Multiple Feature Fusion for Facial Expression Recognition in Video: Survey

Shritika Wayker[1], Mrs. V. L. Kolhe[2]

[1]P.G. Student, Department of Computer Engineering, DYPCOE, Pune, Maharashtra, India
[2]Assistant Professor, Department of Computer Engineering, DYPCOE, Pune, Maharshtra, India

Abstract - *The most common way humans can express their emotions is through facial gestures/motions. Recently facial expression recognition based on video has been a long standing issue. The key for an effective facial expression recognition framework is to explore the various media modalities and different features to successfully describe the facial appearance and identify the changes caused due to different facial gestures.*

This survey exploits the existing methods for automatically segmenting and recognizing human's facial expression from video sequences.

Key Words: Facial expression recognition, Facial gestures, videos, multiple feature fusion, HOG- Histograms of oriented gradients.

1. INTRODUCTION

Facial expression is a capable nonverbal channel that plays a key role for individuals to pass on feelings and transmit messages. Automatic facial expression recognition is an interesting and challenging problem which has important applications in many areas like human-computer interaction [4]. It helps to build more intelligent robots which have the ability to understand human emotions.

Automatic facial expression recognition (AFEC) can be broadly connected in numerous fields, for example, restorative evaluation, lie location and human PC interaction. Ekman in early 1970s discriminated that there are six universal emotional expressions across all cultures, namely disgust, anger, happiness, sadness, surprise and fear. These expressions could be identified by observing the face signals [7]. Due to the importance of facial expression recognition in designing human–computer interaction systems, various feature extraction and machine learning algorithms have been developed. Most of these methods deployed hand-crafted features followed by a classifier such as local binary pattern with SVM classification Haar SIFT, and Gabor filters with fisher linear discriminant and also Local phase quantization (LPQ).

The recent success of convolutional neural networks (CNNs) in tasks like image classification [4] has been extended to the problem of facial expression recognition [8]. Unlike traditional machine learning and computer vision approaches where features are defined by hand, CNN learns to extract the features directly from the training database using iterative algorithms like gradient descent. CNN is usually combined with feed-forward neural network classifier which makes the model end-to-end trainable on the dataset [5].
Facial expression recognition by aggregating various CNN and SIFT models that achieves a state of art results on different datasets like FER-2013 CK+, and etc.

Fig - 1: Examples of CK+ and FER-2013 datasets [5].

2. LITERATURE REVIEW

Junkai Chen[1], proposed "Facial Expression Recognition in Video with Multiple Feature Fusion", introduces the visual modalities (face images) and audio modalities (speech). A new feature descriptor called Histogram of Oriented Gradients from Three Orthogonal Planes (HOG-TOP) is invented to extract dynamic textures from video sequels to characterize facial appearance changes. A new effective geometric feature derived from the warp transformation of facial vantage point is used to capture facial configuration changes. Moreover, the lead of audio modalities on recognition is also explored. The multiple feature fusion to implement the video-based facial expression detection problem under workshop environment and in the wild.

Mundher Al-Shabi[5], proposed "Facial Expression Recognition Using a Hybrid CNN–SIFT Aggregator", deriving an effective facial expression recognition component is important for a successful human-computer interaction system. Nonetheless, recognizing facial expression remains a challenging task. This paper describes a novel approach towards facial expression detection task. This method is motivated by the success of Convolutional Neural Networks (CNN) on the face recognizance problem. Unlike other works, the main focus is on achieving good definiteness while requiring only a small sample data for training. Scale Invariant Feature Transform (SIFT) features are used to increase the accomplishment on small data as SIFT does not require expanded training data to generate useful features. Both Dense SIFT and regular SIFT are examined and compared when merged with CNN features.

