Affect Recognition using Key Frame Selection based on Minimum Sparse Reconstruction

Mehmet Kayaoglu, Cigdem Eroglu Erdem
Department of Electrical and Electronics Engineering
Bahcesehir University
34343, Besiktas, İstanbul, Turkey
Phone: +90 212 381 0895
mehmet.kayaoglu@stu.bahcesehir.edu.tr, cigdem.eroglu@eng.bahcesehir.edu.tr

ABSTRACT
In this paper, we present the methods used for Bahcesehir University team's submissions to the 2015 Emotion Recognition in the Wild Challenge. The challenge consists of categorical emotion recognition in short video clips extracted from movies based on emotional keywords in the subtitles. The video clips mostly contain expressive faces (single or multiple) and also audio which contains the speech of the person in the clip as well as other human voices or background sounds/music. We use an audio-visual method based on video summarization by key frame selection. The key frame selection uses a minimum sparse reconstruction approach with the goal of representing the original video in the best possible way. We extract the LPQ features of the key frames and average them to determine a single feature vector that will represent the video component of the clip. In order to represent the temporal variations of the facial expression, we also use the LBP-TOP features extracted from the whole video. The audio features are extracted using OpenSMILE or RASTA-PLP methods. Video and audio features are classified using SVM classifiers and fused at the score level. We tested eight different combinations of audio and visual features on the AFEW 5.0 (Acted Facial Expressions in the Wild) database provided by the challenge organizers. The best visual and audio-visual accuracies obtained on the test set are 45.1% and 49.9% respectively, whereas the video-based baseline for the challenge is given as 39.3%.

Categories and Subject Descriptors
I.5.4 [Pattern Recognition]: Applications, computer vision, signal processing; I.4.m [Image Processing and Computer Vision]: Miscellaneous

General Terms
Algorithms, Human Factors, Experimentation.

ICMI'15, November 09-13, 2015, Seattle, WA, USA
©2015 ACM. ISBN 978-1-4503-3912-4/15/11 $15.00
DOI: http://dx.doi.org/10.1145/2818346.2830594

Keywords
Emotion Recognition; EmotiW 2015 Challenge; Video Summarization; Affect Recognition; Affective Computing

1. INTRODUCTION
In daily human-to-human interactions, our facial expressions convey non-verbal messages about our emotions and mental states that complement our verbal messages. In the future, human-computer interaction scenarios are also expected to have the ability to recognize emotions to provide more natural man-machine interaction and ubiquitous computing applications such as health care [13], education, psychology [20] and security [18].

In order to test and benchmark automatic affect recognition algorithms, affective databases are needed. Early databases used were mostly acted and recorded in laboratory conditions under controlled head pose and illumination variations. Recently, more spontaneous and close-to-real world databases have been collected [1], [4], [7], [8], [26], several of which have been used in challenges, such as FERA 2015 [22], AVEC 2014 [23], and EmotiW 2015 [5].

The third Emotion Recognition in the Wild (EmotiW 2015) challenge consists of categorical audio-video based emotion recognition based on the Acted Facial Expression in Wild database (AFEW 5.0). The AFEW database contains short audio-visual clips collected from movies and labeled using a semi-automatic approach described in [4]. This challenge is a continuation of the EmotiW 2013 and 2014 challenges and the task is to assign a single emotion label to the video clip from the seven emotions (Anger, Disgust, Fear, Happiness, Neutral, Sad and Surprise). The AFEW database is quite challenging since the video clips contain variability in illumination and head pose, as well as severe occlusion and complex background.

The audio-visual emotion recognition methods in the literature have shown the advantages of fusing audio and video modalities [1], [3], [12], [14], [24], [28]. In this paper, we also present a multimodal affect recognition method using facial expressions and the speech signal. Given a video with an emotional expression, the frames in the video generally reflect the emotion with different intensities. Moreover, some parts of the video might have little motion, which makes subsequent frames to be very similar to each other. Therefore, we aim to select key frames, which will summarize the content of the video effectively. We adopt a recent video summarization method [15] for the problem of selecting key frames from affective videos. Then, static appearance-based facial features are extracted from the selected

key frames and averaged to describe the visual content of the whole video. We also intend to capture the temporal variations of facial expressions using LBP-TOP features. We use the audio features extracted by several approaches (OpenSMILE [8], [10], RASTA-PLP [11] etc.) and fuse with the video-based features at the score level. The experimental results on the AFEW 5.0 test database gave an accuracy of 45.1% using visual features and 49.9% using audio-visual features whereas the video based baseline was given as 39.3% [5] .

