Recognition of Emotions using Energy Based Bimodal Information Fusion and Correlation

Krishna Asawa, Priyanka Manchanda[1]

[1]Department of Computer Science Engineering and Information Technology
Jaypee Institute of Information Technology, Noida, India

***Abstract* — Multi-sensor information fusion is a rapidly developing research area which forms the backbone of numerous essential technologies such as intelligent robotic control, sensor networks, video and image processing and many more. In this paper, we have developed a novel technique to analyze and correlate human emotions expressed in voice tone & facial expression. Audio and video streams captured to populate audio and video bimodal data sets to sense the expressed emotions in voice tone and facial expression respectively. An energy based mapping is being done to overcome the inherent heterogeneity of the recorded bi-modal signal. The fusion process uses sampled and mapped energy signal of both modalities's data stream and further recognize the overall emotional component using Support Vector Machine (SVM) classifier with the accuracy 93.06%.**

***Keywords* — Bimodal Fusion, Emotion Recognition, Intelligent Systems, Machine Learning, Energy Mapping**

I. INTRODUCTION

MULTI-SENSOR information fusion is a rapidly developing area of research and development which forms the foundation of intelligent robotic control. It comprises of methods and techniques which collect input from multiple similar or dissimilar sources and sensors, extract the required information and fuse them together to achieve improved accuracy in inference than that could be achieved by the use of a single data source alone. In this contribution, we discuss a novel approach to fuse heterogeneous datasets obtained from multiple sensors with the aim of analyzing the human's emotional behavior.

Emotions play an important role in human-to-human communication and interaction, allowing people to express themselves beyond the verbal domain. The ability to understand human emotions is desirable for the computer in some applications such as computer-aided learning or user-friendly on-line help. During an interaction, an individual uses multiple modalities such as eye gaze, hand gestures, facial expressions, body posture, and tone of voice. Human behavior is thus, inherently multimodal. In addition to its multimodal nature, the emotional state of an individual is also an integral component of human experience and plays a significant role in developing intelligent systems for human computer communication. It influences numerous phenomenons such as cognition, perception, learning, creativity and decision-making. Besides the problem solving, reasoning, perception and cognitive tasks, emotion recognition also plays a pivot role in functions which are essential for artificial intelligence.

Considering these two aspects of human behavior, we have designed and developed a technique to analyze and correlate bimodal data sets and further recognize the emotional component from these fused data sets. This new technology ensures a proper balance between emotion recognition and cognition tasks.

The existing fusion methods as listed in the section- related work, do not address how to bridge the heterogeneity present in the captured data, which corresponds to the individual modality. The energy based mapping method inspired from how the different sensed stimuli signals by humans, mapped to the corresponding energy onto designated areas of the brain. This method brings homogeneity among heterogeneous emotional cues by transforming them onto their corresponding energy levels. The achieved fusion accuracy of 93.06% can ensures a proper balance between emotion recognition and cognition tasks.

The rest of the paper has been organized in the following manner: in Section II, along with existing fusion approaches, we discuss an energy based method for fusion of multimodal data sets. In Section III, we explain the architectural framework of our model. The implementation of the solution is delineated in Section IV. Section V outlines the applications of this model. Lastly, we conclude the research study in Section VI.

II. RELATED WORK

The wide use of multimodal data fusion technologies in versatile areas of application has invoked an ever increasing interest of researchers all over the globe. Multimodal data fusion techniques are used in numerous areas such as intelligent systems, robotics, sensor networks, video and image processing and many more. Multimodal data fusion can be performed at three levels: feature, decision and hybrid level fusion.

Feature level fusion has been used in [1] for fusing range of spatial cues with the relative assignment of linear weight to them. But they have unable to resolve the issue of how weights

DOI: 10.9781/ijimai.2014.272

should be assigned to justify relevance and importance of different cues.

