# Recognizing Expressions from Face and Body Gesture by Temporal Normalized Motion and Appearance Features

Shizhi Chen[1]
schen21@ccny.cuny.edu
YingLi Tian[1],*
ytian@ccny.cuny.edu
Qingshan Liu[2]
qsliu@nuist.edu.cn
Dimitris N. Metaxas[3]
dnm@cs.rutgers.edu

* Corresponding author

[1] Department of Electrical Engineering,
The City College of New York, USA
[2]School of Information & Control
Engineering, Nanjing University of
Information Science and Technology,
China
[3]DepartmentofComputer Science,
Rutgers University,
USA

## Abstract

Recently, recognizing affects from both face and body gestures attracts more attentions. However, it still lacks of efficient and effective features to describe the dynamics of face and gestures for real-time automatic affect recognition. In this paper, we combine both local motion and appearance feature in a novel framework to model the temporal dynamics of face and body gesture. The proposed framework employs MHI-HOG and Image-HOG features through temporal normalization or Bag of Words to capture motion and appearance information. The MHI-HOG stands for Histogram of Oriented Gradients (HOG) on the Motion History Image (MHI). It captures motion direction and speed of a region of interest as an expression evolves over the time. The Image-HOG captures the appearance information of the corresponding region of interest. The temporal normalization method explicitly solves the time resolution issue in the video-based affect recognition. To implicitly model local temporal dynamics of an expression, we further propose a bag of words (BOW) based representation for both MHI-HOG and Image-HOG features. Experimental results demonstrate promising performance as compared with the state-of-the-art. Significant improvement of recognition accuracy is achieved as compared with the frame-based approach that does not consider the underlying temporal dynamics.

## 1 Introduction

Automatic affective computing has attracted increasingly attention from psychology, cognitive science, and computer science communities due to its importance in practice for a wide range of applications, including intelligent human computer interaction[2, 24, 25], law enforcement, and entertainment industries [32] etc.

Human affective state is complicated and sometimes can be very subtle which may not be detected just from the facial expressions. Fortunately, we observe affective state naturally through multiple modalities, such as facial expression, body gesture, audio signal etc. These observations through different modalities provide complementary information on the affective states. Moreover, the affective behaviors are continuous. Hence, how these affective behaviors evolve over the time can also provide additional information on the observed affective state. This paper focuses on how to efficiently extract the dynamic information of continuous affective behaviors from multiple modalities, which are directly related to this special issue on affect analysis in continuous input.

Many algorithms and systems have been proposed in the past for automatic facial expression recognition [14, 30, 41]. Generally, these methods can be categorized into two categories: image-based approaches [17, 21, 36] and video-based approaches [6, 15, 28, 34, 38, 40, 50, 51].

Lanitis *et al.* [21] performed statistical analysis on static face images to model complicated facial expression. The model captures both shape and appearance features of facial expressions by considering different sources of variations, such as lighting changes, different person identity etc. Guo and Dyer [17] applied Gabor filter and large margin classifiers to recognize facial expressions from face images as well. Both papers classify face images into six basic universal expressions, *i.e.*, “Disgust”, “Fear”, “Happiness”, “Surprise”, “Sadness” and “Anger” [12]. Tian *et al.* [36] combined both geometry and appearance features to recognize action units (AUs) of the Facial Action Coding System (FACS), which are proposed by Ekman and Friesen [13].

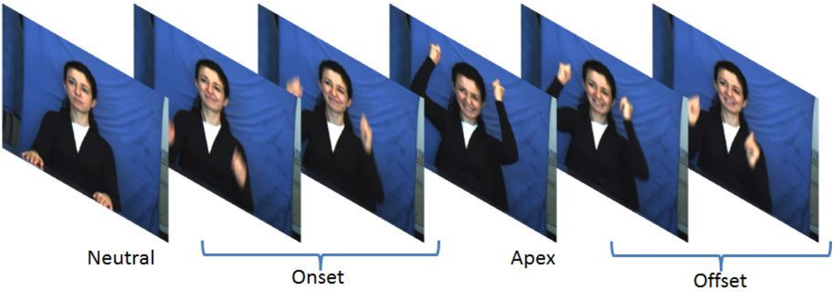


**Figure 1**: The temporal evolution of an expression of “Happiness”**.**

Video-based approaches usually incorporate temporal evolution of facial expressions to improve affect recognition performance. Figure 1 shows the temporal evolution of a “Happiness” expression over the four temporal phases. As compared with the affect recognition using apex frames alone, temporal dynamics in whole expression cycle can capture more subtle changes in a person’s affective state. Hence the affect recognition can be more robust to the noise and improve the performance. Yang *et al*. [50, 40] utilize dynamic Haar-like features and similarity features to model temporal variations of facial events. Zhao and Pietikainen [51] extended 2D local binary pattern (LBP) to volume local binary pattern (VLBP) to explicitly represent the local temporal evolution of facial expressions. Tong *et al.* [38] employ a dynamic Bayesian network (DBN) to systematically account for temporal evolutions for facial action unit recognition.

However, none of these approaches utilizes the body gestures for affect recognition [43, 45, 49]. As the psychology studies [1, 26] suggest, both face and body gesture carry significant amount of affect information. Gunes and Piccardi [15] recently proposed a framework to incorporate both face and body gesture together for affect recognition. They apply the HMM (Hidden Markov Model) video based approach and the maximum voting of apex frames approach for the affect recognition through both face and gesture modalities. Although excellent performance has been achieved, the design of feature extraction is quite complicated for real-time processing, which involves optical flow, edginess, geometry features, and comparison with the neutral frame etc. The feature extraction also involves tracking of several facial components, hands, and shoulders.

In this paper, we propose several types of features which are efficient and effective to describe both appearance and motion information of face and body gesture. We further model the temporal variations in an expression cycle through a temporal normalization approach.

Figure 2 shows an overview of our proposed framework for the affect recognition. The framework includes three main components: (a) feature extraction for face and body gestures; (b) Feature representation; and (c) affect recognition. We use the bi-modal face and body benchmark database (FABO) [16] to evaluate our framework. Overall, the work introduced in this paper offers the following main contributions to effectively and efficiently recognize affects by combining face and body gestures.

- In feature extraction (see Figure 2(a)), we develop and employ several types of simple features which can be extracted in real-time: MHI-HOG and Image-HOG, to capture both motion and appearance information of expressions. MHI-HOG stands for Histogram of Oriented Gradients (HOG) on the Motion History Image (MHI) [3, 35, 44]. It captures motion direction and speed of a Region of Interest (ROI) as an expression evolves over the time. Image-HOG [11] captures the appearance information of the corresponding ROI. By combining only MHI-HOG and Image-HOG, our features achieve comparable performance with the-state-of-the-art.
- To effectively represent the features, we propose two approaches: (1) temporal normalization algorithm (TN), and (2) the Bag of Word algorithm (BOW), as shown in Figure 2(b). The TN algorithm is applied over a complete expression cycle, i.e. from onset, apex to offset frames, to explicitly describe the dynamics of facial expression. The Bag of Words representation models one complete expression cycle as probability distributions over the MHI-HOG and the Image-HOG words. We construct one Bag of Word histogram for each expression cycle, i.e., group features in the same expression cycle to the same bag.
- For affect recognition shown in Figure 2(c), we extract features of both face and body gesture modalities from a single camera to capture both facial features and body gesture features, rather than the conventional approaches which use multiple cameras to extract different modalities. For example, Gunes and Piccardi [15] extract facial features and body gesture features from two cameras respectively.

