2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Depth Videos for the Classification of Micro-Expressions

Ankith Jain Rakesh Kumar*, Bir Bhanu*, Christopher Casey†, Sierra Grace Cheung†, Aaron Seitz†
*Department of Electrical and Computer Engineering, University of California, Riverside
† Department of Psychology, University of California Riverside
Email: {arake001@ucr.edu, bhanu@vislab.ucr.edu, ccase004@ucr.edu, scheu015@ucr.edu, aseitz@ucr.edu}

***Abstract*—Facial micro-expressions are spontaneous, subtle, involuntary muscle movements occurring briefly on the face. The spotting and recognition of these expressions are difficult due to the subtle behavior, and the time duration of these expressions is about half a second, which makes it difficult for humans to identify them. These micro-expressions have many applications in our daily life, such as in the field of online learning, game playing, lie detection, and therapy sessions. Traditionally, researchers use RGB images/videos to spot and classify these micro-expressions, which pose challenging problems, such as illumination, privacy concerns and pose variation. The use of depth videos solves these issues to some extent, as the depth videos are not susceptible to the variation in illumination. This paper describes the collection of a first RGB-D dataset for the classification of facial micro-expressions into 6 universal expressions: Anger, Happy, Sad, Fear, Disgust, and Surprise. This paper shows the comparison between the RGB and Depth videos for the classification of facial micro-expressions. Further, a comparison of results shows that depth videos alone can be used to classify facial micro-expressions correctly in a decision tree structure by using the traditional and deep learning approaches with good classification accuracy. The dataset will be released to the public in the near future.**

I. Introduction

Facial expressions are vital nonverbal gestures that convey human thoughts in our social and non-social life. The automatic facial expressions recognition has a myriad of applications such as human behavior analysis [1], medical applications [2], driver emotion recognition [3], online-learning, game-playing and human-computer interfaces. Facial expressions are classified into two categories, namely, facial macro-expressions and facial micro-expressions. Macro-expressions are the expressions that we tend to see in our daily interactions with people. Micro-expressions (MEs) are special expressions that are spontaneous, involuntary muscle movements, brief and subtle. These expressions appear spontaneously on a human face and they represent person's true emotions [4]. The time scale of MEs is short, and they can sustain only for about 0.5 seconds or less in duration [5]. As a result, these expressions are difficult to spot and recognize.

Traditionally, researchers use RGB images/videos to spot and recognize the facial macro and micro-expressions [6], [7], [8], [9]. However, in real applications, spotting and recognizing facial expressions are challenging tasks. Although current methods have demonstrated to reap appropriate results, they are still prone to some of the issues such as illumination changes, pose variation and privacy concerns. Therefore, researchers have considered facial macro-expression recognition using RGB-D data [10], [11], [12], [13] which are not susceptible to the illumination variation. The paper [13] shows that the depth information alone can be used to classify the facial macro-expressions, and can achieve equal or better performance than the RGB images. This work was one of the motivations for us to work on depth videos alone to classify facial micro-expressions. The advantage of using the depth videos over the RGB videos is that the pixel intensities in the depth are based on the distance of the face to the camera that provides new information about the facial features. Also, the personal identity cannot be obtained with ease from the depth videos that would help to resolve privacy issues whereas it is not the same when applied to RGB videos.

In this paper, we collect a new RGB-D facial micro-expressions dataset using the Intel RealSense D415 camera and classify the facial micro-expressions into 6 classes: anger, happy, sad, disgust, surprise and fear using depth videos. Our idea is to show that depth videos alone can be used to classify facial micro-expressions. We use classical and deep learning approaches to illustrate the importance of depth in the classification of facial micro-expressions. The classical method uses the histogram of oriented gradients in 3D (HOG3D) features in a video format, and these features are classified using the Support Vector Machines (SVM) and K-Nearest Neighbor (KNN) in a decision tree classification structure for both full face and part based approach (left eye, right eye, nose and mouth). Similarly, Convolutional Neural Networks (CNN) and the CNN-LSTM approaches are used to classify the videos in a similar decision tree structure in both parts based and on full face. A decision tree classification structure is used to classify the expressions in a one versus all format. To overcome the problem of data imbalance in a one versus all format, we add the motion magnified videos to a single class folder with one class to balance the number of videos from the remaining classes in a decision tree structure of classification.

The rest of this paper is organized as follows. We introduce the related works and our contributions in Section 2. In section 3, we explain the protocol for the collection of our VISME database. The technical approach for the classification of facial micro-expressions for the depth videos is given in Section 4. The experimental results are described in Section 5. Finally, Section 6 provides the conclusions and future work.

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9412976

II. Related Work and Contributions

Classification of facial expressions has been an important topic in the field of human computer interaction (HCI). In recent few years, spotting and classification of facial micro-expressions has gained a lot of attention, but very limited work has been done until now. Some of the related work in the field of facial micro-expressions are as follows: Liong *et al.* [14], proposed a technique to use only the apex frame of a video to recognize the micro-expression. The feature extractor, Bi-Weighted Oriented Optical Flow (Bi-WOOF) is used to enhance the apex frame feature. Li *et al.* [8] used a technique to detect the apex frames in the frequency domain, as the apex frame of the video has a relationship correlated with the amount of change in the amplitude in the frequency domain. Therefore, to classify the facial micro-expressions apex frames are used. Gan *et al.* [15] (OFF-ApexNet) used a divide and conquer technique to identify the apex frame. The optical flow features are extracted from the apex frame and further classified using CNN.

