IJESRT: 7(5), May, 2018 ISSN: 2277-9655

International Journal of Engineering Sciences &Research Technology

(A Peer Reviewed Online Journal)
Impact Factor: 5.164

IJESRT

Website:www.ijesrt.com Mail:editor@ijesrt.com

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

RECOGNITION OF AUDIO-VISUAL EMOTIONS USING VIDEO CLIPS

Pragya Singh Tomar[*1]& Brahma Datta Shukla[2]
[*1&2]Institute of Computer Science, Vikram University, Ujjain

DOI: 10.5281/zenodo.5184752

ABSTRACT

This research describes a multimodal emotion identification system that uses auditory and visual inputs to recognize emotions. Mel-Frequency Cepstral Coefficients, Filter Bank Energies, and prosodic characteristics are retrieved from the audio channel. Two techniques are being investigated for the visual element. First, the geometric relationships between face landmarks, such as distances and angles, are calculated. Second, we condense each emotional movie into a smaller collection of key-frames that may be used to visually distinguish between different emotions. To accomplish so, key-frame summary films are fed into a convolutional neural network. Finally, in a late fusion/stacking approach, the confidence outputs of all the classifiers from all the modalities are utilized to build a new feature space to be trained for final emotion label prediction. Experiments on the SAVEE, eNTERFACE'05, and RML databases reveal that our proposed solution performs significantly better than current options, defining the current state-of-the-art in all three databases.

KEYWORDS: Multimodal Emotion Recognition, Classifier Fusion, Data Fusion, Convolutional Neural Networks.

1. INTRODUCTION

One of the most significant components in allowing robots to communicate with people is the ability to detect human intents and sentiments. For choosing the appropriate robot/machine/computer reaction, the identified emotional mood will be taken into account [1]–[4]. However, deciding on a reply based on an individual's emotional state necessitates the ability to recognize human emotions. Mobile computing [5], [6], robotics [7], health monitoring [8], [9], and gaming [10], to name a few, have all benefited from the analysis. Several difficulties can impair the performance of an algorithm developed utilizing computer vision techniques. For example, different people may express the same feeling in various ways. Furthermore, different points of view result in unequal representations of emotion. Furthermore, occlusions and variations in lighting may cause the identification technique to be deceived. If the emotion must be recognized by speech, ambient noise and variances in the voices of different participants are important elements that might impair the final recognition result. Humans utilize both aural and visual clues to correctly understand emotions. Humans utilize coverbal signals to underline the meaning of their speech, according to [11]. Body, finger, arm, and head movements, as well as face emotions like gaze and speech prosody, are examples. This is because nonverbal communication, which includes facial expressions, body language, and voice tone, accounts for 93 percent of human communication. Face detection and tracking, feature extraction, and recognition are all part of Computerized Facial Expression Recognition (FER) [12].

The face is first recognized and tracked over a series of photos that make up a video sequence. The (spatial) ratio template tracker [13], the upgraded Kanade-Lucas-Tomasi tracker [14], the AdaBoost learning algorithm [15], the ro-bust face identification algorithm [16], and the piecewise Bezier volume deformation tracker [17], among others, are instances of this method. Because facial expressions are affected by head translation, scaling, and rotation, both motion-based and model-based representations, such as geometric normalization and segmentation, are examined.

The next stage is to extract data from the identified face that will aid in determining the desired emotion [18]. Geometric and appearance traits, such as distances between two determined face landmarks or angles, are the two basic kinds of face characteristics. The geometric aspects of the face include the forms of certain portions of the face, such as the eyes, eyebrows, and mouth, as well as the placements of facial points, such as the

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

corners of the eyes and the corners of the mouth. The characteristics of the face are based on the entire face or a specific portion of it. Texture filters like Gabor can be used to extract them. They are concerned with the skin's textures, which are influenced by wrinkles, furrows, and bulges [19].

We offer an approach for recognizing emotions based on audio-visual data in this research. We use multiclass classification, in which each sample is assumed to reflect just one emotion. We employ Mel-Frequency Cepstral Coefficients (MFCCs), Filter Bank Energies (FBEs), and statistics and acoustics characteristics to analyze audio data [20], [21]. We use key-frames to encode the data in the movie, followed by face geometric relations and convolution. To learn each feature space individually, we employ state-of-the-art classifiers. In order to obtain the final classification prediction, the final prediction is TABLE 23: Fusion by using the RF–PCA on the eNTERFACE'05 database.

