Configural Representation of Facial Action Units for Spontaneous Facial Expression Recognition in the Wild

Nazil Perveen[a] and Chalavadi Krishna Mohan[b]
Department of Computer Science and Engineering, IIT Hyderabad, Hyderabad, India

Keywords: Facial Expression Recognition, Configural Features, Facial Action Units, Facial Action Coding System.

Abstract: In this paper, we propose an approach for spontaneous expression recognition in the wild using configural representation of facial action units. Since all configural features do not contribute to the formation of facial expressions, we consider configural features from only those facial regions where significant movement is observed. These chosen configural features are used to identify the relevant facial action units, which are combined to recognize facial expressions. Such combinational rules are also known as coding system. However, the existing coding systems incur significant overlap among facial action units across expressions, we propose to use a coding system based on subjective interpretation of the expressions to reduce the overlap between facial action units, which leads to better recognition performance while recognizing expressions. The proposed approach is evaluated for various facial expression recognition tasks on different datasets: (a) expression recognition in controlled environment on two benchmark datasets, CK+ and JAFFE, (b) spontaneous expression recognition on two wild datasets, SFEW and AFEW, (c) laughter localization on MAHNOB laughter dataset, and (d) recognizing posed and spontaneous smiles on UVA-NEMO smile dataset.

1 INTRODUCTION

Humans communicate in multiple ways like verbal, spoken, non-verbal, and unspoken. Facial expressions are considered to be the best way of communicating the message in non-verbal and unspoken way. Among all the body expressions, face is considered to play the vital role in communicating with others through its expressions. Ekman (1957) categorize human facial expression into seven universal categories of angry, disgust, fear, happy, neutral, sad, and surprise. Facial expression recognition is used ubiquitously in different areas of human life like psychology, autism, consumer neuro-science, neuro-marketing, media testing and advertisement, investigations, etc. Bartlett et al. (2003); Perveen et al. (2012); Perveen et al. (2018); Perveen et al. (2016); Zhan et al. (2008). However, with growing requirements, the need of relating it with human computer interaction (HCI) system becomes one of the most extensive area of research in pattern recognition and computer vision.

Human behavior is highly dependant on the signal that brain emits as these signals result in movement of one or more combination of muscles for the necessary action. Similarly, facial expression recognition are the results of the movement of facial muscles triggered by the single nerve known as a facial nerve in human psychology Snell (2008). However, owing to the difficulties in tracking of facial muscles, the other widely used approach for the facial expression recognition is the implementation of multiple coding systems of different facial parts whose combination results to one of the facial expressions. This technique is known as facial action coding system (FACS), introduced in 1969, which was later implemented and further improved by Ekman and Friesen (1976). Facial action coding system is basically the combination of multiple facial action units (FAUs), where each FAUs corresponds to a single facial muscle activity. Due to subjective observations of facial action units (FAUs), one FAU can be mis-interpreted as another FAU. Also, manual labelling of the FAUs in datasets are laborious, time-taking, and error prone. But, with the current methodologies like, active appearance model (AAM), active shape model (ASM), and constrained local model (CLM), the automatic formation of FAUs have become less error prone, which reduces the task of manual labelling to some extent.

The proposed work is motivated by Kotsia and

[a] https://orcid.org/0000-0001-8522-7068
[b] https://orcid.org/0000-0002-7316-0836

Perveen, N. and Mohan, C.
Configural Representation of Facial Action Units for Spontaneous Facial Expression Recognition in the Wild.
DOI: 10.5220/0009099700930102
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 4: VISAPP*, pages 93-102
ISBN: 978-989-758-402-2; ISSN: 2184-4321

Pitas (2007) and Benitez-Quiroz et al. (2016). Using geometric distances, Kotsia and Pitas (2007) recognize facial expressions in one of the two ways: (a) by tracking the facial features and its deformation throughout the videos (b) by detecting the facial action units (FAUs) and combining them with the multiple rules for evaluating expressions. In both the ways, authors use multi-class support vector machine to classify the facial expressions where user has to manually define the candide-grid Ahlberg (2001) onto the face for facial feature tracking. This approach achieves a recognition performance of 99.7% on CK+ dataset with feature tracking and 95.1% using FAUs detection. Benitez-Quiroz et al. (2016) proposes the automatic facial action unit recognition from any given face image in real time by considering all categories of geometric difference, angles, and triangles that exist in the face. All possible distances are extracted among the 66 facial landmark points and Delaunay triangle distance is calculated with all possible angles (≤ 360) imposed on the facial landmarks and resulting in the features vector of dimension $\mathbb{R}^{2466}$. Along with facial action unit recognition, the basic expressions and compounded expressions generated during the experimentation are listed. However, an approach described by Irene is semi-automatic, as for the given frame, the user has to manually define the candide-grid Ahlberg (2001) over the face for proper tracking of the facial features. In our approach, we do not require such manual grid for tracking of the facial features. And contrary to the approach described by Benitez-Quiroz et al. (2016), we do not require large dimensions of the feature vector. The onset form of videos is sufficient for the complete pipeline of the proposed method.

The main objective of the proposed framework is to develop a simple and efficient approach for recognizing facial expression in spontaneous environments. We use the facial action coding system (FACS) for recognizing facial expressions, as humans also recognize facial expression through facial templates and its multiple combinations Ekman and Friesen (1976). Following are the observations from the reviewed works presented above:

1. Existing video based facial expression recognition systems generally consider all possible distances and angles among facial landmark points distributed over the facial regions to capture the configural features. Most of these configural features do not contain relevant information about facial expressions but rather result in high inter-expression similarity. Moreover, calculating such features adds extra computational overhead.
2. The existing facial action coding system (FACS) incurs significant overlap in facial action units (FAUs) across different expressions, which leads to misclassification.

