International Journal Of Multidisciplinary Research In Science, Engineering And Technology (IJMRSET)

Volume 1, Issue 1, April 2020

Facial Expression Recognition System using Multimodal Sensors

Harsih Kamar R J, Akash N, Gokul R

Department of Computer Science and Engineering, R.M.K College of Engineering and Technology, Chennai, India

ABSTRACT: Human emotion recognition plays an important role in the interpersonal relationship. The automatic recognition of emotions has been an active research topic for several eras. Hence extracting and understanding of emotion has a high importance of the interaction between human and machine communication. Facial Expression Recognition(FER) can be widely applied to various research areas, such as mental diseases and human social/physiological interaction detection. Although laboratory-controlled FER systems achieve very high accuracy, around 97%, moving from laboratory to real-world applications the accuracy drops to 50% due to various factors such as, illumination variation, head pose and subject dependencies. We focus on three different sensors to help improve the accuracy of FER in both laboratory as well an real world expressions. First group is detailed-face sensors, which detect a small dynamic change of a face component, such as eye-trackers, which help differentiate the background noise and feature faces. The second is non-visual sensors, such as audio, depth and EEG which help in illumination variation and position shift. The third is target focused sensor such as infrared thermal sensors, which can facilitate the FER systems to filter useless visual contents and may help resist illumination variation. In this paper we use a novel facial expression database, Real-World Affective Face Database(RAF-DB), which contains approximately 30,000 facial images with uncontrolled poses and illuminations from thousands of individuals of diverse ages and race. The main drawback faced with this database is being controlled by the above mentioned sensors. Deep Locality Preserving Convolutional Neural Network (DLP-CNN) which categorizes the expressions into 7 basic categories: disgust, anger, fear, sadness, happiness, surprise and contempt.

KEYWORDS*:* Facial expression recognition, Real-world Affective Faces DataBase, deep locality-preserving convolutional neural network, Expectation maximization algorithm (EM)

I.INTRODUCTION

Today a majority of time is spent by interacting with computers, mobiles and other electronic devices [1]. These devices play an important role in our lives. These devices when being able to recognize expressions of the user can change the human-computer interaction. Humans give expressions thorough various ways such as facial expression, oral communication. Among this facial expression is used the most 55%. Hence robots can interact with humans in a better way and can be used in health care systems to detect mental stress, depression, anxiety and others and improve the quality of life. FER is not only used to detect ones expression but also used in fields like Virtual Reality(VR) and Augmented reality(AR).

There are two types of FER systems [1]: spontaneous and pose-based. In spontaneous based system images of people from various social gatherings and events are collected in unconstrained environment such as while talking dialogues and watching movies. In pose-based images of people under controlled environment with frontal faced and uniform illumination and posed expressions are taken. An ordinary FER system works in three main stages: face=processing, feature extraction and detection. At first the face region is taken as an input and landmarks like eyes, nose are recognized. The necessary details are obtained from the face to enable the recognition phase. After the recognition phase the results are classified into the seven major categories of expression as the final stage.

Although many approaches on FER have been made using depth video approaches, and mostly on data caught on camera and have not concentrated on using sensors like electrocardiogram (ECG), electromyography (EMG), electroencephalograph (EEG) and eye-trackers. Since most of experiments using FER were done using video based facial recognition and 3-D facial expression recognition and does not work efficiently in real life unconstrain environments [2].

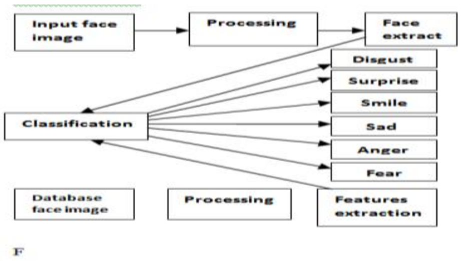

Figure 1. Sequential process of facial expression recognition

Figure 1 show the sequential process of facial expression recognition. In order to make a generic FER system this paper aims to combine existing systems and multimodal sensors to suppress various problems faced during facial recognition in order to obtain apt results for any unconstrained real world expression. The introductions of the multimodal sensors provide more information and aid to improve the accuracy and reliability of the FER systems. The common challenges faced in unconstrained real world environments like illumination variation, subject dependencies and head pose can be solved by these multimodal sensors. We can make use of the available public datasets like RAF-DB. By integrating the sensors along with the FER systems we will be able to produce a generic Facial Expression System that can work in extreme wild environments.

