Segment and Recognize Expression Phase by Fusion of Motion Area and Neutral Divergence Features

Shizhi Chen and YingLi Tian
Department of Electrical Engineering
The City College of New York
New York NY, USA
{schen21, ytian}@ccny.cuny.edu

Qingshan Liu and Dimitris N. Metaxas
Department of Computer Science
Rutgers University
Piscataway NJ, USA
{qsliu, dnm}@cs.rutgers.edu

***Abstract*—An expression can be approximated by a sequence of temporal segments called neutral, onset, offset and apex. However, it is not easy to accurately detect such temporal segments only based on facial features. Some researchers try to temporally segment expression phases with the help of body gesture analysis. The problem of this approach is that the expression temporal phases from face and gesture channels are not synchronized. Additionally, most previous work adopted facial key points tracking or body tracking to extract motion information, which is unreliable in practice due to illumination variations and occlusions. In this paper, we present a novel algorithm to overcome the above issues, in which two simple and robust features are designed to describe face and gesture information, i.e., motion area and neutral divergence features. Both features do not depend on motion tracking, and they can be easily calculated too. Moreover, it is different from previous work in that we integrate face and body gesture together in modeling the temporal dynamics through a single channel of sensorial source, so it avoids the unsynchronized issue between face and gesture channels. Extensive experimental results demonstrate the effectiveness of the proposed algorithm.**

Keywords-temporal segment; motion area; neutral divergence;

I. INTRODUCTION

Automatic expression recognition can potentially open doors to many applications including lie detection, human computer interaction (HCI), video surveillance etc. It allows machine to interpret and respond to human's emotion states similar to human beings. Nevertheless, the problem is challenging if we want to achieve satisfied recognition rate to be used in practice [2, 6]. One of the major challenges is to model temporal dynamics of an expression, which is a critical factor in interpreting human expression [11].

An expression is a dynamic event, which evolves from neutral, onset, apex to offset, as shown in Figure 1. Among these four temporal phases, features during the apex phase result in maximum discriminative power to expression recognition. Thus, successful temporal segmentation can not only help to analyze the dynamics of facial expression, but also improve the performance of expression recognition. Despite its usefulness, there exists very limited number of studies on the temporal segmentation of affective behavior [8, 9, 13].

In paper [9], Pantic and Patras temporally segmented facial action units (AUs) using geometric features of 15 facial key points from profile face images. These 15 facial key points are carefully selected so that they are discriminative enough to distinguish temporal phases of facial action unit. For example, they chose tip of the nose and top of the forehead as the facial points because of their stability with respect to face movement. Then they employed particle filtering [10] to track these 15 facial key points and calculate the corresponding geometry features. However, in practices, it is difficult to obtain accurate detection of these facial key points and track them robustly due to illumination variations and occlusions (see examples in Figure 3).

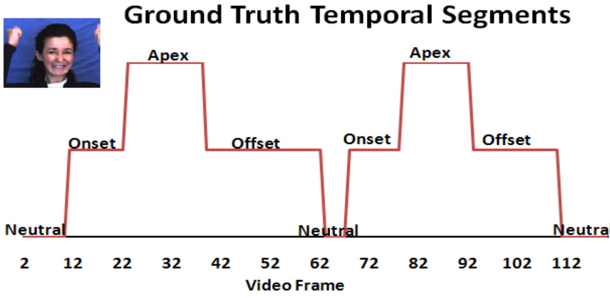


Figure 1: The temporal segments of the expression of "Happiness"**.**

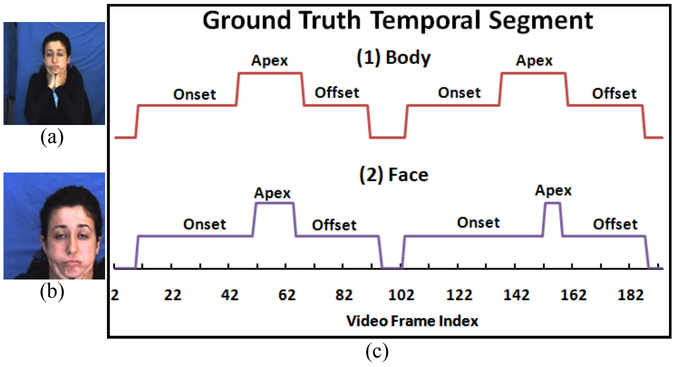


Figure 2: (a) Sample image of "Boredom" expression to extract body gesture feature (Body camera); (b) Sample image of "Boredom" expression to extract facial feature (Face camera); and (c) The corresponding temporal segments from body gesture and facial features respectively.