Neti [2] have been performed decision level fusion for speaker recognition and speech event detection. They have analyzed audio features (e.g. phonemes) and visual features (e.g. visemes) independently to arrive at recognized decision according to single modality. Thereafter they have employed a linear weighted sum strategy to fuse these individual decisions. The authors have used the training data to determine the relative reliability of the different modalities and accordingly adjusted their weights. Where as in [3], for speaker identification, they have considered the results of different classifier at decision level fusion. From the speech corpora, a set of patterns are identified for each speaker on the basis of predefined features by two different classifiers. The majority decision regarding the identity of the unknown speaker is obtained by fusing the output scores of all the classifiers using a late integration approach.

A multimodal integration approach using custom defined rules has been suggested by, Holzapfel et al. [4]. They have shown smooth human - robot interaction in the kitchen setting by fusing results of speech and 3D pointing gestures. This multimodal fusion which is performed at the decision level based on the n-best lists generated by each of the event parsers. A close correlation in time of speech and gesture has been proved by this approach, but this is leading to the process time overhead to determine the best action based on n-best fused input.

In [5] two techniques viz (1) Gradient-descent-optimization linear fusion (GLF) and (2) the super-kernel nonlinear fusion (NLF) are suggested. Each of which does the optimal combination of multimodal information for video concept detection. In GLF, an individual kernel matrix is first constructed and then fused together based on a weighted linear combination scheme. Unlike GLF, the NLF method does nonlinear combination of multimodal information.

In [6], the authors have used NLF method and first construct an SVM for the individual modality as a classifier. Thereafter, for optimal combination of the individual classifier models a super kernel non-linear fusion is applied. Experiments conducted on TREC-2003 Video Track benchmark shows NLF has on average 3.0% better performance than GLF.

To classify image, Zhu et al. [7] have given a hybrid level multimodal fusion framework. They have used SVM to classify the images with embedded text within their spatial coordinates. The fusion process is done in two steps. Firstly, on the basis of low-level visual features, a bag-of-words model [8] is used to classify the given image. At the same time, the text detector records the existence of text in the image using text color, size, location, edge density, brightness, contrast, etc. In the second step, for fusing the visual and textual features together a pair-wise SVM classifier is used.

A time-delayed neural network employed by Cutler and Davis [9] for feature level multimodal data fusion in for locating the speaking person in the scene. This is being done by identifying the correlation between audio and visual streams.

In another work, related to detecting human activities Gandetto et al. [10] have used the Neural Network decision level fusion method to combine sensory data. An environment equipped with a heterogeneous network of state sensors for sensing CPU load, login process, and network load and cameras for sensing observation along with computational units working together in a LAN is considered for the experiment. Human activity is monitored by fusing the data from these two types of sensors at the decision level.

A framework is given in [16] which fuse textual and visual information. Author has proposed additional preprocessing before combining these modalities in a linear weighted fashion at the feature and scoring levels. The pre-processing called as latent semantic mixing, takes care about overlapping information among both modalities by mapping the bimodal feature space onto low dimensional semantic space.

In this paper, we propose a feature level linear weighted fusion model based on a human-inspired concept of brain energy mapping model. Humans collect sensory data via human biological senses (sight, hearing, touch, smell and taste) and map this data as energy stimuli onto designated regions of the brain. The brain then fuses them together to obtain an inference. This analogy is employed in designing the architectural framework of our work. This phenomenon is depicted in Fig 1.

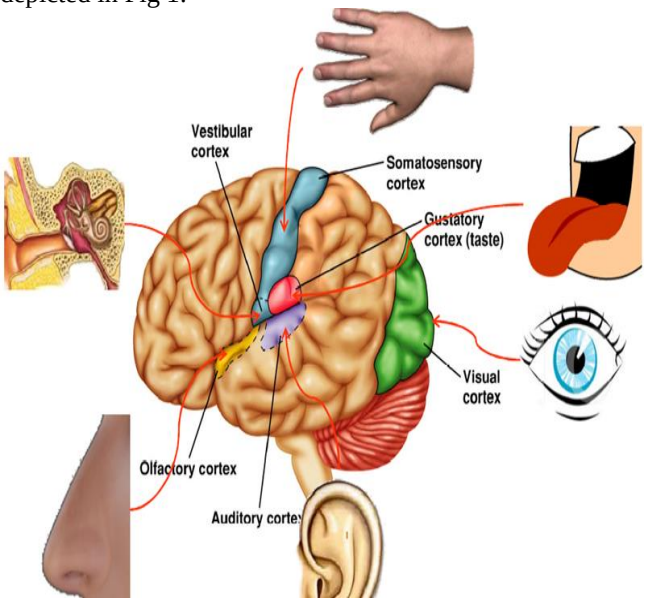

Fig. 1. The Brain Energy Mapping Model

III. ARCHITECTURAL FRAMEWORK

Figure 2 shows the overall architectural framework and computation stages as listed below.