Experimental results indicate the effectiveness and efficiency of the proposed approaches to combine MHI-HOG and Image-HOG for the affect recognition. Different from most existing approaches, which usually extract apex frames from the temporal segmentation results for frame-based affect recognition, we use the whole expression cycle, i.e., onset, apex, and offset for video-based affect recognition by applying the temporal normalization method or Bag of Words model. Intuitively, the dynamics captured from the complete expression cycle can help affect recognition. Our experimental results confirm this intuition.

Compared to the Histogram of Optical Flow (HOF) feature, the MHI-HOG is more computationally efficient due to the underlying technology, i.e., optical flow feature and the gradients of the motion history image. The paper [44] provides computation comparison between several widely used optical flow methods and the gradient of MHI method. Their experimental results suggest that the optical flow based HOF feature extraction can be 10 times slower than the MHI gradient calculation.

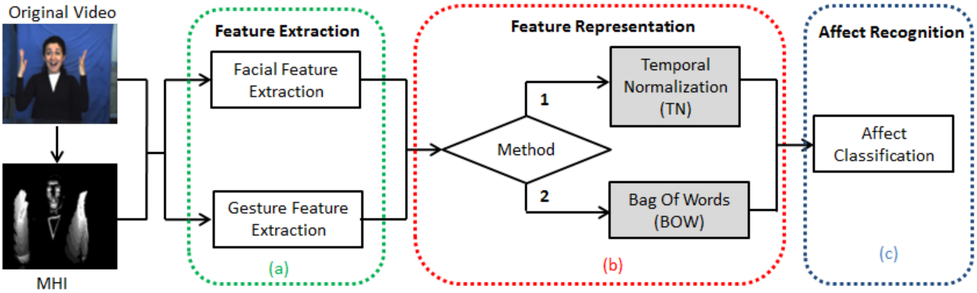


**Figure 2**: An overview of the proposed framework for affect recognition. (a)Feature extraction for face and body gesture; (b) Feature representation using (1) temporal normalization (TN), and (2) Bag of Words (BOW) approach; (c) Affect recognition.

This paper is organized as follows. The next section describes related work. Section 3 summarizes the proposed method including feature extraction for both face and body gestures, the two approaches we proposed for feature representation (*i.e*. TN and BOW). Section 4 describes experiments to evaluate the effectiveness of the proposed framework and presents the results. Section 5 presents conclusions and discussion.

# 2 Related Work

This section provides related work on facial expression recognition through the multiple modalities, particularly on face and body gesture [6, 15, 34]. We also review the related approaches which utilize temporal dynamics of the expressions [8, 15, 27, 29].

## 2.1 Expression Temporal Segmentation

Temporal dynamics of facial expressions is crucial for facial behavior interpretation [33]. An expression is a dynamic event, which evolves from the temporal phases of onset, apex, offset and back to neutral. Among these four temporal phases, it is generally agreed that the apex carries the maximum discriminative power because an expression reaches the maximum spatial extension at this temporal phase [15, 40, 51].

Both Cohen [8] and Otsuka [27] apply Hidden Markov Model (HMM) to temporally segment expression. The transition probabilities from one temporal phase to another are learned during training process. Then these transition probabilities are employed to predict the temporal phases for an expression cycle. Inspired by the Ekman's proposal that each facial expression can be composed of 46 action units in the Facial Action Coding System (FACS), Otsuka and Ohya [27] employ motion information around the right eye and the mouse as input to the HMM model. The motion is estimated by the gradient-based optical flow algorithm [18].

Pantic and Patras [28, 29] have temporally segmented facial action units (AUs) by tracking 15 facial key points from the profile image followed by the description using geometric features. These 15 facial landmarks are chosen such that they are discriminative enough to distinguish temporal phases of facial action unit. For example, the tip of the nose and the top of the forehead are chosen as the facial points due to their stability during tracking. Then they have employed particle filtering [31] to track these 15 facial key points and calculate the corresponding geometry features. However, the tracking algorithms are usually sensitive to lighting, rotation variance, which can potentially degrade the temporal segmentation results.

Chen *et al*. [5] recently propose Motion Area and Neutral Divergence features to temporally segment expression into neutral, onset, apex and offset phases. Neutral divergence feature measures the pixel intensity different between the current image frame and the frame with the neutral expression [5]. The proposed features do not require tracking. Hence it is more robust to the variance of lighting and poses. Motion Area captures the motion intensity of an expression during different temporal phases. Intuitively, the onset and the offset phases generate larger number of motion pixels as compared with the neutral and the apex phase. Neutral Divergence feature is employed to further separate the onset from the offset, and the neutral from the apex phases in the video.

Temporal dynamics are also proven to be important for gesture recognitions [42, 47]. Alon *et al*. [42] proposed a unified framework for simultaneously performing temporal segmentation and recognition of hand gestures. The method is able to automatically detect when a gesture begins and ends. Kim *et al*. [47] proposed differential observation probability to automatically detect starting and ending points of upper body gesture. Based on these temporal segments of body gestures, accumulative hidden Markov model is employed to recognize body gestures.

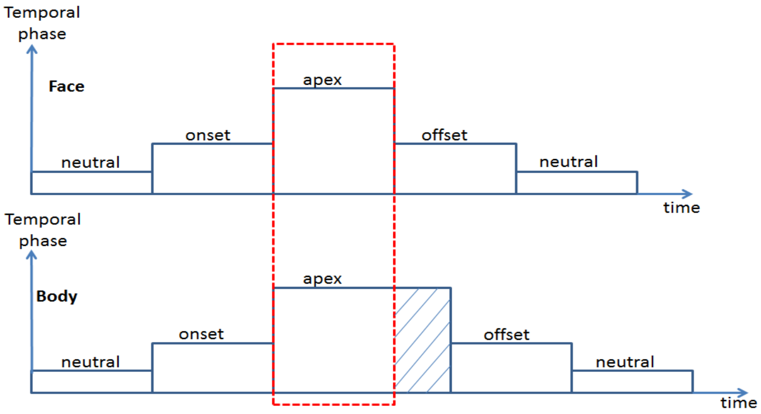


**Figure 3**: Selective fusion method [15] to combine both face and body gesture modalities. However, some apex frames from body gesture modality (bottom shaded area) are discarded since the corresponding frames from the face modality do not have the same temporal phase, *i.e*., apex.

## 2.2 Expression Recognition by Multi-Modalities

As human communicate with each other naturally through multiple modalities, the future human computer interaction (HCI) is generally agreed to be based on multiple modalities [30, 41], such as auditory signal, visual signal of facial expression, and visual signal of body gesture. Several approaches on affect recognition based on multi-modalities have been proposed [15, 16, 34, 46, 48].

Kapoor and Picard [46] proposed a multi-modal Gaussian Process approach to classify interest or disinterest in children trying to solve a puzzle problem in a computer. The approach integrates multi-modal sensory information of facial expression, postural shift, and learner's activities in the computer. The reported multi-modal performance significantly outperforms that of any individual modality.

Gunes and Piccardi [15, 16] have collected a benchmark database, *i.e.*, FABO (Face and Body gesture) database, for the evaluation of face and body gesture on the affect recognition. They select the video frames which share the same temporal phase, i.e., the apex phase, in both modalities for the expression recognition. This selective fusion method is a simple and effective way to address the misalignment issue on the temporal dynamics of multiple modalities. However, the selective fusion method has to discard some apex frames even though these apex frames are equally informative as the other apex frames. As shown in Figure 3, some apex frames from body gesture modality (bottom shaded area) are discarded because the corresponding frames from the face modality do not have same temporal phase, *i.e.*, apex. Gunes and Piccardi [15] also utilize a few hundred of different features to describe both modalities and several trackers to track different body parts such as hands, shoulders, eyebrow etc.