In order to address the above issues, we propose the framework, which comprises of the following:

- To reduce inter-expression similarity, we consider configural features for only those facial regions where significant movement is observed during expressions.
- To reduce overlap in the facial action units across expressions, we introduce a coding system, which are least computational expensive.
- Since existing approaches address expression recognition and expression localization in isolation, we evaluate the efficacy of the proposed approach on (a) spontaneous expression recognition in wild (b) laughter localization, and (c) recognition of posed and spontaneous smiles.

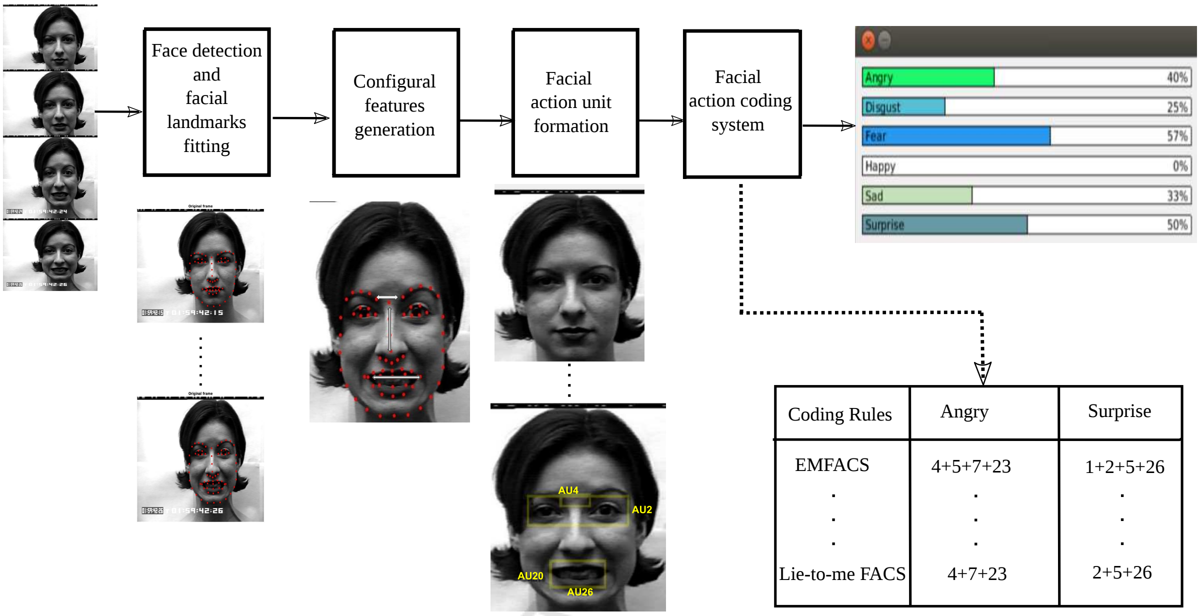
The rest of the section is organized as follows: Section 2 presents the proposed framework. Section 3 and 4 discuss the experimental evaluation of proposed approach in constrained datasets and unconstrained datasets, respectively. Section 5 list experiments of the proposed approach on laughter localization and posed versus spontaneous smile classification. Section 6 gives the conclusion.

2 PROPOSED WORK

In this section, we present the proposed approach for automatic spontaneous facial expression recognition in the wild. Figure 1 presents the block diagram of the proposed approach. The framework consists of four major steps: (i) face and landmark detection, (ii) configural feature generation, (iii) facial action unit (FAUs) recognition, and (iv) facial action coding system (FACS) and expression recognition, which are detailed below:

2.1 Face and Landmarks Detection

In order to recognize facial expressions, the most important step is to detect the face in the video and localize it for further processing. We use discriminative response map fitting method (DRMF) by Asthana et al. (2013), for face detection, localization, and landmarks fitting. Figure 2 shows some of the examples from each dataset used in our experimentation. In general, two modeling approaches are commonly used for facial landmark fitting: 1) holistic modeling and, 2) part-based modeling. Because of certain


Face detection and facial landmarks fitting
Configural features generation
Facial action unit formation
Facial action coding system
Angry 40%
Disgust 25%
Fear 57%
Happy 0%
Sad 33%
Surprise 50%
AU4
AU2
AU20
AU26

Coding Rules	Angry	Surprise
EMFACS	4+5+7+23	1+2+5+26
.	.	.
.	.	.
.	.	.
Lie-to-me FACS	4+7+23	2+5+26

Figure 1: Block diagram of the proposed spontaneous facial expression recognition in wild environment.

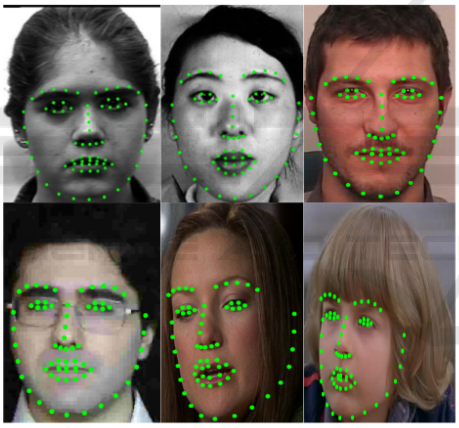


Figure 2: Face detection and landmarks fitting over the facial parts using discriminative response map fitting method Asthana et al. (2013) on different datasets, which are used in experimental evaluation.

drawbacks in holistic modeling like facial features are extracted using facial texture and warping techniques, due to, which variations in the facial movement are not captured properly. Also, general issues like occlusions and 3-D shape of face is not easily modelled using holistic modeling Albrecht et al. (2008). Therefore we use DRMF, as it follows part-based modelling by detecting the facial parts and then using these facial part parameters response maps are created. These response maps with the help of weak learners and a regression technique learn the robust functions to update the shape parameters. This updation goes on till it obtains the best fitting score. Face detection and proper landmark fitting are the most crucial and important stage of our whole approach, as the next stages is heavily dependent on the correct and accurate detection of face and landmark localization on it.