- Step1: Obtain Bi-Modal Input stream
- Step 2: Split Bi-Modal Input into Audio and Video Components
- Step 3: Synchronized Sampling and processing of Audio and Video Components

- Step 4: Run two parallel process, each for audio and video.
- Audio thread performs segmentation and feature extraction for audio sample using praat tool.
- Video thread performs segmentation and facial feature extraction for video sample DAFL library.
- Step 5: Estimation of Audio and Video features energy.
- Step 6: Perform one of the following depending on user's input:
- Train SVM
- Test an unknown sample with trained SVM to predict emotion
- Step 7: Display emotion to the user

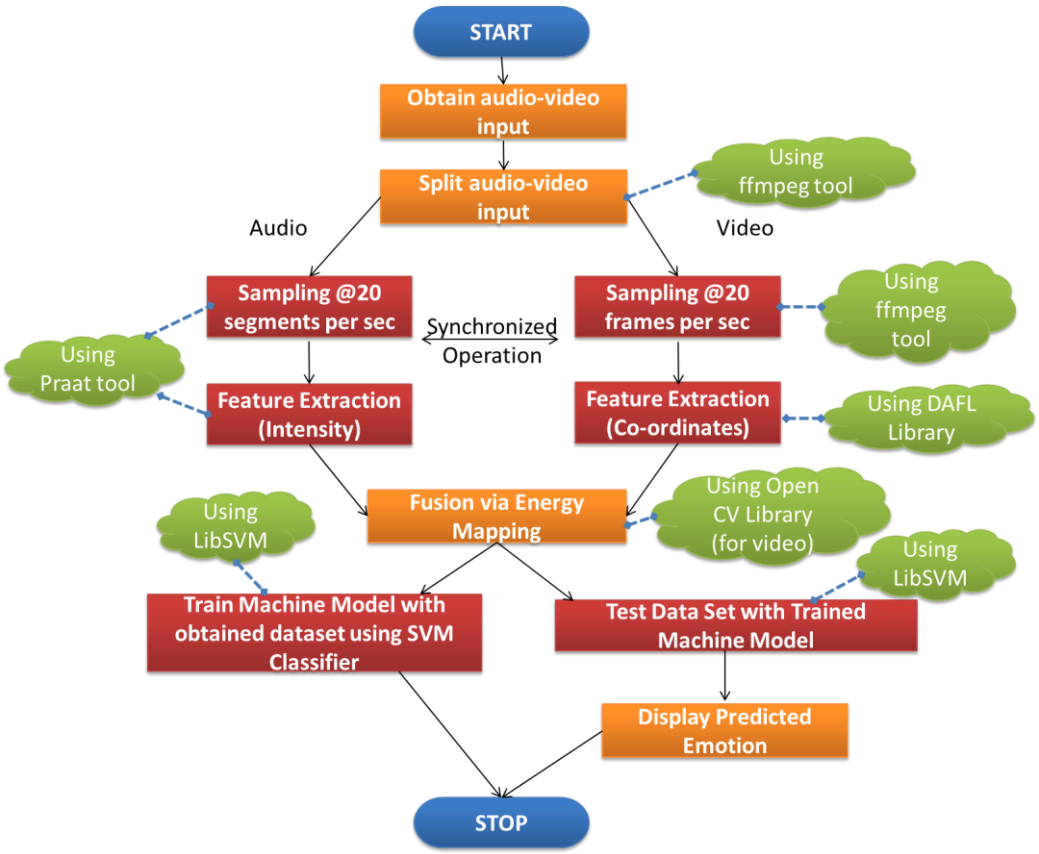

Fig 2. Architectural Framework of Bimodal Energy Based Fusion Model

Stage I – Data Pre-Processing

The bimodal inputs obtained and then split into two components – audio and video. Thereafter, audio processing and video processing is performed simultaneously and in synchronization. The synchronization is necessary to ensure that no data is lost and the audio and video samples at any particular instance are processed simultaneously.