The organization of the paper is as follows. In Section 2, we give the details of the proposed multimodal emotion recognition system focusing on the key frame selection method using minimum sparse reconstruction. In Section 3, experimental results using various fusion cases are presented. Finally, in Section 4, conclusions are given.

2. EMOTION RECOGNITION SYSTEM

In this section, we first present the methods used for feature extraction from video and speech. Then, we present the classification and fusion methods utilized.

2.1 Feature Extraction from Video

An audio-visual video with an emotional expression consists of many frames, where each frame represents the emotion with a different intensity. Therefore, one way of recognizing the emotion in the video is to select and classify the facial expression on the "key frames" selected from the video. The key frames are selected so that they summarize the content of the video in the best way. This is based on the assumption that there is a single emotion in the sequence, which is true for most databases in the literature including the AFEW 5.0 database.

As a preprocessing step, the face regions in all frames of a video are detected and aligned. We use the aligned and cropped face provided by the challenge organizers. Face detection is achieved using the mixture of parts framework of Zhu and Ramanan [29], which is followed by face tracking using IntraFace method [21]. The face regions are aligned and cropped using the tracked feature points.

Below, in Section 2.1.1 we give the details of the key frame selection method used in this paper. Once the key frames are selected, the LPQ features (explained in Section 2.1.2) are computed and averaged over them to represent the visual component of the whole video. Another feature set that we tested is the LBP-TOP feature, which captures the temporal variation of the facial feature expression as well, which is explained in Section 2.1.3.

2.1.1 Key Frame Selection using Minimum Sparse Reconstruction

The assumption behind the utilized key frame selection method is that set of key frames are the most representative frames in the sequence among others and also the conceptual information of the sequence is mostly covered by key frames [15]. Therefore, we view the key frame selection as a similar problem to video summarization. In [15], video summarization is formulated as a problem of selecting the minimum number of frames to reconstruct the entire video as accurately as possible. Video summarization is actually a ranking process of the frames considering how well they represent the video. We adopted this video summarization method to emotional videos to select the key frames from a video with an emotional expression.

Given a video with n frames, each frame is a candidate to be a key frame. Let $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n] \in R^{d \times n}$, where $\mathbf{f}_\mathrm{i} \in \mathrm{R}^\mathrm{d}$ denotes the feature vector corresponding to frame i. The goal of key frame selection is to select an optimal subset $\boldsymbol{F_K} = [\mathbf{f}_{k_1}, \mathbf{f}_{k_2}, \dots, \mathbf{f}_{k_m}] \in R^{d \times m}$ such that $k_1, k_2, \dots, k_m \in [1,2,\dots,n]$. There are two goals when the subset is formed: i) The original video is reconstructed accurately and ii) the number of key frames is as small as possible. That is, the following minimum sparse reconstruction (MSR) expression is minimized:

$$\min_{\mathbf{S}} \frac{1}{2}\|\mathbf{F} - \mathbf{F}_K\mathbf{A}\|_2 + \lambda\|\mathbf{S}\|_0 \quad (1)$$

s.t. $\mathbf{F}_K = \mathbf{FS}$

$\mathbf{A} = f(\mathbf{F}, \mathbf{F}_K)$

where $\mathbf{S}$ is a diagonal selection matrix that models the selection of keyframes from the original video:

$$S_{ij} = \begin{cases} 0, & i \neq j \\ 0 \text{ or } 1, & i = j \end{cases} \quad (2)$$

and $\|\mathbf{S}\|_0$ is the L_0 norm of the selection matrix, which is the number of nonzero elements indicating the number of key frames selected. Therefore, sparsity is ensured by the L_0 norm. In (1), $\mathbf{A}$ represents the reconstruction coefficients of $\mathbf{F}$ by the matrix $\mathbf{F_K}$ which are computed using the reconstruction function $f(.,.)$, $\|.\|_2$ represents the L_2 norm, and λ is a weighting coefficient. The first term in (1) tries to minimize the least-square reconstruction error (LSRE), while the second term minimizes the number of key frames selected.