Liong *et al.* [16] used two sets of features: optical strain and optical strain magnitudes to classify facial micro-expressions on the two datasets CASME II and SMIC.

Liu *et al.* [17] uses a feature called Main Directional Mean Optical-flow (MDMO) feature, for spontaneous facial micro-expression recognition. The paper uses optical flow to get the textural part of images and process it using affine transformation to remove any sensitivity to lighting conditions and head movements. Further, the facial areas are divided into ROIs. They use SVM classifier to classify the micro-expressions.

Zhao *et al*. [18] used a traditional approach known as Local Binary Pattern with Three Orthogonal Planes (LBP-TOP) to extract features for classification of facial micro-expressions. LBP-TOP helps in differentiating the local texture feature information by translating a vector code into histograms on three planes (XY, XT, YT) and finally concatenating the features of three planes into a single histogram feature which is robust to illumination changes.

Davison *et al.* [6], proposed a approach known as temporal feature extractor, i.e. 3D Histogram of Oriented Gradient (3DHOG) method, which extracts features in all three directions of motion (XY, XT, YT) for classification of facial micro-expression.

Kumar *et al.* [7] used a technique to eliminate the low-intensity expression frames in the frequency domain. The low intensity frames are the video frames that have very small variation in texture. The remaining high-intensity frames are transformed into motion magnified emotion avatar image to classify the facial micro-expression using CNN.

Peng *et al.* [19] proposed a approach called Dual Temporal Scale Convolutional Neural Network (DTSCNN), which is a two-stream 3-D CNN model. To overcome the problems of different frame-rates of facial micro-expression datasets and the overfitting problem due to small data size, two streams of the framework were designed.

Khor *et al.* [20] uses a spatio-temporal CNN-LSTM approach to classify the facial micro-expressions. The spatial dimension enrichment is done by channel stacking and the temporal dimension by deep feature stacking. The input to the CNN is the optical flow X, optical flow Y, optical flow magnitude and optical strain (normal strain and shear strain) feature images.

For a recent review on classification of facial micro-expression, we recommend the survey by Merghani *et al.* [21].

In recent years, researchers have considered using RGB-D information in classifying facial macro-expressions [10], [11], [12], [13] [22]. Shao *et al.* [22] used RGB-D data and extracts LBP-TOP features for each part of the face. Later, they initialized the codebooks via K-means clustering and merged all the features using spatial-pyramid pooling. Finally, Conditional Random Field (CRFs) are used to classify the facial macro-expressions.

Uddin *et al.* [12] uses RGB-D images to classify facial macro-expressions by extracting Local Directional Position Pattern (LDPP) features from the depth images. Further, Principal Component Analysis (PCA) is applied for dimensionality reduction. Furthermore, the face features are classified by Global Discriminant Analysis (GDA) to make them more robust. Finally, the features are applied to train a Deep Belief Network (DBN) to classify facial macro-expressions.

Aly *et al.* [11] collected a facial macro-expression dataset using kinect sensor. The LBP features were extracted from the images and then fusion of RGB-D features takes place to classify the facial expressions.

A. Existing Datasets

Micro-expressions are subtle, rapid, and are difficult for humans to spot and recognize. It usually requires special training for humans to be able to recognize these expressions. In recent years, research conducted in this field has resulted in many approaches for extracting features from micro-expressions. However, there are very few publicly available facial micro-expression datasets as shown in Table I. Existing micro-expression datasets are limited in the number of videos and ethnicities. Therefore, to overcome all the challenges in terms of ethnicity, broader participants, and single modality of data (RGB/gray-scale), we collect our dataset which has the RGB and depth data for the classification of facial micro-expressions.

Contributions. The contributions of this paper are:

- Collection of a new RGB-D video based facial micro-expression dataset, called VISME.
- We develop the baseline approaches for the VISME dataset for the classification of facial micro-expressions for the depth videos.
  - An automatic spatio-temporal feature extraction method is used to extract the features such as HOG3D in a decision tree structure for the classification of facial micro-expressions

  - Classification of facial micro-expressions using the CNN approach in a decision tree structure.
  - Classification of facial micro-expressions using the CNN-LSTM approach in a decision tree structure.
- A comprehensive evaluation is performed for the classification of facial micro-expression for RGB and Depth videos using the CNN-LSTM method.

III. Data Collection

The experiments for data collection were conducted in a laboratory condition. The interaction with the participants was kept minimum during data collection. The IRB release document was given to participants at the beginning of the experiment, and a set of instructions were provided to them. They were asked to read the release document and sign the agreement if they were willing to participate in the experiment. They were taken to their respective computer desks where the experiments were conducted.

We selected 30 video clips from various Hollywood movies. To determine which videos would give us a better stimulus, we ran a pilot study on 10 participants to get a fair idea about the videos to be selected, ground truth for the movie clip, and length of the experiment to be conducted, so that participants are not bored. These participants' video data was not included in the dataset. Finally, out of 30 videos, the experiment comprised of 19 video clips that attempt to elicit emotion from the participants.

The participants were given a questionnaire before starting the experiment. The questionnaires were simple such as age, ethnicity, and how active they have been before the start of the experiment, and their sleep level. Also, the participants were given a single questionnaire after each video they watched, what emotion they felt during the video. This allowed us to give a short break to the participants before they start a new video. The participants were also given final questionnaire after the completion of the experiment to determine their alertness level, sleep level in the range of 0 to 7 and how much did they enjoy the experiment.

