RESEARCH ARTICLE OPEN ACCESS

Recognition of Facial Emotions Based on Sparse Coding

A. Sunitha[1], P. Ajay Kumar Reddy[2], S.Nanda Kishore[3], G.N Kodanda Ramaiah[4]
[1]*pg Scholar, Dept. Of Ece, Kuppam Engineering College, Kuppam, Chittoor, Ap*
[2]*research Scholar, Dept. Of Ece, Kuppam Engineering College, Kuppam, Chittoor, Ap*
[3]*associate Professor, Dept. Of Ece, Kuppam Engineering College, Kuppam, Chittoor, Ap*
[4]*professor& Head, Dept. Of Ece, Kuppam Engineering College, Kuppam, Chittoor, Ap*

Abstract: This paper deals with acknowledgment of characteristic feelings from human countenances is a fascinating subject with an extensive variety of potential applications like human-PC communication, robotized mentoring frameworks, picture and video recovery, brilliant situations, what's more, driver cautioning frameworks. Generally, facial feeling acknowledgment frameworks have been assessed on lab controlled information, which is not illustrative of the earth confronted in genuine applications. To vigorously perceive facial feelings in genuine regular circumstances, this paper proposes a methodology called Extreme Sparse Learning (ESL), which can mutually take in a word reference (set of premise) and a non-direct grouping model. The proposed approach consolidates the discriminative force of Extreme Learning Machine (ELM) with the reproduction property of meager representation to empower exact arrangement when given uproarious signs and blemished information recorded in common settings. Moreover, this work exhibits another neighborhood spatio-worldly descriptor that is particular what's more, posture invariant. The proposed structure can accomplish best in class acknowledgment precision on both acted what's more, unconstrained facial feeling databases.

Index Terms: Emotion recognition, Facial emotion, Pose-invariance, Dictionary learning, Sparse representation, Extreme learning machine, Extremesparselearning.

I. INTRODUCTION

Recent advances in technology have enabled human users to interact with computers in more efficient ways such as voice and gesture. However, one essential factor for natural interaction is still missing, and that is the emotion. Emotion plays an important role in human communication and interaction; allowing people to express themselves beyond the verbal domain. Changes of person’s affective state usually emphasise the transmission of a message in human-human interaction. People are able to sense these changes and used them to improve their communication. This fact has motivated a huge number of researches to enable machines to recognize human emotions. There are several applications of human affective computing to facilitate human-computer interaction [1].

For example, a computer may become a more effective tutor if it can sense the user’s affective state. On the other hand, emotional behaviour of the user can be used as feedback on their teaching and how well the students understand them. Another application is to warn drivers by monitoring their emotional condition. For instance, in Japan, there are cars equipped with a camerainstalled in the dashboard to detect whether the driver is angry, sleepy and generally whether he/she is in dangerous emotional situation or not. Another application example is to use computer agents that could understand the user’s preference through the affect-sensitive indicator to support advertisements that would target only things that the specific audience has shown interest in and not any generic product to any audience. Essentially, audio and visual information are considered as the most important cues to assess human affective behaviour. However, the complementary relations of these two cues lead researchers to integrate audio and visual data for better performance. Physiological measurement of emotional state is also considered a reliable representation of human innerfeeling. In this research, we have focused on facial-based emotion recognition as a key part of the way that humans communicate with each other.

For the representation part, we propose a novel spatio-temporal descriptor based on Optical Flow (OF) components, which is very distinctive and also pose-invariant. For the recognition part, the idea of sparse representation is combined with Extreme Learning Machine (ELM) to learn an efficient classifier that can handle noisy and imperfect data. The main objective of the present work is to develop a facial-based emotion recognition system that is able to handle variations in facial pose, illumination and partial occlusion. In other words, the system aims to robustly represent and recognize the facial expressions in real-life situations.

The rest of this paper is organized as follows. Section II first reviews existing method. Our proposed method is described in Section III. Then experimental results are reported in Section IV to demonstrate the superior performance of our framework. Finally, conclusions are presented in Section V.