Let $V = \left\{ f_1, f_2, \cdots, f_i, \cdots, f_n \right\}$ be the video consisting of n frames. From each frame f_i a set of t landmarks points are extracted, which is represented as

$$P_i = \left\{ \vec{p}_{i,1}, \vec{p}_{i,2}, \cdots, \vec{p}_{i,k}, \cdots, \vec{p}_{i,t} \right\}, \tag{1}$$

where $\vec{p}_{i,j} = \{x_j, y_j\}$ and $t = 66$. All landmark points in the training images are normalized as mentioned in Asthana et al. (2013).

2.2 Configural Feature Generation

Once we obtain the facial landmarks points, the next step is to calculate the distance among these landmark points to generate configural features

$$C_i = \left\{ c_{i,1}, c_{i,2}, \cdots, c_{i,l}, \cdots, c_{i,s} \right\}. \tag{2}$$

Here, $C_i \in \mathbb{R}^{i \times s}$ where s is the number of configural features. And each $c_{i,l}$ is generated from the pair of landmark points $\vec{p}_{i,k}$ defined in equation 1, i.e

$$c_{i,l} = \parallel \vec{p}_{i,j_1} - \vec{p}_{i,j_2} \parallel, \tag{3}$$

where $j_1, j_2 \in \{1, 2, \cdots, t\}$ and $j_1 \neq j_2$.

Once the set of configural features C_i within the given frame f_i are obtained, the next step is to determine facial action units (FAUs) from these configural features.

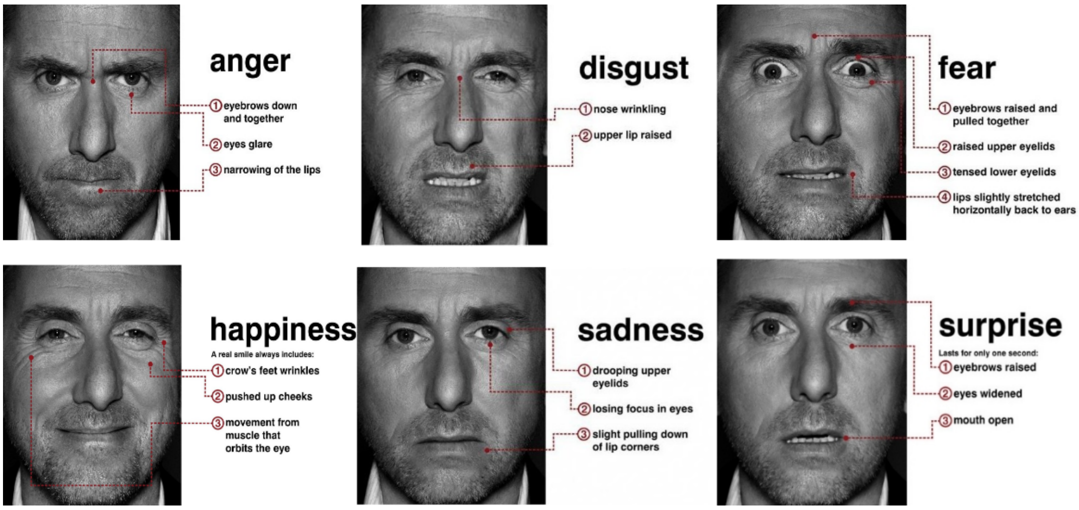


Figure 3: Facial action units considered in Lie-to-me series for expression evaluation Ekman (2009).

2.3 Facial Action Unit (FAU) Formation

In this stage, the configural features C_i obtained from above stage are used to determine the facial action units A_i. Table 1 shows the landmarks and corresponding configural features involved in the FAUs formation. In our approach, we consider 12 basic FAUs for six universal expressions.

Some configural features are evaluated by calculating average of the landmark points, this is because those landmark position are very near to each other and resembles in the similar movements during facial movements. Once FAUs are obtained, the final step is to combine these FAUs to form facial expressions. Let us consider A_i be the set of facial action units (FAUs) for frame f_i represented as

$$A_i = \left\{ a_{i,1}, a_{i,2}, \cdots, a_{i,q}, \cdots, a_{i,r} \right\}, \tag{4}$$

where $r \in \left\{1, 2, \cdots, 12\right\}$. From equation 2 each $a_{i,q}$ is generated from $c_{i,s}$ such that

$$a_{i,q} = \begin{cases} 1, \text{if } \forall \parallel c_{1,s} - c_{i,s} \parallel > \text{T and } i > 1 \\ 0, \text{otherwise.} \end{cases} \tag{5}$$

In the above equation, T is the threshold value that holds the amount of displacement occurring for configural features across the frames with respect to neutral frame, which aids in formation of facial action units. As specified in equation 4, each FAU is the combination of multiple configural features. This threshold value T is determined empirically. In next sub-section we describe how to combine the facial action units obtained for facial expression recognition.

Table 1: Facial action units and their respective landmarks used for computing distances within a frame.