The Bag of Words model is a popular model in the text analysis literature [19, 37]. The basic idea is to represent a document with a collection of order-less words. In other words, each document is represented as a histogram over words. Then a classifier is trained and tested on the histograms. The spatial order between different words in the document is not captured in the resulted histogram. That is why the method is only an order-less collection of words.  It has also recently become popular in computer vision area, especially in the image retrieval and the image classification field [7, 10, 22, 23]. They usually have extracted local features of key points as words, and treat an image as a document. The image is represented by a bag of local features, *i.e.*, the histogram of features. Despite its simplicity, the Bag of Words model usually achieves good performance.

Shan *et al*. [34] apply the Bag of Words model over the spatial temporal interest points features to describe the body gesture for video based affect recognition. Then the authors combine both face and body gesture modalities using the Canonical Correlation Analysis (CCA) in order to maximize the mutual correlation over the two modalities. Due to the order-less characteristic and the gradient features extracted within a cuboid around each spatial temporal interest point, the Bag of Words model can still capture local temporal dynamics of an expression. The histogram of gradients features from the spatial-temporal cuboid represents the local movement in both spatial and temporal domain.

# 3 Proposed Method

As shown in Figure 2, our proposed framework of the expression recognition from both face and body gesture modalities includes three main components: feature extraction for face and body gestures, feature representation, and affect recognition. This section describes the three components in details. Note that we only consider the upper body gesture instead of the full body.

Figure 4 shows the detailed flowchart of our proposed framework. For each input video stream, the MHI is first calculated. As shown in Figure 4(a), Facial feature extraction includes facial landmark points tracking and histogram of gradients (HOG) descriptor over the tracked points. To extract body gesture features, we first track hand and head in each frame. Then their position, motion, and appearance information are extracted to form the body gesture's feature.

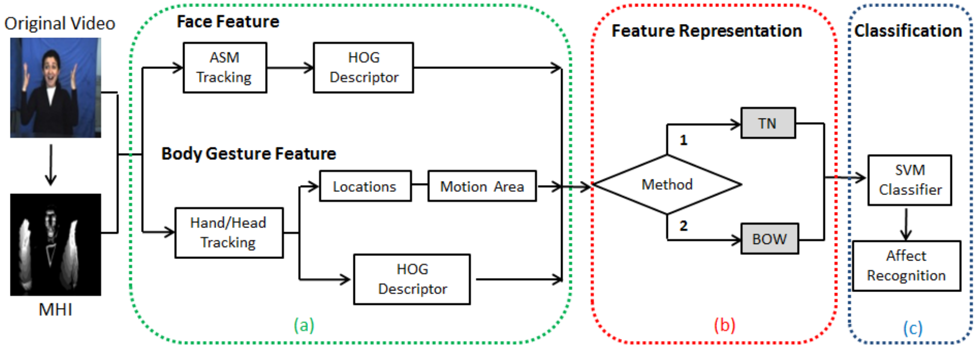


**Figure 4**: Flowchart of the proposed approaches for expression recognition. (a) Feature extraction for face and body gesture modalities. (b) Feature representation using temporal normalization (TN) and the bag of words (BOW). (c) Classification using SVM;

In order to effectively represent both face and body gesture for expression recognition, we employ two approaches: (1) temporal normalization (TN) approach; and (2) Bag of Words (BOW) approach, as shown in Figure 4(b). The TN approach explicitly incorporates the temporal dynamics from both face and gesture modalities for expression recognition. The second approach utilizes the Bag of Words model to represent HOG features.

Finally, we concatenate both face feature and body gesture feature together to the SVM classifier for the affect recognition as shown in Figure 4(c). In our system, the extraction and representation of both face and body gesture features are very simple and efficient. The temporal normalization of these features, i.e., the position, the appearance, and the motion, can efficiently describe the dynamics of facial expression for the affect recognition.

ASM facial landmark points tracking, skin color detection, MHI images as well as the HOG descriptors can all be executed in real time.

## 3.1 Feature Extraction

### 3.1.1 Facial Feature

As shown in Figure 4(a), there are two steps to extract the facial features. The first step is to track the facial landmark points using the ASM (Active Shape Model) model [9, 39] on the original video as shown in Figure 5(a). The ASM model applies principal component analysis (PCA) to constraint the global shape of a face. The shape of a face is represented by the concatenated *x* and *y* coordinates of every facial landmark point. During tracking, an iteration approach is used to fit the face under the global shape constraint. Interested readers can refer to the paper [9] for the details of the ASM model. We directly used the shape model provided by Wei [39] to track the face. The shape model has 53 facial landmark points which are not at the face boundary. Since the face boundary points are not discriminative over different facial expressions, we extract features only from these 53 points.

The second step is to extract the Image-HOG and the MHI-HOG descriptors of the selected facial landmark points. Histogram of Gradients (HOG) has been successfully employed in human detection [11]. The key idea is that the appearance usually can be characterized well with the distribution of local intensity gradient and edge directions over spatial and orientation domain. HOG feature is implemented by dividing a local patch to several "cells" or sub-patches. Within each sub-patch, a histogram over gradient directions is then extracted. The concatenated orientation histograms from all sub-patches form the HOG feature.

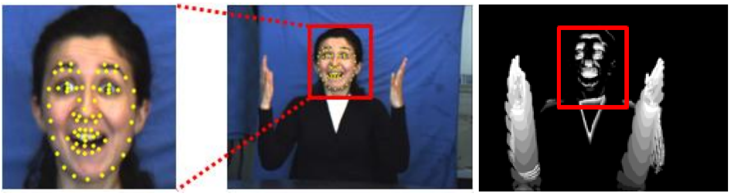


**Figure 5**: (a) ASM facial landmark points tracking; (b) MHI Image

We extract a fixed size ($n$ by $n$) patch around each interest point as the patch on the original image and the MHI image to calculate the Image-HOG and the MHI-HOG feature respectively. The patch is then divided into $m$ by $m$ sub-patches with the number of orientation bin equal to $b_i$ and $b_m$ for the Image-HOG and the MHI-HOG. Hence, the feature dimension of the Image-HOG and the MHI-HOG descriptors are $m^2 * b_i$ and $m^2 * b_m$ respectively. Following the HOG design guideline in [11] and our experimental observations, we set $n$=48, $m$=3, $b_i$=6, and $b_m$=8.

Motion History Image (MHI) is a compact representation of a sequence of motion movement in a video [3, 5, 35]. Pixel intensity is a function of the motion history at that location, where larger values correspond to more recent motion. The intensity at pixel ($x$,$y$) decreases gradually until a specified duration $\tau$. The construction process of the MHI image can be best described using the equation (1) below.

$$MHI_\tau(x,y,t) = D(x,y,t) \times \tau + [1 - D(x,y,t)] \times U[MHI_\tau(x,y,t-1) - 1] \times [MHI_\tau(x,y,t-1) - 1], \qquad (1)$$

Where $U[x]$ is a unit step function. $t$ represents the current video frame index. $D(x,y,t)$ is a binary image of intensity difference between the current frame and the previous frame. $D(x,y,t)$ is assigned to 1 if the intensity difference is larger than a threshold $V_{TH}$. Otherwise, it is 0. $V_{TH}$ should be slightly larger than 0 to remove intensity noise of pixels. In our experiments, $V_{TH}$ =25. $\tau$ is the maximum motion duration. That means only previous $\tau$ frames are used in constructing the current MHI image frame. From the observation, the MHI formed with $\tau$=10 shows clear motion trajectory of body gestures. The MHI image is then scaled to an 8-bit gray image. Figure 5(b) shows an example of the MHI image.

While the MHI image captures motion information, the original image provides the corresponding appearance information. The MHI-HOG and the Image-HOG can provide complementary information for the expression recognition.