Varun Kumar Singhal[3], proposed "Multiple Feature Fusion for Facial Expression Recognition in video", introduces video based facial expression recognition has been a long standing issue and pulled in developing consideration as of late. Thekey to a fruitful facial expression recognition framework is to abuse the possibilities of varying media modalities and plan vigorous features to successfully portray the facial appearance and setup changes caused by facial movements. The investigation, of both visual modalities (confront pictures) and sound modalities (discourse) are utilized. Another feature descriptor called Histogram of Oriented Gradients from Three Orthogonal Planes (HOG-TOP) is proposed to extract dynamic surfaces from video successions to portray facial appearance changes.

Xiaohua Huang[14], proposed "Dynamic Facial Expression Recognition Using Boosted Component-based Spatiotemporal Features and Multi-Classifier Fusion", introduces the component-based facial expression recognition method by utilizing the spatiotemporal features extracted from dynamic image sequences, where the spatiotemporal features are extracted from facial areas centered at 38 detected fiducially interest points. Considering that not all features are important to the facial expression recognition, the Ada Boost algorithm to select the most discriminative features for expression recognition. Moreover, based on median rule, mean rule, and product rule of the classifier fusion strategy, also present a framework for multi-classifier fusion to improve the expression classification accuracy.

CigdemTuran, describes an emotion-based feature fusion method using the Discriminant-Analysis of Canonical Correlations (DCC) for facial expression recognition. There have been many image features or descriptors proposed for facial expression recognition. For the different features, they may be more accurate for the recognition of different expressions. The four effective descriptors for facial expression representation, namely Local Binary Pattern (LBP), Local Phase Quantization (LPQ), Weber Local Descriptor (WLD), and Pyramid of Histogram of Oriented Gradients (PHOG), are treated. Supervised Locality Preserving Projection (SLPP) is applied to the corresponding features for dimensionality reduction and manifold learning. Analysis shows that descriptors are also sensitive to the conditions of images, such as race, lighting, pose, etc. Thus, an adaptive descriptor selection algorithm is used, which determines the best two features for each expression class on a given training set. These two features are dissolve, so as to achieve a higher recognition rate for each expression.

WeiFeng Liu[15], proposed "Facial Expression Recognition Based on Fusion of Multiple Gabor Features", accomplish subject-independent facial expression recognition task, a multiple Gabor features situated facial expression recognition method. Different channels of Gabor filters have different contributions on the facial expression recognition and analytical combination of these features can improve the performance of a facial expression recognition system. NN based data fusion method is designed for facial expression recognition. Analysis show that the facial expression recognition rate can be modernized by using multiple channel features and neural network fusion.

3. TECHNIQUES USED

3.1 Histogram of Oriented Gradients from Three Orthogonal Planes:-

Histograms of oriented gradients (HOG) [9] were first determined for human discovery. The fundamental thought of HOG is that local protest appearance and shape can often be described somewhat well by the assignment of local force gradients or edge headings. HOG is touchy to question misshapenness. Outward appearances are caused by facial muscle developments. For instance, mouth opening and cocked eyebrows will produce an unexpected outward appearance. These developments could be viewed as kinds of misshapenness. HOG can successfully catch and speak to these deformation. Be that as it may, the first HOG is deminished to manage a static image [3].In request to display dynamic surfaces from a video succession with HOG, stretch out HOG to 3-D to process the oriented gradients on three orthogonal planes XY, XT, and YT (TOP),i.e. HOG-TOP.

3.2 Geometric Warp Feature

Facial expressions are caused by facial muscle developments. These movements result in the relocations of the facial landmarks. Here assumption is done that each face picture comprises of numerous sub-regions. These sub-districts can be framed with triangles with their vertexes situated at facial landmarks, as appeared. The displacements of facial landmarks cause the impairment of the triangles. To use the distortions to represent facial arrangement changes.[1] Facial demeanor can be considered as a dynamic process including beginning, pinnacle and offset. The displacement of the comparing facial landmarks between onsets(neutral face) and pinnacle (expressive face).

