), 221–231.

[12] Yong Jin Liu, Jin Kai Zhang, Wen Jing Yan, Su Jing Wang, Guoying Zhao, and Xiaolan Fu. 2016. A Main Directional Mean Optical Flow Feature for Spontaneous Micro-Expression Recognition. *IEEE Transactions on Affective Computing* (2016).

[13] Ling Lo, Hong Xia Xie, Hong Han Shuai, and Wen Huang Cheng. 2020. MER-GCN: Micro-Expression Recognition Based on Relation Modeling with Graph Convolutional Networks. In *International Conference on Multimedia Information Processing and Retrieval*.

[14] Tomas Pfister, Xiaobai Li, Guoying Zhao, and Matti Pietikäinen. 2011. Recognising spontaneous facial micro-expressions. In *Proceedings of the IEEE International Conference on Computer Vision*.

[15] Senya Polikovsky, Yoshinari Kameda, and Yuichi Ohta. 2009. Facial micro-expressions recognition using high speed camera and 3D-Gradient descriptor. In *IET Seminar Digest*.

[16] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. 2011. Random *k*-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering* (2011).

[17] Yandan Wang, John See, R. Raphael, and Yee Hui Oh. 2015. LBP with six intersection points: Reducing redundant information in LBP-TOP for micro-expression recognition. In *Lecture Notes in Computer Science*.

[18] Hong-Xia Xie, Ling Lo, Hong-Han Shuai, and Wen-Huang Cheng. 2020. AU-assisted Graph Attention Convolutional Network for Micro-Expression Recognition. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 2871–2880.

[19] Feng Xu, Junping Zhang, and James Z. Wang. 2017. Microexpression Identification and Categorization Using a Facial Dynamics Map. *IEEE Transactions on Affective Computing* (2017).

[20] Wen Jing Yan, Xiaobai Li, Su Jing Wang, Guoying Zhao, Yong Jin Liu, Yu Hsin Chen, and Xiaolan Fu. 2014. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. *PLoS ONE* (2014).

[21] Wen Jing Yan, Qi Wu, Yong Jin Liu, Su Jing Wang, and Xiaolan Fu. 2013. CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces. In *IEEE International Conference on Automatic Face and Gesture Recognition*.