The participants were allowed to take a break in between if they wanted to use the restroom or if they were not interested in continuing the experiment. The participants were given a full information on how the data would be used and it was well explained to them. A few participants were compensated with cash and others were research assistants and received academic credit.

A. Camera

The RGB-D camera used for the experiment is Intel RealSense D415 with RGB and Depth sensor, set to record at 30fps. The resolution was set to 1280x720.

B. Selection of expressions and labeling

Three coders were used in the analysis and selection of facial micro-expression to increase the accuracy of the dataset. We processed the raw depth videos and used the following steps for the labeling of the facial micro-expressions.

- Removing the unwanted facial movements such as head movements, eye blinking, swallowing saliva, and many similar movements which are irrelevant for the facial micro-expressions selection were removed to have a better and reliable datasets.
- The micro-expression videos were selected and saved based on the subtle behavior and the time frame of the video to be less than 1 second. The onset and offset frames are coded on the videos.
- The videos were converted into a sequence of frames and marked as onset and offset frames for precise labeling of the dataset. The videos were further evaluated and selected into the final dataset if the time frame of the video was less than 0.6 seconds and the process was repeated for other videos. The emotion labels were given to the videos by the three coders to improve the reliability of the ground-truth of the dataset. The majority voting of the coders was considered to be the label for the video to be included in the final dataset.

C. Dataset

The dataset has a total of 21 participants. Few examples of VISME database are shown in Fig. 1, where the RGB images and their corresponding depth images are shown for all 6 expressions such as angry, fear, sad, happy, disgust and surprise. A wide variety of participants were recruited to have a diversification of the emotional responses. The participants recruited for the experiments were the students from the university with a mean age of 22.3. The gender split of the participants are 13 female and 8 male participants and ethnicities of the participants are from different parts of the world such as Chinese (4), Indian (2), African American (2), Caucasians (3), Vietnamese (3), Hispanic/Latinos (7). The total number of videos used for this paper are 238 videos divided into 6 categories: Anger (60), Happy (55), Surprise (52), Sad (22), Disgust (26) and Fear (23) videos. The dataset will be made available to the public.

IV. Technical Approach

In this section, we present our approach for classifying the facial micro-expressions on both the full face and part-based face regions for depth video samples in a decision tree structure. We use a state-of-the-art approach and deep learning approaches as shown in Fig. 2. The key aspects of the approach are: (1) HOG3D represents the state-of-the-art approach, (2) CNN based approach and (3) CNN-LSTM based approach. All these approaches are applied on both full face and part based facial parts. The approach comprises three steps: preprocessing, feature extraction, and classification of facial micro-expression into 6 expressions (anger, happy, sad, fear, surprise, and disgust).

A. Preprocessing

For the raw depth videos, the first step is to obtain the face region from the entire scene of the video. The processing steps are: Firstly, remove the background and other irrelevant

TABLE I: Existing Facial Micro-Expressions datasets

Datasets	Subjects	Expressions	Videos	Resolution	fps	Mean Age Group	Activities	Data Format	Duration of Video (sec)	Comments
Polikovsky [23]	10	6	42	640x480	200	N/A	N/A	Grayscale	N/A	Posed Expressions, Very few videos, no FACS coding and No pose variations
USF-HD [24]	N/A	4	100	1280x720	30	N/A	N/A	RGB	0.66	Posed, only 4 expressions considered, no FACS coding and No pose variations
SMIC [25]	16	3	164	640x480	100	26.7	Movies, YouTube	RGB	0.5	Spontaneous, no FACS coding, and No pose variations
CASME II [26]	35	5	247	640x480	200	22	Movie, YouTube	RGB	0.5	Spontaneous, only 5 expressions and No pose variations
SAMM [9]	32	7	159	2040x1088	200	33.24	YouTube	Grayscale	0.5	Spontaneous and No pose variations
VISME	21	6	238	1280x720	30	22.3	Movie Clips	RGB +Depth	0.6	Spontaneous, Depth Videos included and Various poses of the face