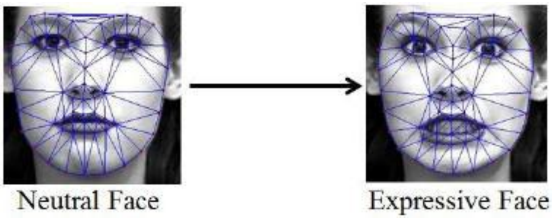

Fig- 2 : Facial landmarks describe the shape of a face [3].

3.3 Convolutional Neural Networks (CNNs):-

The network consists of six convolution layers, three Max-Pooling layers, followed by two crammed fully connected layers. Each time Max-Pooling is added, the number of the next convolution filtrate doubles. The number of convolution filters is 64, 128, and 256, respectively. The window size of the filters is 3x3. Max pooling layers with a stride of size 2x2 is placed after each of two convolutional layers. Max-Pooling is used to encapsulate the filter area which is considered as a type of non-linear down-sampling. Max-Pooling is helpful in providing a form of translation invariance and it reduces the computation for the deeper layers. To contain the spatial size of the output volumes, zero-padding is added around the borders. The output of the convolution layers is formed and fed to the dense layer. The dense layer consists of 2048 neurons linked as a fully connected layer. Each of the Max-Pooling and dense layers is pursued by a dropout layer to reduce the risk of network over-fitting by preventing co-adaptation of the feature extractor. Finally, a softmax layer with seven outputs is placed at the final stage of the network.

$$f(x) = \max(x, \frac{x}{20})$$

The threshold value 20 is selected using the FER 2013 validation set. Leaky ReLU is chosen over ordinary ReLU to solve the dying ReLU problem. Instead of giving zero when x < 0, leaky ReLU will provide a small negative slope. Besides, its de-rivatives is not zero which make the network learns faster than ReLU.

3.4 Multiple Feature Fusion:-

Highlights from various modalities can make distinctive contributions. Traditional SVM connects diverse features into a solitary component vector and constructed a solitary portion for all these distinctive highlights. In any case, developing a portion for each kind of highlights and incorporating these bits optimally can upgrade the discriminative energy of these features.

The consider in demonstrated that utilizing numerous pieces with different sorts of highlights can enhance the execution of SVM. A numerous portion SVM is intended to learn both the decision limits between information from various classes and the bit mix weights through a solitary optimization problem.

3.5 Scale Invariant Feature Transform (SIFT):-

SIFT features are used to increase the execution on small data as SIFT does not require extensive training data to generate useful features. SIFT key point of objects are first extracted from a set of reference images in order to avoid from computing all points in an image[2]. It is thus found that the search of interest points or lands in facial images is more important to component-based approach. The SIFT descriptor is resolute by partitioning the image into 4x4 squares. For each of the 16 squares, a vector length of 8. By merging all the vectors, a vector of size 128 for every key-point. To use the key-point descriptors in classification, a vector of fixed-fixed-size is needed [5].

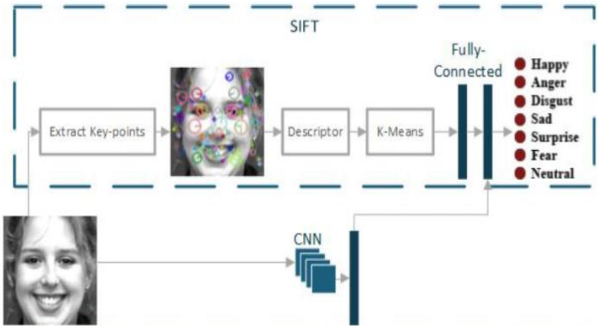

Fig- 3: - SIFT and its key-points [2].

4. EXISTING SYSTEM

4.1 Facial Expression Recognition from Videos:-

Traditionally, video-based recognition problems used per frame features such as SIFT, dense-SIFT, HOG and recently deep features extracted with CNNs have been used. The per-frame features are then used to allocate score to each individual frame. Summary statistics of such perframe features are then used for facial expression recognition. Modification of Inception architecture to capture action unit activation which can be beneficial for facial expression recognition. Various techniques to capture the secular evolution of the per-features. For example, LSTMs have been successfully employed with various names such as CNN-RNN, CNN-BRNN etc.

