Detection and Analysis of Facial Micro-Expression Using Image Processing

Ravi kr. Bhall, Saurabh Singh, Kriti Jain, Vishal Sharma

***Abstract*— In psychology facial expressions are used to analyze the behavioral aspects of a subject to know the suppressed feelings such as anger, sadness, happiness and other more. These expressions are proven to be more successful in identifying the mood and the real intentions of the subject but the main problem with facial expressions is that they can be faked which Leeds to misjudging the subject. Recent studies show that there are some leaked micro-expressions which occur for very small duration i.e. 1/3 to 1/25 second and can't be controlled thus can't be faked. These micro expressions are nearly impossible to detect by naked eyes and without special training. The system will detect these facial micro-expressions which help to reveal the true feelings of the subject. The method is capable of spotting both macro and micro expressions which are typically associated with emotions such as happiness, sadness, anger, disgust, and surprise, and rapid micro-expressions which are typically, but not always, associated with semi-suppressed macro-expressions.**

***Index terms* – facial micro-expression, deception detection, suppressed emotion detection.**

I. INTRODUCTION

Human brain is capable of recognizing facial expressions which provides a vast source of important and affective information. After thirty years of research in micro-expressions by Ekman, Frank and O'Sulliva [1] and a depended group of Portet [2] these micro-expressions were found an important behavioral source for detecting deception and danger demeanor detection as well [1]. These facial micro-expressions are brief, involuntary expression shown by the human face when they are trying to hide or fake an emotion. Micro-expressions usually occur at high-stakes moments, where people have something to gain or lose [3]. Theses micro-expressions are fast involuntary facial expressions which gives a brief reaction to feelings that people undergo but try to hide the feelings.

From the technical point, the detection of facial micro-expressions is a hard task using the traditional approaches. Duration of micro-expression is 1/25th to 1/3rd of a second and with appearance of low muscle intensity. For detecting these micro-expressions requires only a use of a high-speed camera. Only highly trained people are able to detect micro-expressions with naked eyes.

There are number of potential applications of micro-expressions such as police can use them to detect abnormal behavior, and in medical field doctors can detect micro-expressions showing suppressed emotions to know when additional reassurance is needed, in education field teachers can recognize student's unease and make the lecture more affective, business negotiators can use the micro-expressions to know when they have to propose a suitable price.

The main objective in detecting facial micro-expression involve the short duration and there occurrence as they are involuntarily. The occurrence is limited to very short duration and low number of frames with a 25fps camera. To obtain a high detection rate the best practice is to use a high frame rate camera such as 100fps or 200fps camera.

In this paper proposed is a system for detecting facial micro-expression that achieves very good results. This system detects facial micro expression in three steps: 1) video acquisition, 2) feature extraction, 3) analysis of extracted features.

II. RELATED WORK

In facial data extraction and representation for expression analysis, two main approaches exist: geometric feature-based methods and appearance-based methods. A review can be found on bhall et. al.[4]

The geometric facial features are presented by the shape and location of facial components (such as mouth, eyes, eyebrows, and nose). The facial components and facial feature points are extracted by some computer vision techniques that form a feature vector that represents the face geometry.

Superior research results were reported on Active Appearance Model (AAM) by Kanade [5] group. However, there are two disadvantages of AAM. First, this approach requires extensive dataset with large amount of manually tagged points of the face. Second, the accuracy of facial feature tracking significantly decreases in the faces that were not included in the training set.