RF-PCA	Anger	Disgust	Fear	Sadness	Surprise	Happiness	Recognition rate (%)
Anger	204	2	0	1	3	2	96.23
Disgust	0	204	2	0	0	5	96.68
Fear	0	0	197	9	4	0	93.81
Sadness	0	0	4	204	3	0	96.68
Surprise	2	2	4	1	197	4	93.81
Happiness	3	3	0	0	1	199	96.60
						Average rate (%)	95.64

TABLE 24: Comparison of all the fusion results for the three databases.

Fusion result	SVM	SVM-PCA	**RF**	RF-PCA
SAVEE	98.1%	99.52%	**100%**	99.88%
RML	98.47%	98.89%	**99.72%**	99.58%
eNTERFACE'05	94.92%	98.33%	**98.73%**	95.64%

TABLE 26: Comparison of all the fusion methods' recogni-tion rates based on the eNTERFACE'05 database

Emotion recognition system	**Recognition rate** (%)
Hidden Markov model [31]	56.30
Neural networks [32]	67.00
Unified hybrid feature space [55]	71.00
SVM [56]	71.30
KCFA and KCCA [26]	76.00
Bayesian network models [33]	66.54
Combinational method [34]	61.10
Local Phase Quantization [57]	76.40
Our result by SVM	94.92
Our result by SVM with PCA	98.33
Our result by RF	98.73
Our result by RF with PCA	95.64

- with the same kind as the first level. SAVEE, eNTERFACE'05, and RML databases were used to generate the experimental findings. Over all of the databases, the RF classifier performed the best. The recognition rates on the mentioned databases were 99:72%, 98:73% and 100%, respectively. They improved by 0:72 percent, 22:33 percent, and 9:17 percent, respectively, as compared to earlier state-of-the-art findings on the same datasets and modalities. Fear was the most often mislabeled term. We intend to expand the list of key-frames in future study to allow the study to cover other aspects of the emotion movies in order to better distinguish between fear and happiness, as well as rage and disgust. Similarly, we want to use 3D convolutions and RNN-LSTM to augment the CNN section of the model to include extra temporal information [65].

TABLE 29: The number of times the label combo-nations with the greatest misclassification rates were repeated in each of the three datasets.

Label combination (%)	F + H	A + D	D+ F	F+ SU	SU+H	SA+ SU	F+SA
Repetition	3	2	2	2	1	1	1

TABLE 30: The total number of label repeats in the combinations with the greatest misclassification rates, in each database.

	Fear	Happiness	Anger	Disgust	Sadness	Surprise
SAVEE	2	1	1	2	1	1
RML	3	2	0	1	0	2
eNTERFACE'05	3	1	1	1	1	1
Summation	8	4	2	4	2	4

2. CONCLUSION

We demonstrated an audio-visual emotion recognition system. Prosodic features, MFCCs, and FBEs were among the audio features. Estimated key-frames representing each video material in terms of representative face expressions were used to compute visual characteristics. Both geometric characteristics and a CNN-based model were used to characterize visual data. Four different classification techniques were used: multiclass SVM, RF with and without PCA, and RF with and without PCA. The output confidence values of the first classifier were fused to define a new feature vector that was learned by a second level classifier after each set was trained individually.

The Estonian Research Grant (PUT638), the Spanish projects TIN2013-43478-P and TIN2016-74946-P, and the European Commission Horizon 2020 grant project SEE.4C under call H2020-ICT-2015, as well as the Estonian Centre of Excellence in IT (EXCITE) funded by the European Regional Development Fund and the European Network on Integrating Vision and Language (iV&L Net) ICT, have all contributed to this work. The authors are also appreciative for NVIDIA Corporation's support in the form of a Titan X Pascal GPU donation.

REFERENCES

[1] K. Kim, Y.-S. Cha, J.-M. Park, J.-Y. Lee, and B.-J. You, "Providing services using network-based humanoids in a home environ-ment," IEEE Transactions on Consumer Electronics, vol. 57, no. 4,1628–1636, 2011.