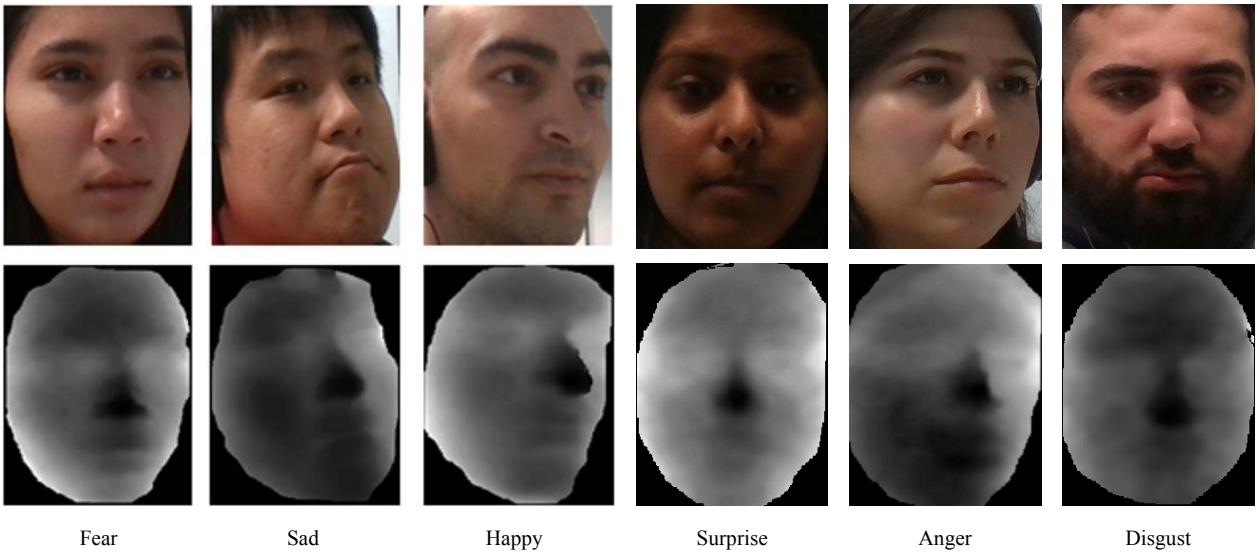

Fig. 1: Examples of VISME database with RGB-D images consisting of 6 expressions such as Anger, Happy, Sad, Disgust, Fear and Surprise. The first row consists of RGB images and second row belongs to their respective depth images

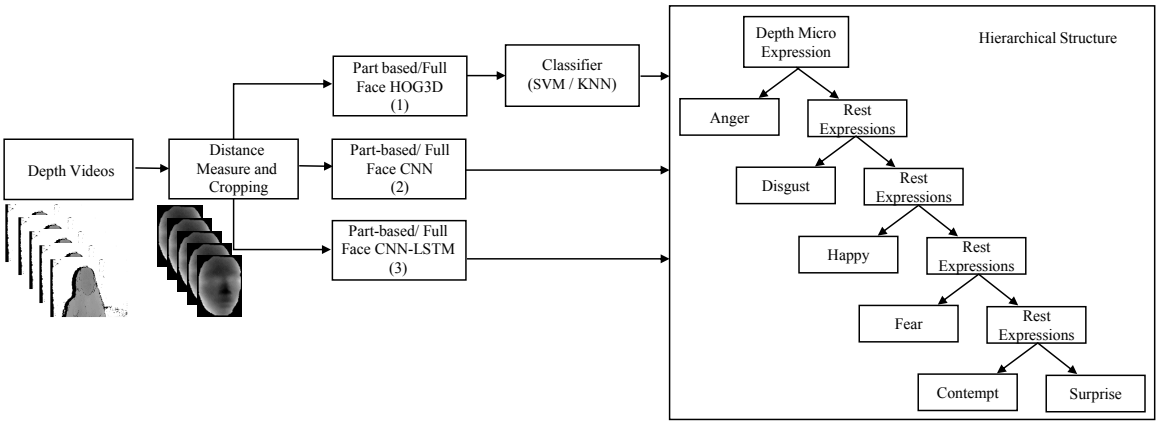

Fig. 2: Overall Architecture of our approach.

regions based on the distance from the camera. Secondly, manually crop the regions of the face to have an accurate feature extraction. The cropped face region is then normalized between 0 to 1 based on the distance from the camera.

For the classification of facial micro-expressions using the part-based approach, we crop the face regions manually based on the Action Units (AUs) into four parts: left eye, right eye, mouth, and nose region in a depth video for accurate extraction of these regions from the image. These regions help in obtaining precise spatio-temporal features for the classification.

B. Balancing the Unbalanced dataset

As the dataset has imbalance, we try to balance the dataset by using the Eulerian motion magnification. Here, we use different amplification factor (α) value to balance a particular class of the dataset.

1) Selection of Amplification factor for Eulerian Motion Magnification: Eulerian motion magnification [27] amplifies the small variations in videos by integrating spatial and temporal processing to focus attention on the subtle facial micro-expression region in a video. These magnified videos are used to solve two purposes: firstly, the micro-expressions are subtle, and magnifying helps in amplifying the signals and makes it easier for us to recognize the expressions. Secondly, using different magnitudes of motion magnified videos and augmenting these samples in the training set helps in solving the class imbalance problem of the dataset.

The selection of the different magnitudes of motion magnification depends on the amplification factor (α) value. Higher the value of α, higher is the artifacts in the video due to the distortion and amplification of noise. To select the best amplification factors for our experiment, we magnify the videos with α *value from 2 to 10*. The alpha value above 10 adds significant noise and these noisy images were difficult to classify. Therefore, we chose a preset value from 2 to 10. Next, we select the center frame from each video and calculate the Peak-Signal-to-Noise-Ratio (PSNR ratio) using Eq. (1) with the respective original video frame for each expression separately.

$$PSNR = 10\log_{10}(\frac{MAX_i^2}{MSE}) \tag{1}$$

where, MAX_i is the maximum possible pixel value of an image, MSE is the mean squared error.

We take the average value of the PSNR ratio for each expression from these videos. Furthermore, for the selection of α value for our experiment, we see if the *PSNR ratio* $\geq$ *average value of the PSNR ratio for each expression*, we use those amplified videos in our experiment.