II.RELATED WORKS

Many available images produced were in constrained environments, the people in it were asked to pose for a picture under strict rules and when the expressions in these pictures were tested with the facial expression algorithm it showed near-perfect performance[1].

Images captured under unconstrained environments usually give complex expression rather than simple or basic expressions.

The number of facial expression annotators working on this database is very small making it unreliable

SFEW 2.0: Images collected from movies using key extraction method had unconstrained expressions with varied head poses, different age group people. However it contains only 1600 images that were labeled by two annotator's detection algorithms [2].

Christopher Pramerdorfer, Martin Kampel proposwd a paper which contains 35,000 images annotated by google image search API unfortunately this didn't solve the purpose as it failed to provide information on facial landmark location [3].

BP4D: It contains many images in HD of 41 people. All the images were captured using 3-D Face Capturing system. But these were pictures taken in unconstrained environment and all the 41 people in it were young adults[4].

AM-FED [5]: It contains 242 candid facial videos. They were captured when the people were watching TV advertisements. Without any expressions it was more suited for research.

EmotioNet [6]: Has more than 1 million facial expressions most samples were annotated by automated facial expression algorithms and 10% were labeled by manual facial expression algorithms. EmotioNet has 6 basic expressions and 17 compound expressions and these emotions were categorized based on facial recognition algorithms.

AffectNet [6]: Has 1 million images downloaded from internet. It has 4, 50,000 images which were annotated by 12 face detection algorithms. It also had images of high emotions and low emotions but were labeled only by 1 facial expression algorithm due to time and budget constraint. In contrast RAF-DB satisfies all these above requirements.

III.PROPOSED WORK

RAF-DB and DLP-CNN

Real-world Affective Faces DataBase (RAF-DB) is a multi-label facial expression dataset with around 5K great-diverse facial images downloaded from the Internet with blended emotions and variability in subjects' identity, head poses, lighting conditions and occlusions. To address the recognition of multi-modal expressions in the wild, we propose a new deep locality-preserving convolutional neural network (DLP-CNN) method that aims to enhance the discriminative power of deep features by preserving the locality closeness while maximizing the inter class scatter. An experiment was carried out with RAF-DB and DLP-CNN. During the experiment 315 well-trained facial annotators were engaged to label faces within one of the seven categories.

Enhance readability of expression: Each image was annotated for around 40 times in a single experiment. The expectation maximization algorithm(EM) is used to select the best algorithm. By analyzing 1.2 million expressions from

29,000 different faces. The results show that the expressions can be categorized into two categories expressions with single-modal distributions and compound emotions with bimodal distributions. Find difference between the expressions captured in constrained and unconstrained environment: A cross-study performed between CK+ and RAF-DB. The results revealed that unconstrained expressions and constrained expressions were very different. Due to variation in pose, illumination and others, handcrafted and focused algorithms could not categorize unconstrained facial expressions.

Improve the CNN based expression recognition: A deep learning based frame work can be used (DLPCNN). A practical back propagation algorithm that takes sin the data of local neighbors from “shallow” learning to deep feature learning approach. By doing so a locality preserving loss (LP loss) occurs which aims to extract the locally neighboring faces belonging to the same class. Locality preserving loss drives the intra-class local clusters of each class to become compact. To achieve this trained with the classical softmax loss which forces different classes to stay apart thus enhancing the discriminative power of the deeply learned features.

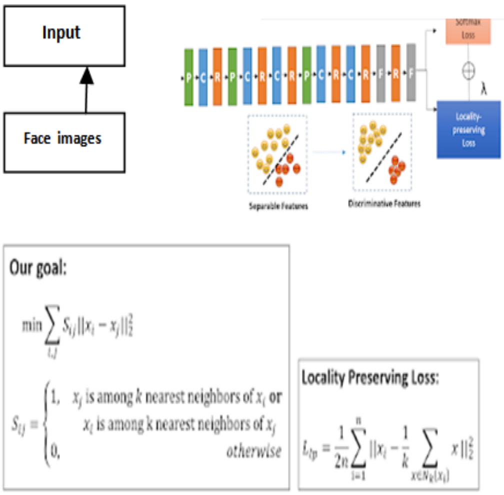
Input
Face images
Softmax Loss
λ
Locality-preserving Loss
Separable Features
Discriminative Features
Our goal:
$\min \sum_{i,j} S_{ij}\|x_i - x_j\|_2^2$
$S_{ij} = \begin{cases} 1, & x_j \text{ is among } k \text{ nearest neighbors of } x_i \text{ or } x_i \text{ is among k nearest neighbors of } x_j \\ 0, & otherwise \end{cases}$
Locality Preserving Loss:
$L_{lp} = \frac{1}{2n}\sum_{i=1}^{n}\|x_i - \frac{1}{k}\sum_{x\in N_k(x_i)} x\|_2^2$