Another approach is based on direct tracking of 20 facial feature points (e.g. eye and mouth corner, eyebrow edges) by particle filter [6]. This approach delivers good results for some facial motions, but fails in detecting subtle motions, that can be detected only by observing skin surface. The performance of this and similar approaches strongly rely on the accuracy of the facial feature points tracking. In practice,

Mr. Ravi Kr. Bhall, M.Tech,Department of computer science engineering, DIT University, Dehradun, Uttarakhand, India.M.No-09756127820

Mr. Saurabh Singh, M.Tech, Department of computer science engineering, DIT University, Dehradun, Uttarakhand, India. M.No-09720518418

Kriti Jain, M.Tech, Department of computer science engineering, DIT University, Dehradun, Uttarakhand, India. M.No- 09997172472

Mr. Vishal Sharma,Associate Professor, Department of computer science engineering, DIT University,Dehradun, Uttarakhand, India.
M.No- 09045958401

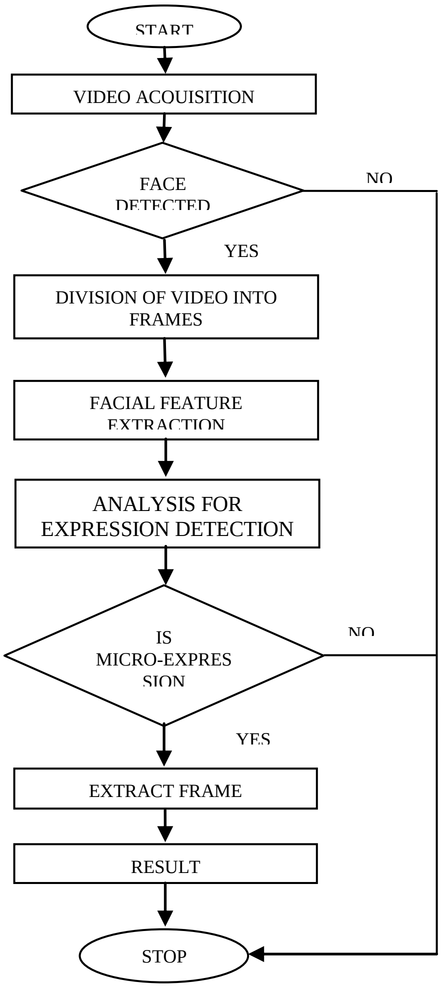
facial feature points tracking algorithm cannot deliver the necessary accuracy for micro-expression recognition task.

In Appearance-Based Methods, image filter, such as Gabor Wavelets are applied to either the entire face or specific regions in the face, to extract a feature vector. This method was applied for spontaneous facial motion analysis and considered to be the most popular [7]. However, this method is based on analyzing the video frame by frame, without considering correlation between frames. In addition, applying this approach for facial surface analysis requires large datasets for training an enormous number of filters.

Using Spatio-Temporal Strain: In this method [8] for the automatic spotting (temporal segmentation) of facial expressions in long videos comprising of macro- and micro-expressions. The method utilizes the strain impacted on the facial skin due to the non- rigid motion caused during expressions. The strain magnitude is calculated using the central difference method over the robust and dense optical flow field observed in several regions (chin, mouth, cheek, and forehead) on each subject's face. This approach is able to successfully detect and distinguish between large expressions (macro) and rapid and localized expressions (micro).

III. METHOD

3.1 Flowchart for Micro-Expression detection and analysis:-


START
VIDEO ACQUISITION
FACE DETECTED
NO
YES
DIVISION OF VIDEO INTO FRAMES
FACIAL FEATURE EXTRACTION
ANALYSIS FOR EXPRESSION DETECTION
IS MICRO-EXPRESSION
NO
YES
EXTRACT FRAME
RESULT
STOP