Fig. 3 shows the plot for the Peak-Signal-to-Noise-Ratio vs the Amplification factor α. We choose the amplified videos *if the PSNR ratio* $\geq$ *average value of the PSNR ratio for each expression*. From Fig. 3 we can conclude that the PSNR ratio of the videos decreases as the amplification factor is increased, which suggests as the value of α increases the levels

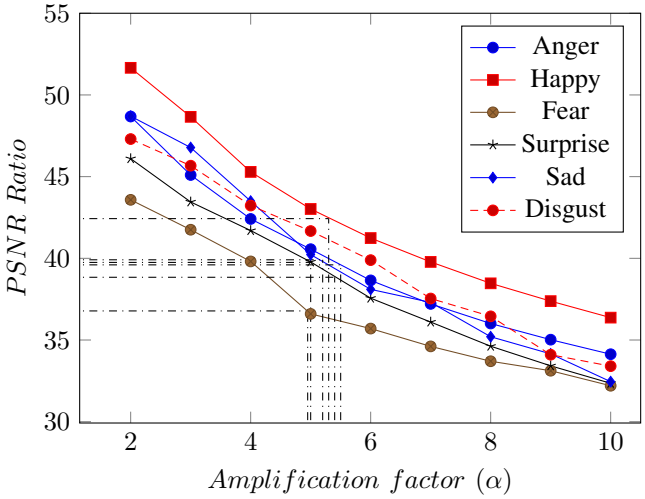

Fig. 3: Amplification factor (α) vs PSNR Ratio for video motion magnification. Here, the dash dotted lines represents the average PSNR ratio for each expression.

of artifacts being added into the video also increases. From Fig. 3, we can conclude that the motion magnified videos with the amplification factor $\alpha < 6$ were used in our experiment depending on the class of expression used.

C. Histogram of Oriented Gradients 3D

Each video in the dataset has a minimum of 8 frames and a maximum of 18 frames. HOG3D [28] uses three planes XY, XT, and YT to extract the spatio-temporal features of a video, and the pixel orientation and magnitude are calculated for these planes. A sliding window with a length of 8 frames is used to generate a subsequences, and these subsequences are considered to be a single video for the classification purpose. Thus, each video consists of 8 frames.

For the part-based approach, each region is uniformly resized to 64x64 patch for the subsequent feature extraction. For a given depth video with 8 consecutive frames, the features extracted for 64x64x8 sequence of patch are further divided into 8x8x1 cuboid cell, as a result each cuboid cell has a dimension of 8x8x8. Therefore, for a given video sequence, every cuboid cell with c(x,y,t), 3D gradients along x,y,t-directions are denoted by partial derivatives w.r.t to c(x,y,t) = $[\frac{\partial c}{\partial x}, \frac{\partial c}{\partial y}, \frac{\partial c}{\partial t}]^T$, and the respective mean gradient is denoted by $\bar{g}_c = [c_x\prime, c_y\prime, c_t\prime]^T$. Calculating the histogram of 3D oriented gradients, a 3-dimensional Euclidean coordinate is placed at the center of the origin. The mean gradient $\bar{g}_c$ passes through the origin and the center points of all n faces of polyhedron.

We choose a 20 sided polyhedron known as icosahedron same as used by Kläser *et al* [28]. The icosahedron has 20 regular triangle faces. In this structure, each pair of opposite face of icosahedron correlate with one histogram bin since they are along the same axis. Therefore, each region of depth patch consists of 10-bin histogram. The HOG3D features can be obtained for each patch by concatenating all the patches for each depth region i.e. 64 cells.

TABLE II: Parameters for the Network

Network	Learning rate	Momentum	Weight Decay	Optimizer
Resnet 34	10^{-3}	0.9	5 x 10^{-4}	SGD

D. Deep Learning Features

1) CNN Based Approach: In our approach, we use state-of-the-art Convolutional Neural Network, to perform feature extraction. We employ the Resnet 34 architecture [29]. The Resnet architecture takes an input size of 224x224 and batch normalization is applied before the use of each convolution layer for faster convergence during the training process. Rectified Linear Unit (ReLu) activation is used after each convolutional layer. Table. II mentions the hyperparameters used for the training the CNN.

For the part-based approach, we use 4 CNN networks to classify the depth videos. We concatenate the output from the last layer and then use the fully-connected layer to classify the micro-expressions in a decision tree structure approach.

For the classification of expressions using CNN for the video, we take the maximum vote from the decision made for the sequence of 8 frames and then classify them into their respective class of expression.

2) CNN-LSTM Based Approach: The input to the LSTM network is the feature vector from the CNN network. Here, the CNN network is the same as mentioned in the section (IV-D1). In our experiment, we use a sliding window of 8 sequence of frames. We chose 8 frames as a sliding window length as this would capture the dynamic changes of the facial micro-expressions. The sequences of a video are chosen with a stride of 1, for example, if the first sequence is v_0 = $\{s_1$, s_2, s_3, s_4, ..., $s_8\}$, the second sequence will be v_0 = $\{s_2$, s_3, s_4, s_5, ..., $s_9\}$.

The CNN-LSTM network uses 4 CNN network for the part-bassed approach and 1 CNN for the full face approach. The network is optimized using the SGD algorithm. The learning rate of the network is 1x 10^{-3} with NLLLoss function with LogSoftmax.

E. Decision Tree Structure for Classification

The six expressions used in this work are anger, happy, sad, fear, disgust and surprise are not mutually exclusive. These expressions are often confused by the humans, such as Anger (Au : 4, 5, 7 and 23) and Fear (AU : 1, 2, 4, 5, 7, 20, and 26) have same Action Units (AUs), and surprise (Au: 1, 2, 5 and 26) and fear (AU : 1, 2, 4, 5, 7, 20, and 26) can also be confused based on the AUs. In order to maximize the classification accuracy we use decision tree structure i.e. one vs all the other expressions. The following steps are use to repeat the experiment for all the expressions until we get the best results in a set. Remove the expression which achieved the best result and continue the procedure until we are left with only two classes of expression at the end. No pre-defined rules are used in this approach.