Assuming that m keyframes have been selected, the next keyframe chosen should maximally decrease LSRE. Therefore, the frame which gives the maximum LSRE at the current iteration should be selected as the next key frame:

$$\mathbf{f}_{k_{m+1}} = \arg\max_{\mathbf{f}_j \in \mathbf{F}/\mathbf{F}_K} \left\| f_j - \mathbf{F}_K\mathbf{a}_j \right\|_2, \quad (3)$$

where $\mathbf{a}_j$ represents the reconstruction coefficient for the j^{th} frame and $\mathbf{F}/\mathbf{F}_K$ represents the set of all non-keyframes. This is equivalent to selecting the worst reconstructed frame, after normalization by the vector magnitude:

$$\mathbf{f}_{k_{m+1}} = \arg\min_{\mathbf{f}_j \in \mathbf{F}/\mathbf{F}_K} \frac{\|\mathbf{F}_K\mathbf{a}_j\|_2}{\|\mathbf{f}_j\|_2}. \quad (4)$$

In order to determine the reconstruction coefficients, the orthogonal subspace projection (OSP) method is used [15] by projecting all frames to the space spanned by $\mathbf{F}_K$, which gives:

$$\mathbf{a}_j = (\mathbf{F}_K^T\mathbf{F}_K)^{-1}\mathbf{F}_K^T\mathbf{f}_j = \mathbf{P}_K\mathbf{f}_j \quad (5)$$

The algorithm continues to select new key frames as long as the percentage of reconstruction error (POR) of any frame is below a predetermined threshold, i.e.:

$$POR_j = \frac{\|\mathbf{F}_K\mathbf{a}_j\|_2}{\|\mathbf{f}_j\|_2} < T_P \quad (6)$$

The overall algorithm that we use for key frame selection can be summarized as follows:

Algorithm 1. Minimum sparse reconstruction based key frame selection algorithm.

Input: The expressive video $\mathbf{F} \in R^{d \times n}$ with *n* frames, where each frame is represented by the LPQ features extracted from the face region.

Output: The key frame set $\mathbf{F}_K \in R^{d \times m}$.

1. Initialize the key frame set using the first frame of the sequence as $\mathbf{F}_K = [\boldsymbol{f}_{k_1}]$ and set $m = 1$.
2. Calculate the POR for all frames in set $\mathbf{F}/\mathbf{F}_K$ using (6).
3. Repeat steps 4-6 while POR of any frame is smaller than T_P.
4. Select the next key frame using (4).
5. Increase m by one.
6. Calculate the POR for all frames in set $\mathbf{F}/\mathbf{F}_K$ using (6).

After the iterations terminate, we discard the first frame selected at the initialization step. Two examples for the key frame selection results are shown in Figure 1 and Figure 2. We can observe that the minimum number of frames that represent the whole video have been selected.

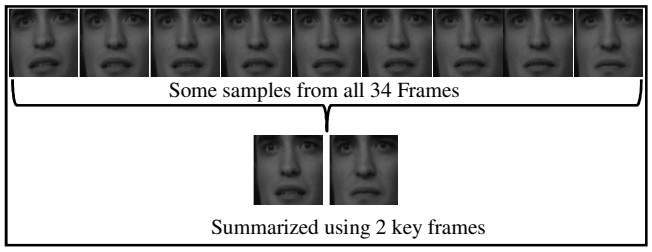

Figure 1: An example of video summarization from the AFEW 5.0 database for a sequence labeled with "sad".

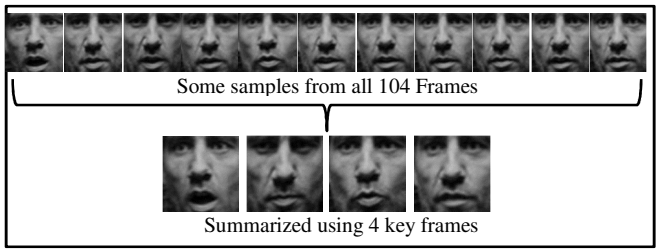

Figure 2: An example of video summarization from the test set of AFEW 5.0 database.