Psychology studies suggest that combining both face and body gesture can be very effective in judging human expressions [1]. Inspired by the study, a few researchers have proven the effectiveness of fusing face and body gesture in expression recognition [8, 12]. Gunes and Piccardi [8] used two cameras (body camera and face camera) to extract body gesture features and facial features respectively, as shown in Figure 2(a) and 2(b). They first performed temporal segmentation operation on face and body channels respectively, as shown in Figure 2(c), and then the temporal segmentations of face and body gesture are combined together by selecting the video frames which share the same temporal phases. However, face and body gesture temporal segments are usually not synchronized. As shown in Figure 2(c), frames 137 to 151 are apex in the body temporal segments, while they are onset in the face temporal segments. Hence, their selective fusion strategy has to discard these frames, even though these frames contains equal amount of body gesture information as the final apex frames do.

Both papers [8, 9] modeled temporal dynamics of face by extracting geometric or appearance features from a set of fixed interesting points. These approaches have two limitations. First, the selection of the fixed interesting points requires experience and mostly needs human intervention. Second, tracking is usually sensitive to occlusions and illumination variations. As shown in Figure 3, the facial point tracking will fail when the hands touch the face. Inaccuracy in tracking will degrade the temporal segmentation performance significantly.

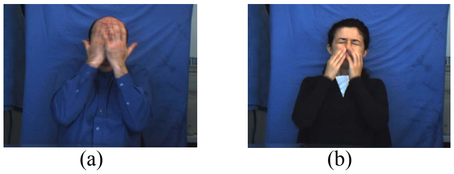


Figure 3: (a) A "Fear" expression with hands covering the whole face; (b) A "Sadness" expression with hands touching the lower part of the face.

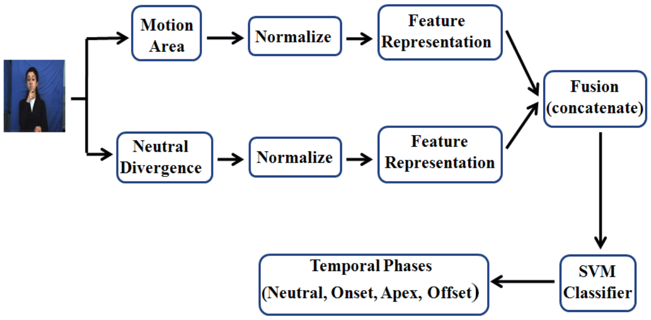


Figure 4: Flow chart of the proposed method to detect temporal segments of an expression.

To handle the above issues, we propose two novel features in this paper, i.e., motion area and neutral divergence, to simultaneously segment and recognize temporal phases of an expression. The motion area feature is calculated by simple motion history image [3, 5], which does not rely on any facial points tracking or body tracking, and the neutral divergence feature is based on the differences between the current frame and the neutral frame. Both of them are simple to calculate and they provide complementary information in segmenting and recognizing the temporal phases of an expression. Moreover, different from previous work, we extract these two information to describe both body gesture and facial appearance dynamics together through a single channel, i.e., only the body camera as shown in Figure 2(a). Hence, the approach eliminates the selective fusion step due to the unsynchronized temporal segments from face and body gesture. Experiments on the benchmark database FABO [7] demonstrate the power of the proposed method.

II. Method for Segmenting and Recognizing Expression Phases

A. Overview

The proposed method is summarized in Figure 4. Unlike exiting methods of using both face and gesture channels, we only use the gesture channel. This channel includes both face and gesture information and is sufficient to perform temporal segmentation of facial expression. We extract the motion area based on the motion history images and the neutral divergence based on the difference between the current image and the neural image respectively. Both features are normalized between 0 and 1. The corresponding feature vector representation of both feature types are generated for each video frame. The motion area and the neutral divergence features are then combined simply by concatenating them together. Finally, we employ Support Vector Machine (SVM) to classify each frame into the different temporal segment in an expression video.

The motion area and the neutral divergence features implicitly incorporate both facial and body gesture information without any motion tracking, so the approach avoids losing informative apex frames due to the unsynchronized face and body gesture temporal phases. Furthermore, both features are efficient and easy to compute.