### 3.1.2 Body Gesture Feature

Figure 4(a) shows the flowchart of body gesture feature extraction. A simple skin color-based hand tracking method is first applied to detect hand regions as shown in Figure 6(a) [20]. The center position of the head is extracted based on the ASM facial landmark points, as shown in Figure 6(b). Then we employ the center points of the hand and head regions, with reference to the neutral frame's corresponding positions, to describe the location of the hands and the head respectively. The neutral frame is the frame in which subject shows neutral expression. The hands and head positions are further normalized with the subject's height, which is measured from the center of the head to the image bottom on the neutral frame.

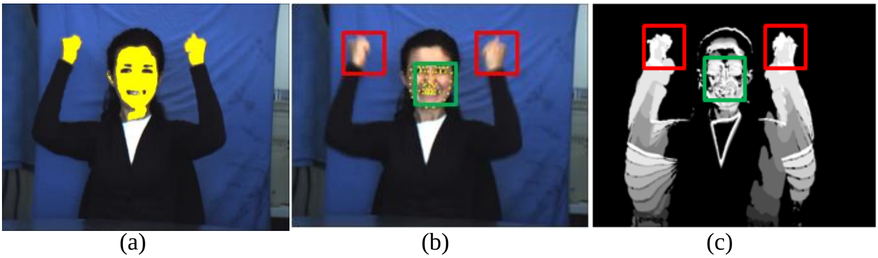


**Figure 6**: (a) Hand tracking by skin color-based tracker; (b) Position of hands using skin color tr acking and position of head using ASM model; (c) Extract motion areas of hand and head regions.

Motion areas of the hands and head regions are measured by counting the number of motion pixels from the MHI image within an *N* x *N* size window at each center, as shown in Figure 6(c). In our implementation, the motion pixel is defined as any non-zero pixel on the MHI image. The window size is chosen so that it is large enough to enclose the hand and head part. From the observation, we set *N=80*.

The MHI-HOG and the Image-HOG of both hand regions are extracted in the following steps. First, we select the uniform grid interest points within both hands' skin regions, which are also within the *N* x *N* size window at each hand's center. Second, we extract the Image-HOG and the MHI-HOG descriptors for each selected skin interest point as described in the facial feature extraction section.

The position and the motion area of the hand and head regions model their trajectories and the motion intensity. The Image-HOG and the MHI-HOG of both hands further describe their appearance, motion direction and speed.

## 3.2 Feature Representation

### 3.2.1 Bag of Words

The Bag of Words [7, 10, 22, 23] models an image or a video as a collection of visual key features without any order. It is represented by the histogram of visual key features, and generally follows three steps, i.e., codebook formation, vector quantization, and histogram generation. The interested readers are referred to the paper [10] for the detailed description. The codebook is a dictionary of visual key features or visual words, which are the representative feature vectors obtained from the training data. These representative feature vectors can be found simply by performing the k-mean clustering. The larger codebook size typically yields better performance [10]. However, as the codebook size increases, the performance eventually saturates. After obtaining the codebook of representative feature vectors, the vector quantization of a feature vector is to find the representative feature vector in the codebook, which has the smallest Euclidean distance to the feature vector as its representation. In other words, every feature can be vector quantized to one of the representative feature vectors in the codebook. Then an image or a video sequence can be represented by the histogram of these representative feature vectors. Such histogram representation is called the Bag of Words (BOW) representation. The BOW model usually achieves good performance over various tasks, such as image retrieval, and image classification *etc*.

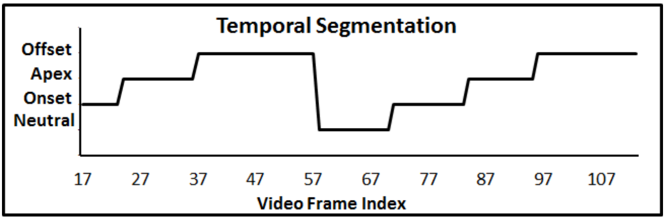


**Figure 7**: A sample temporal segmentation of an expression.

**Face Feature Representation**: we first generate a codebook with the size of 200 visual words for the Image-HOG and the MHI-HOG respectively. Based on our observation, the codebook size of 200 achieves good performance while keeping the feature dimension relatively small. Then we perform vector quantization to assign the Image-HOG descriptor and the MHI-HOG descriptor to the Image-HOG visual word and the MHI-HOG visual word with the smallest Euclidean distance. We construct a histogram over the Image-HOG visual words and the MHI-HOG visual words respectively for one complete expression cycle, i.e., from the onset to the offset. Figure 7 shows a sample manually annotated temporal segmentation of an expression [15, 16]. Finally, we concatenate both histograms together to form the final facial feature descriptor.

**Body Gesture Feature Representation**: we construct a histogram for the Image-HOG visual words and the MHI-HOG visual words respectively over a complete expression cycle. Since the HOG features are used to describe both hand regions, which have relative small variations as compared with the face. So the codebook sizes in our experiments are set as 80 for both Image-HOG and MHI-HOG. Then we concatenate histograms of both MHI-HOG and Image-HOG along with the position and motion area of hands and head as the final body gesture feature representation.

Even though the Bag of Words model is an order-less collection of the MHI-HOG and the Image-HOG features, it can still models the local temporal dynamics. That is because that the MHI image has kept the motion history from previous frames, and the gradient of the MHI image model the local motion direction and speed.

### 3.2.2 Temporal Normalization

Time resolution of expressions can be different for different subjects or even same subject at different time. One way to handle this issue is to apply the temporal normalization, over a complete cycle of an expression, *i.e.*, from the onset to the offset. In addition, the temporal normalization allows us to explicitly incorporate temporal dynamics of the expression into the feature representation.

The temporal normalization can be accomplished through the linear interpolation of each frame's feature vector $\boldsymbol{V}_{fr}$ along the temporal direction uniformly in the whole expression cycle. The expression cycle is not required to be pre-segmented into onset, apex, and offset phases. As long as we know the starting frame of the onset, and the ending frame of the offset, a single linear interpolation can be performed over the whole expression cycle, which is defined by the starting frame of onset and the ending frame of the offset. For the simplicity, we did not perform automatic temporal segmentation. Instead, we use the manual labelling of the starting frame of onset and the ending frame of the offset in the FABO database. Figure 8 illustrates the linear interpolation of frame feature vectors at $j^{th}$ dimension over a complete expression cycle. Note that the red square data illustrate the original values at $j^{th}$ dimension of frame feature vectors, and the blue circle data illustrate the interpolated values at the same dimension. By repeating the same linear interpolation procedures at other dimensions of frame feature vectors, an expression cycle can be temporally normalized to a fixed number of frames $N_{fr}$, from our experimental observation, $N_{fr}$ is set to 30 in our experiments.

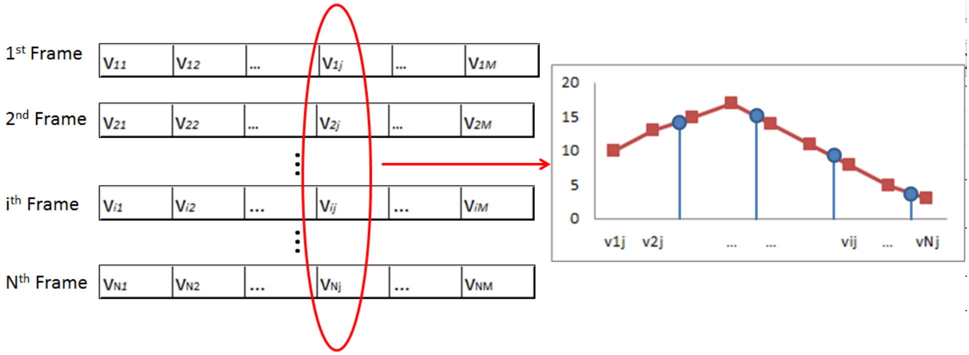


**Figure 8**: Temporally normalize frame feature vectors by showing the linear interpolation of frame feature vector at $j^{th}$ dimension along the temporal direction uniformly over a complete expression cycle. Note that the red squares illustrate the original values at $j^{th}$ dimension of frame feature vectors, and the blue circles illustrate the interpolated values at the same dimension.

