NeuroImage xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Review

Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals

Gyanendra K. Verma *, Uma Shanker Tiwary [1]

Indian Institute of Information Technology Allahabad, Deoghat, Jhalwa, Allahabad 211012, India

ARTICLE INFO

Article history:
Accepted 4 November 2013
Available online xxxx

Keywords:
Multimodal fusion
Multiresolution
Emotion recognition
Wavelet transforms
Discrete wavelet transforms
Physiological signals
EEG
SVM
KNN

ABSTRACT

The purpose of this paper is twofold: (i) to investigate the emotion representation models and find out the possibility of a model with minimum number of continuous dimensions and (ii) to recognize and predict emotion from the measured physiological signals using multiresolution approach. The multimodal physiological signals are: Electroencephalogram (EEG) (32 channels) and peripheral (8 channels: Galvanic skin response (GSR), blood volume pressure, respiration pattern, skin temperature, electromyogram (EMG) and electrooculogram (EOG)) as given in the DEAP database. We have discussed the theories of emotion modeling based on i) basic emotions, ii) cognitive appraisal and physiological response approach and iii) the dimensional approach and proposed a three continuous dimensional representation model for emotions. The clustering experiment on the given valence, arousal and dominance values of various emotions has been done to validate the proposed model. A novel approach for multimodal fusion of information from a large number of channels to classify and predict emotions has also been proposed. Discrete Wavelet Transform, a classical transform for multiresolution analysis of signal has been used in this study. The experiments are performed to classify different emotions from four classifiers. The average accuracies are 81.45%, 74.37%, 57.74% and 75.94% for SVM, MLP, KNN and MMC classifiers respectively. The best accuracy is for 'Depressing' with 85.46% using SVM. The 32 EEG channels are considered as independent modes and features from each channel are considered with equal importance. May be some of the channel data are correlated but they may contain supplementary information. In comparison with the results given by others, the high accuracy of 85% with 13 emotions and 32 subjects from our proposed method clearly proves the potential of our multimodal fusion approach.

Contents

* Corresponding author at: Indian Institute of Information Technology Allahabad, India.
E-mail addresses: vermagkv@gmail.com, gyanendra@iiita.ac.in (G.K. Verma), ust@iiita.ac.in (U.S. Tiwary).
[1] Currently professor at Indian Institute of Information Technology, Allahabad, India.

1053-8119/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.11.007

Introduction

The natural communication among human beings is performed through two ways i.e. verbal and non-verbal. The verbal communication involves voice, speech or audio whereas non-verbal involves facial expression, body movement, sign language, etc. The non-verbal communication plays an important role in deciding the content of the communication. Generally, humans acquire information from more than one modality such as audio, video, smell, touch, etc. and combine these information into a coherent one. The human brain also derives information from various modes of communication so as to integrate various complimentary and supplementary information. In computational systems, information from various modalities, such as audio, video, electroencephalogram (EEG), electrocardiogram (ECG), etc. may be fused together in a coherent way. This is known as multimodal information fusion. In affective systems, multimodal information fusion can be used for the extraction and integration of interrelated information from multiple modalities to enhance the performance (Poh and Bengio, 2005) or to reduce the uncertainty on data classification or to reduce ambiguity in decision making. Several examples are: audio-visual speaker detection, multimodal emotion recognition, human face or body tracking, event detection, etc.

Two or more modalities cannot be integrated in a context free manner. There must be some context dependent model. Information fusion can be broadly classified into three major categories: (i) early fusion, (ii) intermediate fusion, and (iii) late fusion. In early fusion the information is being integrated at the signal or feature level, whereas in late fusion higher semantic level information is to be fused. The key issues in multimodal information fusion are the number of modalities (Pfleger, 2004; Bengio et al., 2002), synchronization of information derivation and fusion process and finding the appropriate level at which the information is to be fused. It is not always necessary that different modalities provide complimentary information in the fusion process; hence it is important to understand the contributions of each modality in reference to accomplishing various tasks. A typical framework of multimodal information fusion for human–computer interaction is illustrated in Fig. 1.

In the area of human–computer interaction (HCI), the information about cognitive, affective and emotional states of a user becomes more and more important as this information could be used to make communication with computers in a more human-like manner or to make computer learning environments more effective (Schaaff, 2008). Emotion recognition is an important task as it is finding extensive applications in the areas of HCI and Human–Robot Interaction (HRI) (Cowie et al., 2001) and many other emerging areas. Emotions are expressed through posture and facial expression as well as through physiological signals, such as brain activity, heart rate, muscle activity, blood pressure, skin temperature, etc. (Schaaff, 2008). Recognizing emotion is an interesting and challenging problem. Generally the emotion is recognized through facial expressions or audio–video data, but facial expressions may not involve all emotions (Ekman, 1993). There are a number of advantages of using physiological signals for emotion recognition (Schaaff, 2008).

- Firstly, bio-signals are controlled by the central nervous system and therefore cannot be influenced intentionally, whereas actors can play emotions on their face intentionally.
- Secondly, physiological signals are constantly emitted and as sensors are attached directly to the body of the subject, they are never out of reach.

In addition to that, physiological data could also be used as complementary to emotional data collected from voice or facial expressions to improve recognition rates. Physiological pattern helps in assessing and quantifying stress, anger and other emotions that influence health (Sebea et al., 2005). The physiological signal contains modalities such as electroencephalogram (EEG), electromyogram (EMG), electrooculogram (EOG), galvanic skin response (GSR), blood volume pressure, respiration pattern, skin temperature, etc. which can provide multiple cues. Emotion recognition from physiological pattern has a vast number of applications in the area of medicine, entertainment and HCI.

The DEAP database (Koelstra et al., 2012) for emotion analysis using physiological signals, contains EEG and peripheral signals in addition to face videos from 32 participants. The EEG signals were recorded from 32 active electrodes (channels), whereas peripheral physiological signals (8 channels) include GSR, skin temperature, blood volume pressure (plethysmography), respiration rate, EMG (zygomaticus major and trapezius) and EOG (horizontal and vertical). We have used the DEAP database in all our experiments. Full description of the database is given in Appendix A.

The purpose of this paper is twofold: (i) to propose a multidimensional 3D affective model of emotion representation and (ii) to develop a multimodal fusion framework to classify and recognize emotions from physiological signals using multiresolution approach. The overall paper is divided into eight sections including the present one as the first. Theories for emotion modeling are described in the second section while the basic approaches used in multimodal fusion framework are described in the third section. The fourth section is about wavelet based multiresolution approach used for feature extraction. The experiments done on the 3D affective model are described in the fifth section while those on emotion classification and recognition from physiological signals in the sixth section. The results and discussions on both types of experiments are mentioned in the seventh section and concluding remarks are given in the last section.

Emotion modeling

Emotion is an affective state of human beings (animals) arising as a response to some interpersonal or other events. It involves an appraisal of the given situation and the response is generally present as some physiological signals and/or some action(s). It is worth mentioning that the functionalist approaches to emotions vary by level of analysis, the individual and dyadic (inter-personal between two people) levels of analysis, and the group and cultural levels of analysis. The group and the cultural level of analysts see emotions as social cultural functions and they believe emotions to be constructed by individuals or groups in social contexts, and they relate them to constructs of individuals, patterns of social hierarchy, language, or requirements of socio-economic organization, etc. (Lutz and Abu-Lughod, 1990). Here we are concerned with the effects (experiences) of emotions within the individual or between interacting individuals (Ekman, 1993; Nesse, 1990) etc.

Please cite this article as: Verma, G.K., Tiwary, U.S., Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2013.11.007

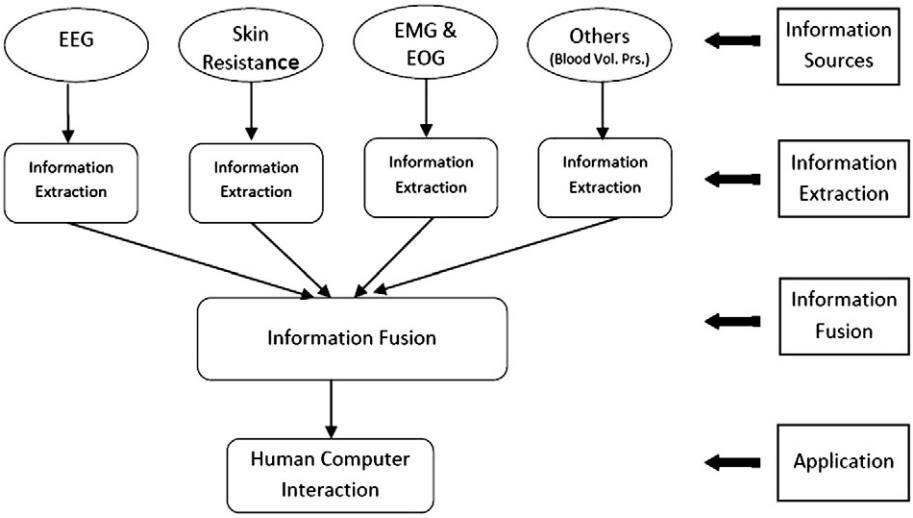

Fig. 1. A typical framework of multimodal information fusion.