II. EXISTING METHOD

The majority of existing researches in this field have focused on facial expression processing via static image data and ignored the temporal information of such a dynamic event. However, the human visual system is demonstrated to have better judgment about an expression when its temporal information is taken into account [2]. Following this fact, some techniques have been developed to deal with dynamic expression recognition. In the case of an image sequence (gesture-oriented approaches), the problem is classified as tracking the face and its features.The temporal information is considered in dynamic processing systems either in the feature extraction part or during the classification stage. Typical examples of such techniques are Hidden Markov Models (HMM) [3], Dynamic Baysian Networks (DBN) [4], dynamic texture descriptors [9] , and geometrical displacement.

There were several reported attempts to track the facial expression over time for emotion recognition via Hidden Markov Models (HMM). A multilevel HMM is introduced by Cohen et al. [5] to automatically segment the video and perform emotion recognition. Their experimental results indicated that multilevel HMM have better performance than the one layered HMM. Cohen et al. introduced a new architecture of HMMs for automatic segmentation and recognition of human facial expression from live videos .

Dynamic Bayesian Networks (DBN) is another successful method for sequence-based expression analysis. Ko and Sim [6] developed a facial expression recognition system based on combining the Active Appearance Model (AAM) for feature extraction and DBN for modelling and analysing the temporal phase of an expression. They claimed that their proposed approach is able to achieve robust categorization of missing and uncertain data and temporal evolution of the image sequences. Optical Flow (OF) is also a widely used approach for facial features tracking and dynamic expression recognition.

Cohn et al. [7] developed an OF based approach to automatically discriminate the subtle changes in facial expression. They considered sensitivity to subtle motion when designing the OF which is crucial for spontaneous emotion detection. Tariq et al. [8] used an ensemble of features including both appearance and motion features for FAUs detection. Their proposed OF based motion features were extracted for seven regions of interest by computing the mean and variance of the OF components for each region. The head motion was also captured at this work from the nose region OF.

Zhao and Pietikainen [9] presented a successful dynamic texture descriptor based on the Local Binary Pattern (LBP) operator and applied it on facial expression recognition as a specific dynamic event. Their proposed dynamic LBP descriptors were calculated on Three Orthogonal Planes (TOP) of the video volume, resulting in LBP-TOP descriptor. Local processing, simple computation and robustness to monogenic gray-scale changes are the advantages of their method. Following their idea, Almaev and Valster [10] developed the LGBP descriptor to spatio-temporal volumes to combine spatial and dynamic texture analysis of facial expressions. Similarly, Bishan et al. proposed an extension of Local Phase Quantization (LPQ) to a dynamic texture descriptor for AU detection. All these work concluded that such kind of dynamic appearance descriptors outperform the static ones.

III. PROPOSED MITIGATION SCHEME

To recognize the emotions in the presence of self-occlusion and illumination variations, we combine the idea of sparse representation with Extreme Learning Machine (ELM) to learn a powerful classifier that can handle noisy and imperfect data. Sparse representation is a powerful tool for reconstruction, representation, and compression of high dimensional noisy data (such as images/videos and features derived from them) due to its ability to uncover important information about signals from the base elements or dictionary atoms. While the sparse representation approach has the ability to enhance noisy data using a dictionary learned from clean data, it is not sufficient because our end goal is to correctly recognize the facial emotion. In a sparse-representation-based classification task, the desired dictionary should have both representational ability and discriminative power.

Since separating the classifier training from dictionary learning may cause the learned dictionary to be sub-optimal for the classification task, we propose to jointly learn a dictionary (which may not be necessarily over-complete) and a classification model. To the best of our knowledge, this is the first attempt in the literature to simultaneously learn the sparse representation of the signal and train a non-linear classifier based on sparse codes.

The key contributions of this paper are as follows:

1. A pose-invariant OF-based spatio-temporal descriptor, which is able to robustly represent

facial emotions even when there are head movements while expressing an emotion. The proposed descriptor is capable of characterizing both the intensity and dynamics of facial emotions.

2. A new classifier called Extreme Sparse Learning (ESL) is obtained by adding the ELM error term to the objective function of the conventional sparse representation to learn a dictionary that is both discriminative and reconstructive. This combined objective function (containing both linear and non-linear terms) is solved using a novel approach called Class Specific Matching Pursuit (CSMP). A kernel extension of the above framework called Kernel ESL (KESL) has also been developed.

Our end goal is to correctly recognize the facial emotion. In a sparse-representation-based classification task, the desired dictionary should have both representational ability and discriminative power. Since separating the classifier training from dictionary learning may cause the learned dictionary to be sub-optimal for the classification task, we propose to jointly learn a dictionary (which may not be necessarily over-complete) and a classification model.