B. Feature Extraction

Motion Area: We extract the motion area based on the motion history image (MHI), which is a compact representation of a sequence of motion movement in a video [3, 5]. Pixel intensity is a function of the motion history at that location, where brighter values correspond to more recent motion. The intensity at pixel (*x, y*) decays gradually until a specified motion history duration τ. The construction process of the MHI image can be best described using the equation (1) below.

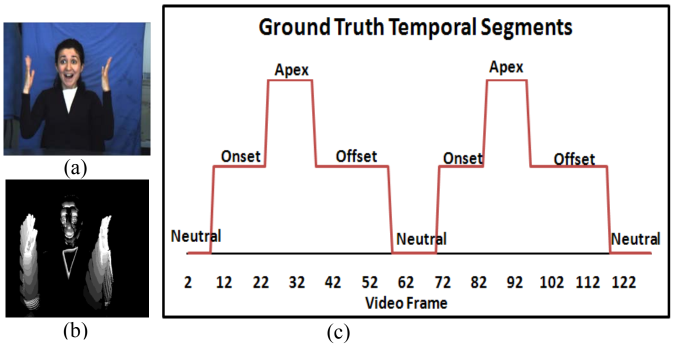
$$\begin{aligned} MHI_{\tau}(x,y,t) = D(x,y,t)\ \times \tau + [1 - D(x,y,t)] \\ \times U[MHI_{\tau}(x,y,t-1) - 1] \\ \times [MHI_{\tau}(x,y,t-1) - 1]\,, \end{aligned} \qquad (1)$$

where *U[x]* is a unit step function. t represents the current video frame index. $D(x,y,t)$ is a binary image of pixel intensity difference between the current frame and the previous frame. $D(x,y,t)$ is 1 if the intensity difference greater than a threshold. Otherwise, it is 0. We use the threshold of 25 in our experiments. τ is the maximum motion duration. In our system, we set τ to 10. The MHI is then scaled to an 8-bit gray image. Figure 5(b) shows the generated motion history image of a "Surprise" expression.

The motion area of each video frame is the total number of the motion pixels in the corresponding MHI image. The motion pixels are defined as the pixels with non-zero intensity in the MHI image. The calculation of the motion area $MA_\tau(t)$ can be described by equation (2).

$$MA_\tau(t) = \sum_{x=1}^{W}\sum_{y=1}^{H} U[MHI_\tau(x,y,t) - \varepsilon]\,, \qquad (2)$$

where $0 < \varepsilon < 1$. *U[x]* is a unit step function. *(x, y)* is pixel position in the MHI image. *t* stands for the current video frame index. *W* and *H* stand for the width and the height of the MHI image respectively.

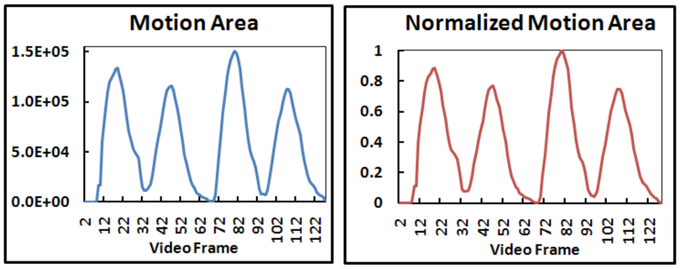


(a) (b) (c)

Figure 5: (a) A “Surprise” expression; (b) Motion History Image of the “Surprise” expression shown in (a); and (c) ground truth temporal segments of the expression.

Figure 6(a) plots the motion area of the “Surprise” expression shown in Figure 5. The expression starts from the neutral frame followed by the onset, the apex, the offset and back to the neutral. During the neutral phase, the subject place hands on desk without any movement. Hence, the motion area $MA_\tau(t)$ is almost 0 at the first few frames up to frame 10, as shown in Figure 6(a). From frames 11 up to 24, the expression enters the onset phase. The subject moves hands up while the facial expression increases gradually, which generates a large number of motion pixels. Therefore, $MA_\tau(t)$ increases and peaks at frame 15. As the expression approaches the apex, the motion begins to slow down, which causes $MA_\tau(t)$ to decrease between frame 15 to frame 24. The apex occurs between frames 25 to 34 in Figure 6(a). During the apex phase, the expression reaches its maximum spatial extent and lasts for some time. Hence, there is relatively small or no motion during the phase. During the offset phase, both facial expression and body gesture are moving from the apex back to the neutral phase. Similar to the analysis of the onset phase, we are expecting that the motion movement increases initially and slows down when the expression approaches the neutral. This is exactly what we see in Figure 6(a) between frames 35 to 54. Finally the expression enters the neutral phase again between frames 55 to 70 with very small motion area.