Figure 2. Working of DLP CNN algorithm

Figure 2. Shows the working of DLP CNN algorithm. A new supervised layer is added to DLPCNN which is locality preserving loss (LP) it is done to improve the discrimination ability of the deep features. It is done by preserving the locality of each sample xi and local neighborhoods within each class as compact as possible.

$\min \Sigma_{i,j}\ S_{ij}\|x_i\text{-}x_j\|_2^2$

Here matrix S is similarity matrix. $X\epsilon R^d$ denotes deep CNN.

{ X_j is among the nearest neighbors(k) of X_i
$S_{ij} =$ In other words,
X_i is the K-nearest neighbors of X_j }

Here X_i , X_j are of the same class of the expression and K is the size of the local neighborhood.

$$L_{ip} = \frac{1}{2n}\Sigma_{i=1}^{n} 1\|X_i - \frac{1}{K}\Sigma_{X\epsilon Ni\{Xi\}} X\|_2^2$$

Here Nk{Xi} shows that the K-nearest neighbors of sample X_i.
The gradient of Lip with respect to X_i is:

$$\frac{\partial L_{IP}}{\partial X_i} = \frac{1}{N}\left(X_i = \frac{1}{K}\ \Sigma_{X\epsilon Nb\{Xi\}}\quad X\right)$$

This method is more convenient when the class conditional distribution is multi-modal. Adopting joint supervision of softmax loss, characterizes the LP loss which characterizes the local scatters within class, to train the CNNs for discriminative feature learning. The softmax loss forces the deep features of different classes to remain apart and the LP losses effectively pull the neighboring deep features of the same class together. Hence by using the joint supervision increases the discriminative power of the deeply learned features and can be easily enhanced.

Algorithm 1. DLP CNN Optimization

```
INPUT: training data: {$X_i$, $Y_i$}i=1 to n and 'N' is the size of mini batch
OUTPUT: Network layer paramaeter
Initializing t=0
Network learning rate $\mu$, Hyperparameter $\lambda$, Network layer parameter W, Softmax loss parameters $\Theta$ Neighboring nodes K.
PROCESS:
    Step 1: t=t+1
    Step 2: Figure out the center of the K nearest neighbor for $X_i$    $C_i^t = 1/k \ \Sigma_{j=1}^{n} X_j^t S_{ij}^t$
    Step 3: Update the softmax loss parameter
    $\Theta^{i+1} = \Theta^i - \mu^t \ \ \partial L/\partial\theta$
    Step 4: update the backpropagation error
    Step 5: compute the network layer parameters
```

HOG AND DENSE SIFT

One of the most common problems faced by FER systems is illumination variation which affects the correctness of the algorithm. Shifting light conditions in various stages can help in reducing the effect caused by it.

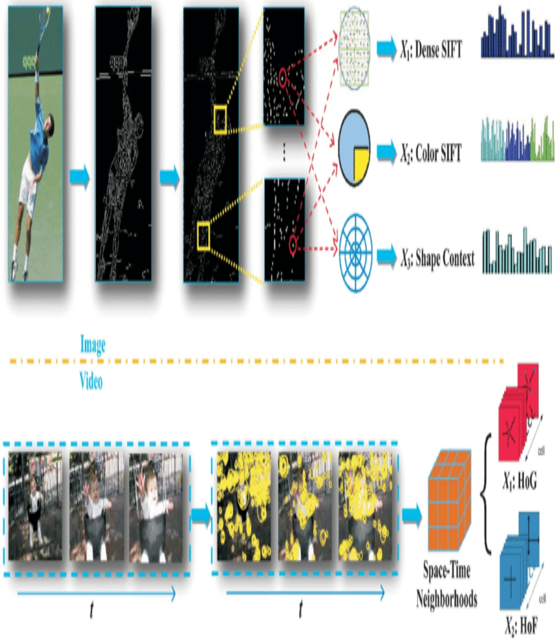

Figure 3. Hog and Dense Shift

The HOG and DENSE SIFT systems enable the system to be resistant to illumination variation. The nine layered model of deep CNN is used to obtain the third group of features. It classifies the expression with an accuracy of 94.19% (CK+ Dataset), 31.73 %(AFEW Dataset) and 75.12 %(MMI Dataset).