	Function	**Landmarks involve (configural features)**
FAU 1	Inner Brow Raiser	$c_{i,1} = \parallel \vec{p}_{i,22} - \vec{p}_{i,40} \parallel$ $c_{i,2} = \parallel \vec{p}_{i,23} - \vec{p}_{i,43} \parallel$ $c_{i,3} = \parallel \vec{p}_{i,18} - \vec{p}_{i,37} \parallel$ $c_{i,4} = \parallel \vec{p}_{i,27} - \vec{p}_{i,46} \parallel$
FAU 2	Outer Brow Raiser	$c_{i,5} = \parallel \frac{\vec{p}_{i,20}+\vec{p}_{i,21}}{2} - \frac{\vec{p}_{i,38}+\vec{p}_{i,39}}{2} \parallel$ $c_{i,6} = \parallel \frac{\vec{p}_{i,24}+\vec{p}_{i,25}}{2} - \frac{\vec{p}_{i,44}+\vec{p}_{i,45}}{2} \parallel$
FAU 4	Brow Lowerer	$c_{i,7} = \parallel \vec{p}_{i,22} - \vec{p}_{i,23} \parallel$
FAU 5	Upper Lid Raiser	Similar to $c_{i,5}$, $c_{i,6}$, and $c_{i,8} = \parallel \frac{\vec{p}_{i,38}+\vec{p}_{i,39}}{2} - \frac{\vec{p}_{i,41}+\vec{p}_{i,42}}{2} \parallel$ $c_{i,9} = \parallel \frac{\vec{p}_{i,44}+\vec{p}_{i,45}}{2} - \frac{\vec{p}_{i,47}+\vec{p}_{i,48}}{2} \parallel$
FAU 7	Lid Tightener	Similar to $c_{i,8}$ and $c_{i,9}$
FAU 9	Nose Wrinkler	$c_{i,10} = \parallel \vec{p}_{i,28} - \vec{p}_{i,30} \parallel$
FAU 10	Upper Lip Raiser	$c_{i,11} = \parallel \vec{p}_{i,61} - \vec{p}_{i,66} \parallel$ $c_{i,12} = \parallel \vec{p}_{i,63} - \vec{p}_{i,64} \parallel$ $c_{i,13} = \parallel \vec{p}_{i,33} - \vec{p}_{i,55} \parallel$ $c_{i,14} = \parallel \vec{p}_{i,34} - \vec{p}_{i,52} \parallel$ $c_{i,15} = \parallel \vec{p}_{i,35} - \vec{p}_{i,53} \parallel$ $c_{i,16} = \parallel \vec{p}_{i,42} - \vec{p}_{i,49} \parallel$ $c_{i,17} = \parallel \vec{p}_{i,47} - \vec{p}_{i,55} \parallel$
FAU 12	Lip Corner Puller	$c_{i,18} = \parallel \vec{p}_{i,49} - \vec{p}_{i,55} \parallel$ $c_{i,19} = \parallel \frac{\vec{p}_{i,40}+\vec{p}_{i,41}+\vec{p}_{i,42}}{3} - \vec{p}_{i,49} \parallel$ $c_{i,20} = \parallel \frac{\vec{p}_{i,43}+\vec{p}_{i,47}+\vec{p}_{i,48}}{3} - \vec{p}_{i,55} \parallel$
FAU 15	Lip Corner Depressor	Similar to $c_{i,19}$ and $c_{i,20}$
FAU 20	Lip Stretcher	$c_{i,21} = \parallel \vec{p}_{i,49} - \vec{p}_{i,55} \parallel$
FAU 23	Lip Tightener	$c_{i,22} = \parallel \vec{p}_{i,52} - \vec{p}_{i,58} \parallel$
FAU 26	Jaw Drop	Similar to $c_{i,22}$ and $c_{i,23} = \parallel \vec{p}_{i,51} - \vec{p}_{i,59} \parallel$ $c_{i,24} = \parallel \vec{p}_{i,53} - \vec{p}_{i,57} \parallel$

2.4 Facial Action Coding System (FACS) and Expression Recognition

Facial action coding system (FACS) is very old technique, which researchers and psychologists have used over the years for recognition of expressions. In this work, we propose to use a coding system, namely, "Lie-to-me" FACS (L-FACS). Lie-to-me is the television drama series guided by Ekman (2009), one of the pioneer in the area of facial expression recognition. The FACS used in this drama series are shown in Figure 3.

In this drama series, an investigation agency tries to solve the investigating cases with the help of body language, facial expressions, and human psychology. The commonly used approaches to combine these facial action units are known as emotional FACS (EMFACS) and FACS-AID (FACS affect interpretation dictionary). Also, different researchers explored multiple combinations based on their applications needs Kotsia and Pitas (2007)-Benitez-Quiroz et al. (2016). However, we observed significant overlap among the FAUs across the expressions in EMFACS and FACS-AID. Therefore, we follow the subjective interpretation of facial movements during expression formation such that overlapping among the facial action units across the facial expressions is as low as possible. For example, from Figure 3 and Table 2, it can be observed that FAU-10 (upper lip raiser) is more suitable than FAU-15 (lip corner depressor) for recognizing disgust expression. Thus, by forming a new coding rule "9+10" for disgust instead of "9+15" reduces the overlap of FAU-15 that exists between the disgust and sad expressions in EM-FACS. Similarly, for every subjective interpretation, corresponding FAU combinations are assigned in L-FACS to recognize facial expressions. Table 2 gives the complete description of FAUs combination used in EM-FACS and L-FACS. The frame level decisions obtained using L-FACS are then combined to give the decision for a video. The combination is based on most occurring expression across the frame level decisions to determine the expression class for a particular video.

Table 2: FAUs combination from two different coding schemes.

Expressions	Facial action units (EKMAN FACS)	Facial action units (L-FACS)
Angry	4+5+7+23	4+7+23
Disgust	9+15	9+10
Fear	1+2+4+5+7+20+26	2+4+5+20
Happy	6+12	12
Sad	1+4+15	1+15
Surprise	1+2+5+26	2+5+26

3 EXPERIMENTAL EVALUATION OF CONSTRAINED DATASETS

In this section, we evaluate the proposed approach on benchmark datasets, namely, Cohn-Kanade-extended and japanese female facial expression dataset.