**Face Feature Representation:** we concatenate the Image-HOG of all 53 facial landmark points, which results the feature dimension of 2862 for each frame. Similarly, the MHI-HOG is also concatenated together for each frame, which results the feature dimension of 3816. We then reduce the feature dimension of the concatenated Image-HOG and the concatenated MHI-HOG down to $D_I$ and $D_M$ respectively by employing principal component analysis (PCA) before concatenating these two feature types together as the feature vector for each frame. From experimental observations, we set $D_I$ and $D_M$ to 40. The principal space of the concatenated Image-HOG and the concatenated MHI-HOG are obtained separately from the training videos.

The final concatenated feature vector is frame feature vector from face modality. Then we temporally normalized the face's frame feature vectors in an expression cycle by performing linear interpolation as discussed previously.

**Body Gesture Feature Representation:** we construct a histogram over the Image-HOG visual words and the MHI-HOG visual words respectively for each frame. Then we perform the PCA to reduce the dimension of the Image-HOG histogram to $D_I$ and the dimension of the MHI-HOG histogram down to $D_M$. Based on our observation, when $D_I$ is greater than 4 and $D_M$ is greater than 1, the recognition performance remains same. Hence, we set $D_I$ to 4, and $D_M$ to 1. In order to eliminate the variance caused by different subjects, we further subtract the neutral frame's MHI-HOG and Image-HOG histograms from that of each frame in an expression cycle.

Finally, we concatenate the dimensionally reduced Image-HOG and MHI-HOG histograms with the position and the motion area of the hand and head regions as the frame feature vector for body gesture modality. The linear interpolation method is then employed to temporally normalize the frame feature vectors to a fixed size.

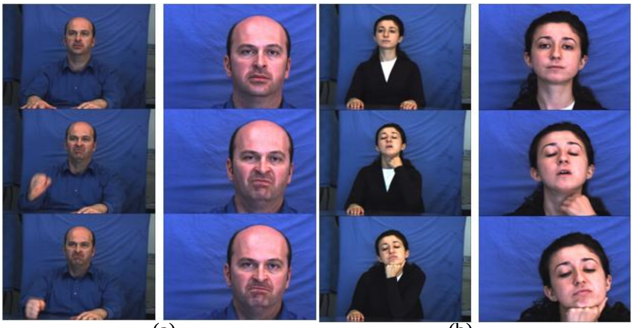
### 3.3 Expression Classification from Face and Body Gesture Modalities

We employ SVM with the RBF kernel using one vs. one approach as our multi-class classifier [4]. SVM is to find a set of hyper-planes, which separate each pair classes of data with maximum margin, then use the maximum vote to predict an unknown data's class. In our experiments, the feature data, *i.e*. the input features to the SVM, are facial features, body gesture features, or feature concatenation of both modalities of a complete expression cycle.

## 4 Experiments

### 4.1 Experimental Setups

The database we used is a bi-modal face and body benchmark database FABO [16]. The database consists of both face and body recordings using two cameras respectively. Subjects were provided with a short scenario, for instance, "the lecture is the most boring one". They were asked what they would do when performing their expression. The expressions of Face recording and body recording are labelled independently, and voted by six observers for the labels. We only select videos, which have same expression labels from both face and body recordings. And the selected videos need at least 3 votes from the six observers on both face and body recordings. Two sample videos from the database are shown in Figure 9(a) and 9(b).



| category | number of videos |
|---|---|
| Disgust | 27 |
| Fear | 22 |
| Happiness | 28 |
| Surprise | 13 |
| Sadness | 16 |
| Anger | 60 |
| Anxiety | 21 |
| Boredom | 28 |
| Puzzlement | 46 |
| Uncertainty | 23 |

(a) (b) (c)

**Figure 9**: (a) sample images from an "Anger" expression video in FABO database recorded by body (left) and face (right) camera; (b) sample images from a "Boredom" expression video in FABO database recorded by body (left) and face (right) camera; (c) a table shows the number of videos for each category we select for our experiments.

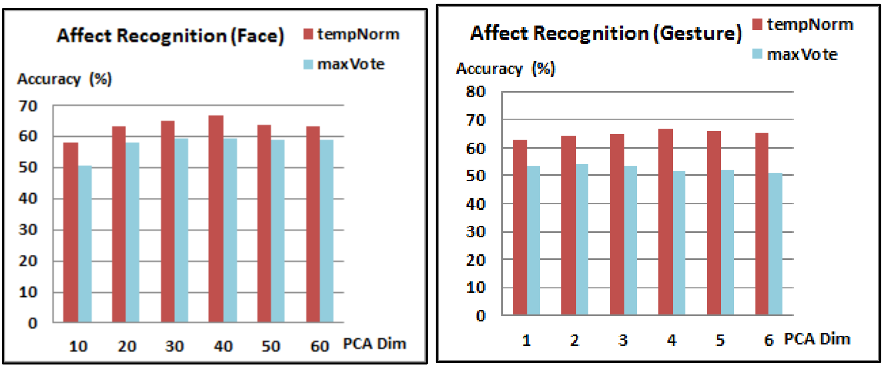
Since it is not practical to use both face and body cameras for the real world applications, we only choose body recording, which contains both face and body gesture information. The total number of videos we selected is 284. These videos include both basic and non-basic expressions. Basic expressions are “Disgust”, “Fear”, “Happiness”, “Surprise”, “Sadness” and “Anger”. Non-basic expressions are “Anxiety”, “Boredom”, “Puzzlement” and “Uncertainty”. The corresponding number of videos for each category is shown in Figure 9(c). Each video contains 2 to 4 expression cycles. Videos in each expression category are randomly separated into three subsets. Two of them are chosen as training data. The remaining subset is used as testing data. No same video appears for both training and testing, but same subject may appear in both training and testing sets due to the random separation process.

Three-fold cross validation is performed over all experiments. The average performances are reported in the paper.

## 4.2 Experimental Results

### 4.2.1 Expression Dynamics

To demonstrate the advantages of expression dynamics in the affect recognition, we compare the temporal normalization approach, which incorporates the expression dynamics, to the apex frame-based approach, which uses the maximum voting of apex frames without considering the expression dynamics. Both face and body gesture modalities are evaluated.

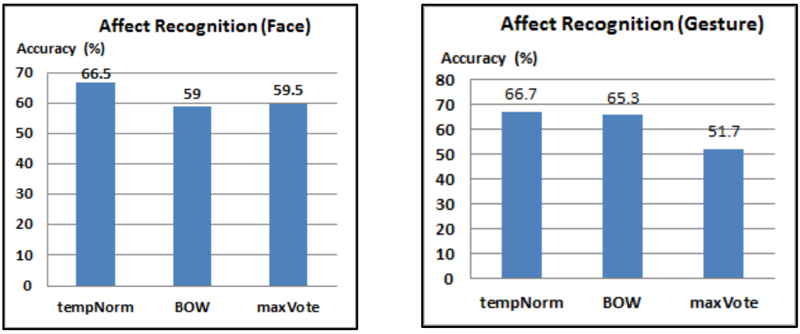


(a) (b)

**Figure 10**: Compare the temporal normalization method (tempNorm) to the maximum voting of apex frames approach (maxVote) in affect recognition through (a) face modality; and (b) gesture modality.