3.2 Algorithm

```
MICRO-DETECT(C)
  I. Detect face, if face not found exit.
  II. Initialise block size $\Gamma= \{8*8*1, 5*5*1, 8*8*2, 5*5*2\}$ and $T=\{10,15,20,30\}$
  III. Convert video into frames.
  IV. For all 1 to number of frames $P_{i,s}.......P_{i,s}$
      i. In first frame $P_{i,1}$ Detect face $F_i$
      ii. Locate and extract facial feature points using ASM
          $\Psi = \{(a_1,b_1).....(a_h,b_h)$
      iii. Compare and normalise face with model face by calculating LWM transformation
          $\zeta = \mathrm{LWM}(\Psi,\omega, P_i,1)$ where $\omega$ is feature point matrix for model face & $P_i$,1 is the first frame
      iv. Apply transformation $\zeta$ to all frames $P_{i,2\ to}P_{i,s}$
      v. Find eyes in each frame and crop the face using ASM
      vi. For each $\theta \in T$ compute Temporal Image Sequence
          $\xi_{i,\theta} = U M f^n(t) + \overline{\xi_{i,\theta}}$
      vii. For all $p \in \Gamma, \theta \in T$ extract set of SLTDs
          $\mu_{i,p,\theta}(\xi_{i,\theta})=\{q_{i,p,\theta},1 .....q_{i,p,\theta},M\}$
          with SLTD feature vector length $M$
  V. Evaluate kernels $K = \{\forall j,k,m,\theta,p.c_j \in \wedge c_k \in C \wedge m=1....M \wedge \theta \in T \wedge p \in \Gamma \wedge r=(m, \theta, p) | \mathrm{HISINT}(q_{j,r},q_{k,r}), \mathrm{POLY}(q_{j,r}, q_{k,r}, 2), \mathrm{POLY}(q_{j,r}, q_{k,r}, 6)\}$
  VI. Micro=MKL-PHASE(K)
```

C is the input data i.e. video recording of the subject's facial movements. Γ is SLTD parameter set where x*y*t are the row, column and temporal block respectively. In these sets features are divided. T is the frame count set in which C_i image sequence is temporally interpolated. LWM(Ψ,ω, P) evaluates the local weighted mean transformation for each frame P by usage of feature points Ψ and the model face feature points ω as in feature point extraction. HISINT($q_{j,r}$,$q_{k,r}$), POLY($q_{j,r}$, d) evaluates the polynomial kernel of degree d and the histogram intersection kernel as in eq. 2 and 3. MKL-PHASE1(K) is the output i.e. detected facial micro-expression.

3.3 Facial Feature Marking

To locate the high variations in the spatial appearances of facial micro-expressions, cropping and normalising the face geometry according to the positions of eye from a Haar eye detector and the feature points are located using an Active Shape Model (ASM) [9] deformation. ASMs are one of the statistical models for the shape of an object that are repeatedly deformed to fit on an example of the object. It initiates the search from a average shape aligned to the location and size of the face stated by a face detector and iterates until convergence. The tentative shape is matched by template identical of image texture around located points to change feature point locations, and Fitting tentative shapes to the global shape model. Using 68 Active Shape Model feature points shown in Figure 1 we evaluate a Local Weighted Mean (LWM) [10] transformation of frame $p_{i,1}$ for sequence *i*. LWM evaluates the weighted mean of every polynomials that passes over each point by setting the values of an arbitrary point *(x, y)* to

$$f(x,y) = \frac{\sum_{i=1}^{N} V(\sqrt{(x-x_i)^2+(y-y_i)^2} \, / \, R_n) S_i(x,y)}{\sum_{i=1}^{N} V(\sqrt{(x-x_i)^2+(y-y_i)^2} \, / \, R_n}$$

EQ. 1

where *Si(x, y)* is polynomial with n parameters which passes through a measurement for control point $(x_i, y_i)$and n – 1 other measurements nearest to it, V is the weight and Rn is the distance of (x_i, y_i) from its $(n-1)^{th}$ nearest control point in the reference image. We then apply the transformation to $p_{i,2}$.... $p_{i,s}$ for an expression with s frames. Figure 1 illustrates the LWM transformation of the facial feature points in an example face compared to a model face. Haar eye detection outcomes were checked against Active Shape Model (ASM) feature points and these points were used to crop the image. Then spatiotemporal local texture descriptors (SLTD) are applied to the video for feature extraction.