3.1 Results on Cohn-Kanade-Extended (CK+)

CK+ Lucey et al. (2010) dataset consist of total 593 posed expression sequences from 123 subjects captured in the duration of 10-60 frames. It contains videos where the expression formation is from neutral to apex expressions under controlled settings. This dataset consist of 69% females, 81% Euro-american, 13% Afro-american, and 6% other groups. The poses ranges from fully frontal to 30 degree facial views of resolution 640 × 480 with 8-bit grayscale. The entire dataset has been categorize into seven facial expressions with their corresponding emotion labels and FAUs labels. Out of these seven expression, we have consider following six expressions, namely, angry, disgust, fear, happy, sad, and surprise in the proposed approach. Table 3 presents comparison of the proposed approach with state of the art approaches. The confusion matrices for the proposed approach using EMFACS and L-FACS are presented in Figure 4 and Figure 5, respectively. It can be observed that proposed approach performs better than existing EMFACS as disgust is misclassified by sad using EMFACS, which is due to common FAU-15 but using L-FACS this misclassification is handled easily. Due to such proper selection of configural features and by reducing the overlap of facial action units across the expressions, we achieve the better results than existing coding systems.

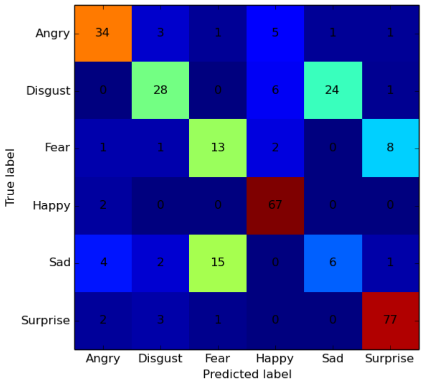
Table 3: Performance comparison with the recent approaches on CK+ datasets.

Methods	Accuracy (%)
Cross-lingual discriminative learning with posterior regularization Ganchev and Das (2013)	74.4
Joint Patch and Multi-label Learning with SVM Zhao et al. (2016)	78.0
Distance-weighted manifold learning Jing and Bo (2016)	80.7
Sparse representation based emotion recognition Lee et al. (2014)	84.4
Deep CNN with multi-layer restricted boltzmann machine Liu et al. (2013)	92.0
3D-CNN with deformable action parts Liu et al. (2014b)	92.4
Configural features with lie-to-me (with $T>3$)	**92.88**

It is to be noted that we used only L-FACS for facial expression recognition in below mentioned datasets.

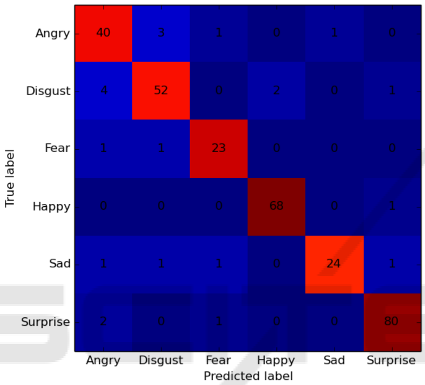
3.2 Results on Japanese Female Facial Expression (JAFFE)

JAFFE Lyons et al. (1998) dataset is posed facial expression dataset consist of 6 basic expressions and neutral expressions. It contains 213 gray scale images from 10 subjects mostly consisting of frontal facial



	Angry	Disgust	Fear	Happy	Sad	Surprise
Angry	34	3	1	5	1	1
Disgust	0	28	0	6	24	1
Fear	1	1	13	2	0	8
Happy	2	0	0	67	0	0
Sad	4	2	15	0	6	1
Surprise	2	3	1	0	0	77

Figure 4: Confusion matrix for CK+ dataset using EMFACS mentioned in Table 2. Using EMFACS, the proposed work obtain 71.19% accuracy.



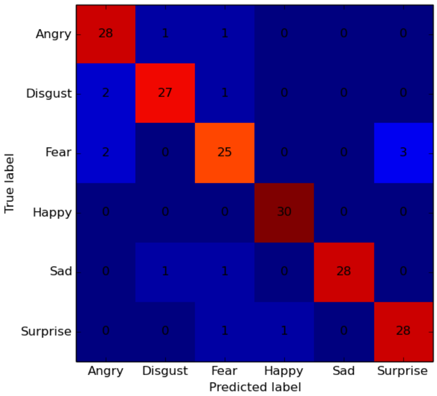
	Angry	Disgust	Fear	Happy	Sad	Surprise
Angry	40	3	1	0	1	0
Disgust	4	52	0	2	0	1
Fear	1	1	23	0	0	0
Happy	0	0	0	68	0	1
Sad	1	1	1	0	24	1
Surprise	2	0	1	0	0	80

Figure 5: Confusion matrix for CK+ dataset using L-FACS mentioned in Table 2. Using L-FACS, the proposed work obtains 92.88% accuracy.

views. Our method mainly design for the evaluation of expression in videos. As this dataset contain neutral expression image for each subject, we are able to evaluate our proposed framework performance on it. Table 4 presents the performance comparison of the proposed approach with recent state of the art methods. Figure 6 gives the confusion matrix evaluated using relevant configural features with L-FACS coding system. As FAU-2 and FAU-5 are the common facial action units in fear and surprise expressions, it can be observed that fear examples are mostly mis-classified as surprise.

Table 4: Performance comparison with the recent approaches on JAFFE dataset.

Methods	Accuracy (%)
Active shape model and support vector machineLei et al. (2009)	89.5
Modified classification and regression tree using LBP and supervised descent methodHappy and Routray (2015)	90.72
Distance weighted manifold learning Jing and Bo (2016)	91
Local binary pattern (LBP) and salient patches extraction with SVMHappy and Routray (2015)	91.78
Configure features with L-FACS (with $T > 3$)	**92.22**



	Angry	Disgust	Fear	Happy	Sad	Surprise
Angry	28	1	1	0	0	0
Disgust	2	27	1	0	0	0
Fear	2	0	25	0	0	3
Happy	0	0	0	30	0	0
Sad	0	1	1	0	28	0
Surprise	0	0	1	1	0	28

Figure 6: Confusion matrix for JAFFE dataset using L-FACS mentioned in Table 4. Using L-FACS, the proposed work obtain 92.22% accuracy in recognizing facial expressions (with T>3).