2.1.2 LPQ

Local Phase Quantization (LPQ) features have been originally used for blur-insensitive texture classification [17], which has also been used for facial expression recognition with success [6], [25]. LPQ features are based on computing the 2D short–term DFT of the image in a window around each pixel. Then the DFT is sampled at four frequencies, which are then decorrelated and quantized. The quantized local phase vectors are represented as integer values between 0-255. Finally, 256 bin histograms of these numbers in sub-blocks of the face region are concatenated and used as a feature vector (see Figure 3).

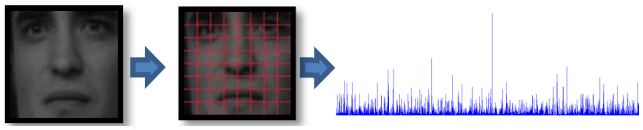

Figure 3: Sub-regions used for extraction of LPQ feature vectors.

2.1.3 LBP-TOP

Local Binary Pattern (LBP) operator considers pixels in a neighborhood and labels them by thresholding to encode local primitives such as uniform regions, edges, corners and spots etc. [16]. It has also been used for facial expression recognition [19]. LBP static descriptor was extended to describe dynamic textures and has been demonstrated to be successful for facial expression recognition [27]. LBP-TOP utilizes three orthogonal spatio-temporal planes around a pixel and computes the LBP features on the planes to obtain XY-LBP, XT-LBP and YT-LBP features, which are then concatenated to obtain a single feature vector.

2.2 Feature Extraction from Audio

In order to extract audio features for emotion recognition, we used Mel-Frequency Cepstral Coefficients (MFCC) and relative spectral features (RASTA) based on perceptual linear prediction (PLP) [11]. Before extraction of audio features, first we detected the endpoints of speech signals to estimate the starting and ending points of the speech in a given audio file. Then, we calculated the MFCC and RASTA-PLP features using filters with an order of 12 and 20, respectively, using a window of length 25msec and 50% overlap ratio. Then, we appended the 12 MFCC and 13 RASTA-PLP coefficients with their first and second time derivatives. The final audio-based feature vector was extracted by applying nine statistical functions (max, min, maximum position, minimum position, mean, variance, range, kurtosis and skewness) to the 75 elements of the MFCC and RASTA-PLP vector.

We also tested the audio features extracted using the OpenSMILE toolbox [9], [10], which were provided by the challenge organizers [5].

2.3 Classification and Fusion

We used a Support Vector Machine (SVM) classifier using chi-square and exponential chi-square kernels with one-vs-all strategy to classify the video and audio features [2], [29].

We used a score level fusion technique, where we combine the probabilities for each class, which are estimated using each modality separately. We tested several approaches for combining the probabilities [1] estimated using the SVM classifiers and the best results were obtained using the product rule [24], in which the probabilities obtained from the classification of each modality is multiplied for a given test vector and we predict the final label as the one which gives the maximum product:

$$P(\omega_k \mid x) = \prod_{i=1}^{2} P(\tilde{\omega}_k \mid x, \lambda_i) \quad , k = 1,...,6 \tag{7}$$

$$\omega^* = \max_k \{P(\omega_k \mid x)\}, \quad k = 1,...,6 \tag{8}$$

where x represents the audio-visual feature vector of the test video, ω and $\tilde{\omega}$ represent the predicted output labels after and before fusion, $P(\tilde{\omega}_k \mid x, \lambda_i)$ is the probability of class k for each individual classifier λ_i and ω^* is the final estimated class of the test video. In order to fuse the LPQ and LBP-TOP based features for the video modality, we utilized a similar approach as well (see Figure 4 and Figure 5).