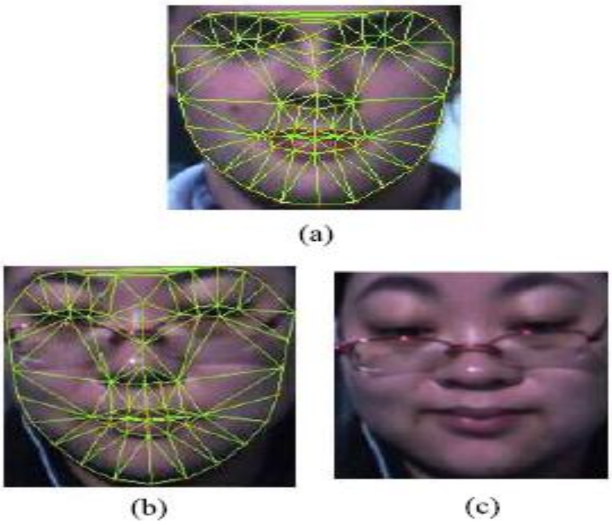


Fig: 1 facial feature point detection using a model face a) is the model face or example face b) is the facial points detected in the frame c) is the face after feature point detection.

All micro-expression to a given set of frames are further normalized temporally θ∈T For every micro-expression image sequence *i* we evaluate a temporally interpolated image sequence $\xi_{i,\theta} = UMF^n(t) + \overline{\xi_{i,\theta}}$ for all *θ∈T*. where *U* is the decomposition matrix for singular value, ***M*** denotes square matrix, $F^n(t)$ is a curve and $\overline{\xi_i}$ is a mean vector.

Then apply SLTD (Spatiotemporal Local Texture Descriptors) to the video for the process of feature extraction. SLTD requires a input video of a short length. In this case LBP-TOP[11] is used with radius of R=3 and the block size is shown in the Algorithm. These parameters needs to remove first and last 3 frames because descriptor can’t be placed here. To enabling at least 1 frame extraction for a segment there come’s a need of at least 7 frames of data. With a camera having 25fps framerate will generate a 1/3 to 1/25 second video which will be having 1 to 8 frames. It is important to derive more frames therefore SLTD is used for longest micro expressions. However it is expected to achieve more statistically stabilized histogram with high number of frames. This is demonstrated in the next section.

To improve the classification results Multiple Kernel Learning (MKL) [12] given training set $H = \{(x_1, l_1) \ldots\ldots (x_n, l_n)\}$ and set of kernels $K_1 \ldots\ldots K_M$ where $K_k \in \mathbb{R}^{n^*n}$and K_k is positive semi-definite, Multiple Kernel Learning (MKL) learns linear/non-linear combinational weights of kernels over different domains by optimizing a cost function *Z(K,H)* where K is basic kernels combination. As shown in algorithm in another section. combine polynomial kernels POLY of degrees 2 and 6 histogram-intersection kernel HISINT with different SLTD parameters P∈Γ over different temporal interpolations θ∈T where

$$POLY(q_{j,r}, q_{k,r}, d) = (1 + q_{j,r} q_{k,r}^T)^d$$

Eq. 2

$$\text{HISINT}(q_{j,r}, q_{k,r}) = \textstyle\sum_{a=1}^{b} \min\{q_{j,r}^a, q_{k,r}^a\}$$

Eq. 3

And *r=(m,θ,p)* and *b* is the no. of bins in $q_{j,r}, q_{k,r}$ Random Forest and SVM is used as alternative classifiers. Our classification system is single phased. MKL-PHASE1(K) detects the occurrence of a facial micro-expression. The IInd phase is to classify the micro expression which is a part of future scope for our project.

If MKL-PHASE1(K) = micro, classifies the facial micro-expression into arbitrary set of classes $L=\{l_1 \ldots l_n\}$. The task is divided into two pipelined phases which enables us to

1) Optimizing phase’s separately.
2) Tailor L for phase II for a given application while retaining the original optimised phase I.

Further for labelling data for phase II require a further deeper analysis, which is subject to many labelling error. By separating the two phases we avoided the subjective labelling of expressions (Phase 2) which affect the whole detection process.

3.4 Analysis for Expression Detection

In this step the features extracted in each frame is the compared using Temporal Interpolation Method. This method is used previously proposed by Zhou et al. [11] for synthesise tracking movements of talking mouth. This method allows inputting sufficient number of frames in feature descriptor even for shortest expressions having smallest number of frames used for extraction and also enables good extraction results on increasing frame numbers used for extraction.