4 EXPERIMENTAL EVALUATION OF UNCONSTRAINED DATASETS

The viability of the proposed approach is also shown on spontaneous and wild datasets, namely, static facial expression in wild (SFEW) and acted facial expression in wild (AFEW).

4.1 Results on Static Facial Expression in Wild (SFEW)

SFEW Dhall et al. (2011) is the collection of static images of six basic facial expressions and one neutral expression. It consists of three level of subject dependency for facial expressions evaluations, i.e. strictly person specific (SPS), partial person independent (PPI), and strictly person independent (SPI). The proposed approach requires neutral images, which can only be obtained from onset expression videos. As onset videos are a part of the SPS level of SFEW dataset, we evaluate our approach on same. Table 5 shows the performance of SFEW dataset compared to recent approaches. The confusion matrix on SFEW SPS dataset is shown in Figure 7.

Table 5: Performance comparison with recent approaches on SFEW dataset.

Methods	Accuracy (%)
Transfer learning Ng et al. (2015)	48.56
Multiple deep network learning Yu et al. (2015)	55.96
Hierarchial committee of deep convolution neural networkKim et al. (2015)	52.8
Configural features with L-FACS (with T >3)	**67.0**

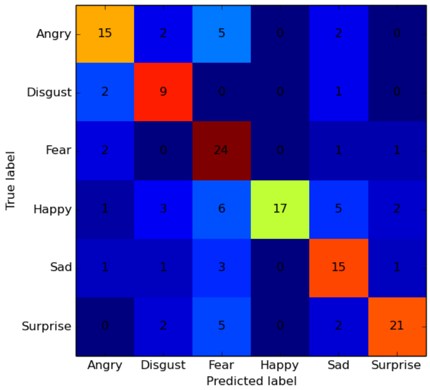


Figure 7: Confusion matrix for SFEW dataset using L-FACS mentioned in Table 5. Using L-FACS, the proposed approach obtains 67.0% accuracy in recognizing facial expressions (with T>3).

4.2 Results on Acted Facial Expression in Wild (AFEW)

AFEW Dhall et al. (2012) is the dynamic movie video corpus consisting of six universal facial expressions and one neutral expression in the wild environment. The training set consists of 723 movie clips and validation set contains 383 movie clips. Figure 8 shows subjects labelled as sad but it can be observed that they do not exhibit any facial actions (FAU-1 and FAU-15), which can be classified as sad. Table 6 shows the performance comparison of existing work with the proposed approach on AFEW dataset. The confusion matrix on AFEW dataset on both training and validation dataset is shown in Figures 9.

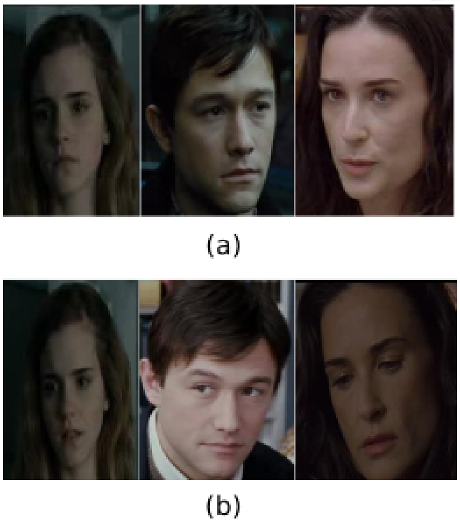


Figure 8: (a) represents the subjects from sad expressions (b) represents the corresponding subjects from neutral expressions. It can be easily observed that, no facial movements is present on both the expressions, due to which sad is also mis-classified as neutral.

Table 6: Performance comparison with the recent approaches on AFEW dataset.

Methods	Accuracy (%)
TOP-k HOG Feature Fusion with Multiple Kernel Learning Chen et al. (2014)	40.2
Combining Multimodal Features with Hierarchical Classifier FusionSun et al. (2014)	42.32
Combining Multiple Kernel Methods on Riemannian ManifoldLiu et al. (2014a)	48.52
Combining Modality Specific Deep Neural Network ModelsKahou et al. (2013)	49.49
Contrasting and Combining Least Squares Based LearnersKaya et al. (2015)	52.30
AU-aware facialfeature relations (two face scales) with Audio fusionYao et al. (2015)	53.80
Recurrent Neural Network Ebrahimi Kahou et al. (2015)	68.463
Configural features with L-FACS (with T >3)	**71.54**

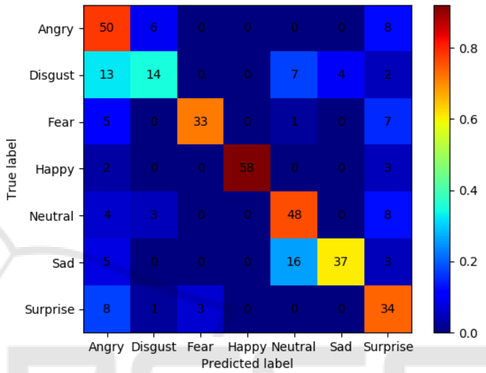


Figure 9: Confusion matrix for AFEW dataset using L-FACS. The proposed work obtains 71.54% accuracy.

5 EXPERIMENTAL EVALUATION OF LAUGHTER LOCALIZATION AND POSED VERSUS SPONTANEOUS SMILES RECOGNITION

The efficacy of the proposed approach is also extended for laughter localization for untrimmed videos on MAHNOB laughter dataset and for posed and spontaneous smile recognition on UVA-NEMO smile dataset.