As Figure 10 shows, our video-based temporal normalization approach achieves significant improvement as compared with the maximum voting of apex frames approach, *i.e*. frame-based, for both face and body gesture modalities. The average accuracy gained is more than 5% and 12% respectively for the face and the body gesture.

For both Image-HOG and MHI-HOG features, we also investigate the effects of PCA dimension on the affect recognition performance. For the face modality, the PCA dimension in Figure 10(a) is the reduced dimension of both Image-HOG and MHI-HOG features. The best performance using the facial features is achieved when the PCA dimension equal to 40 for the Image-HOG and the MHI-HOG respectively, as shown in Figure 10(a). For the body gesture modality shown in Figure 10(b), the PCA dimension is referring to the reduced dimension of the Image-HOG, while the MHI-HOG's dimension is always reduced to 1. The best body gesture performance is achieved when the PCA dimension is 4.



(a) (b)

**Figure 11**: Compare the temporal normalization method (tempNorm), Bag of Words method (BOW) and the maximum voting of apex frames approach (maxVote) on the affect recognition through (a) face modality; and (b) body gesture modality.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Anger | 44 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 2 | | 92% |
| Anxiety | 0 | 10 | 0 | 3 | 0 | 0 | 0 | 4 | 1 | 0 | | 56% |
| boredom | 3 | 0 | 10 | 0 | 0 | 0 | 0 | 5 | 0 | 0 | | 56% |
| Disgust | 3 | 0 | 0 | 18 | 0 | 0 | 0 | 1 | 0 | 0 | | 82% |
| Fear | 3 | 0 | 0 | 0 | 6 | 1 | 2 | 1 | 1 | 0 | | 43% |
| Happiness | 1 | 0 | 0 | 0 | 0 | 17 | 0 | 0 | 0 | 0 | | 94% |
| Surprise | 0 | 0 | 0 | 0 | 0 | 6 | 2 | 0 | 0 | 0 | | 25% |
| Puzzlement | 7 | 2 | 6 | 1 | 0 | 0 | 0 | 28 | 0 | 1 | | 62% |
| Sadness | 3 | 1 | 2 | 0 | 0 | 2 | 0 | 2 | 2 | 0 | | 17% |
| Uncertainty | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 8 | | 50% |

(a)

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Anger | 44 | 0 | 0 | 0 | 2 | 0 | 0 | 1 | 0 | 1 | | 92% |
| Anxiety | 0 | 8 | 0 | 0 | 1 | 1 | 0 | 8 | 0 | 0 | | 44% |
| boredom | 0 | 0 | 12 | 0 | 0 | 0 | 0 | 3 | 3 | 0 | | 67% |
| Disgust | 1 | 1 | 0 | 8 | 5 | 3 | 0 | 3 | 0 | 1 | | 36% |
| Fear | 0 | 0 | 2 | 0 | 6 | 4 | 0 | 0 | 2 | 0 | | 43% |
| Happiness | 4 | 1 | 2 | 0 | 0 | 11 | 0 | 0 | 0 | 0 | | 61% |
| Surprise | 0 | 0 | 2 | 0 | 0 | 2 | 3 | 0 | 0 | 1 | | 38% |
| Puzzlement | 0 | 0 | 0 | 1 | 3 | 0 | 0 | 41 | 0 | 0 | | 91% |
| Sadness | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 7 | 2 | | 58% |
| Uncertainty | 3 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 0 | 9 | | 56% |

(b)

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Anger | 44 | 0 | 0 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | | 92% |
| Anxiety | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 11 | 0 | 0 | | 39% |
| boredom | 0 | 0 | 11 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | | 61% |
| Disgust | 5 | 1 | 0 | 11 | 2 | 2 | 0 | 1 | 0 | 0 | | 50% |
| Fear | 2 | 0 | 0 | 1 | 8 | 2 | 0 | 1 | 0 | 0 | | 57% |
| Happiness | 3 | 0 | 0 | 0 | 0 | 14 | 0 | 1 | 0 | 0 | | 78% |
| Surprise | 1 | 0 | 0 | 0 | 0 | 5 | 2 | 0 | 0 | 0 | | 25% |
| Puzzlement | 6 | 6 | 3 | 1 | 1 | 0 | 0 | 27 | 0 | 1 | | 60% |
| Sadness | 0 | 1 | 2 | 1 | 1 | 0 | 0 | 4 | 3 | 0 | | 25% |
| Uncertainty | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 7 | | 44% |

(c)

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Anger | 41 | 0 | 0 | 0 | 2 | 0 | 0 | 2 | 0 | 3 | | 85% |
| Anxiety | 1 | 4 | 0 | 0 | 0 | 2 | 0 | 11 | 0 | 0 | | 22% |
| boredom | 0 | 0 | 9 | 0 | 0 | 0 | 0 | 5 | 4 | 0 | | 50% |
| Disgust | 3 | 1 | 0 | 6 | 5 | 4 | 0 | 3 | 0 | 0 | | 27% |
| Fear | 0 | 0 | 2 | 1 | 5 | 4 | 0 | 2 | 0 | 0 | | 36% |
| Happiness | 3 | 0 | 2 | 1 | 0 | 11 | 0 | 1 | 0 | 0 | | 61% |
| Surprise | 1 | 0 | 0 | 0 | 0 | 4 | 2 | 1 | 0 | 0 | | 25% |
| Puzzlement | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 42 | 0 | 0 | | 93% |
| Sadness | 2 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 5 | 2 | | 42% |
| Uncertainty | 4 | 2 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 8 | | 50% |

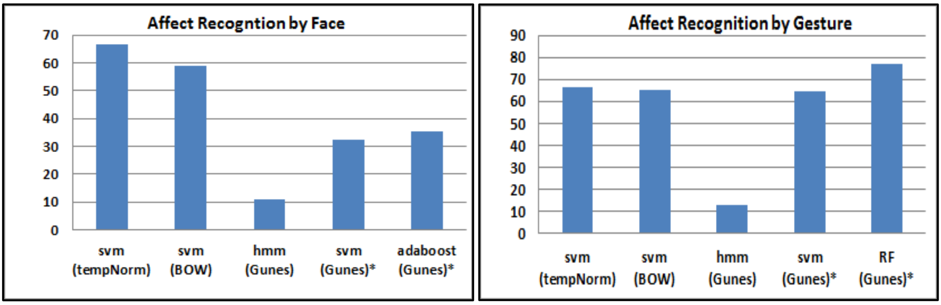
(d)

**Figure 12**: Sample confusion matrix for (a) facial features with temporal normalization method; (b) body gesture features with temporal normalization method; (c) facial features with Bag of Words method; (d) body gesture features with Bag of Words method; The row is the ground truth category, and the column is the classified category. The last column indicates the true positive rate for each class of expressions.

In Figure 11, we further compare the results with the Bag of Words over the complete expression cycle approach for the face and the body gesture. The temporal normalization achieves the best results on both modalities. On the other hand, the Bag of Words approach also significantly outperforms the maximum voting approach on the body gesture modality, while it achieves comparable performance on the face modality.

Figure 12 shows a sample confusion matrix of both temporal normalization method and Bag of Words method for face and body gesture modalities. The class specific true positive rate is presented in the last column. As we can see from the confusion matrices, the "boredom" category is most confused with the "puzzlement" category for both methods. The recognition rate for the "sadness" and the "surprise" categories are significantly lower than that of other categories. That is due to the small number of training samples for these two categories as compared with the other categories.

### 4.2.2 Compare to the State of the Art



(a) (b)

**Figure 13**: Compare our approach with the state of the art using (a) facial features; (b) body gesture features; Note that the performance cited from (Gunes)* [15] is frame-based accuracy instead of video-based accuracy used in our paper.