Theory of emotion

Grandjean et al. (2008) reported three major approaches for describing emotions: (1) the categorical model, (2) the appraisal-based model and (3) the dimensional model. According to the categorical approach, there exist a small number of emotions that are basic, hard-wired in our brain, and recognized universally. While appraisal based model presents emotion as a response to the cognitive appraisal of the (emotional) situations and the response can be a physiological pattern leading to the communication of the emotion and the appropriate action. The dimensional approaches represent emotion (facial expression or physiological patterns) or affective variations through few independent dimensions on continuous or discrete scales.

Darwinian evolutionary view of emotion (basic emotions and other emotions)

Ekman (1999) has described six basic emotions (i.e. happy, sad, anger, fear, surprise and disgust) based on the evolutionary approach taken by Charles Darwin (Darwin, 1872/1988) and others (Tomkins, 1962). As per this approach the core components of emotions are biologically based and genetically coded. Initially the basic emotions evolved in species and then the other emotions are developed in the process of evolution through refinement of these emotions. Ekman added eleven other emotions (i.e. amusement, contempt, contentment, embarrassment, excitement, guilt, pride in achievement, relief, satisfaction, sensory pleasure and shame) in 1990s and claimed that these emotions cannot be recognized through facial expressions.

Plutchik (2008) created a new concept of emotion known as Wheel of emotion in 1980. He suggested eight primary emotions i.e. joy, sadness, anger, fear, trust, disgust, surprise and anticipation. After Robert Plutchik, Parrot proposed a classification of emotion in 2001. In Parrots theory (Parrott, 2001) a hundred plus emotions are identified. Parrots theory comprises six primary emotion, twenty five secondary emotions and more than a hundred tertiary emotions. Plutchik et al. (1997) developed the Psycho-evolutionary Theory of Emotions. According to this theory; emotions are primitive and exist to satisfy the reproductive fitness of animals.

Cognitive appraisal and physiological theory of emotions

According to Schachters Theory of Emotion (Schachter and Singer, 1962) an emotional state is a physiological arousal when an individual cognizes a situation to be emotional and causally attributes arousal to the cognized source. This is in confirmation to the other cognitive appraisal theories (Lazarus et al., 1980; Plutchik, 2008). Smith and Ellsworth (1985) has proposed six dimensions of cognitive appraisal of situations leading to emotional experience, namely, pleasantness, responsibility/control, certainty, attentional activity, anticipated effort and situational control.

According to the James–Lange theory (James, 1950) the experience of emotion is a response to physiological changes in our body. Every emotion is an interpretation of the preceding arousal and it is important to know whether the physiological reactions are different for each emotion for the analysis of emotions. Sundberg et al. (2011) reported a physiology-driven approach adopted for the analyses of emotion. They have proposed an action unit model that collaborates in creating a particular facial expression.

Cacioppo and Tassinary (1990) coined a big question in emotion theory. According to them, it is an open question, whether physiological patterns are sufficient to accompany each emotion as physiological muscle movements. What looks to be a facial expression to an outsider may not always correspond to a real underlying emotional state.

Dimensional approaches to emotions

Dimensional approach is based on the certain degree of emotion intensity represented by some emotional primitives. Human beings not only feel the emotion but they also feel its intensity such as very sad or a little happy. There are several works reported in literature based on two dimensional emotional primitives i.e. valence and arousal. Few works based on 3D emotion primitives are also reported in literature.

According to the dimensional approach, affective states are not independent from one another; rather, they are related to one another in a systematic manner (Mihalis et al., 2011). In this approach, the majority of affect variability is covered by two dimensions: valence and arousal. The valence dimension (V) refers to how positive or negative the affect is, and ranges from unpleasant feelings to pleasant feelings of happiness. The arousal dimension (A) refers to how excited or apathetic the affect is, and it ranges from sleepiness or boredom to frantic excitement. Oliveira et al. (2006) and Lewis et al. (2007) have reported a correlation between valence and arousal. Fontaine et al. (2007) reported four dimensions (valence, potency, arousal and unpredictability) to describe various emotion related phenomena. Their results show that the power/control dimension explains a larger percentage of the variance than the arousal dimension. Few researchers linked more acoustic parameters to an underlying arousal dimension ranging from highly alert and exited to relaxed and calm (Juslin and Scherer, 2005; Pereira, 2000).

Proposed model of emotion

Estimating emotion on a continuous valued scale is an important alternative to emotion categories for computational community to

Please cite this article as: Verma, G.K., Tiwary, U.S., Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2013.11.007

describe humans' affective states because it is able to describe the intensity of emotion, which can be used for recognizing dynamics, and allows for adaptation to individual moods and personalities (Schuller, 2011). In this work, we have used three emotion dimensions valence, arousal and dominance on a continuous scale from 1 to 9. The valence scale ranges from unhappy or sad to happy or joyful. The arousal scale ranges from calm or bored to stimulated or excited. The dominance scale ranges from submissive (or without control) to dominant (or in control, empowered).

The major contribution of this work is as follows:

- It presents a three dimension emotion model in terms of valence, arousal and dominance.
- It proposes a multimodal fusion framework for affective prediction based on physiological signals (not facial expressions).

Multimodal fusion framework

Multimodal fusion refers to integration of two or more than two modalities in order to improve the performance of the system. There are several multimodal fusion techniques reported in literature; however one major category of fusion is early, intermediate and late fusion. In an early fusion, the features obtained from different modalities needs to be combined in a single representation before feeding them to the learning phase. The intermediate fusion is able to deal with the imperfect data together with reliability and asynchrony issues among different modalities. Late fusion, also known as decision level fusion is based on the semantic information of the modalities. The major issue in multimodal data processing is that the data should be processed separately and should be combined only at the end. The general fusion architectures and joint processing of modalities has been discussed in Huang and Suen (1995), Bolt (1980), and Blattner and Glinert (1996)). In order to accomplish a human-like multimodal analysis of multiple input signals acquired by different sensors, the signals cannot always be considered mutually independent. hence, they might not combine in a context-free manner at the end of the intended analysis, but on the contrary, the input data might preferably be processed in a joint feature space and according to a context-dependent model (Karray et al., 2008). The major problems faced in multimodal fusion are dimensionality of joint feature space, different feature formats, and time-alignment. A potential way to achieve multisensory data fusion is to develop context-dependent versions of a suitable method such as the Bayesian inference method as proposed by Pan et al., (1999).

Multimodal systems usually integrate signals at the feature level (early fusion), at a higher semantic level (late fusion), or something in between (intermediate fusion) (Turk, 2005; Corradini et al., 2003). In this section we are presenting the above fusion techniques and related work reported in the literature.

Early fusion

When different modalities are combined into single representation prior to learning phase, the fusion is said to be early fusion. In this class of fusion, features are integrated in the beginning (Cees et al., 2005). In the early fusion framework the recognition process at signal level in some particular mode influences the course of recognition process in the remaining modes. Eventually, this kind of fusion is found to be more appropriate for highly temporally synchronized input modalities. Audio-visual integration might be one of the most suitable examples of the early fusion, where one simply concatenates the audio as well as visual feature vectors to get a combined audio-visual vector. The length of the resulting vector is contained by using the dimensionality reducing approaches like linear discriminant analysis (LDA), prior to feeding the feature vectors to the recognition engine. The most preferred classifier for early integration system is very often the conventional Hidden Markov Model (HMM) trained for the mixed audiovisual feature vector. Pitsikalis et al. (2006) proposed a Bayesian inference method to fuse the audio-visual features obtained from Mel-frequency cepstral coefficients (MFCC) and texture analysis respectively. The joint probability was computed by taking combined features. Mena J.B. and Malpica J. (Mena and Malpica, 2003) proposed a Dempster–Shafer fusion approach for the segmentation of color images. The authors extracted the information from terrestrial, aerial or satellite images. The information was based on the location of an isolated pixel, a group of pixels, and a pair of pixels. The information was fused using the Dempster–Shafer fusion approach. Nefian et al. (2002) used coupled HMM (CHMM) to combine audiovisual features for speech recognition. The authors have modeled the state asynchrony of the features while preserving their correlation over time. Magalhaes and Ruger (2007) used the maximum entropy model for semantic multimedia indexing. They combined the text and image features for image retrieval. The authors reported better performance of maximum entropy model based fusion compare to Naive Bays approach.