3.1. Optimal flow Correction

Since we are only interested in the local motion of facial components resulting from the act of expressing an emotion, the global motion of the head is subtracted from flow vector.

$OFexp=OFtot-OFh\ ead$ (3.1)

Where$OFexp$ is the expression-related OF that we aim to measure, $OFtot$ is the overall OF, and $OFh\ ead$ is the OF representing the global head movement. To measure OFh , we divide the face into a few regions and compute the average flow vector in each region. If the angle difference between the flow vector at individual pixels and the corresponding average flow vector is less than a threshold for a majority of the pixels in the region, the average flow vector is considered as OFh . Otherwise, $OFh\ ead$ is set to zero for that region. Note that in all the subsequent processing steps, OF($P;$ ti) indicates only $OFexp$ and not the overall OF.

3.2. Spatio-Temporal Descriptor

A spatio-temporal descriptor is obtained by accumulating the spatio-temporal features extracted at each pixel. Two types of histograms are used to accumulate the features in each cell. A Weighted Histogram (WH) is used to characterize the magnitude of emotion. The Un-Weighted Histogram (UWH) ignores the magnitude of the emotion and attempts to characterize its dynamics. WH and UWH are computed for each cell based on the four spatio-temporal features (*Div*, *Curl*, *Proj*, and *Rot*). The concatenation of both these histograms is considered as the final descriptor of the corresponding cell. The concatenation of all the cell descriptors results in the final spatio-temporal descriptor representing the given video sequence.

Four pose-invariant features are proposed for encoding the motion information of facial components. The first descriptor is the divergence of the flow field that measures the amount of local expansion or contraction of the facial muscles.

$(P,ti)=\partial u(P,ti)/\partial x+\partial v(P,ti)/\partial y$ (3.2)

The second descriptor that captures the local spin around the axis perpendicular to the OF plane is named *Curl*. It is useful to measure the dynamics of the local circular motion of the facial components.

$Curl(P,ti)=(\partial v(P,ti)/\partial x-\partial u(P,ti)/\partial y)Z$(3.3)

The third descriptor is the scalar projection of the OF vector $U^{\rightarrow\rightarrow}$ in the direction of $\hat{P}$,

$P(P,ti)=U.^{\rightarrow\rightarrow\rightarrow}\hat{P}$(3.4)

This *Proj* feature captures the amount of expansion or contraction of each point with respect to the nose point.

The final descriptor called *rotation* is the defined by the cross product of the unit position vector $\hat{P}$ and OF vector $U^{\rightarrow\rightarrow}$ as follow:

$(P,ti)=\hat{P}\times U^{\rightarrow\rightarrow}$ (3.5)

This feature is a vector perpendicular to the plane constructed by $\hat{P}$ and $U^{\rightarrow\rightarrow}$ that measures the amount of clockwise or anti-clockwise rotation of each facial point movement with respect to the position vector.

Before introducing the ESL, we briefly present the ELM as well as the concepts underlying sparse representation and Dictionary Learning (DL) in the following sub-sections.

3.3. Extreme Learning Machine (ELM)

Since ELM has a successful performance especially for multi-class classification problems, it is a good choice for our problem to be jointed with sparse representation and dictionary learning for further improvement. Additionally, ELM requires fewer optimization constrains in compare to SVM which results in simple implementation, fast learning, and better generalization performance.

ELM learning algorithm is trying to minimize the training error and the norm of output weightsas well.The objective function of the ELM is summarized as follow:

$\{\|(X)\beta-T\|22+\|\beta\|22\}$(3.6)

where X denotes the training samples, β is the output weight, and T is the target vector.

3.4. Extreme Sparse Learning (ESL)

Separating the classification training from dictionary learning may lead to a scenario where the learned dictionary is not optimal for classification. We propose to jointly learn a dictionary and classification model for better performance.

