Recognizing Facial Expressions Using Novel Motion Based Features

Snehasis Mukherjee
IIIT Chittoor, SriCity
snehasis.mukherjee@iiits.in

Bandla Vamshi
IIIT Chittoor, SriCity
vamshi.b14@iiits.in

K.V. Sai Vineeth Kumar Reddy
IIIT Chittoor, SriCity
vineet.k14@iiits.in

Repala Vamshi Krishna
IIIT Chittoor, SriCity
vamshi.r14@iiits.in

S.V.S. Harish
IIIT Chittoor, SriCity
harish.s14@iiits.in

ABSTRACT

This paper introduces two novel motion based features for recognizing human facial expressions. The proposed motion features are applied for recognizing facial expressions from a video sequence. The proposed bag-of-words based scheme represents each frame of a video sequence as a vector depicting local motion patterns during a facial expression. The local motion patterns are captured by an efficient derivation from optical flow. Motion features are clustered and stored as words in a dictionary. We further generate a reduced dictionary by ranking the words based on some ambiguity measure. We prune out the ambiguous words and continue with key words in the reduced dictionary. The ambiguity measure is given by applying a graph-based technique, where each word is represented as a node in the graph. Ambiguity measures are obtained by modelling the frequency of occurrence of the word during the expression. We form expression descriptors for each expression from the reduced dictionary, by applying an efficient kernel. The training of the expression descriptors are made following an adaptive learning technique. We tested the proposed approach with standard dataset. The proposed approach shows better accuracy compared to the state-of-the-art.

CCS Concepts

•**Computing Methodologies → Artificial Intelligence;** *Computer Vision;* Activity Recognition and Understanding;

Keywords

Facial expression; Optical flow; GWOF; WOF; Adaptive learning

1. INTRODUCTION

Facial expression analysis and recognition are gaining much interest of the computer vision researchers due to the wide spectrum of applications ranging from criminology, psychiatry and medical science to advertising and market analysis [1]. Automatic recognition of facial expressions can be useful in emerging application areas like Human Computer Interaction and Robotics [2]. The recognition of facial expressions in videos is a challenging task due to the complex style of movements of facial components, different skin-tone and illumination condition, different positions and allignments of the faces in the video frames and different scales of the faces. In addition to the above challenging factors, variation of the style of expressions for the same emotion makes the facial expression recognition problem even harder. For example, the style of laughing may vary for different persons. Also, different emotions (e.g., anger and disgust) may have similar appearances [3]. Modeling the intra-class diversity and inter-class similarity of the appearance of the facial expressions is challenging.

Facial Action Coding System (FACS) is a tool for measuring facial expressions of a human [4]. FACS is an anatomy based system for analysing and representing facial movements. Each facial movement is called an Action Unit (AU). Any facial expression can be broken down into a series of AU. The FACS system came up with six basic facial expressions: Anger, Disgust, Fear, Happiness, Sadness and Surprise. This paper attempts to recognize the six basic expressions.

We propose two novel motion features based on the optical flow and image gradient and applied the two features separately in recognizing facial expressions, and compare the efficacy of the two features in different kind of videos. Both the proposed features are robust to changes in skin-tone. We construct face pose descriptors in the form of vectors, from the proposed features separately and find the expression descriptors following the bag-of-words model [1]. The vector representation of the proposed motion based features make the facial expression recognition system scale-invariant. We further modify the adaptive learning mechanism proposed in [1], for better recognition of facial expressions. Since, the proposed features track the movement of each individual pixels, they are also capable of dealing with the complex non-rigid movements of different facial components.

The proposed motion features are based on optical flow and image gradient. We inherit the concept of Warp Optical Flow (WOF) of each frame of the video [5]. For finding WOF, the derivatives are calculated on both x and y direc-

ICVGIP, December 18-22, 2016, Guwahati, India
© 2016 ACM. ISBN 978-1-4503-4753-2/16/12...$15.00
DOI: http://dx.doi.org/10.1145/3009977.3010004

tions over the optical flow matrix of the frame. The magnitudes and amplitudes of the derivatives are considered for finding WOF. As a result of this calculation, the motion estimation at the background pixels are minimized. We inherit another way of minimizing the background motion estimations. We pointwise multiply the optical flow with the magnitude of gradient for each frame of the video, to get the Gradient-Weighted Optical Flow (GWOF) [6]. We propose a new motion feature by taking the derivatives in both x and y directions of each pixel of the GWOF matrix (instead of the optical flow matrix as done in [5]). We call the new motion feature as Warp-GWOF. We represent GWOF, WOF and Warp-GWOF for each frame by 168 dimensional vectors following [6]. We concatenate the vectors representing GWOF and WOF measures to get a 336 dimentional vector, which is one of our proposed feature vector, which we call as Gradient Weighted Warp Optical Flow (GWWOF). The other proposed feature vector is the 168 dimensional vector representing Warp-GWOF. We apply the features GWWOF and Warp-GWOF separately for recognizing facial expression recognition and compare the results.

We apply the bag-of-words model following [1] with the two proposed features. The motion features (GWWOF and Warp-GWOF separately) are clustered and stored as facial pose words in a dictionary. We further generate a reduced dictionary by ranking the facial poses based on some ambiguity measure, derived from a facial pose graph, where facial poses as represented as nodes. We prune out the ambiguous facial poses and continue with key facial poses in the reduced dictionary. We followed the average centrallity measure [7] to rank the facial poses according to their ambiguity. We form expression descriptors for each expression from the reduced dictionary, by applying an efficient kernel. The training of the expression descriptors are made following a modified version of the adaptive learning technique proposed in [1]. Figure 1 shows the overall procedure.

