Seminar DIVA Group[*]
Using multimodal features to assess emotion

Simon Brunner
University of Fribourg
Department of Computer Science
boulevard de Perolles 90, 1700 Fribourg, Switzerland
simon.brunner@unine.ch

ABSTRACT

This paper presents an overview of systems using multimodalities to recognize automatically emotions improves over unimodal systems. Indeed, lot of studies have been done to detect emotions with one modality but only few tried to join multiple modalities into one system. This paper ties to clarify the obstacles someone might face by doing a multimodal emotion recognition system. What are the gain over unimodal, the challenges, the most pertinent modalities to take into account, what kind of framework can be used to verify if there are significant differences over unimodal systems. How to fuse the information coming from different channels. And finally a presentation of the results of two studies using multimodal.

1. INTRODUCTION

If we take a look at literature, movies or any story telling media, we can notice a certain fascination toward exploring human experiences through real humans and androids. Blade runner, The terminator, I, Robot, A.I. and many others have imagined, designed what could look like robots that can't be differentiated from real humans. For years, humans have created machines that can look like humans with skin, face, hairs, limbs, etc., to interact with us. But it doesn't matter how close to a human being the resemblance is, because if the actions and reactions don't fit the situation, the illusion is lost. A crucial element is essential to achieve human-computer interaction: emotions. That's right, to feel more like a human a machine has to recognize and show emotions[7]. So to have a realistic human-like interaction the machine has to learn to identify emotions and understand them as well. Of course all emotions might not be useful in most cases like an automated teller machine but it shows some kind of emotional intelligence from the machine in human-computer interaction[5]. Emotion recognition is more useful for applications like instructor, guide, helper, virtual friend etc. The goal is to reach an interaction that feel like a human-human interaction according to the human counterpart's emotional state, mood and feelings. Affective computing[7] tries to recognize, interpret, process, and simulate human emotions. The machine could recognize sadness and show compassion to the user, try to calm the user down when it identify anger. In a human-human interaction, the human sensors works simultaneously to analyze the counterpart, gather all the information and do a synthesis to evaluate the emotion perceived. The idea would be to do the same with a machine using multiple physical sensors and motors combined with artificial intelligence. Most of the experiences done so far focus on one modality. Most of the time, facial expression and speech are the more common modalities used. However, whether it is speech, face recognition, body gesture, etc., those channels of information are integrated independently from each other. If we want a more complete and more accurate system to detect human expressions we shouldn't take into account only one modality at the time. If we considerate how human detect or perceive emotions, it is clear that they take multiple information from different channel at the same time and combined them as a whole to select the right expression. So, the idea is to use several channels of information like facial expression and speech simultaneously, integrate them and analysed them. It would theoretically do a better job and enhance the final result than taken separately. So complementary and diversity of information channels is a good motivation, but there's another point to use multimodal system instead of unimodal. It is to be able to detect emotions when one or more modalities are not available via an alternate mode. For example the subject is a mute, he/she is not well sited in front of the camera or at all, It's too dark, he/she is paraplegic, too much surrounding sounds, etc. With multiple ways to detect emotions the system could work if one modality is still available.

2. PERTINENT MODALITIES

Modalities allow us to observe specific characteristics on individuals. Those characteristics can be heard, observed or measured. A Human body can translate the emotions by its shape, gestures and specially with the facial expression. It is called the body language. It can be captured with a video camera. Speech can also be important to recognise

[*] The Paper is a delivrabled of the Msc Research Seminar Emotion Recognition in 2011, DIVA Group University of Fribourg, supervised by Dr. D. Lalanne, Dr. F. Ringeval

emotions. The way a subject is speaking can revealed a lot of information about his actual mood. Internal changes in the human body can also help us to detect the emotional state of a person. Those changes are the heart beat rate, the pressure, the sweating, the temperature. These physiological characteristics are harder to capture and need more evolved probes.

2.1 Speech emotion recognition system

There are five different groups of features commonly used when it comes to characteristics [4]. the first group focus on fundamental frequency in half-tons characteristics. It measures voiced speech generation mechanism like the pitch, Mel Frequency Cepstral Coefficients (MFCC). the second group, acoustic, analyzes the speech production process. In other words, the energy in frequency bands of an expression (utterance). The third group takes care of temporal features such as the utterances duration/rhythm and the pauses. This process is related to behavioural speech production. The fourth group focuses on the voice quality with formants, articulated based features. The last group is loudness, energy filtered by human perception models. Some computations like the means, standard deviations, maximum and minimum of the obtained values in each group can be done. As the five groups are significantly different they can form a three dimensional feature spaces for characterizing the emotions. For multimodal experience the audio must be synchronize with the other modalities. To collect audio features there's the PRAAT toolkit for C++. It extract 26 features. There are two features that can't be used for instantaneous purposes. Speech rate and pausing structure take more time to deliver a result, about 2-3 seconds, and they focus more on long-term (prosody) analysis [3] on the speech. Prosody can be useful to get the emotional state from the speech.