Learning a discriminative dictionary for sparse representation of Y can be accomplished by solving the following problem:

$$min_{X,D,\beta}\{\|Y-DX\|_2^2+\gamma_1(\|K(X)\beta-T\|_2^2+\|\beta\|_2^2)+\gamma_2\|X\|_1\}$$

Where D is the learned dictionary, X denotes the sparse codes of the input signals, and $\|.\|1$ is the $l1$ norm that simply sums up the absolute value of the elements. The first term $\|Y-DX\|^{22}$ denotes the reconstruction error, second term $(\|(X)\beta-T\|^{22}+\|\beta\|^{22})$ is related to ELM optimization constraints, and third term is related to the sparsity criterion. The framework in Eq. (5-34) is referred to as ESL. When kernels are incorporated in the above framework, we refer to it as Kernel ESL (KESL). In other words, the framework formulated as follow is referred as KESL:

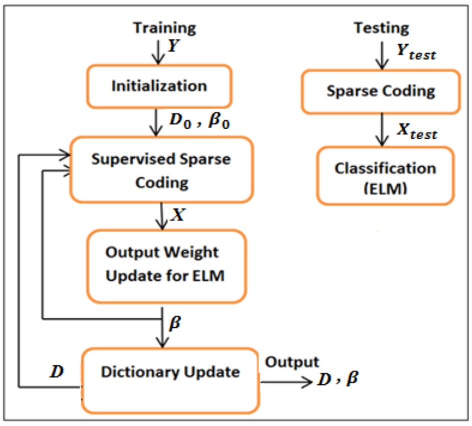

Figure 3.1: Proposed method for recognition framework.

The associated training and classification algorithms are presented in Algorithm 1 and Algorithm 2, respectively.

Algorithm1: Steps for Training ESL

Step-1. Input: Signal set Y, class labels (T) of Y, regularization terms $\gamma1$ and $\gamma2$, stopping criterion for outer and inner loops (ε1,η1, ε2, η2)

Step-2. Initialize the dictionary (0) and the ELM output weight vector (0)

Step-3. Repeat until the stopping criterion defined by (ε2,η2) is met

3.1. Repeat until the stopping criterion defined by (ε1,η1) is met

3.1.1. Supervised sparse coding: find the sparse matrix X(i) by approximating the solution of $minX\{\|Y-DX\|22+\gamma1(\|H(X)\beta-T\|22)+\gamma2\|X\|1\}$

3.1.2. ELM output weight optimization: find the ELM output weight (i) by approximating the solution of $m\{\|H(X)\beta-T\|22+\|\beta\|22\}$

3.2. Dictionary update: find D by approximating the solution of $\{\|Y-DX\|22\}$

Step-4. Output: D, β

Algorithm2: Steps for Classification ESL

Step-1. Input: test signal y, learned dictionary D and the ELM output weight β, regularization term γ

Step-2. Find the sparse code of test data by approximating the solution of $\{\|y-Dx\|22+\gamma\|x\|1\}$

Step-3. Find the output function of ELM : $f(x)=h\,(x)\,\beta$

Step-4. Estimate the class label of the test data: $l(x)=argmax(fi(x))$

Step-5. Output: $l(x)$

IV. RESULTS

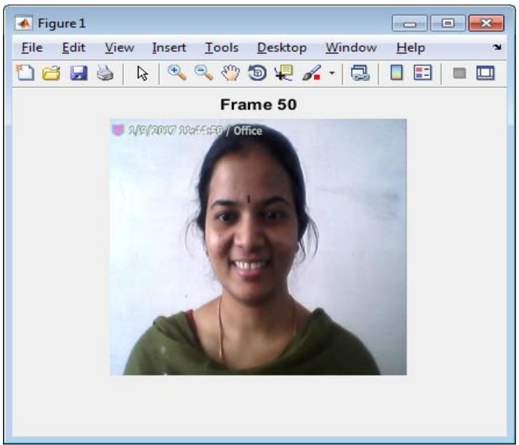

Fig.4.1: 50th Frame from the input video

The figure 4.1 gives the 50th frame which was taken from the input video. Here the input video is taken and divided into frames.

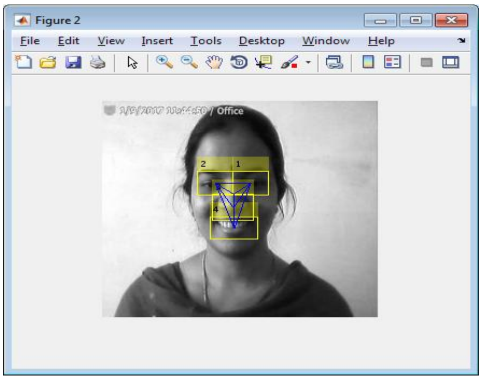

Fig.4.2: Representing face parts with bounding boxes

The figure 4.2 describes the representation of the face parts with bounding boxes by taking nose tip as the reference point. Based on the reference point the centers of other face parts are taken.