To effectively represent the motion area, we normalize it to 0 and 1 with the maximum motion area in the expression corresponding to 1. Figure 6(b) shows the corresponding normalized motion area of Figure 6(a). The normalization is to handle variation of the motion area due to different expressions or subjects.



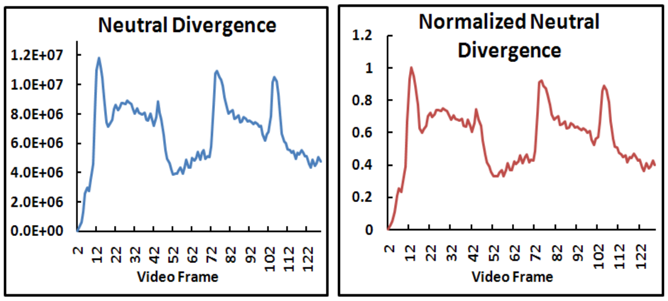
(a) (b)

Figure 6: (a) Motion area of the “Surprise” expression shown in Figure 5; and (b) The corresponding normalized motion area.

Neutral Divergence: The neutral divergence feature measures the degree of difference between the current frame and the neutral frame of an expression video. From observations of FABO database [7], all expression videos in the bi-modal database start from the neutral position. Hence, we choose the starting frame as our neutral frame. Then the current frame’s neutral divergence *ND(t)* is calculated by summing up the absolute intensity difference between the current frame image *I(x,y,d,t)* and the neutral frame image *I(x,y,d,t₀)* over three color channels, as shown in equation (3).

$$ND(t) = \sum_{d=1}^{3}\sum_{x=1}^{W}\sum_{y=1}^{H} abs[I(x,y,d,t) - I(x,y,d,t_0)]\,, \quad (3)$$

where *W* and *H* are the width and the height of a video frame respectively. *d* is the number of color channels of the frame.



(a) (b)

Figure 7: (a) Neutral divergence of the “Surprise” expression shown in Figure 5; and (b) the corresponding normalized neutral divergence.

Figure 7(a) plots the neutral divergences of the “Surprise” expression shown in Figure 5. Similar to the motion area normalization, the neutral divergence is also down scaled between 0 and 1. Figure 7(b) shows the corresponding normalized neutral divergence of each frame.

Since we assume each video starts from the neutral, the neutral divergence is 0 at the starting frame. During the onset phase, the intensity deviation from the neutral frame increases.

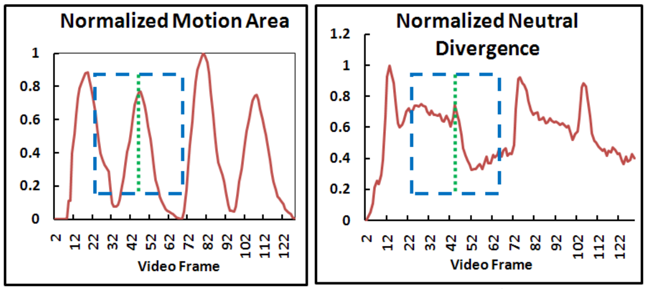
Hence, the neutral divergence increases as we observe in Figure 7(a). During the apex, the neutral divergence remains relative stable with a large neutral divergence value. That is because that there is little movement of facial expression or body gesture while the apex frame is quite different from the neutral frame.

Facial expression and body gesture relax back to the neutral position during the offset phase. Therefore, the neutral divergence decreases since the difference between the offset frames and the neutral frame becomes smaller.

However, when the expression enters the neutral phase again between frames 55 to 70 as shown in Figure 7(a), the neutral divergence does not return back to 0 as expected. Instead, it is roughly half of the apex neutral divergence, which implies that the neutral frames from 55 to 70 are still quite different from the starting neutral frame. In another word, the face and body parts do not return back to the exact position of the starting neutral frame. There is certain shift in position, which is not noticeable by us. Nevertheless, the difference between the returned neutral frames and the apex frames on the neutral divergence is still recognizable from Figure 7(a).