3. 5 Temporal Interpolation Method

Video of a micro-expression is viewed as a sequence of sampled images along a curve [13] to create a continues function in a low-dimensional manifold by micro-expression represented video as graphical path P_n having *n* vertices as in figure 2. Vertices correspond to the video frame and edges correspond to the adjacency matrix $W \epsilon \{0,1\}^{n \times n}$ with $W_{i,j} = 1$ if |*i-j*| =1 and 0 otherwise.

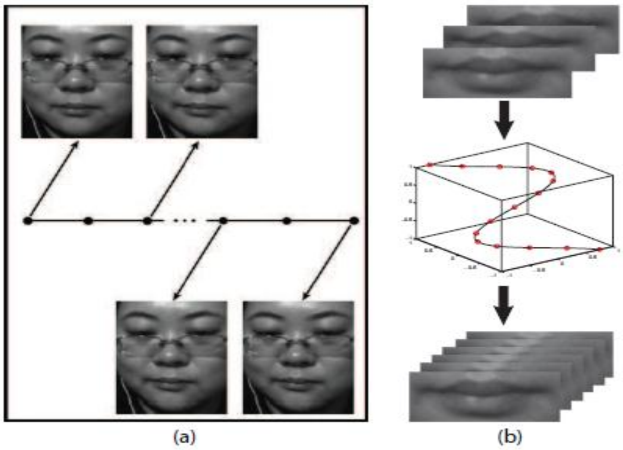


Fig 2: a) graphical representation of facial micro expression, b) shows the temporal interpolation method mapped video along a curve.

For embedding the manifold in the graph mapping P_n to the line which minimises the distance between joined vertices. Let $y = (y1,y2,.........,y_n)^T$ be the map. Minimised to

$$\sum_{i,j} (y_i - y_j)^2 W_{ij}\,, \quad i,j = 1,2,\ldots\ldots,n$$

Eq. 4

Obtaining *y*, which is equal to calculating the Eigen vectors of the laplacian graph of $P_{n.}$ After computing the laplacian graph such a way that it has Eigen vectors i.e. $\{y_1, y_2 \ldots\ldots y_{1n-1}\}$ and it enables to view y_k as sets of points described as

$$f_k^n(t) = \sin(\pi kt + \pi(n-k) \;/\; (2n)), t \epsilon [1 \,/\, n, 1]$$

Eq. 5

Sampled at t=1/n, 2/n,.....,1. the resulting curve can be used to temporary interpolate the images at arbitrary position within a micro-expression.

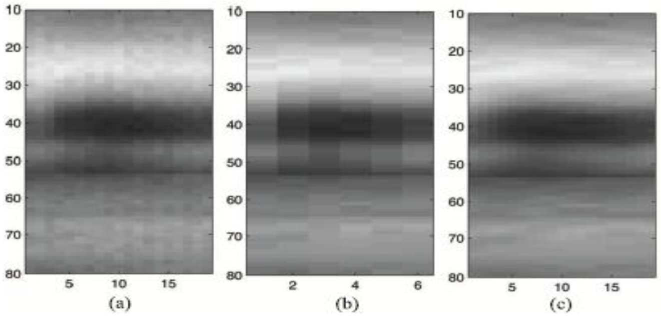


FIG 3: Temporal interpolation. The vertical temporal patterns.

$$f^n(t) = \begin{bmatrix} f_1^n(t) \\ f_2^n(t) \\ . \\ . \\ f_{n-1}^n(t) \end{bmatrix}$$

Eq. 6

To find correspondence for curve f^n within the image space, by mapping the image frames to points defined as $f^n(1/n), f^n(2/n), \ldots\ldots, f^n(1)$ and using the linear extension of graph embedding for learning a transformation vector *W* which will minimise

$$\sum_{i,j} (w^T x_i - w^T x_j)^2 W_{ij}\,, \quad i,j = 1,2,\ldots\ldots\ldots,n$$

Eq. 8

Where $x_i = \xi_i - \bar{\xi}$ is a removed mean vector and ξ_i is the vectorised image. X. He et al. solved this eigen value resulting problem

$$XLX^T w = \lambda XX^T w$$

Eq. 8

By using singular value decomposition with X = U∑V^T. Zhou etal. Proved that a new image can be interpolated by

$$\xi = UMf^n(t) + \bar{\xi}$$

Eq. 9

Where M is square matrix

The validity depends upon assuming linear independency of ξ_i the assumption held for SMIC database.