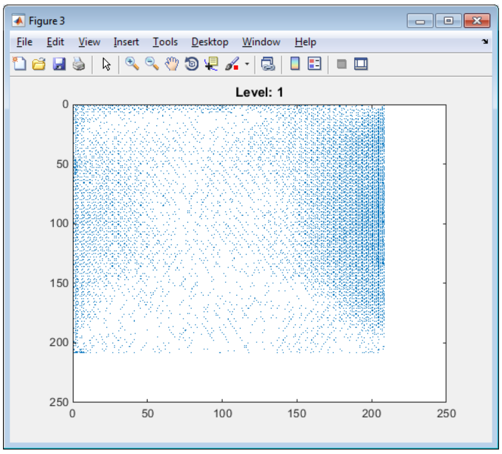

Fig.4.3: Optical Flow

The fig 4.3 reveals the information of the optical flow. The optical flow is defined as the movement of pixels from frame to frame.

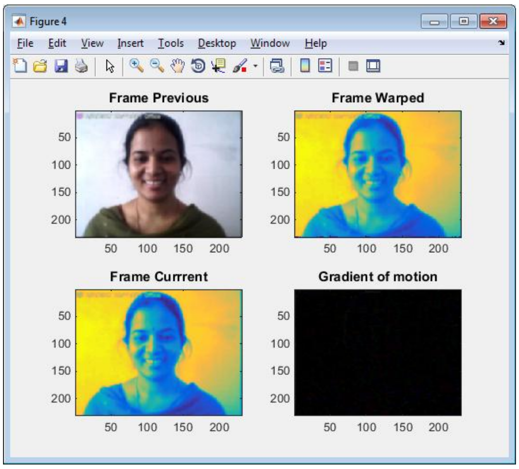

Fig.4.4: Representation of gradient of motion

The fig.4.4 describes the representation of the gradient of motion. Here the previous ,warped and current frame is taken to calculate the gradient of motion. for the movement of pixels gradient is calculated.

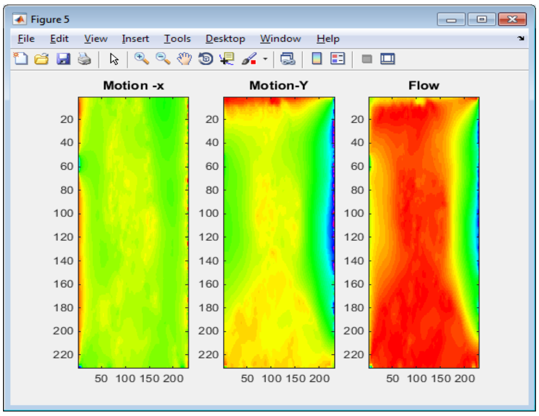

Fig.4.5: Representation of flow of pixels

The fig.4.5 describes the representation of flow of pixels. Here the direction of pixels in X-direction and Y-direction are calculated. Then the overall flow i.e. direction of pixels is calculated and represented.

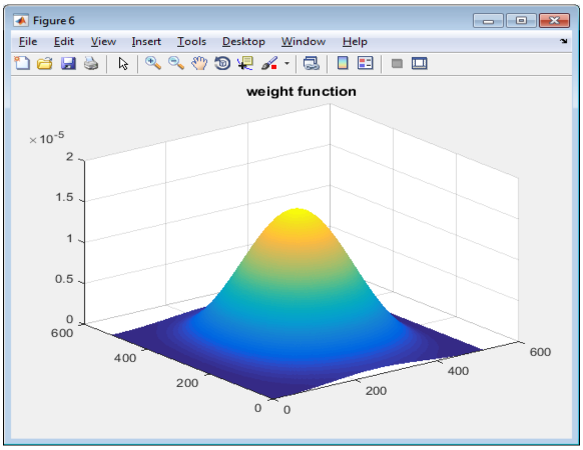

Fig.4.6: Weight function

The weight function represented in fig.4.6 gives the weights of the frame in a 3d format.