Intermediate fusion

The basic short coming of the early fusion technique was its inability to deal with the imperfect data. Besides, early fusion avoids the explicit modeling of the different modalities also resulting in the failure to model the fluctuation in the signal. Asynchrony problem among the different streams is also a major problem for the early fusion. An approach to overcome these issues is to consider the features of the corresponding stream at various time instances. Hence by comparing the previously observed instances with the current data of some observation channel, one can make some statistical prediction with certain derivable probability for the erroneous instances. Therefore the probabilistic graphical models like HMM (including the hierarchical variants), Bayesian network and dynamic Bayesian networks are the most appropriate framework for fusing the multiple source of information in such a situation (Sebe et al., 2005). They can handle not only the noisy feature, but using the probabilistic inference, also the temporal information and missing feature values. Application of the hierarchical HMM has been demonstrated to recognize the facial expression (Cohen et al., 2003). Dynamic Bayesian network and HMM variants (Minsky, 1975) have shown their caliber to fuse several sources of the information to recognize intention, office activities or other events in video using both audio and video signals (Carpenter, 1992).

Late fusion

Multimodal system integrates common meaning representations derived from different modalities to arrive at some common interpretation. It necessitates for a common framework for common meaning representation of all the modalities being used besides well defined operations for integrating the partial meaning. Late integration models generally use independent classifiers, which can be trained separately for each stream. The ultimate classification decision is made by combining the partial outputs of each unimodal classifiers. The cognizance of the correspondence between the channels is made during the integration step only. There are some obvious advantages of the late fusion. Now the inputs can be recognized separately, independent of one another, hence they need not occur simultaneously. It also simplifies the software development process (Rabiner and Juang, 1993). Late integration system uses recognizers trainable on unimodal data set but the scalability in the number of modes as well in vocabulary is an important issue. For example, in case of audiovisual recognition, one can explore for a good heuristic and extract the promising hypothesis from audio only and then rescoring can be done on the basis of the visual evidences (Kittler et al., 1998). Multimodal human–computer interaction is one such application where we need advance techniques and models for cross model information fusion to integrate the audio and video features. This is an active area of research and many paradigms are being

Please cite this article as: Verma, G.K., Tiwary, U.S., Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2013.11.007

proposed for addressing the issues arising in audiovisual fusion. However they can be metamorphosed into a general framework addressing the integration of other modalities. Some of those that are based on late fusion are discussed here.

Aguilar et al. (2003) proposed a rule-based fusion and learning based fusion strategy for combining the scores of face, fingerprint and online signatures. The sum rule and Radial Basis Function Support Vector Machine (RBF SVM) are being used for comparison. The experimental results demonstrate that the learning-based RBF SVM scheme outperforms the rule-based scheme based on some appropriate parameter selections. Meyer et al. (2004) proposed a decision level fusion by taking speech and visual modalities. The speech features were extracted using MFCC algorithm and the lip contour features from the speakers face in the video. HMM classifiers were used in order to obtain the individual decisions for both. Then fusion was performed using Bayesian inference method to estimate the joint probability of a spoken digit. Singh et al. (2006)) used D–S theory to fuse the scores of three different finger print classification algorithms based on the Minutiae, ridge and image pattern features. The results obtained from D–S theory of fusing proved better. Beal et al. (2003) proposed a graphical model to fuse audio-visual observations for tracking a moving object in a noisy environment. They modeled audio and video observations jointly by computing their mutual dependencies. The Expectation–Maximization algorithm was used to learn the model parameters from a sequence of audio-visual data. The results were demonstrated in a two microphones and one camera setting. Guironnet M. et al. (Guironnet et al., 2005) proposed a Neural Network based fusion method to detect human activities. In this work, the authors combined sensory data obtained by CCD cameras and inputs from computing machine working in a LAN environment. The fusion was performed at decision level by taking inputs from computational units i.e. CPU load, network load and observation sensors (e.g. cameras) in order to detect the human activities in regard to usage of laboratory resources. An overview of the fusion approaches in terms of modalities used and their applications is given in Table 1.

Wavelet based multiresolution approach

The wavelet transform

The Wavelet Transform is suitable for multiresolution analysis, where the signal can be analyzed at different frequency and time scales. As EEG signals contain information at different frequency bands, we have used Daubechies Wavelet Transform coefficients for feature extraction. The Wavelet is a mathematical transformation function that divides the data into various different frequency components, and then they are analyzed individually with a different resolution matched to its scale. Wavelet transform is the representation of a function by mother wavelets. The one dimensional continuous wavelet transform denoted by $W_f(s,t)$ of a one dimensional function $f(t)$ is defined by Eq. (1) (Bhatnagar et al., 2012)

$$W_f(s,\tau) = \int_{-\infty}^{\infty} f(t)\varphi_{s,\tau}(t)dt. \tag{1}$$

where ψ is the mother wavelet function, s is the scale parameter and t is the translation parameter. Now, to reconstruct the original signal back from transformed signal, the inverse continuous wavelet transform is defined as

$$f(t) = \frac{1}{C_\varphi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W_f(s,\tau)\,\varphi_{s,\varphi}(t)\frac{dsd\tau}{s^2} \tag{2}$$

such that $C_\varphi = \int_{-\infty}^{\infty} (|\varphi(u)|^2/|u|)\,du$ where $\psi(u)$ is the Fourier Transform of $\psi(t)$.

Multiresolution approach

Many emotion theories reveal that physiological signals are important for emotion. Ekman et al. (1987) demonstrated that characteristic physiological pattern can be associated with a particular emotion. In this work, two types of signals i.e. EEG and peripheral physiological signals are used for feature extraction. EEG can be described by frequency and amplitude. The following frequency bands are included in EEG signal (Stickel et al., 2007):

- Delta: 1–4 Hz.
- Theta: 4–8 Hz.
- Alpha: 8–12.5 Hz.
- Beta: 12.5–28 Hz.
- Gamma: 30–40 Hz.

In DEAP database, EEG was recorded at a sampling rate of 512 Hz using 32 active AgCl electrodes, placed according to international 10–20 system. We have applied the Daubechies Wavelet Transform at five levels and analyzed the high frequency wavelet coefficients at each level. These coefficients at five levels represent the physiological signal in five frequency bands as shown above. The peripheral physiological signals are GSR, respiration amplitude, skin temperature, electrocardiogram and blood volume by plethysmograph, electromyograms of zygomaticus and trapezius muscles, and electrooculogram (EOG).

Table 1
An overview of the fusion approaches in terms of modalities used and their applications.

Fusion approach	Level of fusion	Work	Modalities	Application
Support Vector Machine	Decision	Aguilar et al. (2003)	Video, audio (MFCC) and textual cues	Semantic concept detection
	Hybrid	Bredin and Chollet (2007)	Audio (MFCC), video (DCT of lip area)	Biometric identification of talking face
		Zhu et al. (2006)	Low level visual features, text color, size, location, edge density, brightness, contrast	Image classification
Bayesian inference	Feature	Pitsikalis et al. (2006)	Audio (MFCC), video (Shape and texture)	Speech recognition
	Decision	Meyer et al. (2004)	Audio (MFCC) and video (lips contour)	Spoken digit recognition
	Hybrid	Xu and Chua (2006)	Audio, video, text, web log	Sports video analysis
		Atrey et al. (2006)	Audio & video	Event detection for surveillance
Dempster–Shafer theory	Feature	Mena and Malpica (2003)	Video (trajectory coordinates)	Segmentation of satellite images
	Decision	Guironnet et al. (2005)	Audio (phonemes) and visual (visemes)	Video classification
		Singh et al. (2006)	Audio, video and the synchrony score	Finger print classification
Dynamic Bayesian networks	Feature	Nefian et al. (2002)	Audio and visual (2D-DCT coefficients of the lips region)	Speech recognition
	Decision	Beal et al. (2003)	Audio and video	Object tracking
	hybrid	Town (2007)	Video (face and blob), ultrasonic sensors	Human tracking
		Xie et al. (2005)	Text, audio, video	clustering in video
Neural networks	Feature	Zou and Bhanu (2005)	Audio and video	Human tracking
	Decision	Gandetto et al. (2003)	CPU load, login process, network	Human activity monitoring
	hybrid	Ni et al. (2004)	Load, camera images	Image recognition
Maximum Entropy Model	Feature	Magalhaes and Ruger (2007)	Text and image	Semantic image indexing

Please cite this article as: Verma, G.K., Tiwary, U.S., Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2013.11.007

Experiments on 3D affective model

DEAP emotion database

Recent advances in emotion recognition have motivated many researchers to create emotion databases. Some of the emotion databases are MIT (Healey and Picard, 2005), MMI (Pantic et al., 2005), HUMAINE (Douglas-Cowie et al., 2007), VAM (Grimm et al., 2008), SEMAINE (McKeown et al., 2010), MAHNOB-HCI (Soleymani et al., 2012a) and DEAP (Koelstra et al., 2012). DEAP database is being used in this study. These databases contain speech, visual or audio-visual and physiological emotion data. Further DEAP dataset also contains values of valence, arousal, dominance, liking and familiarity, for various emotions for various subjects.