For the HOG3D features, we use SVM and KNN approach to classify the facial micro-expressions. For the deep learning approaches , we use CNN and CNN-LSTM to classify the micro-expressions.

V. Experimental Results

A. Experimental Setup

VISME database is the first micro-expression database which has the depth videos. It involves 21 subjects (8 male and 13 female) of various ethnicities. The total number of videos used in this paper are 238 videos divided into 6 categories: Anger (60), Happy (55), Surprise (52), Sad (22), Disgust (26) and Fear (23) videos.

We use 16 subjects for the training and 5 subjects for the testing for the 5 fold cross-validation . We experiment with two settings: one with full face and the other with part-based approach where only part of the face are available. To increase the number of sequences for classification in the database, we use 8 frames in a sliding window manner to represent it as a video. Also, we use motion magnified videos for the training and testing purposes, to overcome the class imbalance problem during the decision tree structure for the classification of data. The class imbalance problem is overcome here in this paper by using different factors of motion magnification to increase the number of videos for the particular class of the dataset.

B. Results

The results for the classification of facial micro-expressions from the depth video using HOG3D features, CNN and CNN-LSTM are shown in Table III. Also, the decision tree/ hierarchical structure for the classification of facial micro-expressions from the depth videos using CNN and CNN-LSTM method for the part-based approach is shown in Fig. 4, 5. The results from the Table III shows for both full face and part-based approach for the depth videos. The results are obtained after balancing the dataset before applying any approach. For the HOG3D features, we had used both SVM and KNN classifier to determine the best classifier for the full Face and part-based approach for the classification, but the KNN classifier gave us the best results compared to the SVM classifier.

From the Table. III, we see that the part-based approach gives us the better results when compared to the full face in both HOG3D and CNN classifier. However, for the CNN-LSTM approach full face approach gives better results when compared to the part-based approach. We can also notice from Table III, that the happy expression can be distinguished easily from the rest of the expressions, as the Action Units (AUs) that contribute to the happy expressions (AU6 - Cheek Raiser and AU12 - Lip Corner Puller) are different than the rest of the other expressions AUs. The sad expression has muscle movements similar to happy, but not all sad expressions have the same Action Units. Thus, the happy expression can be recognized prominently and it is easier than other expressions at the top of the tree. The results from all the approaches helps in evaluating our new VISME database only using depth

TABLE III: Overall Results for the 5 Fold Cross-Validation for HOG3D, CNN and CNN-LSTM approaches for both part-based and full face. MA : Mean Accuracy (S.D.) and MFAR : Mean False Alarm Rate (S.D.)

Method	Happy		Surprise		Disgust		Anger		Fear		Sad		Overall	
	MA	MFAR	MA	MFAR	MA	MFAR	MA	MFAR	MA	MFAR	MA	MFAR	MA	MFAR
HOG3D (Full Face)	0.6211 (0.0211)	0.2073 (0.0110)	0.6366 (0.0216)	0.2566 (0.0150)	0.6978 (0.0241)	0.1000 (0.0143)	0.6650 (0.0147)	0.2957 (0.0267)	0.7342 (0.0402)	0.2529 (0.0274)	0.8714 (0.0119)	0.2813 (0.0207)	0.7044 (0.0223)	0.2323 (0.0192)
HOG3D (Part-Based)	0.6400 (0.0232)	0.2026 (0.0159)	0.6595 (0.0113)	0.2270 (0.0142)	0.6825 (0.0194)	0.2653 (0.0122)	0.6917 (0.0184)	0.2731 (0.0247)	0.8480 (0.0310)	0.0984 (0.0273)	0.8914 (0.0296)	0.1347 (0.0263)	0.7355 (0.0221)	0.2002 (0.0201)
CNN (Full Face)	0.6569 (0.0273)	0.1219 (0.0107)	0.6973 (0.0161)	0.1812 (0.0180)	0.6667 (0.0159)	0.1745 (0.0157)	0.7476 (0.0359)	0.2236 (0.0198)	0.9413 (0.0218)	0.2898 (0.0326)	0.7371 (0.0163)	0.0780 (0.0466)	0.7412 (0.0222)	0.1782 (0.0239)
CNN (Part-Based)	0.6442 (0.0190)	0.1112 (0.0138)	0.7000 (0.0123)	0.1497 (0.0103)	0.7021 (0.0288)	0.2429 (0.0193)	0.7389 (0.0174)	0.2702 (0.0167)	0.9461 (0.0267)	0.2631 (0.0283)	0.7400 (0.0156)	0.0742 (0.0437)	0.7452 (0.0200)	0.1852 (0.0220)
CNN-LSTM (Full Face)	0.6800 (0.0253)	0.1069 (0.0193)	0.7108 (0.0130)	0.1170 (0.0143)	0.7206 (0.0181)	0.1359 (0.0174)	0.7452 (0.0201)	0.1342 (0.0208)	0.8902 (0.0340)	0.1287 (0.0297)	0.8743 (0.0120)	0.1347 (0.0412)	0.7702 (0.0204)	0.1262 (0.0238)
CNN-LSTM (Part-Based)	0.6695 (0.0147)	0.1009 (0.0213)	0.7014 (0.0088)	0.1082 (0.0177)	0.7275 (0.0235)	0.2020 (0.0227)	0.7159 (0.0124)	0.1217 (0.0192)	0.8671 (0.0457)	0.1287 (0.0273)	0.8771 (0.0128)	0.1382 (0.0457)	0.7598 (0.0196)	0.1333 (0.0257)

TABLE IV: Overall Results for the 5 Fold Cross-Validation for CNN-LSTM approach for both RGB and Depth videos for part-based approach. MA : Mean Accuracy (S.D.) and MFAR : Mean False Alarm Rate (S.D.)