2.2 Facial emotion recognition system

Facial recognition is essential to detect emotions. A lot of human emotions are revealed through facial expression. A smile, a long face, an risen eye brow give many clues of the emotional state. It is the one of the best way to communicate implicitly with others. To be able to detect those emotions, features need to be established. Those features can be determined by Facial animation parameters (FAP)[4]. MPEG-4 FAP needs a neutral face to calibrate called the neutral frame. Then it can measure the deformation of the features point (fig.1). There are two kind of measurement. One on coordinate features and the other based on distances features. One important point with facial recognition is the face position and orientation. To estimate where the face is pointing and its orientation, detectors focus on the face, the mouth and the eyes.

2.3 Gesture recognition system

Body gesture can be tracked with a video camera. Software like EyeWeb with its EyesWeb Expressive Gesture Processing Library can provide many of the information needed to extract the body features. It takes into account the silhouette of the body. There are five main features extracted, the fluidity of the gestures, the acceleration, the quantity of motion, the velocity and the contraction index of the body. The recorded data are then normalized with minimum and maximum values of each features to be compared with all subjects.

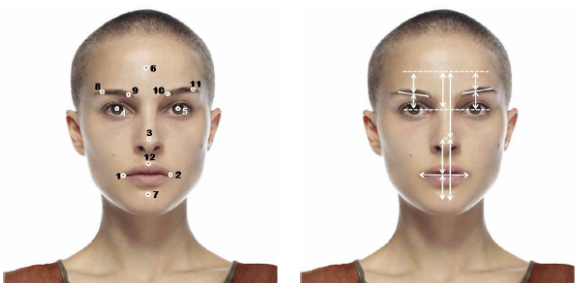

Figure 1: Facial recognition features

2.4 Physiological recognition system

It consists into observing and collecting information on internal body features on the central nervous system and the peripheral nervous system. For example the respiration rate, the temperature, the electroencephalogram (EEG) for the brain's electro activity, the electroocculogram (EOG) for the eyes' electro activities, the electromyogram (EMG) measuring the electrical impulses of muscles, the electrocardiogram (ECG) checks the activity of the heart, Blood volume pulse (BVP) detects changes in tissue blood volume and the pulse.

3. SYNCHRONIZATION OF MODALITIES

If we observe human behaviours it is noticeable that some modalities are closely coupled, like speech and gesture. Gestures occur at the same time or just a little after the speech units. Speech is also closely coupled with the mouth movements. According to multiple studies, when it comes to human-computer those observations still stands[8]. Using different modalities implies to deal with different kind of devices and sensors. The data coming from those sources like the signals forms and rate can be very different from one another. An important task is to successfully integrate and synchronize the signals. A simple way is to use event based synchronization like a common device/computer which can insert timestamps for each data frame of each modality, SMD file format and "Motu Timepiece" provide this kind of feature for example. It allows linearizing captured data and synchronized them with other streams like the audio clap used to synchronize audio and video. A solution is to use data fusion with canonical correlation analysis[6]. It reduces the difficulties of analyzing and utilizing common and unique information of very different modalities. The data fusion techniques combined multiple modalities in a combined analysis. This combined analysis provides the possibility to work with different data types. Each modality is reduced to a feature which can be from different dimensionality and nature and even recorded asynchronously.

4. FUSING THE INFORMATION

When only one modality is used, the data process is trivial. But when it comes to multiple modalities, the data processing must be organized. The signals from each modalities mustn't be treated separately but in a complementary way or redundant way. To approach a real human like analysis, the signals must be considered dependant to each other and the context of the situation should be also considered. Of course the latter is hard to evaluate and to transcribe but it should be in the process with the joint features. As for the final decision, three approaches can be used. The first one take the decision at the feature-level, early fusion. It means that there's one classifier for the three modalities. The data are analyzed to get the features to be then fused together. The second approach is decision-level also called late fusion. In this case, there's a classifier for each modality. The integration is done afterwards considering the decisions made by each classifier. The idea is to select the emotion that has the best probability (Max, Min, average or weight) in the three classifiers. A third solution could also be consider. It consists in adding another classifier which can be trained on the decisions provided by the classifiers. Actually there's another solution, data-fusion level, but the this one a rarely used due to its complications to use like a high level of synchronization needed between the modalities, highly susceptible to noise and the need to have information of the same nature like multiple cameras focusing on an object[8].