The audio component is segmented at the rate of 20 samples (utterances) per second. Video Sampling is done at the rate of 20 frames per second.

Stage II – Feature Extraction

The prosodic feature mean intensity of the audio component between time 't1' and 't2' is computed as:

$$\frac{1}{(t2-t1)}\int_{t1}^{t2} x(t)dt \qquad (1)$$

where x(t) is intensity as function of time (in dB).

To compute the intensity, the values in the sound are first squared, then convolved with a Gaussian analysis window (Kaiser-20; sidelobes below -190 dB). The effective duration of this analysis window is 3.2 / (minimum_pitch), which guarantee that a periodic signal is analysed as having a pitch-synchronous intensity ripple not greater than 0.00001 dB.

The processing of video frames is done in two steps:

- Facial Feature extraction using the Discrete Area Filters (DAF) Library used to extract coordinates of 15 facial feature points. [13]
- Energy (gradient) computation (Fig 3) of extracted facial features co-ordinates using OpenCV Library.

A	B	C
D	E	F
G	H	I

$$energy(E) = \sqrt{xenergy^2 + yenergy^2}$$
$$xenergy = a + 2d + g - c - 2f - i$$
$$yenergy = a + 2b + c - g - 2h - i$$

Fig 3. Energy (Gradient) Computation

In Fig. 3, each lowercase letter represents the brightness (sum of the red, blue, and green values) of the corresponding pixel. To compute the energy of edge pixels, we consider that the image is surrounded by a 1 pixel wide border of black pixels (with 0 brightness).

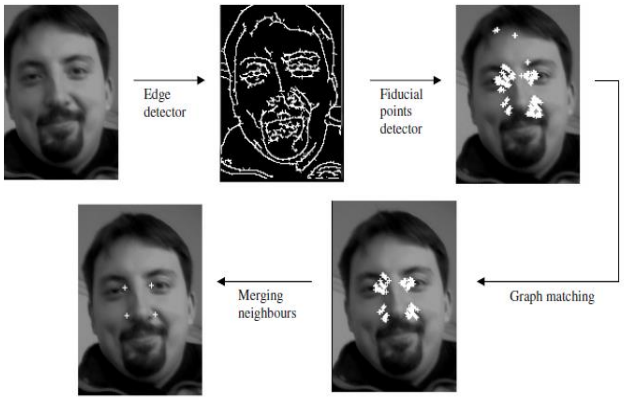

Fig 4. Facial Feature Detection using Discrete Area Filters

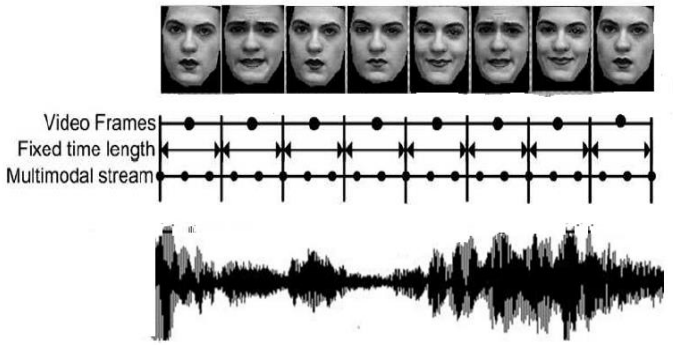

Fig 5. Bi-modal Input Processing

Stage III – Fusion via Energy Mapping

Our framework uses the technique of feature level linear weighted fusion. Consider a feature set <Ev, Ea>, where Ev is the total energy of video features and Ea is the total energy of audio features. The feature set is computed at intervals of 1 second for the bi-modal input.

$$E_a = wE_{a1} + wE_{a2} + \ldots\ldots\ldots + wE_{an} \quad (2)$$

where,

$E_a$is the audio energy of the audio sample of 1sec duration.