On computing a temporally interpolated frame image sequence $\xi_{i,\theta} = UMf^n(t) + \xi_{i,\theta}$ for all $\theta \epsilon T, c_i \epsilon C$,

Compute all possible combination of them with different SLTD block parameter Γ and choosing the number of frames θ∈T and parameters p∈Γ this will help maximizing the accuracy for given C.

IV. RESULT

4.1 Dataset

SMIC Database : The database [14] was recorded in an indoor bunker environment designed to resemble an interrogation room. Indoor illumination was controlled stable through the whole data recording period with four lights from the four upper corners of the room. 16 carefully selected movie clips, which can induce strong

emotions, were shown to participants on a computer monitor together with a speaker for audio output. Participants sat about 50cm from the computer monitor. While participants were watching the film clips, a camera fixed on top of the monitor recorded their facial reactions. The setup is illustrated in Figure 1. 20 participants participated in the recording experiment.

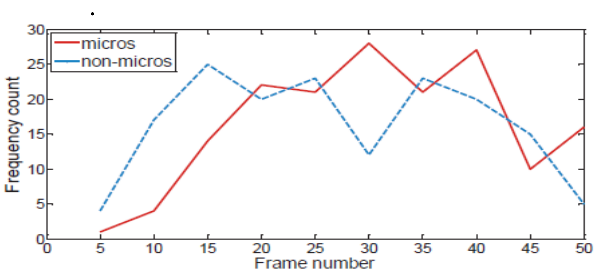
For the recording of the first ten participants, a high speed (HS) camera (PixeLINK PL-B774U, 640×480) of 100fps was used to record the short duration of micro-expressions. In addition to the high speed camera another integrated camera box was added which consists of a normal visual camera (VIS) and a near-infrared (NIR) camera, both with 25 fps and resolution of 640×480. The

VIS and NIR cameras were added for two reasons: first, to improve the diversity of the database; second, to investigate whether the current method can also be used on normal speed cameras of 25 fps. In contrast to a down-sampled version of the 100 fps data, the 25 fps data yields data similar to standard web cameras, including their limitations such as motions blurs. When multiple cameras were used, they were put parallel to each other and fixed on the middle top of the monitor to ensure frontal view recording. Due to technical issues there was a time delay about 3-5 seconds between the starting points of the three cameras. VIS and NIR clips were manually synchronized with the reference to the HS data.

4.2 Implementation

On implementing proposed algorithm detection is done on the basis of leave one subject out on database i.e. the first frame of every dataset video is taken as a baseline frame for that detection process.

The final SMIC database contain 164 micro-expression video clips from 16 participants, the clips are recorded at HS (high speed) camera rate. The frame distribution of the recorded video clips is shown in figure 4.



F

ig 4: The frame distribution for the video recordings.

As the SLTD (Spatiotemporal Local Texture Descriptors) uses LBP-TOP. Thus in this system block sizes are used for MKL (Multiple Kernel Learning). Non-MKL classifications results are listed with SLTD$_{8*8*1}$, where the image is divided into 8*8 blocks in spatial domain. The proposed system is tested on the SMIC database. The input video is in a .AVI format. The video is a short duration recording of a subject responding to the hidden emotion but that emotion is deliberately suppressed by the subject. Objective of this system is to detect facial micro-expression.

The output of this system is in a Gray scale image format showing the particular output frame of the video which has the highest peek value for micro-expression detection as shown in figure 5.