3D convolutional neural networks have also been used for facial expression recognition. However, performance of a single 3D-ConvNet was poor than applying LSTMs on per-frame features [9]. State-of-art result reported in was obtained by score fusion of multiple models of 3D-ConvNets and CNN-RNNs. Covariance matrix representation was used as one of the summary statistics of per-frame features in[10] Kernel based partial least squares (PLS) were then used for recognition. Here, the methods in as baseline and use the SPD Riemannian networks instead of kernel based PLS for recognition and obtain slight improvement.

5. CONCLUSION

In this work, a survey of several methods for expression recognition from video. A new feature descriptor called Histogram of Oriented Gradients (HOG) is used to extract dynamic textures from video sequences to characterize facial appearance changes. The different methodologies like Geometric Warp Feature, CNN, Multiple Feature Fusion and etc are briefly described. Multiple Feature Fusion has highest feature for segmenting and recognizing human facial expression from video sequences.

REFERENCES

[1] Junkai Chen, Zenghai Chen, Zheru Chi, and Hong Fu, “Facial Expression Recognition in Video with Multiple Feature Fusion” 2016 IEEE.

[2] Berretti, S. et al.: “A Set of Selected SIFT Features for 3D Facial Expression Recog-nition”. In: 2010 20th International Conference on Pattern Recognition (ICPR). pp. 4125–4128 (2010).

[3] Varun Kumar Singhal “Multiple Feature Fusion for Facial Expression Recognition in video” International Journal of Electronics Engineering2018.

[4] He, K. et al.: “Deep Residual Learning for Image Recognition”. ArXiv151203385 Cs. (2015).

[5] Mundher Al-Shabi, Wooi Ping Cheah, Tee Connie “Facial Expression Recognition Using a Hybrid CNN–SIFT Aggregator”

[6] Ilke C¸ ugu, Eren S¸ ener, EmreAkbas “MicroExpNet: An Extremely Small and Fast Model For Expression Recognition From Frontal Face Images” arXiv:1711.07011v2 [cs.CV] 13 Aug 2018.

[7] Ekman, P., Friesen, W.V.: “Constants Across Cultures in the Face and Emotion”. J. Pers. Soc. Psychol. 17, 2, 124–129 (1971).

[8] Kim, B.-K. et al.: “Hierarchical Committee of Deep Convolutional Neural Net-works for Robust Facial Expression Recognition”. J. Multimodal User Interfaces. 10, 2, 173–189 (2016).

[9] N. Dalal and B. Triggs,”Histograms of Oriented Gradients forHuman Detection,” IEEE Conference on Computer Vision and PatternRecognition, 2005, pp. 886-893.

[10] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen.”Combining multiple kernel methods on Riemannian manifold for emotion recognition in the wild”. In Proceedings of the 16th International Conference on Multimodal Interaction, ICMI ’14, pages 494–501, New York, NY, USA, 2014. ACM.

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

[12] T. Kanade, J. F. Cohn, and Y. Tian. “Comprehensive database for facial expression analysis”. In Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), pages 46–53, 2000.

[13] Dinesh Acharya, Zhiwu Huang, DandaPaniPaudel, Luc Van Gool “Covariance Pooling for Facial Expression Recognition” IEEE 2018.

[14] Xiaohua Huang, Guoying Zhao, MattiPietik¨ainen, Wenming Zheng “Dynamic Facial Expression Recognition Using Boosted Component-based Spatiotemporal Features and Multi-Classifier Fusion”

[15] WeiFeng Liu, ZengFu Wang “Facial Expression Recognition Based on Fusion of Multiple Gabor Features” 2006 IEEE.