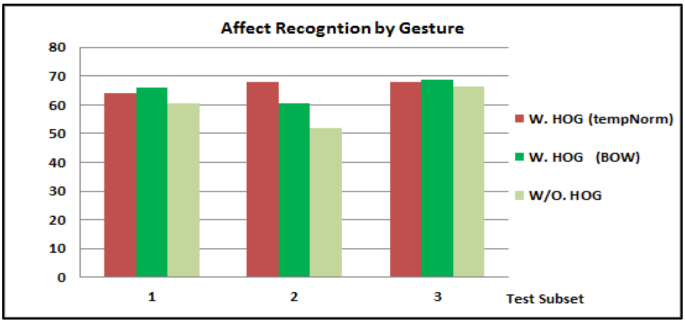


**Figure 14**: Compare affect recognition accuracy of gesture feature using MHI-HOG and Image-HOG to that without the MHI-HOG and the Image-HOG.

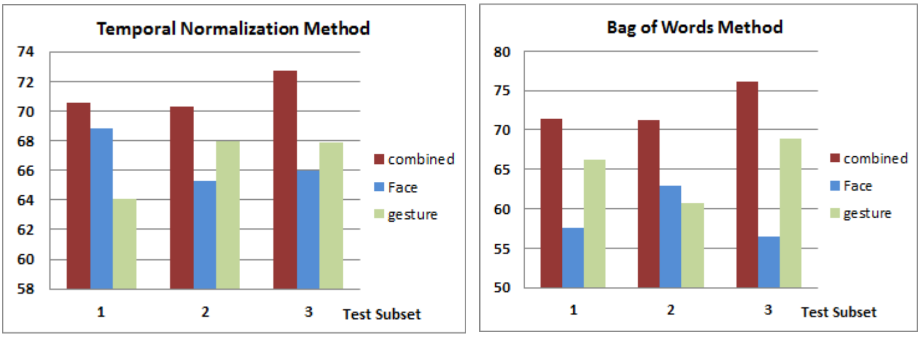
In the face modality, both temporal normalization approach and the Bag of Words approach significantly outperform the state of the art reported in [15], as shown in Figure 13(a). Note that the performance cited from (Gunes)* [15] in Figure 13 is frame-based accuracy. HMM method from Gunes and Piccardi [15] is the video-based accuracy result. Gunes and Piccardi report that the maximum voting of apex frames approach performed better than the video-based approach, which is the opposite from our conclusion in Figure 11.

For the body gesture modality, our method achieves the comparable performance with the paper in [15], as shown in Figure 13(b). Gunes and Piccardi [15] reported 76% accuracy with the Random Forest (RF) classifier. However, they use more complex features which include optical flow, edginess, geometry features, and comparison with the neutral frame etc. The feature extraction also involves several facial components tracking, hand tracking and shoulder tracking.

In order to evaluate the effectiveness of the MHI-HOG and the Image-HOG features, we also compare the performance of gesture modality using both MHI-HOG and Image-HOG features to that without the MHI-HOG and the Image-HOG features, as shown in Figure 14. Without the MHI-HOG and the Image-HOG features, body gesture features also include face and hands positions and the motion area around the face and hand regions, as described in section 3.1.2. The average accuracy gain with the MHI-HOG and the Image-HOG features is more than 6% and 5% for the temporal normalization method and the Bag of Words method respectively.

### 4.2.3 Fusion of Face and Body Gesture

We also evaluate the affect recognition by fusing both face and body gesture modalities. As compared with the individual modalities, *i.e.* face and body gesture, the fusion of face and body gesture modalities improves performance over all three testing subsets for both temporal normalization approach and Bag of Words approach, as shown in Figure 15. This conclusion is consistent with findings in [15]. Face and gesture modality achieves comparable performance in our experiments, while Gunes and Piccardi report that the single modality of body gesture has significantly better performance as compared with the face modality.



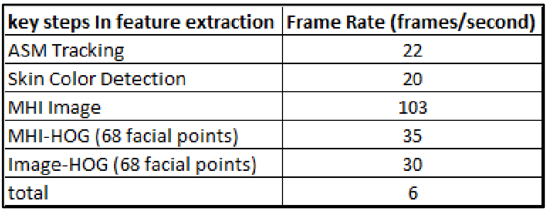
(a) (b)

**Figure 15**: Affect Recognition by the fusion of face and body gesture using simple concatenation for (a) temporal normalization method; (b) Bag of Words Method;

### 4.2.4 Algorithm Efficiency Analysis

The proposed feature extraction and representation are programmed in C++ without optimization. Table 1 lists the average frame rate for the key steps of the feature extraction for videos with resolution at 1024x768 pixels. The speed of each key step is 22 frames per second for ASM tracking, 20 frames per second for skin color detection, 103 frames per second for MHI image calculation, 35 frames per second for MHI-HOG and 30 frames per second for Image-HOG. The frame rates of the MHI-HOG and the Image-HOG descriptors are based on the total time of all 68 ASM facial landmark points. The testing is performed on a computer with multi-core CPU (2.13GHz) with 15.9 GB memory. The speed of the whole core algorithm of feature extraction (all key steps) is about 6 frames per second.

**Table 1**: Average frame rate for key steps in feature extractions for videos with resolution at 1024x768.



| key steps In feature extraction | Frame Rate (frames/second) |
|---|---|
| ASM Tracking | 22 |
| Skin Color Detection | 20 |
| MHI Image | 103 |
| MHI-HOG (68 facial points) | 35 |
| Image-HOG (68 facial points) | 30 |
| total | 6 |

# 5 Conclusion

We have proposed novel approaches, which combine the MHI-HOG and the Image-HOG features through the temporal normalization method and the Bag of Words model, to describe expressions using a complete expression cycle. Despite the simplicity of features used, the proposed approaches show promising results as compared with the state of the art. Face and body gesture modalities achieve comparable performance in our experiments. We also experimentally demonstrate that the expression dynamics can help affect recognition by comparing with the maximum voting of apex frames approach, and using both face and body gesture modalities could further improve the affect recognition performance, as compared with each individual modality.

# 6 Acknowledgement

The authors thank Dr. H. Gunes for providing the FABO Database. This article was partially developed under an appointment to the DHS Summer Research Team Program for Minority Serving Institutions, administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and U.S. Department of Homeland Security (DHS). ORISE is managed by Oak Ridge Associated Universities (ORAU) under DOE contract number DE-AC05-06OR23100. It has not been formally reviewed by DHS. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DHS, DOE, or ORAU/ORISE. DHS, DOE and ORAU/ORISE do not endorse any products or commercial services mentioned in this article.

Shizhi Chen is supported by NOAA Cooperative Remote Sensing Science and Technology (CREST) Fellowship.

The authors also would like to thank the anonymous reviewers for their constructive comments and insightful suggestions that improved the quality of this manuscript.

# References

[1] N. Ambady and R. Rosenthal, “Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis,” *Psychol. Bull.*, vol. 11, no. 2, pp. 256–274, 1992.

[2] M. Bartlett, G. Littlewort, I. Fasel, J. Movellan, “Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction”, IEEE Int'l Conf. Computer Vision and Pattern Recognition Workshop, 2003.

[3] A. Bobick and J. Davis, The recognition of human movement using temporal templates. IEEE Trans. PAMI. 23, 257–267, 2001.