5. MULTIMODAL EMOTION RECOGNITION SYSTEMS

The two most common modalities to acquire information to recognize emotions are facial expression recognition and speech recognition. Body gesture can also be consider due to the fact it can use the same camera has the face recognition. To check if there's a significant difference of using multimodal over unimodal, the experiences must also test each modalities individually. It could be difficult to be able to compare an unimodal system to a multimodal system. A framework proposition (fig. 2) is to use a Bayesian classifier (BayesNet) which can be found with free software Weka. It allows the use of one modality at the time or several.

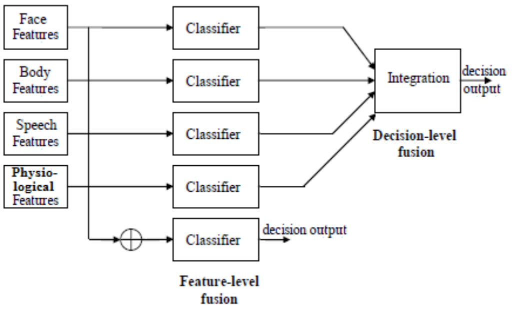

Figure 2: unimodal and multimodal system

So each modalities, speech, facial, body gesture and physiological signals have their own classifier. The system can extract the features individually, and a posteriori gives its result.

Let's take a look at two case studies. One [1] in which three modalities are used, speech, facial and gesture. The second study [2] focuses on two modalities, facial and speech expressions.

5.1 First study

This study[1] uses three modalities, facial, speech and gestures. The emotions tested in this study are Anger, Despair, Interest, Irritated, Joy, Pleasure, Pride and Sadness. Ten subjects were tested. They were students from five different nationalities, French, Hebrew, German, Greek and Italian. The balance between boys and girls was evenly done. Their expressions were acted in front of two cameras, one focusing on the face and the other on the body. The recordings were done at 25 frame per second. The subjects were asked to come without moustaches, beards and sunglasses to ease the detection. for the audio recording a micro attached to the subject was used. The voices were directly recorded on a hard drive with a sound editing software.

Facial features are extracted with MPEG-4 FAP (cf. 2.2), which locates and tracks points in the facial area. Feature points obtained from each frame were compared to feature points obtained from the neutral frame to estimate facial deformations and produce the FAP. As a result (cf. Table 1), the facial recognition process achieved to recognize 48% of the expressions with success.

Speech focus on features based on pitch, intensity, MFCC (Mel Frequency Cepstral Coefficient), Bark spectral bands, voiced segment characteristics and pauses length. The full set contains 377 features. The sampling rate of the recording is 44.1 kHz and 16 bit for the quantization. The results (cf. Table 1) are better than the facial recognition with an overall success of 57% with a noticeable 93% to recognize anger.

The body gesture recognition focuses on five main features extracted, the fluidity of the gestures, the acceleration, the quantity of motion, the velocity and the contraction index of the body. The overall performance reach 67% (cf. Table 1).

Now, the modalities are taken together. Two approaches are used. The first one takes the decision at the feature-level. It means that there's one Bayesian classifier for the three modalities. The second approach is decision-level. In this case, there's a Bayesian classifier for each modality. The idea is to select, with BayesNet, the emotion that has the best probability in the three classifiers. At the feature-level the results are quite impressive with 78% of success. With decision-level, the result is a bite under the feature-level with 74% but they are still better than any unimodal results done in this experience.

Table 1: 1st study results in percent

	Anger	Despair	Interest	Irritated	Joy	Pleasure	Pride	Sad	Mean	Std. dev.
Facial	56.67	40	50	53.33	53.33	53.33	33.33	46.67	48	7.45
Speech	93.33	23.33	60	50	43.33	53.33	56.67	76.67	57	19.68
Gesture	80	56.67	56.67	63.33	60	66.67	96.67	56.67	67	13.37
Features-level	90	53.33	73.33	76.67	93.33	70	86.67	83.33	78	12.13
Decision-level	96.67	53.33	60	60	86.67	80	80	80	74	14.13

This study concluded that multimodal improves the overall performance by minimum 10%. We also can notice feature-level has the best average score and a small standard deviation which increases its position to be the best solution for efficiency. The other advantage to use several modalities is to compensate the missing features from one another.