In this paper, our contribution is three fold, over [1]. First, we propose new features to minimize background motion. Second, for constructing the reduced dictionary of facial pose words, we rank the words by average centrallity measure as proposed in [7]. Third, we modified the adaptive learning technique by approximating the kernel by t-distribution, instead of approximating by Gaussian distribution.

The paper is organized as follows. Section 2 presents a brief survey of the state-of-the-art techniques in the facial expression recognition field. Section 3 illustrates the process of extracting the proposed motion features, followed by a detailed description of the process of constructing the final dictionary of facial poses in Section 4. The process of constructing the expression descriptors and the adaptive learning mechanism of the descriptors is illustratedin Section 5. Section 6 discuss the results of applying the proposed approach on standard datasets. Section 7 concludes the paper.

2. RELATED WORKS

Facial expression recognition is an active area of research in the computer vision field during the last couple of decades [8]. However, recognizing facial expression is still far from being a solved problem. In this section we discuss about the recent advances in this problem. A typical facial expression recognition approach has two major steps: first, extracting effective features from the video based on motion and static information and second, applying an efficient classification mechanism to recognize the facial expressions based on the features. The features applied for facial expression recognition can be classified into two categories: Geometry-based features and Appearance-based features.

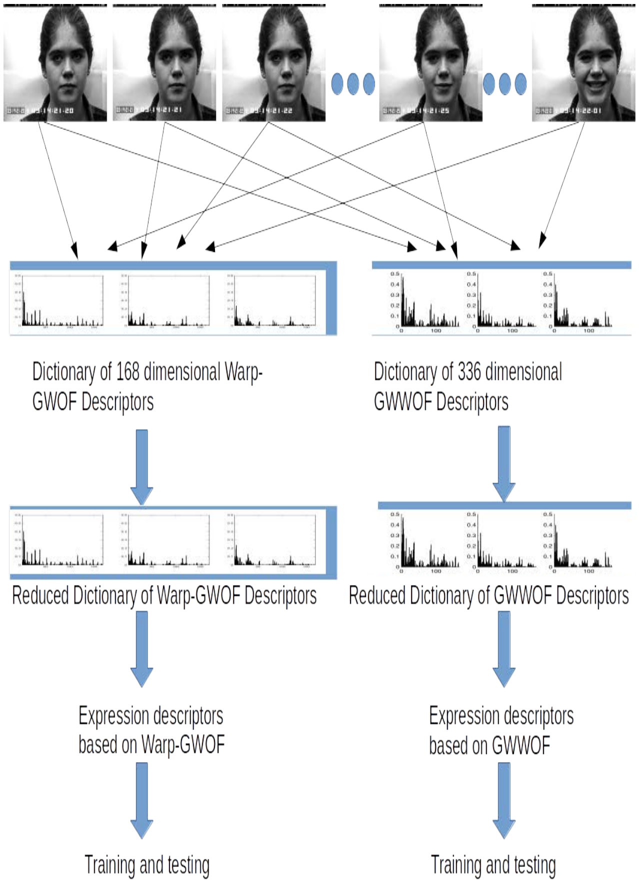
Dictionary of 168 dimensional Warp-GWOF Descriptors
Dictionary of 336 dimensional GWWOF Descriptors
Reduced Dictionary of Warp-GWOF Descriptors
Reduced Dictionary of GWWOF Descriptors
Expression descriptors based on Warp-GWOF
Expression descriptors based on GWWOF
Training and testing
Training and testing

Figure 1: Overall procedure of the proposed scheme for facial expression recognition.

The geometry-based features rely on the geometric information (e.g., shape, length, etc.) about various components of the face during the expression. In a typical geometry-based approach, Jain *et.al.* proposed a shape-based feature for facial expression recognition [9]. In [9], a 2D representation of face shape is proposed using a set of 68 landmark points on the face. The landmark points are located around the contours of different components of the face, e.g., the eyebrows, eyes, nose, chin, inner lips and outer lips. The sequencial change of shape of face during a facial expression, is modelled by Latent Dynamic Conditional Random Field (LDCRF). Hsu *et.al.* have selected some specific points on the face and modelled the change of distances between the points during a facial expression [10]. However, Zhang *et.al.* have shown that Appearance-based features are more effective than Geometry-based features, for facial expression recognition [11].

The appearance based features rely on the changes in texture of the surface of the face during a facial expression. The change in texture on face can be happened due to appearance (or disappearance) of wrinkles, bulges, furrows. Changes in shapes of facial components such as eyes, mouth, nose, chick and forhead also may cause changes in texture. Different features have been applied to capture the information on change of texture of the components of face. Most popular appearance-based features applied for facial expression recognition are Gabor, active appearance model, Local Binary Pattern (LBP) and Local Phase Quantizer (LPQ).

Wu *et.al.* have extracted motion information from difference of frames and used the information to apply spatio-temporal Gabor filters on face videos [12], where the facial expressions are modelled by biologically inspired spatio-temporal Gabor filters. In [13], Martins *et.al.* applied active appearance model to represent the texture and shape of the face and model the interaction between them during a facial expression. Classification is done using Mahalanabis distance between the feature vectors. However, for analyzing finer motion patterns of individual components of faces during facial expression, low level features are more useful.