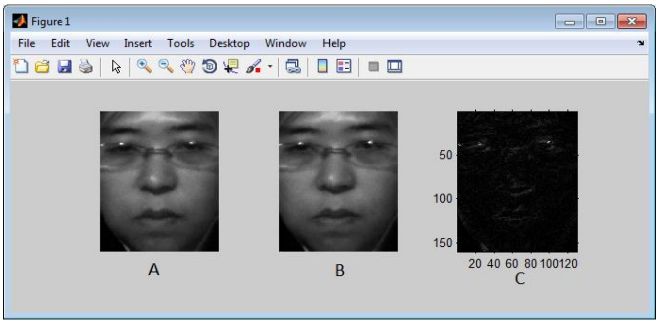


Fig 5: A) Neutral face frame, B) detected frame and C) difference between A and B.

A is the first frame in the video which is a neutral face and taken as a baseline face for applying Face to Model Face mapping through which the facial features are extracted using ASM (Active Shape Model). B is the detected frame having the micro expression. It is hard to see the difference between the two frames that's why the third figure is shown as the difference of A and B.

The graphical representation of the outputs is shown in the figure 6.

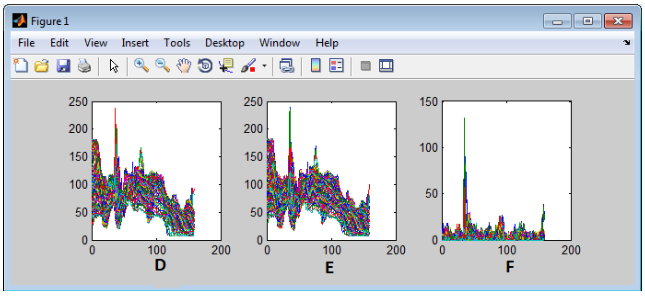


Fig 6: Graphical representation for D) Neutral face frame, E) Detected frame and F) difference between D and E.

As it is clear to see the graph, the difference between the two frames is having the highest peek value for the detected micro-expression.

REFERENCES

[1] P. Ekman, "Telling Lies" 2nd Edition, Norton, (2009).

[2] S. Porter and L. Brinke, "Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions", Psychological Science, vol. 19, pp. 508-514, (2008).

[3] P. Ekman, "Facial Expressions of Emotion: an Old Controversy and New Findings", Philosophical Transactions of the Royal Society, vol. B335, pp. 63-69, (1992).

[4] Bhall R., Sharma V., A review: detection and analysis of facial micro-expression. IJESRR, Volume-2, Issue-1, Feb 2015.

[5] S. Lucey, "Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face", Face Recognition Book. Pro Literatur Verlag, (2007).

[6] M. Pantic, "Detecting Facial Actions and their Temporal Segments in Nearly Frontal-View Face Image Sequences", Systems, Man and Cybernetics, vol. 4, pp. 3358-3362, (2005).

[7] M. Bartlett, "Automatic Recognition of Facial Actions in Spontaneous Expressions", Journal of Multimenia, vol. 1, no. 6, pp. 22-35, (2006).

[8] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar. Macro-and micro-expression spotting in long videos using spatio- temporal strain. In FG, 2011.

[9] T. Cootes, c. Taylor, d. Cooper, and j. Graham. Active shape Models – their training and application. Computer vision and Image understanding, 61(1):38–59, 1995.

[10] A. Goshtasby. Image registration by local approximation methods. IMAVIS, 6(4):255–261, 1988.

[11] G. Zhao and M. Pietik¨ainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. PAMI, 29(6):915–928, 2007.

[12] M. Varma and D. Ray. Learning the discriminative power invariance trade-off. In *ICCV*, pages 1–8, 2007.

[13] T. Pfister, X. Li, G. Zhao, and M. Pietikinen. Recognising spontaneous facial micro-expressions. IEEE International Conference on Computer Vision 2011.

[14] Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, Matti Pietikäinen. A Spontaneous Micro-expression Database: Inducement, Collection and Baselin