C. *Feature Representation and Fusion*

Feature Representation: Based on the above description, we extract the normalized motion area and the neutral divergence of every frame in an expression video. One question arises: how can we use a vector to represent the motion area and the neutral divergence features of each frame.



(a) (b)

Figure 8: (a) Motion area feature representation of current frame is a vector of normalized motion area, which frames are within a temporal window centered at current frame; (b) Neutral divergence feature representation of current frame is a vector of normalized neutral divergence, which frames are within a temporal window centered at current frame.

To recognize the expression phases of the current frame, we employ a fixed-size temporal window with center located at the current frame as shown in Figure 8(a). The normalized motion area of every frames within the temporal window is extracted and forming a vector in the chronological order. The formed vector is the motion area feature representation of the current frame. Therefore, the motion area feature vector has the dimensionality of temporal window size. Similar to the motion area feature representation, we use a fixed-size temporal window with the center located at the current frame to represent the neutral divergence features, as shown in Figure 8(b). The normalized neutral divergence of every frame within the temporal window is extracted and forming a vector in the chronological order. In our experiments, we set the temporal window size as 31 for both the motion area features and the neutral divergence features.

Some frames are truncated at the temporal boundary of an expression video, such that the temporal window can fit into the temporal sequence of the video. The number of the truncated frames at the beginning or ending boundary of an expression video is half of temporal window size.

Feature Fusion: The motion area and the neutral divergence features provide complementary information regarding temporal dynamics of an expression. The motion area is able to separate the onset/offset from the apex/neutral phases, since the onset/offset generates large movement. However, the motion area can neither distinguish the apex from the neutral, nor the onset from the offset.

Nevertheless, the apex has large intensity deviation from the neutral frame. Hence, the neutral divergence is able to separate the neutral from the apex. During the onset phase, the neutral divergence is increasing while the opposite occurs during the offset. Therefore, the neutral divergence is able to separate the onset from the offset as well.

The fusion of both features can be implemented by simply concatenating the motion area feature vector with the neutral divergence feature vector together.

D. *Classifier*

We employ SVM with a RBF kernel 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. The temporal segmentation of an expression phase can be considered as a multi-class classification problem. That is to classify each frame in an expression video to the neutral, onset, apex, and offset temporal phases.

III. EXPERIMENTS

A. *Experimental Setup*

We conduct the experiments on a bi-modal face and body benchmark database FABO [7]. The FABO database consists of both affective face and body recordings using two cameras simultaneously.

There are 10 expression categories in our experiments, including both basic expressions and non-basic expressions. Basic expressions are “Disgust”, “Fear”, “Happiness”, “Surprise”, “Sadness” and “Anger”. Non-basic expressions are “Anxiety”, “Boredom”, “Puzzlement” and “Uncertainty”.

We chose 288 videos that the ground truth expressions from both the face camera and body camera are identical. For each video, there are 2 to 4 complete expression cycles.

Videos of each expression category are randomly separated into three subsets. Then two of the three subsets are chosen as the training data and the remaining is chosen as the testing data. The subjects may overlap between the training and testing sets due to the random separation process. Figure 9 shows some sample images of "Boredom" and "Puzzlement" expression videos from both face and body cameras. As mentioned before, the temporal segments from these two cameras are not synchronized, which requires additional alignment step. Moreover, it is not practical to use both body and face cameras in the real world applications. Therefore, we only use the videos captured from the body camera, which contains both face and body gesture information.

(a) (b)

Figure 9: (a) sample images from a "Boredom" expression video in FABO database recorded by body (left) and face (right) camera; (b) sample images from a "Puzzlement" expression video in FABO database recorded by body (left) and face (right) camera;

B. *Experiment Results*

We first perform a three-fold cross-validation by combining the motion area and the neutral divergence features. That is to choose one of the three subsets of the expression videos as the testing data. The remaining two subsets are used as the training data. The process is repeated three times with each of the three subsets used as the testing data exactly once. The accuracy is calculated by averaging true positive rate of each class of temporal segments, i.e., the neutral, the onset, the apex, and the offset.