Local Binary Pattern (LBP) features are extensively used in facial expression recognition [14]. Zhao *et.al.* analyzed the texture of the face using LBP [3], where a dynamic motion descriptor was proposed based on LBP in three orthogonal planes (LBP-TOP). LPQ based descriptors are also widely used for facial expression recognition [15]. Jiang *et.al.* have followed the concept of LBP-TOP and proposed LPQ-TOP that outperforms LBP-TOP [16]. Efforts have also been made by analyzing the texture of face in frequency domain using curvelet transform [17]. However in order to deal with background motion, some efficient combination of foreground information and the motion information is necessary. In [1], Agarwal *et.al.* have applied GWOF feature by combining the foreground information from the gradient and motion information from the optical flow, which have outperformed LBP-TOP and LPQ-TOP. The WOF features are used in action recognition field to reduce the background motion given by optical flow, but the WOF feature has not been tested yet in the facial expression recognition problem. This paper shows that an efficient combination of the concepts of GWOF and WOF can outperform both of them.

Several methods have been proposed for classification of the features used for recognizing facial expressions. Hsieh *et.al.* have extracted the different components of face (eyes, eye-brows, mouth, nose, etc.) using active shape model [18]. Image-gradient based Gabor features are extracted from the components of faces and classified using a multi-class Support Vector Machine (SVM) classifier. Li *et.al.* have applied a Dynamic Bayesian Network (DBN) mechanism to model the interaction of low, mid and high level features of different components of the face, for facial expression recognition [19]. Efforts have been made to classify the features using multi-class neural network model [20], where the motion features are characterized by a statistical tool called ANOVA. Boughrara *et.al.* proposed a multi layer perceptron (MLP) based classification scheme where features extracted from different components of face are fed into the first layer of the MLP architecture. Contributions from each neuron of first layer are fed into a neuron in the next layer for classification of facial expressions [21].

Recently bag-of-words (BoW) based classification schemes are shown to provide much efficacy over the neural network based models in facial expression recognition field [10, 1]. Hsu *et.al.* have measured the distances between some specific points on the face and considered the distances as words in the BoW model [10]. Agarwal *et.al.* have constructed a reduced dictionary of facial poses as words, by pruning out the ambiguous facial poses based on an ambiguity measure [1]. The ambiguity measure is calculated from the edge weights of the graph where facial poses are represented as nodes. The edge weights are determined by the co-occurrence of the facial poses during the expression. After the pruning, the key facial poses are stored in the reduced dictionary. An adaptive learning technique based on Gaussian kernel, is introduced for recognizing facial expressions. We follow the classification procedure proposed by [1] with two modifications: first, we modified the ambiguity measure for finding the key facial poses and second, we used t-distribution kernel instead of Gaussian kernel, for learning. Next we illustrate the process of extracting the proposed features.

3. FEATURE EXTRACTION

We propose two motion based features: Warp-GWOF and GWWOF features. Both the features are extracted by applying some efficient combination of the method of constructing the GWOF feature and the WOF feature. For each frame (except the first frame) we first calculate the optical flow and the image gradient. We pixelwise multiply the magnitude of optical flow $|\mathbf{F}|$ with the magnitude of gradient $|\mathbf{G}|$, to get the magnitude of the GWOF measure $|\mathbf{V}|$ as follows:

$$|V_x| = |F_x|.*|G_x| \quad and \quad |V_y| = |F_y|.*|G_y|, \tag{1}$$

where the binary operator '.*' between two matrices of same size represents the elementwise multiplication and $|V| = |V_x|^2 + |V_y|^2$. The direction of GWOF at each pixel is kept as the same as the direction of the optical flow at that pixel.

Next we calculate the WOF measure on the same frame. The WOF measure is obtained from the optical flow matrix $|\mathbf{F}|$ by taking the derivatives W_x and W_y of the magnitude of optical flow at each pixel along the x and y directions respectively. The magnitude of the WOF measure for the (i, j)th pixel of the current frame is calculated as,

$$W(i, j) = \sqrt{W_x^2 + W_y^2}, \tag{2}$$

and the direction of WOF measure is given as, $\theta = \tan^{-1} \frac{W_y}{W_x}$. Next we calculate the GWWOF measure $|D|$ for the same frame by taking the derivatives $|D_x|$ and $|D_y|$ from $|V|$ obtained from equation (1). We apply the equation (2) on $|D_x|$ and $|D_y|$ to get the GWWOF measure $|D|$. Next we obtain the feature vectors from $|V|$, $|W|$ and $|D|$.

3.1 Obtaining Feature Vectors

We obtain 168 dimensional feature vectors from each of the $|\mathbf{V}|$, $|\mathbf{W}|$ and $|\mathbf{D}|$ measures separately. The 168 dimensional vectors are obtained by quantizing the directions of the measures, into 8 bins as proposed in [6]. We experimented with different number of bins and observed the accuracy of the approach. We experimented with 4, 5, 6, 8, 9 and 10 bins and found maximum accuracy for 8 bins. For each pixel, the direction of the measure is first observed. If the direction at the pixel falls in the ith octant, then the ith bin of an 8-bin histogram is increased by the magnitude of the measure at that pixel. Then we divide the matrix (of the measure) into 4 parts and obtain 8 bin histograms from each part separately. The frame is further divided into 16 parts

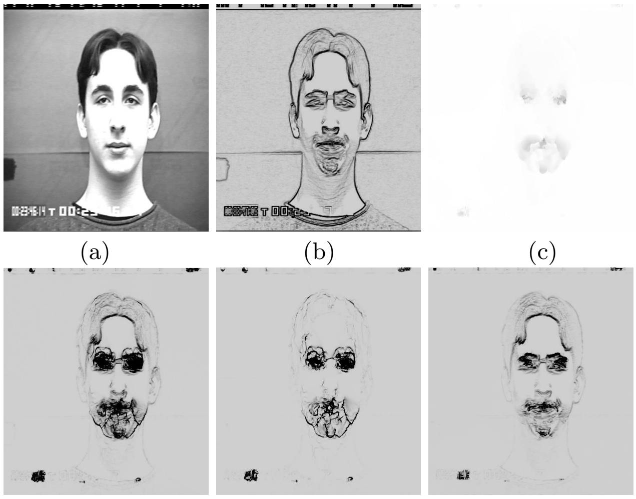

Figure 2: An example showing results of applying various motion based measures on a frame taken from laughing expression of the Extended Cohn Kanade dataset (CK+) [22]. (a) Original frame, (b) After applying image gradient, (c) Optical flow, (d) GWOF, (e) WOF and (f) GWWOF measures.