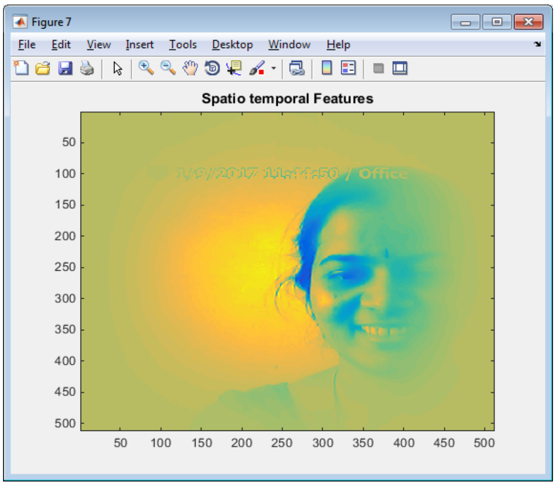

Fig.4.7: Spatio temporal features

The fig 4.7. describes the spatio-temporal features obtained by the spatio temporal feature descriptor.

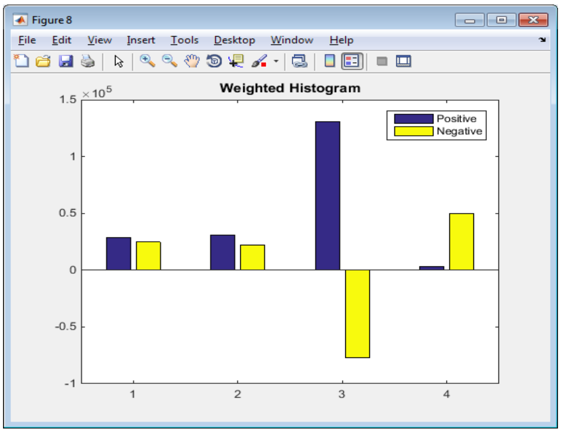

Fig.4.8: Weighted Histogram

The weighted histogram is represented in fig.4.8 characterizes the magnitude of emotion, i.e., it differentiates a subtle emotion from an exaggerated emotion. Each WH consists of two bins - positive and negative bins, and the magnitude of the associated features is used to vote for each bin.

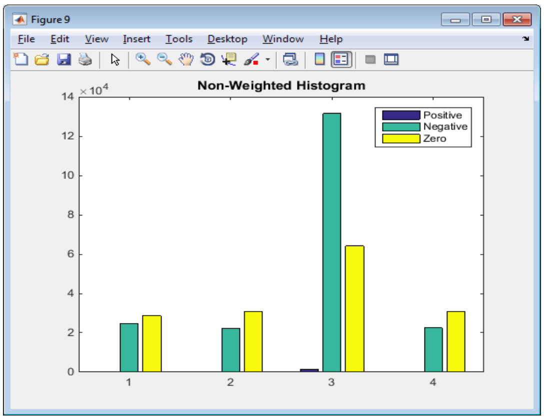

Fig.4.9: Unweighted histogram

The un-weighted histogram mentioned in fig.4.9 ignores the magnitude of the emotion and attempts to characterize its dynamics. It involves three bins related to positive, negative, and zero features. Equal vote is assigned for each bin, which means that the total number of positive, negative, and zero features are counted. The UWH minimizes the effect of changes in the emotion speed3 by considering only the sign (positive, negative, or zero) of the features and ignoring their magnitude.