Data Format	Happy		Surprise		Disgust		Anger		Fear		Sad		Overall	
	MA	MFAR	MA	MFAR	MA	MFAR	MA	MFAR	MA	MFAR	MA	MFAR	MA	MFAR
Depth	0.6695 (0.0147)	0.1069 (0.0213)	0.7014 (0.0088)	0.1082 (0.0177)	0.7275 (0.0235)	0.2020 (0.0227)	0.7159 (0.0124)	0.1217 (0.0192)	0.8671 (0.0457)	0.1287 (0.0273)	0.8771 (0.0128)	0.1382 (0.0457)	0.7598 (0.0196)	0.1333 (0.0257)
RGB	0.6823 (0.0165)	0.1273 (0.0190)	0.7179 (0.0151)	0.1224 (0.0167)	0.7321 (0.0120)	0.2189 (0.0251)	0.7592 (0.0297)	0.1472 (0.0232)	0.8750 (0.0427)	0.1168 (0.0373)	0.8832 (0.0257)	0.1250 (0.0347)	0.7750 (0.0236)	0.1429 (0.0260)

videos, suggest that only depth videos can be used to classify the facial micro-expressions into 6 expressions (anger, happy, surprise, disgust, fear and sad).

We perform 5 fold cross-validation approach on the VISME dataset. The results of the 5-fold cross-validation are shown in the Table. III for all the three methods: HOG3D, CNN, CNN-LSTM.

Based on the results from the Table. III, CNN-LSTM results are clearly higher than the other methods. Therefore, we compare the results of CNN-LSTM method for the RGB and Depth videos. The comparison results for the RGB and Depth videos is shown in Table IV. From Table IV, we can see that the results for RGB and Depth videos vary by 1.52% in accuracy. Therefore, depth videos alone can be used to classify facial micro-expressions.

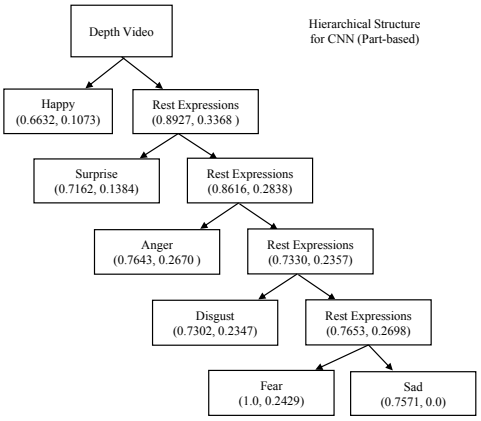

Fig. 4: Decision tree /Hierarchical structure for classification of facial micro-expression using CNN feature for the part-based face regions

VI. Conclusions and Future Work

In this paper, we collected a new depth based video database called VISME for facial micro-expressions, and to the best of our knowledge, there is no other depth based facial micro-expression dataset. We used both classical and deep learning approaches to get the baseline results on the VISME dataset using the 5-fold cross-validation. The methods used are: a histogram of oriented gradients 3D (HOG3D), a convolutional neural network (CNN), and the CNN-LSTM methods to obtain the spatio-temporal information in the depth videos to classify the micro-expressions in a decision tree structure for the full and the part-based approach. The depth features help in accurately classifying the facial micro-expressions. The overall mean accuracy (MA) and mean false alarm rate (MFAR) results are better for the part-based approach using HOG3D when compared to the full face approach by 3.11% (MA) and 3.21% (MFAR). Similarly, for the CNN, the part-based

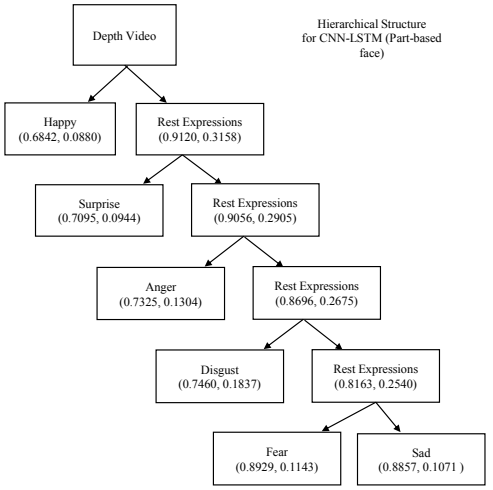

Fig. 5: Decision Tree/Hierarchical structure for classification of facial micro-expression using CNN-LSTM feature for the part-based face regions

approach results are 0.40% (MA) better than the full face approach, but the MFAR results for the full-face approach are better by 0.7%. However, in the CNN-LSTM approach, full face results are better by 1.04% (MA) and MFAR by 0.71%. We also compare depth video results with the RGB videos. The overall MA for the RGB videos is 1.52% higher than the depth videos, but the MFAR result for the depth video is 0.96% better than RGB videos. Therefore, we can use depth videos alone for the classification of facial micro-expressions. In the future, we plan to add more data to our VISME dataset and carry out the fusion of RGB and the depth videos to further improve the performance and release the dataset publicly.