and an 8 bin histogram is generated from each part in a similar way. We normalize all the 21 histograms (all of 8 bins) separately and concatenate them to get the 168 dimensional feature vector corresponding to the measure. These three 168 dimensional vectors obtained from the three measures $|\mathbf{V}|$, $|\mathbf{W}|$ and $|\mathbf{D}|$ are used to construct the proposed motion descriptors. The layered architecture for obtaining the descriptors facilitates emphasizing on both global and local variations of motion features during a facial expression.

3.2 Obtaining the Proposed Descriptors

As discussed earlier, we proposed Warp-GWOF and GWWOF motion descriptors. The 336 dimensional Warp-GWOF descriptor corresponding to a frame is obtained by concatenating the 168 dimensional vectors obtained from $|\mathbf{V}|$ and $|\mathbf{W}|$ measures, following the process discussed in the previous subsection. The 168 dimensional GWWOF descriptor is the feature vector obtained from $|\mathbf{D}|$ following the similar process.

Figure 2 illustrates the results of calculating the various measurements such as optical flow, image gradient, GWOF, WOF and GWWOF, on a sample frame of a video of laughing expression. It is evident from the Figure 2 that, WOF measure plays an important role in measuring the local motion pattern compared to Optical flow and GWOF measures. GWOF measure emphasizes on spatial intensity changes because of the involvement of image gradient in calculating GWOF. Hence, the hair of the person gets emphasis in GWOF measure (which is unnecessary in expression recognition), whereas WOF measure more emphasizes on the intensity changes in the temporal direction only. However, since WOF measure is calculated on the optical flow measure, some unnecessary flow information can be found in WOF measure, which can be removed by an efficient use of image gradient. So, a combination of the GWOF measure and the WOF measure can be helpful in emphasizing on the motion patterns during facial expressions. This is the motivation of proposing the two descriptors in this paper. As discussed, both the proposed Warp-GWOF and GWWOF descriptors are actually some combinations of GWOF and WOF measures.

Our goal is to use the Warp-GWOF and the GWWOF descriptors separately for facial expression classification and compare the results. Henceforth we will discuss the proposed classification approach for facial expression recognition. The same approach is followed with both the Warp-GWOF and GWWOF descriptors. Next we discuss the process of forming the final dictionary of facial poses in the BOW approach.

4. FORMING THE DICTIONARY OF KEY FACIAL POSES

We apply BoW model to classify the expressions. Following the approach of [1], we first construct an initial dictionary of facial poses by clustering the descriptors using Max-diff kd-tree data structure [23], which is shown to provide better clustering than Kmeans and K-nearest-neighbour algorithms [6]. Next we construct a reduced dictionary of key facial poses by ranking the facial poses according to some ambiguity measure and pruning out the ambiguous facial poses.

4.1 Initial Dictionary of Facial Poses

We cluster the motion descriptors related to a facial expression, using Max-diff kd-tree data structure [23]. Let $C_j = \{p_1, p_2, \ldots, p_l\}$ be the descriptors related to a facial expression. Here l is the number of frames required to represent one expression. Our intention here is not to get true partitioning of the facial pose space but to obtain an initial dictionary S the redundancies in the facial pose space is reduced. The Max-diff kd-tree based data condensation technique mines the multi-dimensional facial poses into a kd-tree data structure. The kd-tree is a binary tree, formed with root containing all motion descriptors of an expression. We split the set of motion descriptors at the root node into two subsets to pass the subsets of motion descriptors to each of the two children of the root node. The same procedure of dividing the set of motion descriptors is applied for each of the two children and so on. This splitting mechanism continues until we get at least T leaf nodes. We have taken the value of T as 10. Since we further reduce the number of facial poses at later stage, to build the reduced dictionary, the value of T can be kept sufficiently large (10 can be considered as large number as number of distinct facial poses related to an expression can not exceed 10), so that we can accommodate as many possible key facial poses as possible. The splitting of the set of motion descriptors is done after making a splitting rule using the pivot element and pivot dimension of the node [23]. The pivot dimension at each node is computed by finding out in which dimension of the motion descriptors the separation (between two numbers of the same dimension of two different motion descriptors) is maximum. This dimension-wise separation of facial pose pattern leads to a good clustering of different facial poses related to the expression. The leaf nodes of the kd-tree denote facial pose clusters. One can choose (depending on computational expense) multiple samples from each leaf node to construct the initial dictionary S for the expression. Next we discuss the process of constructing the reduced dictionary from this initial dictionary.

4.2 Reduced Dictionary of Facial Poses

We reduce the size of the initial dictionary by pruning out the ambiguous facial poses related to an expression, based on some ambiguity measure assigned to all the facial poses related to the expression. The ambiguity measures are assigned using a graph-based technique.