E_{ai} is the audio energy of i^{th} sub-sample

w is the weight assigned to each sub-sample.

n is the number of sub samples = 20

$$E\_vi = w\_1\ (E\_vio + E\_vi1 + E\_vi2 + E\_vi3) + w\_2\ (E\_vi4 + E\_vi5 + E\_vi6 + E\_vi7) + w\_3\ (E\_vi8 + E\_vi9 + E\_vi10) + w\_4\ (E\_vi11 + E\_vi12 + E\_vi13 + E\_vi14) \quad (3)$$

where,

E_{vi} is the energy of i^{th} frame of the video sub-sample.

w_1 is weight assigned to left eye fiducial points.

w_2 is weight assigned to right eye fiducial points.

w_3 is weight assigned to nose fiducial points.

w_4 is weight assigned to mouth fiducial points.

$$w_4 > w_1, w_2 > w_3$$

$$E_v = wE_{v1} + wE_{v2} + \ldots\ldots\ldots + wE_{vn} \quad (4)$$

where,

E_v is the combined energy of the video sub-samples.

E_{vi} is the energy of i^{th} video sub-sample.

w is the weight assigned to each sub-sample.

n is the number of sub samples = 20

We further label the feature set with appropriate class (1 – Happy, 2 – Anger, 3 – Fear) depending on the emotional state of the user. This feature set is then used to train the machine model designed for predicting the mood of the user.

Stage IV – Emotion Prediction

The Support Vector Machine (SVM) classifier is used to predict the emotional state of the bimodal input. We use LibSVM[14] to develop the C-SVC (C - Support Vector Classification) SVM having RBF (Radial Basis Function - exp(-gamma*|u-v|^2)) kernel.

We further develop a machine model using C-SVC SVM and train it using the feature sets obtained in Stage III. The feature sets of the input to be tested are then labeled with an arbitrary label. These are then tested using the trained machine model. Finally, the predicted emotional state of the bi-modal input is displayed to the user.

IV. RESULTS AND ACCURACY CALCULATION

The energy based bimodal data fusion model was tested for the eNTERFACE[15] database with 3 discrete emotions that are happy, anger and fear. The specifications of the database are as follows:

- 648 samples
  - 43 subjects enacting 5 sentences of each of 3 emotions (happy, anger and fear)
- Samples having both male and female subject
- Frontal views with moderate lighting conditions
- Single person input

The 80: 20 ratios of the training and testing samples are considered for cross validation. The total samples for each emotion are 215. Three times process has been repeated with

different sets of training and testing in the ratio of 80:20. On average in each round 485 samples are classifies correctly on the basis of 518 samples. The average percentage of classification for the three emotions is shown in the table 1.

TABLE 1. CONFUSION MATRIX (IN %)

Predicted Emotion				
		Happy	**Anger**	**Fear**
Actual Emotion	**Happy**	92.59%	4.17%	3.24%
	Anger	1.85%	94.44%	3.70%
	Fear	1.85%	6.02%	92.13%

The model shows 93.06% accuracy for emotion recognition of Happy, Anger and Fear Emotions using energy mapping model.

V.CONCLUSION

In this research study, we have developed a tool to analyze and correlate bimodal data sets of emotional cues using energy based fusion model and further recognized the emotional component from these bimodal data sets using Support Vector Machine classifier. We have mapped the audio and video features of bimodal input to their corresponding energy levels. The model is tested for eNTERFACE 2005 database and an accuracy of 93.06% is obtained recognition of happy, anger and fear emotions. The tool developed for bimodal energy based fusion model can further be used as a wrapper tool to develop intelligent applications which require multimodal data fusion and emotion recognition, such as Real Time Emotion Recognition, Expressive Embodied Conversational Agent, Virtual Tutor, Questionnaire which analyse verbal and non-verbal behavior.