In this experiment, first we have proposed an emotion representation model consisting of three continuous dimensions viz. valence, arousal and dominance. Each dimension has values ranging from 1 to 9. These dimensions denote various aspects of emotion as follows:

- Valence: The valence scale ranges from unhappy or sad to happy or joyful.
- Arousal: The arousal scale ranges from calm or bored to stimulated or excited.
- Dominance: The dominance scale ranges from submissive (or without control) to dominant (or in control, empowered).

However, we did not consider the two dimensions given in DEAP dataset, liking and familiarity, as it is not so significant for finding emotion in short durations and our purpose is to consider minimum number of dimensions that are sufficient to represent various emotions. The dataset contains a total of 1280 values (40 trials for each 32 objects) for valence (V), arousal (A) and dominance (D). For first type of experiment, we have calculated the emotion centroids in 3D (VAD) space by taking the average value of more than 50 samples of each emotion category. We have then calculated the Euclidean distances of each emotion from the other which is shown in Table 2. Then K-means clustering was done on V, A and D values of all emotion points. It resulted in five optimum number of clusters. Fig. 2 shows the five clusters and their midpoints in black crosses.

Experiments on classification and recognition of emotions from physiological signals

There are different kinds of physiological signals that can be used for emotion recognition. Emotion recognition from the physiological signals from DEAP database involves feature extraction, multimodal fusion and classification. The architecture of emotion recognition from physiological signal is given in Fig. 3.

Feature extraction

Physiological activity is considered as an important component of an emotion. Several emotion studies correlated the physiological patterns with emotions. In this study, EEG and peripheral signals are considered for feature extraction. EEG was recorded using 32 active AgCl electrodes, placed according to international 10–20 system.

The peripheral signal includes Electro dermal Activity (EDA), Galvanic skin response (GSR), Skin conductance response (SCR) and skin temperature. EDA and GSR is a measure of skin conductance and commonly used for automatic affect recognition. Kim and Andre (2008) reported GSR a very reliable physiological measure of human arousal. All the physiological signals were recorded at a 512 Hz sampling rate and down sampled to 256 Hz. From EEG signals, Relative Power Energy (RPE), Logarithmic Relative Power Energy (LRPE), Absolute Logarithmic Relative Power energy (ALRPE) of four frequency bands (Theta, Alpha, Beta and Gamma) was extracted. In addition to Power spectral features, Standard deviation and Spectral Entropy were also calculated from each level of detail coefficients and highest level of approximation coefficients. In total, 25 features were extracted from physiological signals. The extracted features from EEG and Physiological signals are listed in Table 3.

The sub band energy of EEG signal, can be calculated by Eq. (3).

$$E_j = \sum_k \left(d_j(k)\right)^2 \tag{3}$$

where E_j is the sub band energy at jth frequency band.

The Relative Power Energy (E_{RPE}), Logarithmic Relative Power Energy (E_{LRPE}) and Absolute Logarithmic Relative Power Energy (E_{ALRPE}) can be calculated by Eqs. (4), (5) and (6) respectively.

$$E_{RPE} = \frac{E_j}{E_{TOTAL}} \tag{4}$$

where $E_{TOTAL} = E_{Alpha} + E_{Beta} + E_{Gamma} + E_{Theta}$

$$E_{LRPE} = log(E_{RPE}) \tag{5}$$

$$E_{ALRPE} = |(E_{LRPE})|. \tag{6}$$

Multimodal fusion

The primary objective of multimodal fusion is to improve the classification results by exploiting the complementary nature of different modalities. Generally, multimodal fusion can be classified into three broad categories, namely, early fusion, intermediate fusion and late fusion. The

Table 2
Euclidean distances of each emotion from other emotions based on centroids in VAD (valence, arousal and dominance) space.

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12
E1	0	4.7024	2.7903	3.0255	1.1539	1.2791	0.6569	2.0658	1.3311	3.7630	4.1819	2.5741
E2	4.7024	0	2.0866	2.4879	4.3785	4.1845	4.2124	2.9519	3.9027	2.0082	1.4565	2.7661
E3	2.7903	2.0866	0	0.8270	2.4163	2.1952	2.2290	1.6070	2.3260	2.1677	2.2172	1.4238
E4	3.0255	2.4879	0.8270	0	2.6279	2.3719	2.4409	2.1307	2.7152	2.6873	2.7649	1.5790
E5	1.1539	4.3785	2.4163	2.6279	0	0.2699	0.7221	2.3893	2.0717	3.9552	4.2169	2.8017
E6	1.2791	4.1845	2.1952	2.3719	0.2699	0	0.7304	2.2996	2.0712	3.8264	4.0729	2.6373
E7	0.6569	4.2124	2.2290	2.4409	0.7221	0.7304	0	1.8544	1.4175	3.5065	3.8573	2.2583
E8	2.0658	2.9519	1.6070	2.1307	2.3893	2.2996	1.8544	0	0.9643	1.6977	2.1550	1.0230
E9	1.3311	3.9027	2.3260	2.7152	2.0717	2.0712	1.4175	0.9643	0	2.5764	3.0850	1.6562
E10	3.7630	2.0082	2.1677	2.6873	3.9552	3.8264	3.5065	1.6977	2.5764	0	0.6624	1.7175
E11	4.1819	1.4565	2.2172	2.7649	4.2169	4.0729	3.8573	2.1550	3.0850	0.6624	0	2.1546
E12	2.5741	2.7661	1.4238	1.5790	2.8017	2.6373	2.2583	1.0230	1.6562	1.7175	2.1546	0

Note: E1 — happy, E2 — sad, E3 — anger, E4 — hate, E5 — fun, E6 — exciting, E7 — joy, E8 — cheerful, E9 — love, E10 — sentimental, E11 — melancholy, E12 — pleasure.

Please cite this article as: Verma, G.K., Tiwary, U.S., Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2013.11.007

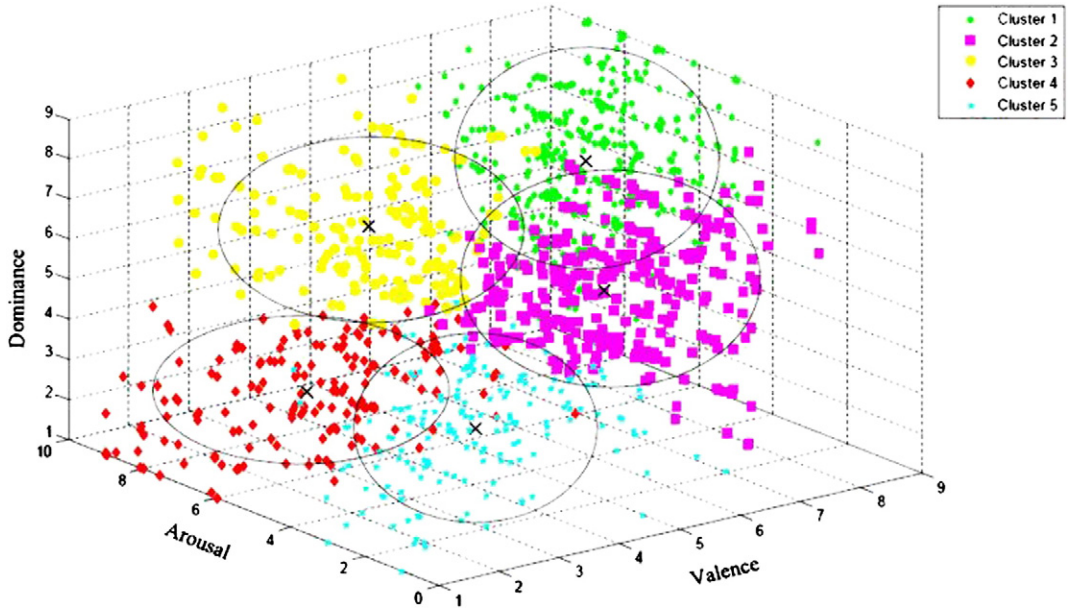

Fig. 2. Clusters of different emotions based on valence, arousal and dominance.

multimodal fusion is described in detail in the Multimodal fusion framework section. The feature level fusion was applied in this study.