TABLE I
Summary of three-fold cross validation temporal segment performance

Testing Data	Subset 1	Subset 2	Subset 3	Average
Accuracy (%)	82	83	84.3	83.1

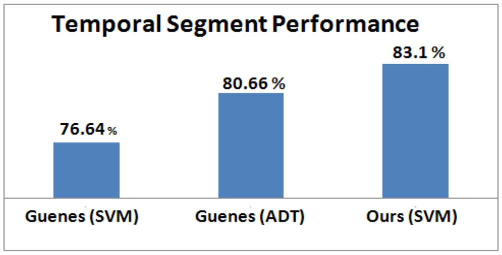


Figure 10: Temporal segment performance comparing to the state of the art reported in [8].

Table I reports the results of the temporal segmentation. The performance over each subset is relatively constant, which shows the robust of the combined feature. The average accuracy of the three-fold cross validation achieves 83.1%. On the same database, Gunes and Piccardi [8] report the state of the art performance of 80.66% on the temporal segmentation. A sophisticated classifier called Adaboost with Decision Tree (ADT) is used in their experiments. Gunes and Piccardi also reported the temporal segmentation accuracy of 76.64% using SVM classifier [8]. Figure 10 shows the summary of comparison. Our results exceed the state of the art performance by almost 3%. With the SVM classifier, our method outperforms the state of the art by almost 7%.

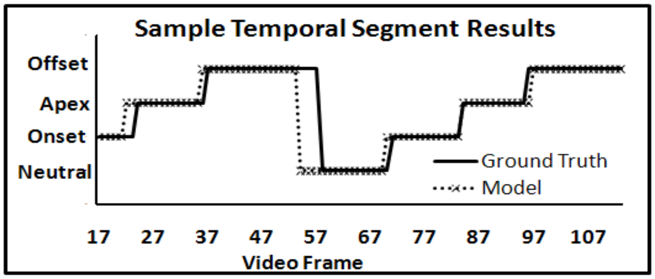
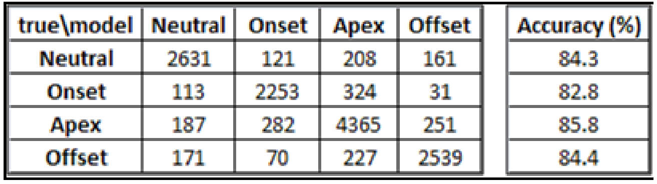


Figure 11: Temporal segmentation results corresponding to the "Surprise" expression video shown in figure 5.

Figure 11 shows the temporal segmentation results of the "Surprise" expression video shown in Figure 5. The ground truth temporal phase of each frame in the expression video is indicated by the solid line, while the corresponding predicted temporal phase is plotted using the dash line with "x". The predicted temporal segmentation of the expression video matches the ground truth temporal phase quite well except at the phase transition frames. For example, frames 22 and 23 are predicted as the apex phase while the ground truth indicates that they are the onset frames right before the apex.

TABLE II
A sample confusion matrix with true positive rate for each class of temporal segments. Rows are the ground truth temporal segments while columns are the classified temporal segments.



true\model	Neutral	Onset	Apex	Offset	Accuracy (%)
Neutral	2631	121	208	161	84.3
Onset	113	2253	324	31	82.8
Apex	187	282	4365	251	85.8
Offset	171	70	227	2539	84.4

Table II shows a sample confusion matrix of expression phase segmentation. Each row is the ground truth temporal segment while the columns are the classified temporal segments. Based on the confusion matrix, both the onset and the offset seem most confused with the apex. The apex phase is temporally adjacent to both the onset and the offset. The temporal boundary between them is not a clear cut, which explains the confusion between the apex and the onset/offset. However, the overall performance of each temporal phase is fairly stable as shown in the last column of Table II.

The other experiment is to evaluate the effectiveness of the fused feature comparing to the individual features, i.e., the motion area and the neutral divergence. The experiment uses the first subset of expression videos as the testing data and the other two subsets as the training data. Table III shows the comparison on the temporal segmentation performance. Using the motion area alone, the temporal phase detection rate is 68.5%. The neutral divergence feature alone achieves 74.1%. By combining both the motion area and the neutral divergence, the expression phase segmentation performance has boosted up to 82%.