4.2.1 *Facial Pose Graph*

We construct a facial pose graph with all the facial poses of the initial dictionary represented as nodes. Facial pose graph is an weighted undirected graph where the weight of the edge e_{ij} between the nodes p_i and p_j is computed from the frequency of occurrances of the facial poses p_i and p_j during the expression.

$$e_{ij} = \frac{1}{n(p_i, p_j)}, \tag{3}$$

where $n(p_i, p_j)$ is calculated as,

$$n(p_i, p_j) = min(n(p_i), n(p_j)), \tag{4}$$

where $n(p_i)$ is the number of occurrances of p_i during the expression. The edge e_{ij} does not exist if at least one of the facial poses p_i and p_j never occur during the expression. Next we assign ambiguity measure to each of the facial poses.

4.2.2 *Ranking of Facial Poses*

According to (3), the lower the edge weight, the stronger is the semantic relationship between the facial poses. Hence, the significant facial poses become central to the facial pose graph. Next, we calculate the average-centrality measure of graph connectivity to measure the ambiguity of a facial poses [7]. We apply the Floyd-Warshall algorithm [24] to compute all-pair-shortest path between the facial pose nodes. If the distance $d(p_k, p_r)$ between the facial poses p_k and p_r is the sum of the edge weights e_{ij} on a shortest path from p_k to p_r in the facial pose graph, then the average-centrality $a(p_k)$ of a pose p_k is given by,

$$a(p_k) = \frac{1}{T-1} \sum_{p_r} d(p_k, p_r) \;\; \forall \;\; p_r \in S, \tag{5}$$

where T is the number of elements in S.

The average-centrality measure is different from the eccentricity measure used in [1]. The eccentricity measure of a node computes the maximum distance from a pose p_k in the pose graph, instead of average distance as shown in (5). As shown in [7], average-centrality measure can better emphasize on the relevance of a facial pose (in terms of less ambiguity) to the facial expression. The higher the average-centrality measure, more ambiguous is the facial pose. We rank the facial poses according to the average-centrality measure and select the t best facial poses as key facial poses. The value of t is experimentally chosen as 4 for all the expressions. The reduced dictionary is the collection of all the key facial poses related to all the expressions. Next we construct the expression descriptors for all the expressions and learn the expression descriptors to classify the expressions.

5. LEARNING EXPRESSION DESCRIPTOR

We construct the expression descriptors for each facial expression. Expression descriptors are vectors of dimensions equal to the cardinality of the reduced dictionary of key facial poses. For each expression video we build the corresponding expression descriptor by finding the relevance of the key facial poses in each frame I_r, $r = 1, 2, \ldots, M$, where M is the number of frames in the video sequence. We extract motion descriptor q_r for each frame I_r and then map it to some key facial pose p_j in the reduced dictionary S_r, where $j = 1, 2, 3, \ldots, x$; x being the cardinality of S_r. Hence we obtain a histogram count that gives number of occurrences of each key facial pose in the video sequence.

We follow the plausibility model for constructing the expression descriptors. According to [1], plausibility model works better than all the other models discussed in [6] because, the plausibility model gives higher weightage to the most relevant motion descriptor. According to the plausibility model, the ith bin (i.e., the bin corresponding to the ith key facial pose) of the expression descriptor, $ED_P(i)$ is calculated as,

$$ED_P(i) = \frac{1}{M} \sum_{r=1}^{M} \begin{cases} K(Distance(p_j, q_r)) \;\; when \\ i = \arg\min_j(Distance(p_j, q_r)), \\ \forall \; j = 1, 2, ..., x \\ 0 \;\; otherwise \end{cases}, \tag{6}$$

where $Distance()$ is a function that finds the Euclidean distance between two vectors and $K()$ is the kernel density function. In [1, 6, 7], Gaussian distribution with 0 mean and standard deviation 1 is used as the kernel. However, as discussed earlier, our goal is to assign higher weightage to the most relevant motion descriptor. Hence, we propose to use t-distribution as the kernel. For the stipness of the density function of t-distribution, it is expected to assign more values to the corresponding bin, if the distance (between p_j and q_r) is low. Hence, kernel density function $K(z)$ is chosen as follows:

$$K(z) = \frac{\Gamma(\frac{x+1}{2})}{\sqrt{\pi x}\Gamma(\frac{x}{2})}(1 + \frac{z^2}{x})^{-\frac{x+1}{2}}, \tag{7}$$

where $x - 1$ is the degrees of freedom for the t-distribution, as the number of key facial poses is x. For infinite degrees of freedom t-distribution converges to Gaussian distribution. But in this problem, degrees of freedom is always finite. As we mentioned earlier, we have selected 4 key facial poses for each expression. We worked with 6 expressions and hence, the degrees of freedom becomes 23. So, t kernel is expected to provide better learning of the expression descriptors.

Next we apply an adaptive learning technique on the expression descriptors following [1]. According to the adaptive learning technique for obtaining the expression descriptor, we do not prefer any hard assignment of the motion descriptor p_j to a facial key pose q_r. Rather, we solicit a proper distribution of p_j into several facial key poses, according to their distances from the motion descriptor. For example, if the ratio of distances of a motion descriptor from three key facial poses a, b and c is 2:3:5, then 0.2 of the kernel value for p_j will be assigned to a, 0.3 to b and 0.5 to c. We assign a zero weightage to all other key facial poses having distance greater than a threshold. If we prefer to work with 95% confidence interval of the t kernel, then we can easily calculate the threshold accordingly. Next we discuss the results of applying the proposed method on benchmark dataset.

6. RESULTS AND DISCUSSIONS

Table 1: The accuracy of the proposed approach with GWWOF descriptor, compared to the GWOF and texture based features.