[2] Zaraki, D. Mazzei, M. Giuliani, and D. De Rossi, "Design-ing and evaluating a social gaze-control system for a humanoid robot," IEEE Transactions on Human-Machine Systems, vol. 44, no. 2,157–168, 2014.

[3] Lusi,¨ J. C. S. Jacques Junior, J. Gorbova, X. Baro,´ S. Escalera, H. Demirel, J. Allik, C. Ozcinar, and G. Anbarjafari, "Joint chal-lenge on dominant and complementary emotion recognition using micro emotion features and head-pose estimation: Databases," in Automatic Face and Gesture Recognition, 2017. Proceedings. 12th IEEE International Conference on. IEEE, 2017.

[4] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbar-jafari, "Fusion of classifier predictions for audio-visual emotion recognition," in Pattern Recognition (ICPR), 2016 23rd International Conference on. IEEE, 2016, pp. 61–66.

[5] Z. Lv, S. Feng, L. Feng, and H. Li, "Extending touch-less interaction on vision based wearable device," in 2015 IEEE Virtual Reality (VR). IEEE, 2015, pp. 231–232.

[6] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech,"IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 115–126, 2015.

[7] E. Russell, A. Stroud, J. Christian, D. Ramgoolam, and A. B. Williams, "Smile: A portable humanoid robot emotion interface,"in 9th ACM/IEEE International Conference on Human-Robot Interaction,

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7
Workshop on Applications for Emotional Robots, HRI14, Bielefeld University, Germany, 2014, pp. 1–5.
[8] J. Torous, R. Friedman, and M. Keshavan, "Smartphone ownership and interest in mobile applications to monitor symptoms of mental health conditions," JMIR mHealth and uHealth, vol. 2, no. 1, p. e2, 2014.
[9] M. S. Hossain and G. Muhammad, "Cloud-assisted speech and face recognition framework for health monitoring," Mobile Networks and Applications, vol. 20, no. 3, pp. 391–399, 2015.
[10] M. Szwoch and W. Szwoch, "Emotion recognition for affect aware video games," in Image Processing & Communications Challenges 6. Springer, 2015, pp. 227–236.
[11] D. Bolinger and D. L. M. Bolinger, Intonation and its uses: Melody in grammar and discourse. Stanford University Press, 1989.
[12] T. Wu, S. Fu, and G. Yang, "Survey of the facial expression recognition research," in International Conference on Brain Inspired
Cognitive Systems. Springer, 2012, pp. 392–402.
[13] K. Anderson and P. W. McOwan, "A real-time automated system for the recognition of human facial expressions," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 36, no. 1, pp. 96–105, 2006.
[14] N. Ramakrishnan, T. Srikanthan, S. K. Lam, and G. R. Tulsulkar, "Adaptive window strategy for high-speed and robust klt feature tracker," in Pacific-Rim Symposium on Image and Video Technology. Springer, 2015, pp. 355–367.
[15] Y. N. Chae, T. Han, Y.-H. Seo, and H. S. Yang, "An efficient face detection based on color-filtering and its application to smart devices," Multimedia Tools and Applications, pp. 1–20, 2016.
[16] C. Ding, J. Choi, D. Tao, and L. S. Davis, "Multi-directional multi-level dual-cross patterns for robust face recognition," IEEE transactions on pattern analysis and machine intelligence, vol. 38, no. 3, pp. 518–531, 2016.
[17] H. Dibeklio˘ glu, F. Alnajar, A. A. Salah, and T. Gevers, "Combining facial dynamics with appearance for age estimation," IEEE Transactions on Image Processing, vol. 24, no. 6, pp. 1928–1943, 2015.
[18] G. Hemalatha and C. Sumathi, "A study of techniques for facial detection and expression classification," International Journal of Computer Science and Engineering Survey, vol. 5, no. 2, p. 27, 2014.
[19] M. Pantic and M. S. Bartlett, Machine analysis of facial expressions. I-Tech Education and Publishing, 2007.
[20] D. Kami´ nska, T. Sapi ´ nski, and G. Anbarjafari, "Efficiency of chosen speech descriptors in relation to emotion recognition," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2017, no. 1, p. 3, 2017.
[21] F. Noroozi, T. Sapi ´ nski, D. Kami´ nska, and G. Anbarjafari, "Vocalbased emotion recognition using random forests and decision tree," International Journal of Speech Technology, pp. 1–8, 2017.
[22] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion(savee) database," 2014.
[23] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audiovisual emotion database," in 22nd International Conference on Data Engineering Workshops (ICDEW'06). IEEE, 2006, pp. 8–8.
[24] zhibing xie, "Ryerson multimedia research laboratory (rml)." [25] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski et al., "Emonets: Multimodal deep learning approaches for emotion recognition in video," Journal on Multimodal User Interfaces, pp. 1–13, 2015.
[26] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel crossmodal factor analysis for information fusion with application to bimodal emotion recognition," IEEE Transactions on Multimedia, vol. 14, no. 3, pp. 597–607, 2012.
[27] M. Wimmer, B. Schuller, D. Arsic, G. Rigoll, and B. Radig, "Lowlevel fusion of audio, video feature for multi-modal emotion recognition." in VISAPP (2), 2008, pp. 145–151.
[28] F. Cid, L. J. Manso, and P. N´ unez, "A novel multimodal emotion recognition approach for affective human robot interaction."
[29] S. Haq, T. Jan, A. Jehangir, M. Asif, A. Ali, and N. Ahmad, "Bimodal human emotion classification in the speaker-dependent scenario," PAKISTAN ACADEMY OF SCIENCES, p. 27.
[30] D. Gharavian, M. Bejani, and M. Sheikhan, "Audio-visual emotion recognition using fcbf feature selection method and particle