5.1 Results on MAHNOB-laughter Dataset and UVA-NEMO-smile Dataset

MAHNOB Petridis et al. (2013) is a audio-visual laughter dataset, where video is recorded at 25 fps and microphone is used for audio data. There are 191 samples of 22 subjects where there are 12 males and 10 females, in total 28 posed laughter videos, 121 spontaneous and 42 speech features are recorded. Also, UVA-NEMO Dibeklioglu et al. (2012) is a smile dataset, which consist of 643 posed and 597 spontaneous smile videos by 400 subjects, under controlled illumination condition at 50 fps. We use both datasets to check the authenticity of our approach through laughter localization and posed versus spontaneous laughter recognition in videos. Table 7 gives the performance measures on the laughter datasets for laughter localization and for detecting posed and spontaneous laughter. For laughter localization, we keep track of the most prominent FAU in laughter(smile), i.e AU-12, and its occurrences are compared with the annotation provided by the author. The recognition performance on MAHNOB laughter dataset is 89.017%. The graph plot in Figure 10 shows the AU-12 activation in spontaneous videos and Figure 11 shows the AU-12 activation in posed videos. However, for smile dataset, laughter localization is not evaluated as the videos are very small and dataset is also not meant for this purpose. For the case of predicting posed and spontaneous laughter in the videos, we keep track of the duration when AU-12 is active in the frame sequences. We notice that in most of the posed videos the duration of activation of AU-12 is very less and they occur frequently as compared to the sustainable activation of AU-12 in spontaneous videos. Along with this, the proposed approach also keeps track of the threshold value T mentioned in section 3 for AU-12. In case of spontaneous laughter as shown in Figure 12, the T value starts from some minimum value ($T>3$) and reaches a peak value where laughter expression is at the apex and sustains across the video without decreasing. And if the value of T fluctuates frequently across the frames then such videos are considered as posed. This is demonstrated using Figures 13 and 14. The T value plot shows that T value starts from some minimum ($T>3$) and reaches a peak value at the apex point of laughter and then decreases gradually across the video.

Table 7: Experimental results on two different laughter datasets for posed and spontaneous facial expression recognition and for expression localization.

	Posed (%)	**Spontaneous (%)**	**Expression Localization. (%)**
MAHNOB **Petridis et al. (2013)**	71.42	86.77	89.017
UVA-Nemo **Smile dataset** **Dibeklioglu et al. (2012)**	80.56	85.69	—

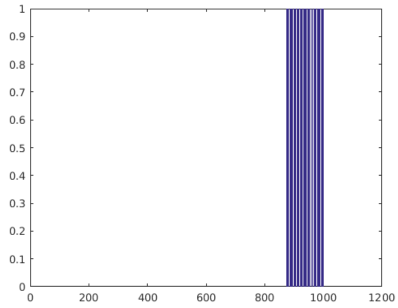


Figure 10: Activation of FAU-12 is plotted for the subject S001_001 (spontaneous laughter) from mahnob laughter dataset. The above plot shows the continuity in the activation of FAU-12 through out the videos once the laughter is started (approximately from frame number 870).

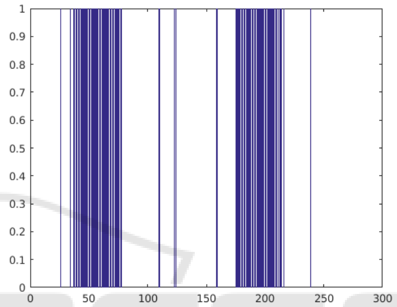


Figure 11: Activation of FAU-12 is plotted for the subject S009_003 (posed laughter) from mahnob laughter dataset. Discontinuity of laughter can be noticed by discontinued activation of the FAU-12 (approximately from frame number 39 and 175).

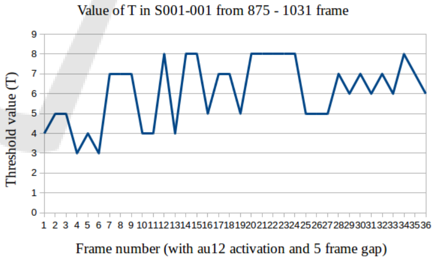


Figure 12: T value plot for predicting the spontaneous laughter (or smiles) in the video.

6 CONCLUSIONS

This paper presents a novel approach for selecting configural features and a subjective interpretation based coding system L-FACS. The efficacy of the proposed approach is demonstrated on onset to apex categories of videos for facial expression recognition on the following tasks: (i) facial expression recognition in controlled environments, (ii) spontaneous expression recognition in wild environments, (iii) laughter

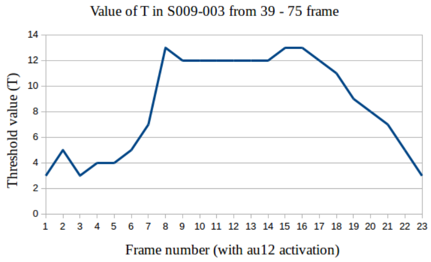


Figure 13: T value plot for posed laughter. Subject S009_003 is tracked from frame number 39-75.

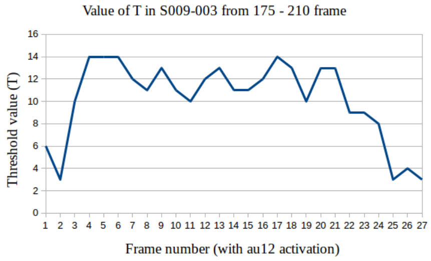


Figure 14: T value plot for posed laughter. Subject S009_003 is tracked from frame number 175-210.

localization on large videos, and (iv) posed and spontaneous smile recognition. The selection of configural features combined with L-FACS is shown to outperform state of the art approaches. By combining frame level decisions to classify a video into a particular expression, the proposed approach handles scaling and pose-related issues that may arise in a few frames of the video. In future, we would like to extend our approach for estimating the intensity of facial expressions for any categories of videos (apex/offset) in an unconstrained environment.

REFERENCES

Ahlberg, J. (2001). Candide-3 - an updated parameterised face. Technical report.

Albrecht, T., Luthi, M., and Vetter, T. (2008). A statistical deformation prior for non-rigid image and shape registration. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8.