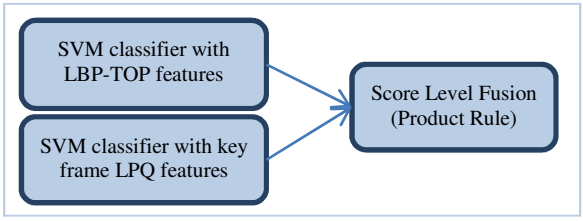

Figure 4: Fusion of LBP-TOP and LPQ video features

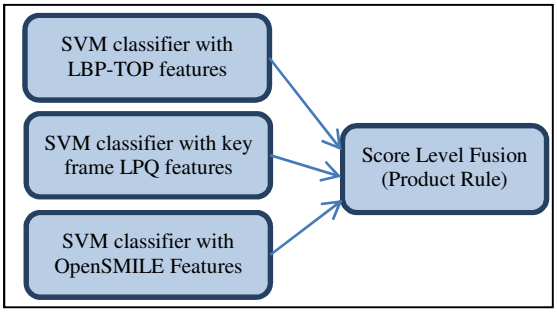

Figure 5: Fusion of LBP-TOP, LPQ based video and OpenSMILE audio features

3. EXPERIMENTAL RESULTS

3.1 EmotiW 2015 Challenge

The challenge is based on the AFEW 5.0 (Acted Facial Expressions in the Wild) database, which is divided into three parts for training, validation and testing. The numbers of samples for each emotion in each set are shown in Table 1. The labels of the test set are not given to the participants of the challenge.

The organizers of the EmotiW 2015 challenge provide the LBP-TOP feature set on the AFEW 5.0 database. After the pre-processing step for aligning and cropping the face region, LBP-TOP features are extracted from non-overlapping spatial 4x4 blocks. The LBP-TOP features from each block are concatenated to create one feature vector of size 1×2832 for each frame of a video.

Table 1: The numbers of samples for each emotion in AFEW 5.0 database for the EmotiW 2015 Challenge.

	Anger	Disg.	Fear	Happy	Neut.	Sad	Surp
Train	118	72	77	145	131	107	73
Valid.	64	40	46	63	63	61	46
Test	79	29	66	108	159	71	27

The video only baseline classification accuracy provided by the challenge organizers using an SVM classifier with Chi square kernel on the validation set is 36.08%. Similarly, baseline classification accuracy on the test set is 39.33% (see Table 2).

We would like to note that while the database comes with Zhu's [29] face tracking results, the difficult and close-to-real-world conditions cause problems even in the early stages of the processing pipeline. Some examples of incorrectly detected and tracked face images are illustrated in Figure 6.

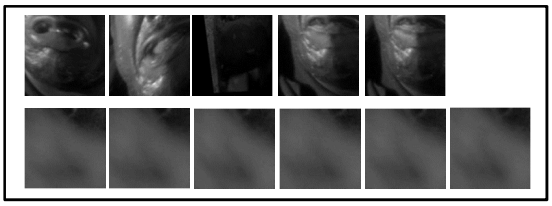

Figure 6: Noisy face detection and tracking results of two different sequences from train and validation sets of AFEW 5.0 dataset.

3.2 Experimental Setup

In our tests, we used the aligned faces provided by the challenge organizers. The size of a face region is 128×128 pixels. We divide the aligned faces into non-overlapping sub-blocks of size 16×16 and obtain 64 sub-blocks for feature extraction as shown in Figure 3. In the created sub-regions, we extract the 256 bin histogram of LPQ features of each region. The LPQ features of the 64 sub-blocks are concatenated into a single feature vector of length 1×16384.

During the key frame extraction phase, LPQ features are used as the feature vector for each frame. The threshold for percentage of reconstruction (POR), i.e. the T_P parameter has been selected as 0.80. After key frames are selected by the method explained in section 2.1.1, average LPQ features of the selected key frames are calculated and used to represent the whole video feature of the sequence.

The audio features include the first 12 MFCC features. As a common practice, delta and double-delta MFCCs (sometimes referred to as first and second derivatives, respectively) are calculated to capture local dynamics, forming a 1×36 dimensional feature vector. Then, nine statistical functions and their first and second time derivatives are applied to the first 12 MFCCs in order to get a vector of length 1×324 representing MFCC features. We also calculated 13 RASTA-PLP coefficients by using a filter of order 20 and augmented by their delta and double delta features. The same statistical parameters as in MFCCs are used for the RASTA-PLPs, giving a 1×351 dimensional feature vector. The MFCC and RASTA-PLP feature vectors are then concatenated to get an audio feature vector of length 1×675 for the a video clip.