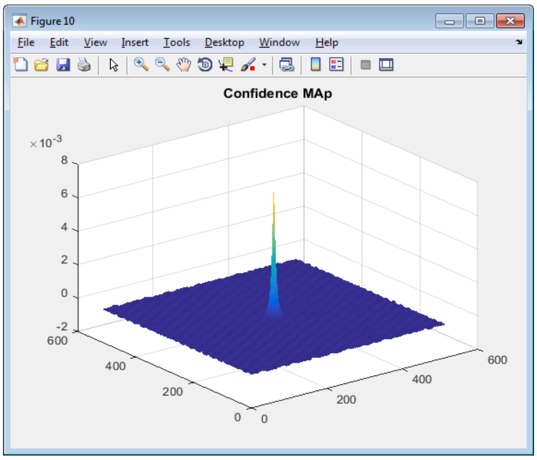

Fig.4.10: Confidence map

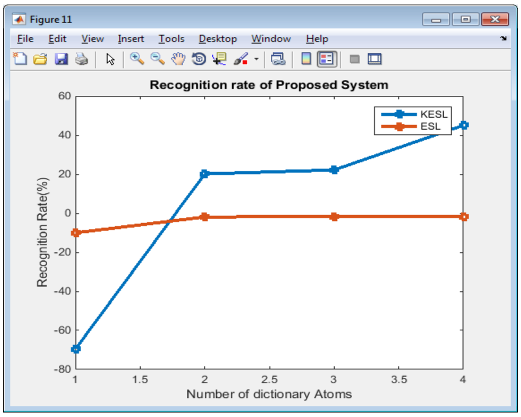

Fig.4.11: Recognition rate of proposed system

The fig4.11 describes the recognition rate comparison between the ESL and KESL. The KESL gives the better performance than ESL.

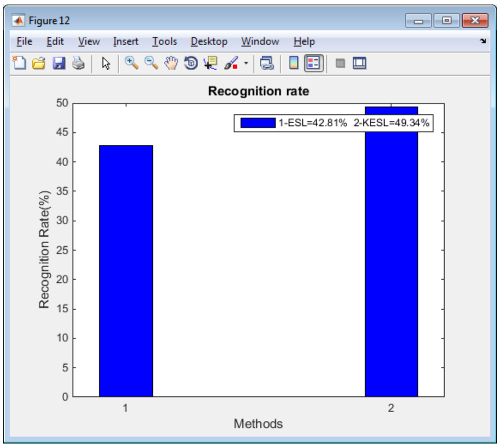

Fig.4.12: Comparison of ESL and KESL

The Fig.4.12describes the comparison of recognition rate in the form of bar graphs. the percentages is given and then theKESL is better than the ESL.

V. CONCLUSION

We have presented an effective approach for facial emotion recognition that can handle challenges such as illumination variation, expression speed variations, different imageresolution, head pose changes, and partial occlusion. The proposed approach has novelty in both the feature extraction and recognition. We have performed extensive experiments on both acted and spontaneous emotion databases to evaluate the effectiveness of the proposed feature extraction and recognition schemes under different scenarios. The results clearly demonstrate the robustness of the proposed emotion recognition framework, especially in challenging scenarios that involve illumination changes, occlusion, and head pose variations. However, possible limitations include higher computational cost for both feature extraction and classification and parameter optimization.

REFERENCES

[1] A. Jaimes and N. Sebe, "Multimodal human-computer interaction: A survey," *Computer Vision and Image Understanding,* vol. 108, pp. 116-134, 2007.

[2] T. Wehrle, S. Kaiser, S. Schmidt, and K. R. Scherer, "Studying the Dynamics of Emotional Expression Using Synthesized Facial Muscle Movements," *Journal of Personality and Social Psychology,* vol. 78, pp. 105-119, 2000

[3] S. Mingli, B. Jiajun, and C. Chun, "Expression recognition from video using a coupledhidden Markov model," *TENCON*, Vol. 1, pp. 583-586, 2004.

[4] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, pp. 699-714, 2005.

[5] I. Cohen, A. Garg, and T. Huang, "Emotion recognition from facial expressions using multilevel *Process. Syst.,* 2000.

[6] K.-E. Ko and K.-B. Sim, "Facial emotion recognition using a combining AAM with DBN," *Control Automation and Systems (ICCAS)*, pp. 1436-1439, 2010.

[7] J. F. Cohn, A. J. Zlochower, J. J. Lien, and T. Kanade, "Feature-point tracking by optical flow discriminates subtle differences in facialexpression," *IEEE Conf. Automatic Face and Gesture Recognition*, pp. 396-401, 1998.

[8] U. Tariq, L. Kai-Hsiang, L. Zhen, Z. Xi, W. Zhaowen, L. Vuong, T. S. Huang, L. Xutao, and T. X. Han, "Recognizing Emotions From an Ensemble of Features," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics,* vol. 42, pp. 1017-1026, 2012.

[9] G. Zhao and M. Pietikainen, "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, pp. 915-928, 2007.

[10] T. R. Almaev and M. F. Valstar, "Local Gabor Binary Patterns from Three Orthogonal Planes for Automatic Facial Expression Recognition," *Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 356-361, 2013.