We have two feature vectors, say $F_i = \{f_{i,1}, f_{i,2}, \ldots, f_{i,n}\}$ feature vector for one modality and $X_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,n}\}$ be the fused feature vector of vectors F_i and H_i. X_i is obtained by augmenting the normalized feature vectors F_i and H_i. and then performing the feature selection on the concatenated vector formed. F_i and H_i. are calculated by applying normal transformation techniques to each of the individual feature values of the feature vectors F_i and H_i. Suppose the feature vectors $\{F_i, H_i\}$ and $\{F_i, H_i\}$ obtained at two different time instances i and j, then the corresponding fused vectors may be denoted as X_i and X_j respectively.

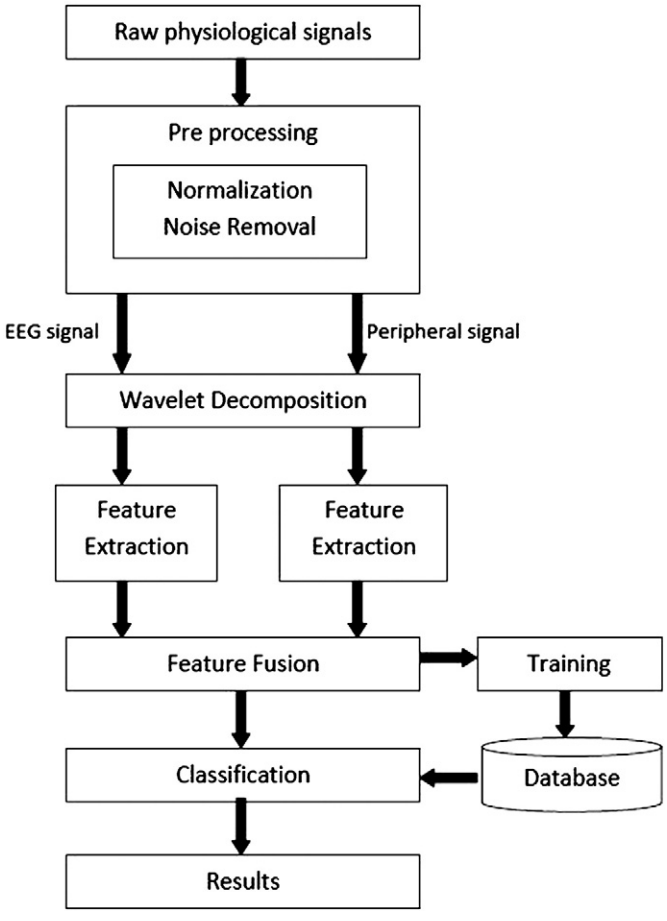

Fig. 3. Architecture of emotion recognition from physiological signal.

Classification

In machine learning, pattern recognition is the assignment of some sort of output value (or label) to a given input value (or instance), according to some specific algorithm. An example of pattern recognition is classification, which attempts to assign each input value to one of a given set of classes. However, pattern recognition is a more general problem that encompasses other types of outputs, such as regression, sequence labeling and parsing.

This study evaluated four classifiers Support Vector Machine (SVM), Multilayer Perceptron (MLP), K-Nearest Neighbor (KNN) and Meta-multiclass (MMC). The selections of these classifiers are based on the success rate obtained for EEG classification. These classifiers have been applied to features of different frequency bands i.e. Theta, Alpha, Beta and Gamma and combination of all.

The MLP consists of an input layer, a hidden layer with sigmoid function and an output layer. The learning rate is 0.3 with a validation threshold of 20. The other classifier used in this study is SVM, a set of related supervised learning method that analyzes data and recognizes patterns. The original SVM algorithm was invented by Vladimir Vapnik and the current standard incarnation (soft margin) was proposed by Cortes and Vapnik (1995). The standard SVM is a nonprobabilistic binary linear classifier. An SVM training algorithm builds a model that predicts whether the new case belongs to one category or the other. The K-NN, an

Table 3
Extracted features from EEG and physiological signals.

Feature category	Extracted features
Relative Power Energy (RPE)	Four band of Alpha, Beta, Gamma and Theta
Logarithmic Relative Power Energy (LRPE)	Four band of Alpha, Beta, Gamma and Theta
Absolute Logarithmic Relative Energy Power energy (ALRPE)	Four band of Alpha, Beta, Gamma and Theta
Standard deviation	All levels of detail coefficients and highest level approximation coefficient
Spectral Entropy	All levels of detail coefficients and highest level approximation coefficient

Please cite this article as: Verma, G.K., Tiwary, U.S., Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2013.11.007

instance-based learning classifier, is also used here for classifying unknown instances based on some distance or similarity function. A MMC classifier, for handling multi-class datasets with 2-class classifiers by building a random class-balanced tree structure, has been also used in this work.

Results and discussion

The goal of analysis and evaluations is twofold: first the validation of 3D emotion model, and second, the evaluation of classification and prediction of various emotions from EEG and other peripheral physiological signals. The DEAP database consists of 40 channels of physiological signals (EEG + Peripheral) with 8056 data with 40 trials each for 32 participants.

All the calculations were carried out in Matlab 7.7.0 (R2008b), on a 32-bit Intel 3.06 GHz processor, with 2 GB RAM. The Wavelet Transformation was done by using Wavelet toolbox available in MatLab.

Results of 3D emotion modeling using continuous VAD (valence, arousal, dominance) dimensions

Table 2 shows the Euclidean distances among various emotions in VAD space. It shows the maximum distance of 4.70 between happy and sad, and the minimum distance of 0.27 between fun and exciting, which explains that 'fun' and 'exciting' are very near as compare to 'anger' and 'sad'. Similarly 'happy' and 'joy', 'love' and 'cheerful' and 'anger' and 'hate' are near to each other compared to other emotions. These results qualitatively validate the model.

Further, we have plotted a graph of 12 emotions as nodes and the distances as edges in Fig. 4. First we have only shown the distances up to 25% of the maximum value (4.7/4 = 1.18) as solid lines. This emotion network (or graph) interestingly groups 12 emotions in 5 groups, which is exactly equal to the number of clusters in Fig. 4. Interestingly these five clusters in the graph are (happy, joy, fun, exciting), (love, cheerful, pleasure), (anger, hate), (sad) and (melancholy, sentimental). However, as the points in the clusters of Fig. 3 have large deviations, one or two emotions might have been clustered in different groups, but the emotion-network very well qualifies the 3D VAD model of emotion. We have also shown the distances more than 25% (of maximum) and less than 40% (of maximum) by dotted lines to show the relative positions of each emotion group. As can be seen (happy, joy, fun, exciting) group is a good distance away from both (sentimental, melancholy) group and (anger, hate) group. The (sad) emotion is further away from (sentimental, melancholy) group. These results are in agreement to the general findings.

These results also demonstrate to some extent the sufficiency of the three continuous dimensions namely valence, arousal and dominance (although in DEAP database the values of liking and familiarity are also indicated). As per the cognitive appraisal theory, the emotions are responses to the cognitive appraisal of the (abnormal) situation. Smith and Ellsworth (1985) have pointed out eight dimensions of cognitive appraisal. These dimensions include pleasantness, attentional activity, control, certainty, goal–path obstacle, legitimacy, responsibility and anticipated effort. Out of this, goal–path obstacles and legitimacy have not been considered by many other researchers. However, here we are not considering the appraisal, but the common quantities which result from the appraisal. Valence gives some kind of direction or a kind of pleasantness or sadness (like hue in color representation). Arousal gives sharpness like very sad or very happy (which may be due to attentional activity or anticipated effort). Dominance quantifies the amount of self or situational control (may be the result of control, certainty and responsibility). As mentioned in the Emotion modeling section, only two dimensions of valence and arousal are not sufficient to explain distinction among many emotions. Hence, we think that these three continuous dimensions, valence, arousal and dominance are sufficient to uniquely represent an emotion.