5.2 Second study

This study[2] considers two modalities, facial and speech recognition. The whole study was done with an actress acting and reading 258 sentences expressing the different emotions. Each modality is individually tested. Then, the modalities are fused together in a third and fourth experiment to test the multimodality performances at feature-level and decision-level.

Speech recognition is provided by the Praat speech processing software. It takes into account prosodic features like pitch and intensity statistics. In addition, the voiced and unvoiced speech ratios are also estimated. The overall performance shows that 71% of the time the emotions were correctly detected (cf. Table 2).

Facial recognition is supported with a camera focusing on markers carefully placed on the face of the subject on five distinguished areas such are the forehead, the eye brows, the eyes and the two cheeks. It collects the data and can evaluate the position and orientation of the face and then compute the distances between the areas to determine the emotion. The overall performance is 85% (cf. Table 2).

The experience now focuses on combining the two modalities. The features-level overall performance is 89%. The decision-level overall performance is also 89%.

Table 2: 2nd study results in percent

	Anger	Sadness	Happiness	Neutral	Mean	Std. dev.
Facial	79	81	100	81	85.2	9.87
Speech	68	64	70	81	70.75	7.27
Features-level	95	79	91	92	89.25	7.04
Decision-level	84	90	98	84	89	6.64

The conclusion of this study makes notice that even if the two multimodal techniques are superior to the unimodal ones and have better results on average; they don't perform equally for the same emotions. Indeed, feature-level performed best for anger and neutral emotions than sadness and happiness. On the other hand, decision-level is way better to detect sadness and specially happiness than anger and neutral emotions. So the best method depends on the application.

To compare the two studies, we can notice that, from a numerical point of view, the second study is better in every way. It has better scores and the standard deviations are smaller than the other study which means less disparity between the results. But it does not especially mean the techniques used were better. If we take a closer look at the way the two studies were conducted, we see that the recording from the second study were made by one single actress. It had half the emotions to recognize. The emotions in the first study were more subtle like joy and happiness. This makes recognition more difficult to distinguish one emotion from another.

6. CONCLUSION

To build a functional multimodal system for emotion recognition, there are many challenges that need to be managed[3]. The system demands robustness and very high accuracy, so it has to define, for the modalities, the main integration processes and functions for emotion recognition. It also need to identify in the recognition processes the best or the more suitable modalities for the different emotions. Another important requirement is the ability to response in short times which is essential for human-computer interaction. finding ways to synchronize the modalities and also to manage the temporal sequences which can be different from a modality to another.

We can conclude that multimodal improves the overall performance of emotion recognition over unimodal systems. It was expected, modalities cover the lack of the others in a complementary way like human beings. Multimodal systems take more resources and are more complex to put in place, but the benefice is worth it. With a minimum gain of 10% in the worst cases over unimodalities, and an average of success that can go up to 89%. The choice of using feature-level or decision-level depends on the desired of the emotions the system has to recognize. Indeed, the performances are better with feature-level for anger and neutral emotions where decision-level perform best with happiness and sadness. Either way, multimodal is the way to go in the future. Technologies like camera, microphones, and sensors of all types are more affordable and smaller than ever and they can be integrated in lots of devices. The power in those devices also increases years over years allowing the developers to use the resources for multimodal purposes.

7. REFERENCES

[1] Castellano G., Kessous L., Caridakis G., *Multimodal emotion recognition from expressive faces, body gestures and speech*, CACII, Lisbon, 2007

[2] Busso C., Deng Z., Yildirim S., Bulut M., Lee C.M., Kazemzadeh A., Lee S., Neumann U., Narayanan, S.*Analysis of emotion recognition using facial expressions, speech and multimodal information* ICMI 2004

[3] Kollias S., Karpouzis, K., *Multimodal Emotion Recognition and Expressivity Analysis*, ICME 2005

[4] Paleari M., Huet B., Chellali R., *Towards Multimodal Emotion Recognition: A New Approach*, MIR 2008

[5] Sebea N, Cohenb I, Geversa T, Huangc T, *Multimodal Approaches for Emotion Recognition: A Survey*, USA, 2005

[6] Rok Gajsek, Vitomir Struc, France Mihelic, *Multimodal Emotion Recognition using Canonical Correlations and Acoustic Features*,Pattern Recognition (ICPR), 2010

[7] R. Picard, *Affective Computing*, Cambridge, MA: The MIT Press, 1997

[8] Rajeev Sharma, Vladdimir I. Pavlovic, Thomas S. Huang, *Toward Multimodal Human Computer Interface*,IEEE, MAY 1998