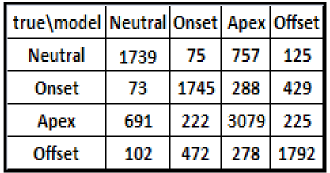
TABLE III

Comparison of combined features with individual feature

	Motion Area	Neutral Divergence	Both Features
Accuracy	68.5%	74.1%	82%

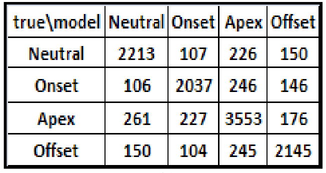
TABLE IV

(a) Confusion matrix using motion area feature only, where rows are ground truth temporal phases while columns are classified temporal phases; (b) Confusion matrix using both motion area and neutral divergence features with identical training and testing video subsets as (a). Rows are ground truth temporal phases while columns are classified temporal phases;



true\model	Neutral	Onset	Apex	Offset
Neutral	1739	75	757	125
Onset	73	1745	288	429
Apex	691	222	3079	225
Offset	102	472	278	1792

(a)



true\model	Neutral	Onset	Apex	Offset
Neutral	2213	107	226	150
Onset	106	2037	246	146
Apex	261	227	3553	176
Offset	150	104	245	2145

(b)

In order to understand why the combined feature significantly improves the performance, we compare the confusion matrices obtained from the motion area alone and the combined feature. Table IV(a) reports the confusion matrix using the motion area feature alone. From the matrix, we can see that the apex frames are most confused with the neutral frames. As an example, there are 757 neutral frames are misclassified into the apex frames, while there are 691 apex frames are misclassified into the neutral frames. Similarly, the onset phase is most confused with the offset. These results are consistent with our previous observations of the motion area feature. The motion area can neither distinguish the apex from the neutral, nor the onset from the offset.

Table IV(b) lists the confusion matrix using the fused feature, i.e., combining both the motion area and the neutral divergence features. The confusion between neutral phase and apex is significantly reduced. As an example, there are only 226 neutral frames are misclassified into the apex frames and 261 apex frames misclassified into the neutral frames. Similarly the confusion between the onset and the offset is also reduced.

These comparisons confirm the effectiveness of combining both the motion area and the neutral divergence features on the expression temporal segmentation. The neutral divergence and the motion area are providing complementary information on the temporal dynamics of an expression.

IV. Conclusion

This paper proposed a novel approach to segment and to recognize the expression temporal phases by the motion area and the neutral divergence features. Both features are global feature modeling the temporal dynamics of an expression. They naturally integrate both face and gesture information together for expression phase detection. Thus, the approach eliminates the selective fusion step due to the unsynchronized temporal segments between face and gesture channels. Neither the motion area nor the neutral divergence requires facial key points tracking or body tracking, which eliminates the complicacy induced by the tracking. The two features provide complementary information on the temporal dynamics of an expression. Experiments on the FABO database showed that the proposed approach outperforms the state of the art performance by almost 3%. If comparing with the SVM classifier, the proposed method outperforms the state of the art by almost 7%.

Acknowledgment

This research was partially performed at DHS *CCICADA Center* under an appointment to the U.S. Department of Homeland Security (DHS) Summer Research Team Program for Minority Serving Institutions. Thanks to Tami Carpenter and Patti Obenour to make this happen. The FABO Database was provided by Dr. Hatice Gunes. Shizhi Chen is funded by NOAA Cooperative Remote Sensing Science and Technology (CREST) Fellowship.

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, M. Frank, C. Lainscsek, I. Fasel and J. Movellan, “Fully automatic facial action recognition in spontaneous behavior”, Int. Conf. on Automatic Face and Gesture Recognition, 2006.

[3] A. Bobick and J. Davis, The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. 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] J. Davis, “Hierarchical motion history images for recognizing human motion”, Proc. Of IEEE Workshop on. Detection and Recognition of Events in Video, 2001.

[6] H. Gu and Q. Ji, “An automated face reader for fatigue detection”, Int. Conf. on Automatic Face and Gesture Recognition, 2004.

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

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

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

[10] M. Pitt and N. Shephard, “Filtering via simulation: auxiliary particle filtering”, J. Amer. Stat. Assoc., vol. 94, pp. 590-599, 1999.

[11] J. Russell and J. Fernandez-Dols, “The psychology of facial expression”, Cambridge University Press, 1997.

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

[13] M. Valstar and M. Pantic, “Fully automatic facial action unit detection and temporal analysis”, CVPR workshop, 2006.