Results of emotion classification using the proposed multimodal fusion approach

For this category of experiment, the EEG (32 channels) and peripheral physiological (8 channels) signals were used to extract the features using multiresolution analysis. The Wavelet Transform, a classical transform for multiresolution analysis, is used to decompose the signals up to

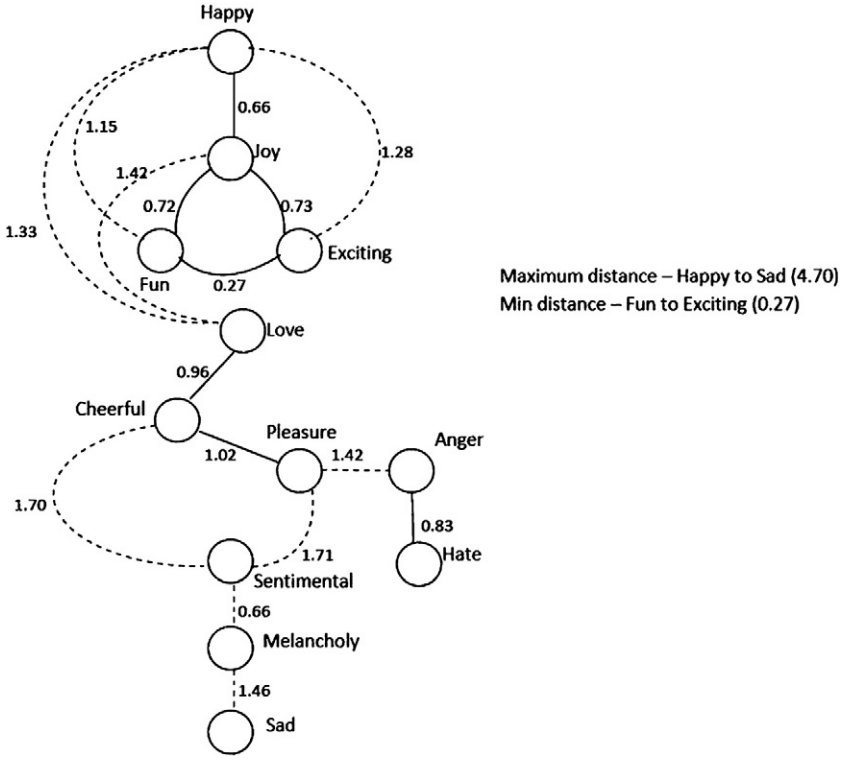

Fig. 4. Emotion graph or network with Euclidean distance ≤ 1.02 (solid lines) shows five clusters. Dotted lines show distance > 1.02 < 2.0.

Please cite this article as: Verma, G.K., Tiwary, U.S., Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2013.11.007

Table 4
Accuracy obtained with a 10-fold cross-validation test over the 640 training instances for each individual classifier and emotion for 32 channels EEG.

Emotions	Accuracy			
	SVM	MLP	KNN	MMC
Terrible	80.93	75.78	55.62	75.00
Love	82.18	75.62	59.06	71.87
Hate	79.84	76.71	59.06	74.84
Sentimental	82.50	77.65	60.15	77.96
Lovely	81.56	75.62	60.00	79.06
Happy	82.81	74.53	58.90	77.03
Fun	79.84	75.78	53.12	75.15
Shock	79.68	72.03	57.18	75.15
Cheerful	80.31	58.90	55.62	73.75
Depressing	85.46	81.25	63.28	79.21
Exciting	83.59	78.28	57.03	76.87
Melancholy	77.65	67.18	56.40	73.90
Mellow	82.50	77.50	55.31	77.50

Table 5
Accuracy obtained for multimodal fusion of EEG and peripheral (GSR, BVP, EMG & EOG etc.) with a 10-fold cross-validation test.

Emotions	Accuracy			
	SVM	MLP	KNN	MMC
Terrible	77.96	76.25	62.81	76.48
Love	77.96	76.87	63.20	75.39
Hate	79.45	76.25	63.67	77.18
Sentimental	78.28	78.04	62.96	75.31
Lovely	79.14	76.25	65.00	75.78
Happy	78.43	76.40	64.14	75.54
Fun	77.96	76.64	62.81	74.14
Shock	78.20	77.26	62.26	75.31
Cheerful	80.28	78.82	59.68	75.07
Depressing	80.15	78.35	65.39	78.04
Exciting	79.21	77.10	64.29	75.01
Melancholy	79.14	77.65	62.03	76.40
Mellow	79.06	76.64	61.01	77.65

five levels in order to obtain the approximation and detail coefficients. As the sampling rate of physiological signal is 512 Hz, the signal decomposition up to five levels is sufficient to capture the information from signals. The Discrete Wavelet Transform (DWT) is used with Daubechies 4 family (DB4) as it provides good results for non-stationary signal like EEG (Karray et al., 2008). Prior to classification, all features have been normalized to the range of [0, 1] using the Min–Max algorithm. We have performed the experiments for thirteen emotions for which sufficient data was available in the DEAP database. A 25 dimensional feature vector was generated using the features listed in Table 3.

All experiments are performed with 10-fold cross-validation with random values of dataset to increase the recognition rate. In 10-fold cross-validation, whole dataset is divided into ten subsets. The nine subsets of feature vectors are used for training and remaining subsets for testing. This procedure is repeated ten times with different subset splits. The classification accuracy is obtained by the correctly classified number of instances and the total number of instances.

The Waikado Environment for Knowledge Analysis Weka (Weka machine learning tool) tool was used for classification purpose as it provides a collection of machine learning algorithms for data classification. From the collection of various algorithms, we have selected four algorithms i.e. Support Vector Machine (SVM), Multilayer Perceptron (MLP), K-Nearest Neighbor (K-NN) and Meta-multiclass (MMC), based on best classification accuracy. A 10-fold cross-validation test over 640 training instances have been performed for each selected classifier. The accuracies obtained with 32 EEG channel data with each classifier for each emotion are shown in Table 4.

Table 4 shows the accuracy rate for different emotions obtained from four classifiers. The average accuracies are 81.45%, 74.37%, 57.74% and 75.94% for SVM, MLP, KNN and MMC respectively. The best accuracy is for depressing with 85.46% using SVM. As can be seen from Table 4, SVM outperformed other classifiers. The SVM was implemented with kernel — PolyKernel and the value of C (200). The tolerance parameter is 0.001 with Epsilon (1.0E — 12). The classification accuracy for MLP is also significant.

Fig. 5 presents the accuracies of SVM classifier for each emotion when the features obtained from various frequency bands (Theta, Alpha, Beta and Gamma) are taken separately and in a combined (fused) manner. It shows that the fusion at the feature level, combining multilevel frequency features, outperform each single band. Hence, in our proposed method the fusion takes place at two levels: one at different frequency band features and other at various channel features. The 32 channels are considered as independent modes and features from each channel are considered with equal importance. Maybe some of the channel data are correlated but they may contain complementary

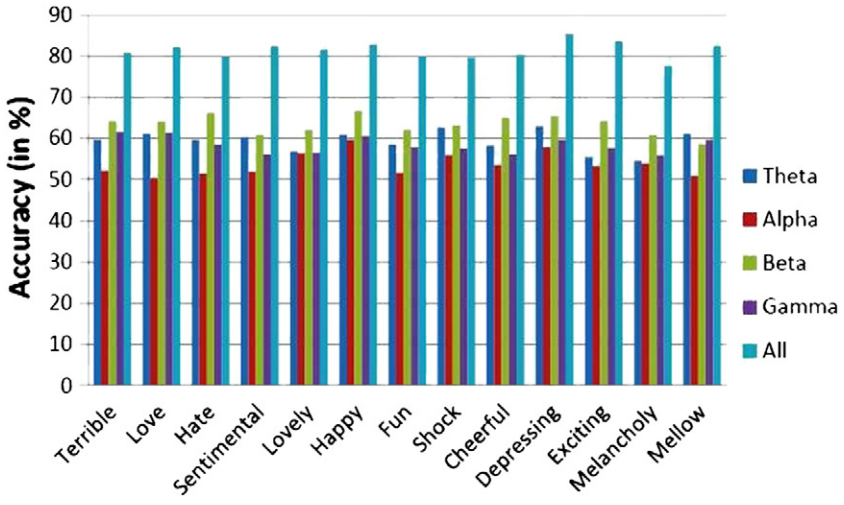

Fig. 5. SVM results for different emotions with EEG frequency band. All represents all bands.