3.3 Test Cases

We evaluated the performance of video features (LBP-TOP and LPQ`s from key frames) and audio features (OpenSMILE and MFCC & RASTA-PLP features) on the validation and test sets. The emotion recognition results of eight combinations on validation and test sets are illustrated in Table 2. The configuration used for each test case is explained below.

Case 1: Video based experiment, where key frames of all video sequences are selected using T_P as 0.80. For each video sequence, average LPQ feature vector of all key frames is used to represent the whole video. An SVM with a chi-square kernel is trained.

Case 2: Video based experiment combining case 1 and a second chi-square kernel SVM, which is trained using LBP-TOP features. Then score level fusion is applied using the product rule.

Case 3: Video based experiment, which is same as case 2 but the parameter T_P has been selected as 0.85.

Case 4: Same as case 3 with the SVM classifier trained using both the training and the validation sets of AFEW 5.0 dataset.

Case 5: Audio based experiment using feature level fusion of MFCC and RASTA-PLP features. An SVM classifier with an exponential chi-square kernel is used.

Case 6: Audio based experiment using the OpenSMILE feature set. An SVM classifier with an exponential chi-square kernel is used.

Case 7: Audio-visual experiment using score level fusion of Case 2 and Case 5.

Case 8: Audio-visual experiment using score level fusion of Case 2 and Case 6.

Table 2: Overall accuracies of the 8 test cases on validation and test sets. Numbers are given in percentages.

Case	Methods		Accuracy	
			Val.	Test
Video Based Baseline			36.08	39.33
1	Video Based	LPQ ($T_P = 0.80$)	40.70	41.37
2		LBP-TOP + LPQ ($T_P = 0.80$)	43.40	45.08
3		LBP-TOP + LPQ ($T_P = 0.85$)	44.47	44.71
4		LBP-TOP + LPQ (trained by train + val. set) ($T_P = 0.80$)	-	43.23
5	Audio Based	MFCC & RASTA-PLP	24.02	33.40
6		OpenSMILE	31.85	33.21
7	Video + Audio Based	LBP-TOP + LPQ + MFCC & RASTA-PLP	41.24	47.68
8		LBP-TOP + LPQ + OpenSMILE	40.70	49.91

The best classification accuracy on the test set is achieved for Case 8, using score level fusion of LBP-TOP, key frame LPQ's and OpenSMILE audio features, which is 49.91%. The confusion matrices of this case for the validation and test sets are shown in Table 3 and Table 4, respectively.

Table 3: Confusion matrix of Case 8 (audio-visual) on validation set. Numbers are given in percentages.

	Ang	Disg	Fear	Hap	Neut	Sad	Surp
Ang	67.8	0.00	0.00	13.5	6.78	8.47	3.39
Disg	20.5	7.69	0.00	23.08	25.64	17.95	5.13
Fear	34.0	0.00	11.3	15.91	20.45	9.09	9.09
Hap	4.7	1.59	0.00	79.37	9.52	4.76	0.00
Neut	4.92	0.00	4.92	26.23	55.74	8.20	0.00
Sad	11.8	3.39	8.47	18.64	32.20	25.42	0.00
Surp	28.2	2.17	10.8	13.04	30.43	6.52	8.70

Table 4: Confusion matrix of Case 8 (audio-visual) on test set. Numbers are given in percentages.

	Ang	Disg	Fear	Hap	Neut	Sad	Surp
Ang	73.42	1.27	3.80	6.33	12.66	2.53	0.00
Disg	13.79	0.00	0.00	37.93	17.24	20.69	10.34
Fear	31.82	3.03	18.18	7.58	10.61	16.67	12.12
Hap	5.56	0.93	0.93	73.15	8.33	9.26	1.85
Neut	5.03	2.52	0.63	15.72	55.35	18.87	1.89
Sad	12.68	4.23	2.82	21.13	14.08	42.25	2.82
Surp	22.22	0.00	7.41	14.81	29.63	18.52	7.41

Table 5: Confusion matrix of Case 2 (visual) on test set. Numbers are given in percentages.