Methods	Hsieh [18]	Agarwal [1]	Proposed
Accuracy	93.6%	94.2%	96.44%

We experimented the proposed facial expression recognition scheme on the extended Cohn Kanade dataset (CK+) [22, 25]. The CK+ dataset contains 593 video sequences from 123 persons (subjects). The video sequences are of varying length and the duration varies from 10 to 60 frames (average around 19 frames). The video sequences are captured in varying lighting conditions. All the video sequences start from the neutral facial pose and the intensity of the expression increases gradually towards the peak formation of the expression at the last frame. The locations of landmark points on the faces are provided along with the dataset. Out of the 593 video sequences in the dataset only 309 are labeled as one of the six basic expressions : Anger (An), Disgust (Di), Fear (Fe), Happiness (Ha), Sadness (Sa) and Surprise (Su). Out of the 309 labeled video sequences, 45 are labeled as Anger (1022 frames), 59 are labeled as Disgust (868 frames), 25 are labeled as Fear (546 frames), 69 are labeled as Happiness (1331 frames), 28 are labeled as Sadness (547 frames) and 83 are labeled as Surprise (1329 frames) expressions. Since the proposed method is learned with supervised learning technique, we only used the 309 labeled video sequences from the CK+ dataset.

Each video sequence of the CK+ dataset is related to only one expression. In each video sequence, the neutral facial expression continues upto a certain number of frames and then the changes due to the facial expression starts. Several state-of-the-art techniques perform some pre-processing tasks by manually labeling the neutral faces in each video [9]. Since we aim to propose a real-time facial expression recognition system, such pre-processing is impossible. Rather, in the proposed method, such pre-processing is not necessary because, the neutral facial poses are automatically pruned out in the process of constructing the reduced dictionary.

We used the binary Support Vector Machine (SVM) classifier with one-against-all mechanism, for classification of the expressions. As Agarwal *et.al.* have shown that for facial expression recognition problem, binary SVM with linear kernel provides better classification compared to the SVM with radial basis function [1], we use SVM with linear kernel, for classification in the proposed approach. We conduct leave-one-out cross validation scheme for measuring the efficacy of the proposed approach. The accuracies of the proposed method and the competing methods are calculated in terms of the number of correct detections of the facial expressions.

Table 1 shows the efficacy of the proposed descriptor compared to the state-of-the-art motion features, where the GWWOF descriptor is considered, as it outperforms the Warp-GWOF descriptor. We compared the proposed approach with the two most recent techniques for facial expression recognition using motion based features. Table 2 shows a comparative measure of accuracy of the proposed approach with the two different motion descriptors introduced in the paper. Also, Table 2 shows the effect of the proposed improvement in the classification procedure over [1].

Table 2: The accuracy of the proposed approach with GWWOF and Warp-GWOF descriptors, compared to the GWOF and texture based features. Also the effect of applying the improvements over [1] is shown in the table.

Descriptors	Accuracy
GWOF [1]	94.2%
Warp-GWOF without improvement	94.5%
Warp-GWOF with improvement	95.15%
GWWOF without improvement	96.12%
Warp-GWOF with improvement	96.44%

Table 3: Confusion matrix for the proposed approach applied to CK+ dataset.

	An	Di	Fe	Ha	Sa	Su
An	44	0	0	0	1	0
Di	2	57	0	0	0	0
Fe	0	0	21	0	4	0
Ha	0	0	0	69	0	0
Sa	1	0	3	0	24	0
Su	0	0	0	0	0	83

The Tables 1 and 2 show that the involvement of WOF feature increases the accuracy of the facial expression recognition system. The reason behind the GWWOF descriptor providing better performance compared to the Warp-GWOF feature is possibly the length of the descriptor. The proposed GWWOF is a 168 dimensional vector whereas, the proposed Warp-GWOF is a 336 dimensional vector. The length of the feature vector may have been the cause of reduction of the efficacy.

Table 3 shows the confusion matrix of the proposed GWWOF descriptor, applied to the CK+ dataset. It is evident from Table 3 that, the proposed facial expression recognition system can recognize the Surprise and Happiness expressions without mistake. Because, these two expressions have no similarity with any other expressions. However, the proposed approach confuses between the Sadness and Fear expressions, as several persons show similar expressions for these two expressions.

Figure 3 shows an example to illustrate the efficacy of the GWWOF feature over the GWOF and the WOF features. We have taken sample frames from the Happiness and Disgust expressions by the same person of the CK+ dataset (the first row). This person depicts almost similar expressions for Hapiness and Disgust. The GWOF feature [1] is unable to identify the minute difference depicted by the person for these two expressions (the second row). However, the WOF feature could successfully find the minute difference between these two expressions (the third row). For Happiness, the movements of the lower portions of the cheeks are slightly upward, whereas, for Disgust expression, the movements of the lower cheeks are slightly downward. After combining the WOF feature with GWOF feature (i.e., the proposed GWWOF feature), the minute movements of the cheeks are accumulated with the involvement of the image gradient (the fourth row).