VII. Acknowledgements

This material is based upon work supported by the National Science Foundation under grant number 1911197.

References

[1] W. J. Yan, Q. Wu, Y-J. Liu, S-J Wang, and X. Fu, "Casme database: A dataset of spontaneous micro-expressions collected from neutralized faces," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, April 2013, pp. 1–7.

[2] V. Bevilacqua, D. D'Ambruoso, G. Mandolino, and M. Suma, "A new tool to support diagnosis of neurological disorders by means of facial expressions," in *IEEE International Symposium on Medical Measurements and Applications*, May 2011, pp. 544–549.

[3] R. Theagarajan, B. Bhanu, and A. Cruz, "Deepdriver: Automated system for measuring valence and arousal in car driver videos," in *24th International Conference on Pattern Recognition (ICPR)*, Aug 2018, pp. 2546–2551.

[4] "What are micro-expressions?," *https://www.paulekman.com/resources/micro-expressions/*.

[5] C. M. Hurley, A. E. Anker, M.G. Frank, D. Matsumoto, and H. C. Hwang, "Background factors predicting accuracy and improvement in micro expression recognition," in *Motivation and Emotion*, 2014, p. 700–714.

[6] A. K. Davison, W. Merghani, and M. H. Yap, "Objective classes for micro-facial expression recognition," *CoRR*, vol. abs/1708.07549, 2017.

[7] A. J. R. Kumar, R. Theagarajan, O. Peraza, and B. Bhanu, "Classification of facial micro-expressions using motion magnified emotion avatar images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[8] Y. Li, X. Huang, and G. Zhao, "Can micro-expression be recognized based on single apex frame?," in *25th IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 3094–3098.

[9] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "Samm: A spontaneous micro-facial movement dataset," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 116–129, Jan 2018.

[10] X. Li, Q. Ruan, and Y. Ming, "3d facial expression recognition based on basic geometric features," in *IEEE 10th International Conference on Signal Processing Proceedings*, Oct 2010, pp. 1366–1369.

[11] S. Aly, A. Trubanova, L. Abbott, S. White, and A. Youssef, "Vt-kfer: A kinect-based rgbd+time dataset for spontaneous and non-spontaneous facial expression recognition," in *2015 International Conference on Biometrics (ICB)*, May 2015, pp. 90–97.

[12] M. Z. Uddin, M. M. Hassan, A. Almogren, A. Alamri, M. Alrubaian, and G. Fortino, "Facial expression recognition utilizing local direction-based robust features and deep belief network," *IEEE Access*, vol. 5, pp. 4525–4536, 2017.

[13] M. Z. Uddin, W. Khaksar, and J. Torresen, "Facial expression recognition using salient features and convolutional neural network," *IEEE Access*, vol. 5, pp. 26146–26161, 2017.

[14] S.T. Liong, J. See, K. Wong, and R. C. W. Phan, "Less is more: Micro-expression recognition from video using apex frame," *Signal Processing: Image Communication*, vol. 62, pp. 82 – 92, 2018.

[15] Y.S. Gan, S. T. Liong, W. C. Yau, Y. C. Huang, and L. K. Tan, "OFF-ApexNet on micro-expression recognition system," *Signal Processing: Image Communication*, vol. 74, pp. 129 – 139, 2019.

[16] S.T. Liong, J. See, R. C.W. Phan, Y.H. Oh, A. C. Le Ngo, K. Wong, and S.W. Tan, "Spontaneous subtle expression detection and recognition based on facial strain," *CoRR*, vol. abs/1606.02792, 2016.

[17] Y. Liu, J. Zhang, W. Yan, S. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 299–310, 2016.

[18] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, June 2007.

[19] M. Peng, C. Wang, T. Chen, G. Liu, and X. Fu, "Dual temporal scale convolutional neural network for micro-expression recognition," *Frontiers in Psychology*, vol. 8, pp. 1745, 2017.

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

[21] W. Merghani, A. K. Davison, and M. H. Yap, "A review on facial micro-expressions analysis: Datasets, features and metrics," *CoRR*, vol. abs/1805.02397, 2018.

[22] J. Shao, I. Gori, S. Wan, and J.K. Aggarwal, "3d dynamic facial expression recognition using low-resolution videos," *Pattern Recognition Letters*, vol. 65, pp. 157 – 162, 2015.

[23] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor," in *3rd International Conference on Imaging for Crime Detection and Prevention (ICDP 2009)*, Dec 2009, pp. 1–6.

[24] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar, "Macro- and micro-expression spotting in long videos using spatio-temporal strain," in *Face and Gesture 2011*, March 2011, pp. 51–56.

[25] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, April 2013, pp. 1–6.

[26] W. J. Yan, X. Li, S. J. Wang, G. Zhao, Y. J. Liu, Y. H. Chen, and X. Fu, "Casme ii: An improved spontaneous micro-expression database and the baseline evaluation," *PLOS ONE*, vol. 9, no. 1, pp. 1–8, 01 2014.

[27] H. Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. T. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Transactions on Graphics (Proc. SIGGRAPH 2012)*, vol. 31, no. 4, 2012.

[28] A. Kläser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," 09 2008.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.