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

swarm optimization for fuzzy artmap neural networks," Multimedia Tools and Applications, pp. 1–22, 2016.

[31] D. Datcu and L. Rothkrantz, "Multimodal recognition of emotions in car environments," DCI&I 2009, 2009.

[32] M. Paleari, B. Huet, and R. Chellali, "Towards multimodal emotion recognition: a new approach," in Proceedings of the ACM International Conference on Image and Video Retrieval. ACM, 2010, pp. 174–181.

[33] D. Jiang, Y. Cui, X. Zhang, P. Fan, I. Ganzalez, and H. Sahli, "Audio visual emotion recognition based on triple-stream dynamic bayesian network models," in International Conference on Affective Computing and Intelligent Interaction. Springer, 2011, pp. 609–618.

[34] K.-C. Huang, H.-Y. S. Lin, J.-C. Chan, and Y.-H. Kuo, "Learning collaborative decision-making parameters for multimodal emotion recognition," in 2013 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2013, pp. 1–6.

[35] A. Gera and A. Bhattacharya, "Emotion recognition from audio and visual data using f-score based fusion," in Proceedings of the 1st IKDD Conference on Data Sciences. ACM, 2014, pp. 1–10.

[36] C. Fadil, R. Alvarez, C. Mart´ınez, J. Goddard, and H. Rufiner, "Multimodal emotion recognition using deep networks," in VI Latin American Congress on Biomedical Engineering CLAIB 2014, Paran´a, Argentina 29, 30 & 31 October 2014. Springer, 2015, pp. 813–816.

[37] K. Seng, L.-M. Ang, and C. Ooi, "A combined rule-based and machine learning audio-visual emotion recognition approach," IEEE Transactions on Affective Computing, 2015.

[38] S.-M. Guo, Y. Pan, Y.-C. Liao, C. Hsu, J. S. H. Tsai, and C. Chang, "A key frame selection-based facial expression recognition system," in First International Conference on Innovative Computing, Information and Control-Volume I (ICICIC'06), vol. 3. IEEE, 2006, pp. 341–344.

[39] A. R. Doherty, D. Byrne, A. F. Smeaton, G. J. Jones, and M. Hughes, "Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs," in Proceedings of the 2008 international conference on Content-based image and video retrieval. ACM, 2008, pp. 259–268.

[40] Q. Zhang, S.-P. Yu, D.-S. Zhou, and X.-P.Wei, "An efficient method of key-frame extraction based on a cluster algorithm," Journal of human kinetics, vol. 39, no. 1, pp. 5–14, 2013.