ACKNOWLEDGMENTS

The work reported in this paper is supported by the grant received from All India Council for Technical Education; A Statutory body of the Govt. of India. vide f. no. 8023/BOR/RID/RPS-129/2008-09.

REFERENCES

[1] Wang, J., Kankanhalli, M. S., Yan, W., & Jain, R. (2003). Experiential sampling for video surveillance. In First ACM SIGMM international workshop on Video surveillance (pp. 77-86).

[2] Neti, C., Maison, B., Senior, A. W., Iyengar, G., Decuetos, P., Basu, S., &Verma, A. (2000). Joint processing of audio and visual information for multimedia indexing and human-computer interaction(pp. 294-301).

[3] Radová, V., & Psutka, J. (1997). An approach to speaker identification using multiple classifiers. In Acoustics, Speech, and Signal Processing, ICASSP-97 (Vol. 2, pp. 1135-1138).

[4] Holzapfel, H., Nickel, K., & Stiefelhagen, R. (2004). Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3D pointing gestures. In Proceedings of the 6th international conference on Multimodal interfaces (pp. 175-182).

[5] Wu, K., Lin, C.K., Chang, E., Smith, J.R. (2004) Multimodal information fusion for video concept detection. In: IEEE International Conference on Image Processing, Singapore (pp. 2391–2394).

[6] Adams, W. H., Iyengar, G., Lin, C. Y., Naphade, M. R., Neti, C., Nock, H. J., & Smith, J. R. (2003). Semantic indexing of multimedia content using visual, audio, and text cues. EURASIP Journal on Advances in Signal Processing (pp. 170-185).

[7] Zhu, Q., Yeh, M. C., & Cheng, K. T. (2006). Multimodal fusion using learned text concepts for image categorization. In Proceedings of the 14th annual ACM international conference on Multimedia (pp. 211-220).

[8] Li, F.F., Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington (vol. 2, pp. 524–531)

[9] Cutler, R., & Davis, L. (2000). Look who's talking: Speaker detection using video and audio correlation. In Multimedia and Expo, 2000. ICME 2000 (Vol. 3, pp. 1589-1592).

[10] Gandetto, M., Marchesotti, L., Sciutto, S., Negroni, D., Regazzoni, C.S. (2003). From multi-sensor surveillance towards smart interactive spaces. In: IEEE International Conference on Multimedia and Expo, Baltimore (pp. I:641–644).

[11] Bellard, F., & Niedermayer, M. (2012). FFmpeg. http://ffmpeg.org

[12] Boersma, Paul &Weenink, David (2014). Praat: doing phonetics by computer [Computer program]. Version 5.3.77, retrieved 18 May 2014 from http://www.praat.org/.

[13] Naruniec, J., &Skarbek, W. (2007). Face detection by discrete gabor jets and reference graph of fiducial points. In Rough Sets and Knowledge Technology Springer Berlin Heidelberg (pp. 187-194).

[14] Martin, Olivier, et al. 2006. The eNTERFACE' 05 Audio-Visual Emotion Database. Data Engineering Workshops, Proceedings.

[15] Chih-Chung Chang and Chih-Jen Lin (2006). LIBSVM : a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[16] [16] Nam Khanh Tran (2012) Multimodal Fusion for Combining

[17] Textual and Visual Information in a Semantic mode, Thesis submitted to Universitat Des Saarlandes.

Dr. Krishna Asawa presently working with Jaypee Institute of Information Technology (JIIT), Deemed to be University, NOIDA, INDIA in the capacity of Associate Professor. Dr. Krishna awarded Doctor of Philosophy (CSE) in 2002 from Banasthali Vidyapith, Deemed to be University, Banasthali, INDIA. Her area of interest and expertise includes Soft Computing and its applications, Information Security, Knowledge and Data Engineering. Before joining to the JIIT she worked with National Institute of Technology, Jaipur, INDIA and with Banasthali Vidyapith.

Ms. Priyanka Manchanda has completed her graduation in Computer Science and Engineering from Jaypee Institute of Information Technology (JIIT), Deemed to be University, NOIDA, INDIA in 2014. She is currently pursuing MS at Columbia University, New York.