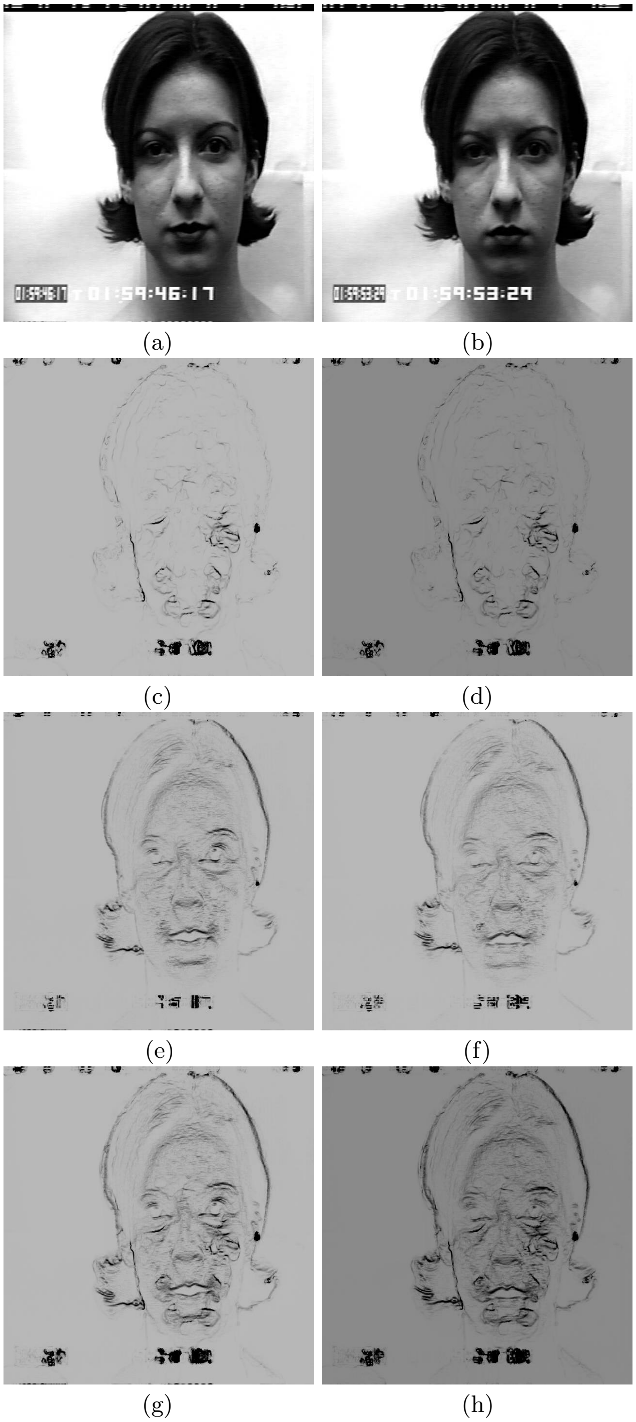

Figure 3: An example showing results of applying various motion based measures on a sample frame taken from happiness (left column) and disgust (right column) expressions of the Extended Cohn Kanade dataset (CK+) [22]. (a,b) Original frames, (c,d) GWOF measures, (e,f) WOF measures and (g,h) GWWOF measures.

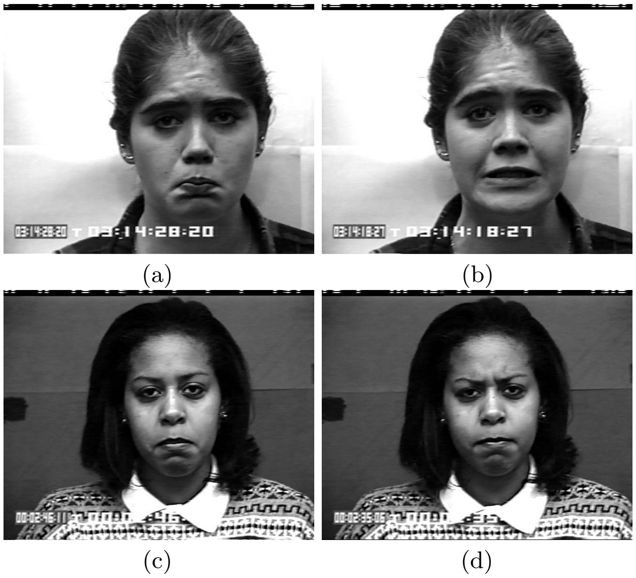

Figure 4: Examples of some failure cases by the proposed approach. First row shows a typical example of confusion of the proposed approach between Fear and Sadness expressions. (a) Sadness, detected as Fear, (b) Fear detected as Sadness, (c) Sadness detected as Anger, (d) Anger detected as Sadness.

Figure 4 shows examples of some failure cases by the proposed approach. The failures are due to the style of expressions of the person which resembles with some other expressions. The first row shows a typical example where the proposed approach confuses between the Sadness and Fear expressions. The second row shows the only examples where Sadness expression is confused with the Anger expression by the same person.

The proposed GWWOF descriptor has the same length as the GWOF feature (168 dimensions). And, [1] and the proposed method follow almost the same technique. So, the computational complexity of the proposed GWWOF descriptor based method is similar to [1]. However, the Warp-GWOF descriptor based method is computationally more complex compared to [1] and the GWWOF descriptor based method. We implemented the proposed method in $MATLAB^{TM}$ version 2008a. Agarwal *et.al.* analyzed the time complexity of the GWOF based approach and have shown that, the method in [1] takes much less time for both training and testing compared to the competing techniques. Since the proposed GWWOF based technique follows almost the same procedure (with slight modifications, which do not create any extra computational overhead) as [1] with a different descriptor of same size, the computation time for the proposed approach is exactly the same as [1]. Hence, the proposed approach provides better accuracy compared to the state-of-the-art methods, with no extra computational overhead.

7. CONCLUSIONS AND FUTURE SCOPE

This paper introduces two motion descriptors and applied them for recognizing facial expressions. We follow the classification procedure of a recent facial expression recognition approach, with two minor modifications. The proposed method has outperformed the state-of-the-art approaches in terms of accuracy of recognition. According to this study, introduction of Warp Optical Flow as a motion feature, enhances the efficacy of facial expression technique. In future, the two newly introduced motion based descriptors Warp-GWOF and GWWOF can be applied to other video analysis problems like human action recognition, event recognition, object tracking, etc. If the proposed supervised learning technique can be replaced by an efficient semi-supervised classification technique, then the proposed scheme can be used for real-time facial expression recognition, which may be useful in automatic classroom activity recognition. The proposed method cannot be applied on half-occluded faces. The proposed method can be enhanced by proposing a mathematical model for predicting the motion pattern at the occluded parts of the face, based on the motion pattern at the visible parts of the face.