	Ang	Disg	Fear	Hap	Neut	Sad	Sur
Ang	73.42	2.53	5.06	6.33	6.33	6.33	0.00
Disg	10.34	17.24	3.45	17.24	13.79	24.14	13.79
Fear	27.27	3.03	15.15	6.06	15.15	18.18	15.15
Hap	9.26	2.78	0.93	68.52	4.63	12.04	1.85
Neut	10.69	3.14	3.14	15.72	43.40	18.87	5.03
Sad	12.68	7.04	5.63	22.54	11.27	33.80	7.04
Surp	37.04	0.00	7.41	14.81	18.52	11.11	11.11

Table 6: Confusion matrix of Case 6 (audio) on test set. Numbers are given in percentages.

	Ang	Disg	Fear	Hap	Neut	Sad	Surp
Ang	55.70	0.00	11.39	20.25	7.59	3.80	1.27
Disg	13.79	0.00	3.45	31.03	34.48	17.24	0.00
Fear	25.76	1.52	18.18	21.21	21.21	10.61	1.52
Hap	14.81	0.93	0.00	43.52	25.93	12.04	2.78
Neut	6.29	0.63	5.66	33.33	36.48	16.35	1.26
Sad	15.49	0.00	5.63	25.35	26.76	23.94	2.82
Surp	11.11	3.70	3.70	18.52	33.33	25.93	3.70

We can see that happiness and anger have the highest accuracies. In Table 5 and Table 6, we give the confusion matrices of Case 2 (video only) and Case 6 (audio only), as well.

4. CONCLUSIONS

We presented a multimodal categorical emotion recognition method based on key frame selection from video. The key frames are selected so that they represent the content of the whole video clip in the best possible way. The video and audio features are fused at the score level using multiplication rule. We tested eight different combinations of audio and visual features on the AFEW 5.0 (Acted Facial Expressions in the Wild) database provided by the challenge organizers. The best visual and audio-visual accuracies obtained on the test set are 45.1% and 49.9% respectively, which are above the baseline accuracy of 39.3%.

5. ACKNOWLEDGMENTS

This work was partially supported by The Scientific and Technological Research Council of Turkey (TUBITAK) under project number EEAG-110E056.

6. REFERENCES

[1] Atrey, P. K., Hossain, M. A., El Saddik, A. and Kankanhalli, M. S. 2010. Multimodal fusion for multimedia analysis: a survey, *Multimedia Systems,* 16 (6), 345-379.

[2] Chang, C. C. and Lin, C. J. 2011. LIBSVM: a library for support vector machines, *ACM Transactions on Intelligent Systems and Technology (TIST),* 2, 27.

[3] Cruz, A., Bhanu, B., and Yang, S. 2011. A psychologically-inspired match-score fusion model or video-based facial expression recognition, *Proc. Int. Conf. Affective Computing and Intelligent Interaction (ACII)*, 341-350.

[4] Dhall, A., Goecke, R., Lucey, S. and Gedeon, T. 2012. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimedia,* 19(3), 34-41.

[5] Dhall, A., Ramana Murthy, O. V., Goecke, R., Joshi, J., and Gedeon, T., 2015. Video and image based emotion recognition challenges in the wild: EmotiW 2015, *ACM Int. Conf. on Multimodal Interaction (ICMI 2015).*

[6] Dhall, A., Asthana, A., Goecke R., Gedeon T. 2011. Emotion recognition using PHOG and LPQ features. In *Proceedings of the workshop on Facial Expression Recognition and Analysis Challenge FERA2011*, *IEEE Automatic Face and Gesture Recognition Conference FG2011*. Santa Barbara.

[7] Douglas-Cowie, E., Cowie, R., Schoder, M. 2000. A new emotion database: Considerations, resources and scope. *In Proceedings ISCA ITRW on Speech and Emotion*, 39–44.

[8] Erdem, C. E., Turan, C., and Aydın, Z. 2015. BAUM-2: A multilingual audio-visual affective face database. *Multimedia Tools and Applications*, 74, 18 (2015), 7429- 7459.

[9] Eyben, F., Wöllmer, M., and Schuller, B. 2010. OpenSMILE - The Munich versatile and fast open-source audio feature extractor. In *Proc. s of the 18th ACM Int. Conf. on Multimedia (ACM'MM 2010)* (Florence, Italy), 1459-1462.