Please cite this article as: Verma, G.K., Tiwary, U.S., Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2013.11.007

Table 6
Accuracy comparison of various studies.

Methods	Accuracy rate (%)
Schaaff and Schultz (2009)	66.70
Soleymani et al 2012b.	68.50
Proposed method	81.45

or supplementary information. The high accuracy of 85% with 13 emotions and 32 subjects clearly proves the potential of our multimodal fusion approach.

The other modalities can be seen to perform moderate complementary. Table 5 shows the results of different classifiers obtained after fusion of 40 channels (32 EEG and 8 peripheral) features. As can be seen there is not much improvement as compared to the fusion results of only 32 EEG channel features given in Table 4. It is interesting to note that there are improvements in the case of the classification accuracies of the K-NN classifier, improvement in few cases with that of MLP and MMC classifiers and only one improvement in melancholy emotion in the case of the SVM classifier. One of the reasons for this may be the actual accuracy of the collected data itself (DEAP database). Collection of emotion data is a tedious task and there are factors like individual physiological differences of subjects, signal–noise and quality of assessments of emotion. Even the feature richness due to large frequency bandwidth of EEG signal may not be present in the case of GSR, EMG, EOG, etc.

The accuracy rate comparison of various studies along with the same of the proposed method is given in Table 6. Schaaff and Schultz (2009) reported 66.7% accuracy based on SVM classifier with three emotional states: pleasant, neutral, and unpleasant. Comparing this to that of our proposed method (81.45% average) for 13 emotions proves the soundness and performance of our framework. They used Short Term Fourier Transform (STFT) for feature extraction from EEG signal. Our results also show that the DWT is more successful in contrast to STFT for emotion recognition from physiological signals. The success rate reported by Soleymani et al., 2012b work is 68.50%. Some EEG emotion recognition studies reported their performance based on valence and arousal measures only, which are not included here for comparison.

Conclusion and future work

This study presents a multimodal fusion framework based on multiresolution approach using Daubechies Wavelet Transform features for emotion recognition from physiological signals. In contrast to emotion recognition through facial expression, we may claim that a large number of (in this paper thirteen) emotions can be recognized accurately through physiological signals. The DEAP, a multimodal database of EEG and peripheral signals with 32 participants were used. As the experience of emotion is a response to physiological changes in our body, according to the James–Lange theory (James, 1950), the measurement of physiological reactions for emotion prediction is a better option than the facial expression reading. The proposed method takes into account those features which are subject independent and can incorporate many more number of emotions. As it is possible to handle many channel features, especially EEG channels which are synchronous, feature level fusion works. The method can be extended for decision level fusion for asynchronous data.

In addition to the fusion framework, we have also proposed a continuous 3D emotion model based on the three emotion primitives: valence, arousal and dominance. A simple clustering of valence, arousal and dominance data for emotions shows that the thirteen emotions can be grouped into five clusters and the possible clusters are validated through a proposed emotion graph considering the Euclidean distances among various emotions in the VAD space. This novel idea of representing emotions can possibly be generalized if more data is available.

Appendix A

Table: Database content summary.

Online subjective annotation	
Number of videos	120
Video duration	1 minute affective highlight
Selection method	60 via last.fm affective tags, 60 manually selected
No. of ratings per video	14–16
Rating scales	Arousal Valence Dominance
Rating values	1–9
Physiological experiment	
Number of participants	32
Number of videos	40
Selection method	Subset of online annotated videos with clearest responses
Rating scales	Arousal Valence Dominance Liking Familiarity
Rating values	Familiarity: discrete scale of 1–5 Others: continuous scale of 1–9
Recorded signals	32-channel 512Hz EEG Peripheral physiological signals Face video (for 22 participants)

References

Aguilar, J.F., Garcia, J.O., Romero, D.G., Rodriguez, J.G., 2003. A comparative evaluation of fusion strategies for multimodal biometric verification. In: Kittler, J., Nixon, M.S. (Eds.), Int. Conf. on Video-Based Biometrie Person Authentication VBPA 2003, LNCS 2688. Guildford Springer -Verlag, Berlin Heidelberg, pp. 830–837.
Atrey, P.K., Kankanhalli, M.S., Jain, R., 2006. Information assimilation framework for event detection in multimedia surveillance systems. Springer/ACM Multimed. Syst. J. 12 (3), 239–253.
Beal, M.J., Jojic, N., Attias, H., 2003. Gaphical model for audiovisual object tracking. IEEE Trans. Pattern Anal. Mach. Intell. 25, 828–836.
Bengio, S., Marcel, C., Marcel, S., Mariethoz, J., 2002. Confidence measures for multimodal identity verification. Inf. Fusion 3 (4), 267–276.
Bhatnagar, G., Wu, Q.M.J., Raman, B., 2012. A new fractional random wavelet transform for fingerprint security. IEEE Trans. Syst. Man Cybern. Syst. Hum. vol. 42 (1), 262–275.
Blattner, M., Glinert, E., 1996. Multimodal integration. IEEE Multimed. 3 (4), 14–24.
Bolt, R., 1980. Put-That-There: voice and gesture at the graphics interface. Comput. Graph. 14 (3), 262–270.
Bredin, H., Chollet, G., 2007. visual speech synchrony measure for talking-face identity verification. IEEE International Conference on Acoustics, Speech and, Signal Processing, vol. 2, pp. 233–236.
Cacioppo, J., Tassinary, L., 1990. Inferring psychological significance from physiological signals. Am. Psychol. 45, 16–28.
Carpenter, R., 1992. The Logic of Typed Feature Structures. Cambridge University Press.
Cees, G., Snoek, M., Worring, M., Smeulders, W.M., 2005. Early versus late fusion in semantic video analysis. Proc. of the 13th annual ACM int. Conf. on Multimedia. ACM, New York.
Cohen, I., Sebe, N., Garg, A., Chen, L., Huang, T.S., 2003. Facial expression recognition from video sequences. temporal and static modeling. CVIU 91 (1–2), 160–187.
Corradini, M., Mehta, N., Bernsen, J., Martin, C., 2003. Multimodal input fusion in human–computer interaction. NATO-ASI Conf. on Data Fusion for Situation Monitoring, Incident Detection, Alert, and Response Management.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20.
Cowie, R., Douglas, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J., 2001. Emotion recognition in human–computer interaction. IEEE Signal Process. Mag. 18, 32–80.
Smith, Craig A., Ellsworth, Phoebe C., 1985. Patterns of cognitive appraisal in emotion. J. Pers. Soc. Psychol. 48 (4), 813–838.
Darwin, Charles, 1872/1988. The Expression of the Emotions in Man and Animals.
Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M., Martin, J.C., Devillers, L., Abrilian, S., Batliner, A., Amir, N., Karpouzis, K., 2007. The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data. Proc. Second Int'l Conf. Affective Computing and Intelligent Interaction, pp. 488–500.
Ekman, P., 1993. Facial expression and emotion. Am. Psychol. 48 (384), 392.
Ekman, P., 1999. Basic emotions. Handbook of Cognition and Emotion. 45–60.
Ekman, P., Friesen, W.V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W.A., Pitcairn, T., Ricci-Bitti, P.E., 1987. Universals and cultural

Please cite this article as: Verma, G.K., Tiwary, U.S., Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2013.11.007