[4] C. Chang and C. Lin, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[5] S. Chen, Y. Tian, Q. Liu and D. Metaxas, “Segment and Recognize Expression Phase by Fusion of Motion Area and Neutral Divergence Features”, IEEE Int'l Conf. Automatic Face and Gesture Recognition , 2011

[6] S. Chen, Y. Tian, Q. Liu, D. Metaxas. Recognizing Expressions from Face and Body Gesture by Temporal Normalized Motion and Appearance Features. IEEE Int'l Conf. Computer Vision and Pattern Recognition workshop for Human Communicative Behavior Analysis (CVPR4HB). 2011.

[7] S. Chen, Y. Tian. Evaluating Effectiveness of Latent Dirichlet Allocation Model for Scene Classification. IEEE Wireless and Optical Communications Conference (WOCC). 2011.

[8] I. Cohen *et al.*, “Facial expression recognition from video sequences: Temporal and static modeling,” Computer Vision and Image Understanding, 2003.

[9] T. Cootes, C. Taylor, D. Cooper and J. Graham, “Active Shape Models – Their Training and Application”, Computer Vision and Image Understanding, 1995.

[10] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, “Visual Categorization with Bag of Keypoints”, European Conference on Computer Vision (ECCV), 2004

[11] N. Dalal, B. Triggs, “Histogram of Oriented Gradients for Human Detection”, IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR), 2005.

[12] P. Ekman and W. Friesen, “Constants across Cultures in the Face and Emotion”, Journal of Personality Social Psychology, 1971.

[13] P. Ekman and W. Friesen, “The Facial ActionCoding System”, Consulting Psychologists Press Inc., 1978.

[14] B. Fasel and J. Luttin, “Automatic Facial Expression Analysis: Survey,” Pattern Recognition, 2003.

[15] H. Gunes and M. Piccardi, “Automatic Temporal Segment Detection and Affect Recognition from Face and Body Display”, IEEE Trans. on Systems, Man and Cybernetics – Part B: Cybernetics, Vol. 39, NO. 1 2009.

[16] H. Gunes and M. Piccardi, “A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior”, International Conference Pattern Recognition, 2006.

[17] G. Guo and C. Dyer, “Learning from Examples in the Small Sample Case: Face Expression Recognition”, IEEE Trans. On Systems, Man, and Cybernetics, 2005.

[18] B. Horn and B. Schunck,“Determiningopticalflow”,Artificial Intelligence, Vol. 17, 1981.

[19] T. Joachims, “Text categorization with support vector machines: Learning with manyrelevant features”, ECML, 1998.

[20] J. Kovac, P. Peer, F. Solina, “Human Skin Color Clustering for Face Detection”, EUROCON – Computer as a Tool, 2003.

[21] A. Lanitis, C. Taylor, and T. Cootes, “Automatic Interpretation and Coding of Face Images Using Flexible Models”, IEEE Trans. PAMI, 1997.

[22] S. Lazebnik, C. Schmid, J. Ponce, “Beyond Bags of Features: SpatialPyramid Matching for Recognizing Natural Scene Categories”, IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR),2006.

[23] L. Li, R. Socher, and L. Fei-Fei, “Towards Total SceneUnderstanding: Classification, Annotation and Segmentation in anAutomatic Framework”, IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR), 2009

[24] P. Lucey, J. Cohn, K. Prkachin, P. Solomon, I. Matthews,“Painful data: The UNBC-McMaster Shoulder Pain Expression Archive Database”, IEEE Int'l Conference on Automatic Face and Gesture Recognition (AFGR), 2011

[25] M. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, “Coding Facial Expressions with Gabor Wavelets”, IEEE Int'l Conference on Automatic Face and Gesture Recognition (AFGR), 1998.

[26] H. Meeren, C. Heijnsbergen, and B. Gelder, “Rapid perceptual integration of facial expression and emotional body language”, Proceedings of the National Academy of Sciences of USA, 2005.

[27] T. Otsuka and J. Ohya, “Spotting segments displaying facial expressionfrom image sequences using HMM,” IEEE Int'l Conference on Automatic Face and Gesture Recognition (AFGR), 1998

[28] M. Pantic and I. Patras, “Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments from Face Profile Image Sequences”, IEEE Trans. Systems, Man, and Cybernetics, 2006.

[29] M. Pantic and I. Patras, “Temporal modeling of facial actions from face profile image sequences”, IEEE Int. Conf. on Multimedia and Expo, 2004.

[30] M. Pantic and L.Rothkrantz, “Automatic Analysis of Facial Expressions: The State of the Art,” IEEE Trans. PAMI, 2000.

[31] M. Pitt and N. Shephard, “Filtering via simulation: auxiliary particle filtering”, Journal of the American Statistical Association, vol. 94, pp. 590-599, 1999.

[32] J. Saragih, S. Lucey, J. Cohn, “Real-time avatar animation from a single image”, IEEE Int'l Conference on Automatic Face and Gesture Recognition (AFGR), 2011.

[33] K. Schmidt and J. Cohn, “Human Facial Expressions as Adaptations: Evolutionary Questions in Facial Expression Research”, Yearbook of Physical Anthropology, 2001.

[34] C. Shan, S. Gong and P. McOwan, “Beyond facial expressions: learning human emotion from body gestures”, British Machine Vision Conference, 2007.

[35] Y. Tian, L. Cao, Z. Liu, and Z. Zhang, “Hierarchical Filtered Motion for Action Recognition in Crowded Videos”, IEEE Trans. on Systems, Man, and Cybernetics--Part C: Applications and Reviews, 2011.

[36] Y. Tian, T. Kanade and J. Cohn, “Recognizing Action Units for Facial Expression Analysis”, IEEE Trans. PAMI, 2001.

[37] S. Tong and D. Koller, “Support vector machine active learning with applications to text classification”, Int’l Conference on Machine Learning (ICML), 2000.

[38] Y. Tong, W. Liao and Q. Ji, “Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships”, IEEE Trans. PAMI, 2007.

[39] Y. Wei, “Research on Facial Expression Recognition and Synthesis”, Master Thesis, 2009, software available at: http://code.google.com/p/asmlibrary.

[40] P. Yang, Q. Liu and D. Metaxas, “Similarity Features for Facial Event Analysis”, European Conference on Computer Vision (EECV), 2008.

[41] Z. Zeng, M. Pantic, G. Roisman, T. Huang, “A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions”, IEEE Trans. PAMI, 2009.

[42] J. Alon, V. Athitsos, Q. Yuan, S. Sclaroff, “A Unified Framework for Gesture Recognition and Spatiotemporal Gesture Segmentation”, IEEE Trans. PAMI, 2009.

[43] D. Bernhardt and P. Robinson, “Detecting Affect from Non-stylised Body Motions”, Lecture Notes in Computer Science, 2007.

[44] G. Bradski, J. Davis, “Motion Segmentation and Pose Recognition with Motion History Gradients”, Int’l Journal of Machine Vision and Applications, 2002.

[45] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, and K. Scherer,“Toward a minimal representation of affective gestures”. IEEE Trans. Affective Computing, 2011.

[46] A. Kapoor, R. Picard, “Multimodal Affect Recognition in Learning Environments”, ACM Multimedia, 2005.

[47] D. Kim, J. Song, D. Kim, “Simultaneous gesture segmentation and recognition based on forward spotting accumulative HMMs”, Pattern Recognition, 2007.

[48] K. Scherer and H. Ellgring, "Multimodal Expression of Emotion:Affect Programs or Componential Appraisal Patterns," Emotion, 2007.

[49] Y. Wang, G. Mori, “Hidden Part Models for Human Action Recognition: Probabilistic versus Max Margin”, IEEE Trans. PAMI 2011.

[50] P. Yang, Q. Liu, D. Metaxas, “Boosting Coded Dynamic Features for Facial Action Units and Facial Expression Recognition”, IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR), 2007.

[51] G. Zhao, M. Pietikainen, “Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions”, IEEE Trans. PAMI, 2007.