Asthana, A., Zafeiriou, S., Cheng, S., and Pantic, M. (2013). Robust discriminative response map fitting with constrained local models. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3444–3451.

Bartlett, M. S., Littlewort, G., Fasel, I., and Movellan, J. R. (2003). Real time face detection and facial expression recognition: Development and applications to human computer interaction. In *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on*, volume 5, pages 53–53. IEEE.

Benitez-Quiroz, C. F., Srinivasan, R., and Martinez, A. M. (2016). Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5562–5570.

Chen, J., Chen, Z., Chi, Z., and Fu, H. (2014). Emotion recognition in the wild with feature fusion and multiple kernel learning. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pages 508–513, New York, NY, USA. ACM.

Dhall, A., Goecke, R., Lucey, S., and Gedeon, T. (2011). Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2106–2112.

Dhall, A., Goecke, R., Lucey, S., and Gedeon, T. (2012). Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 19(3):34–41.

Dibeklioglu, H., Salah, A. A., and Gevers, T. (2012). *Are You Really Smiling at Me? Spontaneous versus Posed Enjoyment Smiles*, pages 525–538. Springer Berlin Heidelberg, Berlin, Heidelberg.

Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R., and Pal, C. (2015). Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 467–474, New York, NY, USA. ACM.

Ekman (2009). Lie-to-me. http://www.paulekman.com/lie-to-me/.

Ekman, P. (1957). A methodological discussion of nonverbal behavior. *The Journal of psychology*, 43(1):141–149.

Ekman, P. and Friesen, W. V. (1976). Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75.

Ganchev, K. and Das, D. (2013). Cross-lingual discriminative learning of sequence models with posterior regularization. In *EMNLP*, pages 1996–2006.

Happy, S. and Routray, A. (2015). Automatic facial expression recognition using features of salient facial patches. *IEEE transactions on Affective Computing*, 6(1):1–12.

Jing, D. and Bo, L. (2016). Distance-weighted manifold learning in facial expression recognition. In *Industrial Electronics and Applications (ICIEA), 2016 IEEE 11th Conference on*, pages 1771–1775. IEEE.

Kahou, S. E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, Ç., Memisevic, R., Vincent, P., Courville, A., Bengio, Y., Ferrari, R. C., et al. (2013). Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 543–550. ACM.

Kaya, H., Gürpinar, F., Afshar, S., and Salah, A. A. (2015). Contrasting and combining least squares based learners for emotion recognition in the wild. In *Proceedings of the 2015 ACM on International Conference* *on Multimodal Interaction*, ICMI '15, pages 459–466, New York, NY, USA. ACM.

Kim, B.-K., Lee, H., Roh, J., and Lee, S.-Y. (2015). Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 427–434, New York, NY, USA. ACM.

Kotsia, I. and Pitas, I. (2007). Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Transactions on Image Processing*, 16(1):172–187.

Lee, S. H., Plataniotis, K. N., and Ro, Y. M. (2014). Intra-class variation reduction using training expression images for sparse representation based facial expression recognition. *IEEE Transactions on Affective Computing*, 5(3):340–351.

Lei, G., Li, X.-h., Zhou, J.-l., and Gong, X.-g. (2009). Geometric feature based facial expression recognition using multiclass support vector machines. In *Granular Computing, 2009, GRC'09. IEEE International Conference on*, pages 318–321. IEEE.

Liu, M., Li, S., Shan, S., and Chen, X. (2013). Au-aware deep networks for facial expression recognition. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6.

Liu, M., Wang, R., Li, S., Shan, S., Huang, Z., and Chen, X. (2014a). Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pages 494–501, New York, NY, USA. ACM.

Liu, P., Han, S., Meng, Z., and Tong, Y. (2014b). Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1805–1812.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101.

Lyons, M., Akamatsu, S., Kamachi, M., and Gyoba, J. (1998). Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 200–205.

Ng, H.-W., Nguyen, V. D., Vonikakis, V., and Winkler, S. (2015). Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 443–449, New York, NY, USA. ACM.

Perveen, N., Gupta, S., and Verma, K. (2012). Facial expression recognition using facial characteristic points and gini index. In *2012 Students Conference on Engineering and Systems*, pages 1–6.

Perveen, N., Roy, D., and Mohan, C. K. (2018). Spontaneous expression recognition using universal attribute model. *IEEE Transactions on Image Processing*, 27(11):5575–5584.

Perveen, N., Singh, D., and Mohan, C. K. (2016). Spontaneous facial expression recognition: A part based approach. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 819–824.

Petridis, S., Martinez, B., and Pantic, M. (2013). The mahnob laughter database. *Image and Vision Computing*, 31(2):186 – 202. Affect Analysis In Continuous Input.

Snell, R. (2008). *Clinical Anatomy by Regions*. Lippincott Williams & Wilkins.

Sun, B., Li, L., Zuo, T., Chen, Y., Zhou, G., and Wu, X. (2014). Combining multimodal features with hierarchical classifier fusion for emotion recognition in the wild. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pages 481–486, New York, NY, USA. ACM.

Yao, A., Shao, J., Ma, N., and Chen, Y. (2015). Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 451–458, New York, NY, USA. ACM.

Yu, X., Zhang, S., Yan, Z., Yang, F., Huang, J., Dunbar, N. E., Jensen, M. L., Burgoon, J. K., and Metaxas, D. N. (2015). Is interactional dissynchrony a clue to deception? insights from automated analysis of nonverbal visual cues. *IEEE Transactions on Cybernetics*, 45(3):492–506.

Zhan, C., Li, W., Ogunbona, P., and Safaei, F. (2008). A real-time facial expression recognition system for online games. *Int. J. Comput. Games Technol.*, 2008:10:1–10:7.

Zhao, K., Chu, W. S., la Torre, F. D., Cohn, J. F., and Zhang, H. (2016). Joint patch and multi-label learning for facial action unit and holistic expression recognition. *IEEE Transactions on Image Processing*, 25(8):3931–3946.