This method is not advisable to in unconstrained environment where there is a limit of resource such as mobile phone.

Local binary patterns can be deployed in FER due to less computational time and better tolerance. Local directional pattern (LDP) represents the gradient features and is demonstrated robust against illumination variation. For betterment double LBP can be used to reduce dimensions, logarithm laplace (LL) can be used to achieve robustness, taylor feature pattern can be used to extract the optimal facial feature from the taylor feature map.

SUBJECT-DEPENDENCE

Subject-Dependency refers to the inability of the FER system to detect facial expression. it can only detect pre-trained human faces. This requires a large dataset with many faces with various natural dissimilarities. To counter this challenge methods like PCA and LBP cannot be implemented as they miss the key features. A part based hierarchical recurrent neural network (PHRNN) can be used to extract temporal geometric features of the face. Multi-signal convolutional network neural network(MSCNN) can be deployed to find the spatial features of the face, and loss functions can be used to maximize the variation of facial expressions. This method achieves 98.5% of accuracy in CK+ DataSet and 81.18% in MMI DataSet. Hence it is a satisfactory way for unconstrained environments.

HEAD POSE

Images taken in constrained environments have the face in the frontal view. But this is not the case in real world environments. Hence this poses an challenge to FER. Geometric features can be extracted from the wrap transformation of facial landmarks to detect facial shape variation and dynamic facial textures must be extracted Histogram of Oriented Gradients from Three Orthogonal Planes (HOG-TOP) features to enable system to track facial movements. This method achieves 46.8% accuracy with (AFEW).when the face is not in frontal view the relation between image patch and smile strength must be characterized conditional to head pose by random regression forests, and several regression trees combined together to improve accuracy by training a small dataset augmented using multi data strategy.

MULTIMODAL SENSORS

Camera is one of the most popular sensor use din FER systems. Using cameras in FER systems pose three main challenges: illumination variation, head pose, subject dependency which affects the performance of the FER system. This can be overcome by adding an extra dimension to the camera feature vector by means of other sensors.

DETAILED-FACE SENSORS

The detailed face sensor works by finding new patters captured from each section of the face. The human eye gives details about the mental status of individuals, it is an integral part of the face. It possible to capture these by means of the eye tracing sensor that provides exact information of the eyes when focused in the right position.

Emotracker is an application which comprises of two softwares (Tobii Studio and Noldus Face Reader). These softwares maintain accuracy In the real-world environment. It works by recording a video of the user's face as well as user's gaze log and user's navigation information. The Face reader analyses the video and emotracker processes the result in addition to the other users;s information. It produces two maps, "emotional heat" and "emotional saccade". By processing these maps, the individuals' emotions are detected, and an emoji (A small digital image or icon to show the emotion) related to the perceived emotion is displayed on the application.

IV.CONCLUSION

Pictures taken in unconstrained environment has various challenges like illumination variation, head pose, subject dependence can be overcomed by using multimodal sensors. This increases the accuracy reliabity of any dataset under any type of environment as the three factors have been solved.

REFERENCES

[1] J. He, J.-F. Hu, X. Lu, and W.-S. Zheng. Multi-task mid-level feature learning for micro-expression recognition. Pattern Recognition, 66:44–52, 2017.

[2] A. Mollahosseini, D. Chan, and M. H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1–10. IEEE, 2016

[3] M. Liu, S. Shan, R. Wang, and X. Chen. Learning expressionlets via universal manifold model for dynamic facial expression recognition. IEEE Transactions on Image Processing, 25(12):5920–5932, 2016.

[4] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality- preserving learning for expression recognition in the wild. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2584–2593. IEEE, 2017.

[5] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. Journal on Multimodal User Interfaces, pages 1–17, 2016.

[6] Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition Shan Li, and Weihong Deng.