[41] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on, vol. 1. IEEE, 1998, pp. 866–870.

[42] S. Hasebe, M. Nagumo, S. Muramatsu, and H. Kikuchi, "Video key frame selection by clustering wavelet coefficients," in Signal Processing Conference, 2004 12th European. IEEE, 2004, pp. 2303–2306.

[43] S. Planet and I. Iriondo, "Comparison between decision-level and feature-level fusion of acoustic and linguistic features for spontaneous emotion recognition," in Information Systems and Technologies (CISTI), 2012 7th Iberian Conference on. IEEE, 2012, pp. 1–6.

[44] S. Haq and P. J. Jackson, "Multimodal emotion recognition," Machine audition: principles, algorithms and systems, pp. 398–423, 2010.

[45] F. Noroozi, T. Sapi ´ nski, D. Kami´ nska, and G. Anbarjafari, "Vocalbased emotion recognition using random forests and decision tree," International Journal of Speech Technology, pp. 1–8.

[46] I. L¨ usi, S. Escarela, and G. Anbarjafari, "Human head pose estimation on sase database using random hough regression forests," in International Workshop on Face and Facial Expression Recognition from Real World Videos. Springer, 2016, pp. 137–150.

[47] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE transactions on acoustics, speech, and signal processing, vol. 28, no. 4, pp. 357–366, 1980.

[48] I. Bocharova, Compression for multimedia. Cambridge University Press, 2010.

[49] S. K. Kopparapu and M. Laxminarayana, "Choice of mel filter bank in computing mfcc of a resampled speech," in Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on. IEEE, 2010, pp. 121–124.

[50] Y. A. K. J. Calder, A.J. and M. Dean, "Configural information in facial expression perception. journal of experimental psychology: Human perception and performance," Journal of Experimental Psychology: Human perception and performance, vol. 26, no. 2, p. 527, 2000.

[51] F.-M. A. Calvo, M.G. and L. Nummenmaa, "Facial expression recognition in peripheral versus central vision: Role of the eyes and the mouth," Psychological research, vol. 78, no. 2, pp. 180–195, 2014.
[52] S. Lindsay, "A tutorial on princial components analysis," 2002.
[53] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273–297, 1995.
[54] L. B. Statistics and L. Breiman, "Random forests," pp. 5–32, 2001.
[55] M. Mansoorizadeh and N. M. Charkari, "Multimodal information fusion application to human emotion recognition from face and
speech," Multimedia Tools and Applications, vol. 49, no. 2, pp. 277– 297, 2010.
[56] V. ˇStruc, F. Mihelic et al., "Multi-modal emotion recognition using canonical correlations and acoustic features," in Pattern Recognition (ICPR), International Conference on. IEEE, 2010, pp. 4133–4136.
[57] S. Zhalehpour, Z. Akhtar, and C. E. Erdem, "Multimodal emotion recognition with automatic peak frame selection," in Innovations in Intelligent Systems and Applications (INISTA) Proceedings, 2014 IEEE International Symposium on. IEEE, 2014, pp. 116–121.
[58] "Surrey audio-visual expressed emotion (savee) database," http: //kahlan.eps.surrey.ac.uk/savee/Database.html, accessed: 2017- 04-08.
[59] "Rml emotion database," http://www.rml.ryerson.ca/ rml-emotion-database.html, accessed: 2017-04-07.
[60] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[61] R. Rothe, R. Timofte, and L. Van Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," International Journal of Computer Vision, pp. 1–14, 2016.
[62] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in Proceedings of COMPSTAT'2010. Springer, 2010, pp. 177–186.
[63] Q. V. Le, "A tutorial on deep learning part 1: Nonlinear classifiers and the backpropagation algorithm," 2015.
[64] Y. Kim and E. Mower Provost, "Say cheese vs. smile: Reducing speech-related variability for facial emotion recognition," in Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014, pp. 27–36.
[65] Q. Li, Z. Qiu, T. Yao, T. Mei, Y. Rui, and J. Luo, "Action recognition by learning deep multi-granular spatio-temporal video representation," in Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval. ACM, 2016, pp. 159–166.