differences in the judgments of facial expressions of emotion. J. Pers. Soc. Psychol. 53 (4), 712–717.
Fontaine, J., Scherer, K.R., Roesch, E., Ellsworth, P., 2007. The world of emotions is not two-dimensional. Psychol. Sci. 18, 1050–1057.
Gandetto, M., Marchesotti, L., Sciutto, S., Negroni, D., Regazzoni, C.S., 2003. From multi-sensor surveillance towards smart interactive spaces. IEEE Int. Conf. Multimed. Expo. I 641–644.
Grandjean, D., Sander, D., Scherer, K.R., 2008. Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. Conscious. Cogn. 17, 484–495.
Grimm, M., Kroschel, K., Narayanan, S., 2008. The Vera am Mittag German audio-visual emotional speech database. Proc. IEEE Int. Conf. Multimed. Expo. 865–868.
Guironnet, M., Pellerin, D., Rombaut, M., 2005. Classification based on low-level feature fusion model. the European Signal Processing Conference. Antalya, Turkey.
Healey, J.A., Picard, R.W., 2005. Detecting stress during real-world driving tasks using physiological sensors. IEEE Trans. Intell. Transp. Syst. 6 (2), 156–166.
Huang, Y.S., Suen, C.Y., 1995. Method of combining multiple experts for the recognition of unconstrained handwritten numerals. IEEE Trans. Pattern. Anal. Mach. Intell. 17 (1), 90–94.
James, W., 1950. The Principles of Psychology, vol. 1. Dover Publications.
Juslin, P.N., Scherer, K.R., 2005. Vocal expression of affect. The New Handbook of Methods in Nonverbal Behavior Research.Oxford Univ. Press 65–135.
Karray, F., Alemzadeh, M., Saleh, j.A., Nours, M., 2008. Human–computer interaction: an overview. Int. J. Smart Sensing Intell. Syst. vol. 1 (1).
Kim, J., Andre, E., 2008. Emotion recognition based on physiological changes in music listening. IEEE Trans. Pattern Anal. Mach. Intell. vol. 30 (12), 2067–2083.
Kittler, J., Hatef, M., Duin, R.P., Matas, J., 1998. On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. 20 (3), 226–239.
Koelstra, S., Muhl, C., Soleymani, M., Yazdani, A., Lee, J.S., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I., 2012. DEAP: a database for emotion analysis using physiological signals. IEEE Trans. Affect. Comput. 3 (1), 18–31.
Lazarus, R.S., Kanner, A.D., Folkman, S., 1980. Emotions. A cognitive–phenomenological analysis. Theories of Emotion.New York Academic Press 189–217.
Lewis, P.A., Critchley, H.D., Rotshtein, P., Dolan, R.J., 2007. Neural correlates of processing valence and arousal in affective words. Cereb. Cortex 17 (3), 742–748.
Lutz, Catherine A., Abu-Lughod, Lila (Eds.), 1990. Language and the Politics of Emotion. Studies in Emotion and Social Interaction. Cambridge University Press, New York.
Magalhaes, J., Ruger, S., 2007. Information-theoretic semantic multimedia indexing. Int. Conf. on Image and Video Retrieval.Amsterdam 619–626.
McKeown, G., Valstar, M.F., Cowie, R., Pantic, M., 2010. The SEMAINE corpus of emotionally coloured character interactions. Proc. IEEE Int. Conf. Multimed. Expo. 1079–1084.
Mena, J.B., Malpica, J., 2003. Color image segmentation using the Dempster–Shafer theory of evidence for the fusion of texture. Int. Arch. Photogram. Rem. Sens. Spatial Inform. Sci. vol. XXXIV, 139–144 (Part 3/W8).
Meyer, G.F., Mulligan, J.B., Wuerger, S.M., 2004. Continuous audiovisual digit recognition using N-best decision fusion. J. Inf. Fusion 5, 91–101.
Mihalis, A., Gunes, H., Pantic, M., 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence–arousal space. IEEE Trans. Affect. Comput. vol. 2 (2).
Minsky, M., 1975. A framework for representing knowledge. The Psychology of Computer Vision.McGraw-Hill, New York.
Nefian, A.V., Liang, L., Pi, X., Liu, X., Murphye, K., 2002. Dynamic Bayesian networks for audio-visual speech recognition. EURASIP J. Appl. Signal Process. 11, 1–15.
Nesse, R.M., 1990. Evolutionary explanations of emotions. Hum. Nat. 1, 261–289.
Ni, J., Ma, X., Xu, L., Wang, J., 2004. An image recognition method based on multiple BP neural networks fusion. IEEE Int. Conf. Inf. Acquis. 323–326.
Oliveira, A.M., Teixeira, M.P., Fonseca, I.B., Oliveira, M., 2006. Joint model-parameter validation of self-estimates of valence and arousal. probing a differential-weighting model of affective intensity. Proc. 22nd Ann. Meeting Int'l Soc. for Psychophysics, pp. 245–250.
Pan, H., Liang, Z., Anastasio, T., Huang, T., 1999. Exploiting the dependencies in information fusion. Proc. Conf. Comput. Vis. Pattern Recognit. 2, 407–412.
Pantic, M., Valstar, M., Rademaker, R., Maat, L., 2005. Web-based database for facial expression analysis. Proc. IEEE Int. Conf. Multimed. Expo. 317–321.
Parrott, W., 2001. Emotions in Social Psychology. Psychology Press, Philadelphia.
Pereira, C., 2000. Dimensions of emotional meaning in speech. Proc. ISCA Workshop Speech and Emotion, pp. 25–28.
Pfleger, N., 2004. Context based multimodal fusion. ACM Int. Conf. Multimodal Interfaces 265–272.
Pitsikalis, V., Katsamanis, A., Papandreou, G., Maragos, P., 2006. Adaptive multimodal fusion by uncertainty compensation. Ninth Int. Conf. on Spoken Language Processing, Pittsburgh.
Plutchik, R., 2008. A psycho-evolutionary theory of emotions. Soc. Sci. Inf. 21 (4–5), 529–553.
Plutchik, Robert, Conte, R., Hope, 1997. Circumplex Models of Personality and Emotions. American Psychological Association, Washington, DC.
Poh, N., Bengio, S., 2005. How do correlation and variance of base experts affect fusion in biometric authentication tasks? IEEE Trans. Signal Process. 53, 4384–4396.
Rabiner, L., Juang, B., 1993. Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs, NJ.
Schaaff, K., 2008. EEG-based Emotion Recognition. Ph.D. thesis Universitat Karlsruhe.
Schaaff, K., Schultz, T., 2009. Towards emotion recognition from electroencephalography signals. 3rd Int. Conf. on Affective Computing and Intelligent Interaction.
Schachter, S., Singer, J.E., 1962. Cognitive, social and physiological determinants of emotional state. Psychol. Rev. 69, 379–399.
Schuller, B., 2011. Recognizing affect from linguistic information in 3D continuous space. IEEE Trans. Affect. Comput. 4 (4), 192–205.
Sebe, N., Cohen, I., Garg, A., Huang, T.S., 2005. Machine Learning in Computer Vision. Springer, Berlin, NY.
Sebea, N., Cohenb, I., Geversa, T., Huangc, T.S., 2005. Multimodal approaches for emotion recognition: a survey. Proc. of SPIE-IS&T Electronic Imaging, vol. 5670, pp. 56–67.
Singh, R., Vatsa, M., Noore, A., Singh, S.K., 2006. Dempster–Shafer theory based finger print classifier fusion with update rule to minimize training time. IEICE Electron. Express 3 (20), 429–435.
Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M., 2012a. A multimodal database for affect recognition and implicit tagging. IEEE Trans. Affect. Comput. vol. 3 (1).
Soleymani, C., Pantic, M., Pun, T., 2012b. Multi-modal emotion recognition in response to videos. IEEE Tran. On Affective Computing 99 (2), 211–223.
Stickel, C., Fink, J., Holzinger, A., 2007. Enhancing universal access — EEG based learnability assessment. Lect. Notes Comput. Sci. 4556, 813–822.
Sundberg, J., Patel, S., Bjo, E., Scherer, K.R., 2011. Interdependencies among voice source parameters in emotional speech. IEEE Trans. Affect. Comput. vol. 2 (3).
Tomkins, S.S., 1962. Affect, imagery, consciousness. The Positive Affects, vol. 1. Springer, New York.
Town, C., 2007. Multi-sensory and multi-modal fusion for sentient computing. Int. J. Comput. Vis. 71, 235–253.
Turk, M., 2005. Multimodal human–computer interaction. Real-time Vision for Human–computer Interaction.Springer, Berlin.
Weka machine learning tool — http://www.cs.waikato.ac.nz/ml/weka/downloading.html.
Xie, L., Kennedy, L., Chang, S.F., Divakaran, A., Sun, H., Lin, C.Y., 2005. Layered dynamic mixture model for pattern discovery in asynchronous multi-modal streams. IEEE Int. Conf. Acoust. Speech Signal Process. 2, 1053–1056.
Xu, H., Chua, T.S., 2006. Fusion of AV features and external information sources for event detection in team sports video. ACM Trans. Multimed. Comput. Commun. Appl. 2 (1), 44–67.
Zhu, Q., Yeh, M.C., Cheng, K.T., 2006. Multimodal fusion using learned text concepts for image categorization. ACM Int. Conf. Multimed. 211–220.
Zou, X., Bhanu, B., 2005. Tracking humans using multimodal fusion. IEEE Conf. Comput. Vis. Pattern Recognit. 4.

Please cite this article as: Verma, G.K., Tiwary, U.S., Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals, NeuroImage (2014), http://dx.doi.org/10.1016/j.neuroimage.2013.11.007