[10] Eyben,F., Weninger, F., Groß, F., and Schuller, B. Recent developments in openSMILE, the Munich open-source multimedia feature extractor, In *Proceedings of the 21st ACM International Conference on Multimedia (ACM'MM 2013)* (Barcelona, Spain, October 2013), 835-838.

[11] Hermansky, H. and Morgan, N. 1994. RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, 2, 4, 578–589. DOI=10.1109/89.326616

[12] Khou, S. E., Pal, C., Bouthillier, X., Froumenty, P., Gulcehre, C., Memisevic, R., Vincent, P., Courville, A., Bengio, Y. et. al.2013. Combining modality specific deep neural networks for emotion recognition in video. *In Proc. of the 15th ACM on Int. Conf. on Multimodal Interaction* (Sydney, December 9-13, ICMI'13), 543-550.

[13] Littlewort, G. C., Bartlett, M. S., and Lee, K. 2009. Automatic coding of facial expressions displayed during posed and genuine pain, *Image and Vision Computing,* 27 (12), 1797-1803.

[14] Liu, M., Wang, R., Li, S., Shan, S., Huang, Z., Chen, X., 2014. Combining Multiple Kernel Methods on Riemannian Manifold for Emotion Recognition in the Wild, *ICMI 2014,* In *Proceedings of the 15th ACM on Int. Conf. on Multimodal Interaction,* 494 - 501.

[15] Mei, S., Guan, G., Wang, Z., Wan, S., He, M., and Feng, D. D. 2015. Video summarization via minimum sparse reconstruction. *Pattern Recognition*, 48, 522-533.

[16] Ojala T., Pietikainen M., Maenpaa T. 2002.Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal Mach. Intell.* 24 (7), 971–987.

[17] Ojansivu V., Heikkil J. 2008. Blur insensitive texture classification using local phase quantization. *Lecture Notes in Computer Science*, 5099:236–243.

[18] Ryan, A., Cohn, J. F., Lucey, S., Saragih, J., Lucey, P., and De la Torre, F. *et al.*, 2009. Automated facial expression recognition system, *Int'l Conf. on Security Technology,* 172-177. DOI=10.1109/AFGR.1998.670980"

[19] Shan, C., Gong, S., and McOwan, P.W. 2009. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27, 6, 803-816. DOI=10.1016/j.imavis.2008.08.005

[20] Sloan, D. M., and Kring, A. M. 2007. Measuring changes in emotion during psychotherapy: Conceptual and methodological issues. *Clinical Psy.: Science and Practice,* 14, 307-322. DOI=10.1111/j.1468-2850.2007.00092.x

[21] Xiong, X. and De La Torre, F. 2013. Supervised descent method and its application to face alignment. *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 532-539.*

[22] Valstar, M. F., Almaev, T., Girard, J. M., McKeown, G., Mehu, M., Yin, L., Pantic, M., Cohn, J. F. 2015. Second facial expression recognition and analysis challenge, *IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition*, Ljubljana, Slovenia.

[23] Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., Pantic, M. 2014. AVEC 2014: 3 dimensional affect and depression recognition challenge. In *Proceedings of the 4th Int. Workshop on Audio/Visual Emotion Challenge (AVEC), ACM Multimedia, 3-10*.

[24] Wang, Y., Guan, L. and Venetsanopoulos, A. N. 2012. Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition, *IEEE Transactions on Multimedia ,* 14 (3), 597-607.

[25] Yang, S. and Bhanu, B. 2011. Facial expression recognition using emotion avatar image, in *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, 866-871. DOI=10.1109/FG.2011.5771364

[26] Zeng, Z. H. , Pantic, M., Roisman, G. I., Huang, T. S., 2009. A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 31(1), 39 – 58.

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

[28] Zhalehpour, S., Akhtar Z., and Erdem, C. E. 2014. Multimodal emotion recognition with automatic peak frame selection, in *IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA) Proceedings*, Alberobello, Italy, June 23-25, 2014, 116-121.

[29] Zhu, X. and Ramanan, D., 2012. Face detection, pose estimation, and landmark localization in the wild. *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR),* pp. 2879-2886.

[30] LibSVM http://wmii.uwm.edu.pl/~ksopyla/libsvm_chi2