8. REFERENCES

[1] S. Agarwal and D.P. Mukherjee. Facial Expression Recognition through Adaptive Learning of Local Motion Descriptor, Multimedia Tools and Applications, Springer, Doi: 10.1007/s11042-015-3103-6, pp.- 1-27, 2015.

[2] M.E. Hoque, M. Courgeon, J.C. Martin, B. Mutlu and R.W. Picard. Mach: My automated conversation coach. Proc. of UbiComp, ACM, pp. 697-706, 2013.

[3] G. Zhao and M. Pietikinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(6), pp.- 915-928, 2007.

[4] P. Ekman, W.V. Friesen and J.C. Hager. Facial Action Coding System: The Manual on CD ROM. A Human Face, Salt Lake City, 2002.

[5] H. Wang, A. Klaser, C. Schmid and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision, Springer, 103(1), pp.- 60-79, 2013.

[6] S. Mukherjee, S.K. Biswas and D.P. Mukherjee. Recognizing Human Action at a Distance in Video by Key Poses. IEEE Transactions on Circuits and Systems for Video Technology, 21(9), pp.- 1228-1241, 2011.

[7] S. Mukherjee, S.K. Biswas and D.P. Mukherjee. Recognizing Interactions Between Human Performers by 'Dominating Pose Doublet'. Machine Vision and Applications, Springer, 25(4), pp.- 1033-1052, 2014.

[8] T. Wu, S. Fu and G. Yang. Survey of the Facial Expression Recognition Research. Proc. of International Conference on Brain Inspired Cognitive System (BICS), pp.- 392-402, 2012.

[9] S. Jain, C. Hu, J.K. Aggarwal. Facial expression recognition with temporal modeling of shapes. Proc. of IEEE International Conference on Computer Vision (ICCV) Workshops, pp.- 1642-1649, 2011.

[10] F.S. Hsu, W.Y. Lin and T.W. Tsai. Facial expression recognition using bag of distances. Multimedia Tools and Applications, Springer, 73(1), pp.- 309-326, 2014.

[11] Z. Zhang, M.J. Lyons, M. Schuster and S. Akamatsu. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. Proc. of IEEE International Conference on Automatic Face and Gesture Recognition (AFGR), pp.- 454-459, 1998.

[12] T. Wu, M.S. Bartlett and J.R. Movellan. Facial expression recognition using Gabor motion energy filters. Proc. of IEEE Computer Vision and Pattern Recognition (CVPR) Workshops, pp.- 42-47, 2010.

[13] P. Martins, J. Sampaio and J. Batista. Facial Expression Recognition Using Active Appearance Models. Proc. of IEEE International Conference on Computer Vision Theory and Applications (VISAPP), pp.- 123-129, 2008.

[14] C. Shan, S. Gong and P.W. McOwan. Facial expression recognition based on Local Binary Patterns: A comprehensive study. Image and Vision Computing, 27(2009), pp.- 803-816, 2009.

[15] A. Vo and N.Q. Ly. Facial Expression Recognition Using Pyramid Local Phase Quantization Descriptor. Proc. of International Conference on Knowledge and Systems Engineering (KSE), pp.- 105-115, 2014.

[16] B. Jiang, M.F. Valstar and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. Proc. of IEEE International Conference on Automatic Face and Gesture Recognition (AFGR), pp.- 314-321, 2011.

[17] M. Tang and F. Chen. Facial expression recognition and its application based on curvelet transform and PSO-SVM. Optik - International Journal for Light and Electron Optics, 124(22), pp.- 5401-5406, 2013.

[18] C.C. Hsieh, M.H. Hsih, M.K. Jiang, Y.M. Cheng and E.H. Liang. Effective semantic features for facial expressions recognition using svm. Multimedia Tools and Applications, Springer, DOI: 10.1007/s11042-015-2598-1, pp.- 1-20, 2015.

[19] Y. Li, S. Wang, Y. Zhao and Q. Ji. Simultaneous facial feature tracking and facial expression recognition. IEEE Transactions on Image Processing, 22(7), pp.- 2559-2573, 2013.

[20] M. Bejani, D. Gharavian and N.M. Charkari. Audiovisual emotion recognition using anova feature selection method and multi-classifier neural networks. Neural Computing and Applications, Springer, 24(2), pp.- 399-412, 2014.

[21] H. Boughrara, M. Chtourou, C.B. Amar and L. Chen. Facial expression recognition based on a mlp neural network using constructive training algorithm. Multimedia Tools and Applications, Springer, 75(2), pp.- 709-731, 2016.

[22] P. Lucey, J.F. Cohn, T. Kanade and J. Saragih. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. Proc. of IEEE Computer Vision and Pattern Recognition (CVPR), pp.- 94-101, 2010.

[23] B.L. Narayan, C.A. Murthy and S.K. Pal. Maxdiff kd-trees for data condensation. Pattern Recognition Letters, Elsevier, 27(3), pp.- 187-200, 2006.

[24] T.H. Cormen, C.E. Leiserson, R.L. Rivest and C. Stein. Introduction to Algorithms. MIT Press, Cambridge, 2003.

[25] T. Kanade, Y. Tian and J.F. Cohn. Comprehensive database for facial expression analysis. Proc. of IEEE International Conference on Automatic Face and Gesture Recognition (AFGR), pp.- 46-53, 2000.