

MULTIMODAL EMOTION ANALYSIS

Lisa Graziani

Febraury 8, 2018

Affective computing

- Affective computing is an emerging field of research that aims to enable intelligent systems to recognize, feel, infer and interpret human emotions.
- It is an interdisciplinary field spanning computer science, psychology, and cognitive science.
- Emotions can be detected in different ways, such as in speech, in facial expression, in images, in hand gesture, in body movements, in text, etc.
- Aim of simulating emotions in conversational agents in order to enrich and facilitate interactivity between human and machine.

Affective computing

Affective computing can be subdivided in

- **Sentiment analysis**
divides text (few works have been made in other channels) into two binary states (positive/negative).
- **Emotion recognition**
classifies data according to a large set of emotion labels (anger, disgust, fear, happiness, sadness and surprise).






Universal emotions

- In the early 1970s, Ekman identified six basic emotions that are universal across different cultures:
anger, **disgust**, **fear**, **happiness**, **sadness**, and **surprise**.
- In 1992, the seventh emotion 'contempt' was added to the universal set of emotions.
- Few tentative efforts to detect non-basic affective states, such as fatigue, anxiety, satisfaction, confusion, or frustration, have been also made.

Action Units

- Ekman and Friesen (1978) developed a Facial Action Coding System (FACS) by deconstructing a facial expression into a set of Action Units (AUs).
- AUs are characteristics of the face change, by comparison with neutral expression.
- AUs are defined via specific facial muscle movements (30 AUs).
- Later details on head movements and eye positions were also added (in total there are 44 AUs).
- An AU consists of three basic parts: AUnumber, FACS name, and muscular basis.

Action Units

AUNumber	FACS name	Muscular basis	Example image
1	Inner Brow Raiser	Frontalis, Pars Medialis	
2	Outer Brow Raiser	Frontalis, pars lateralis	
4	Brow Lowerer	Corrugator supercilii, Depressor supercilii	
5	Upper Lid Raiser	Levator Palpebrae Superioris	
9	Nose Wrinkler	Levator labii superioris alaeque nasi	

Russell's circumplex

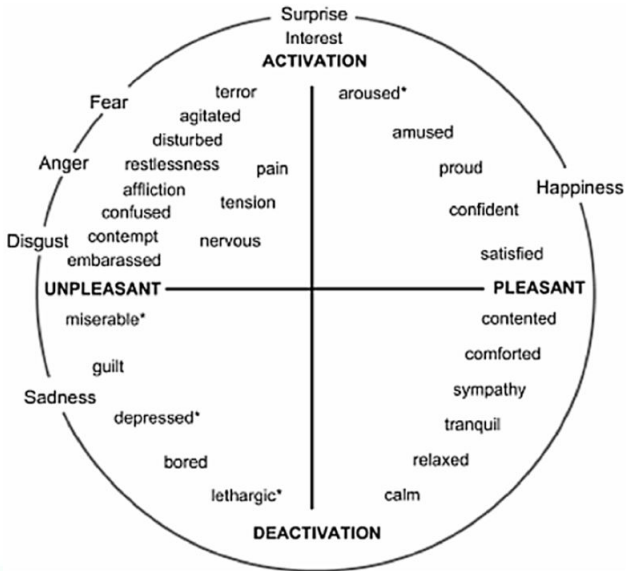
There are dimensional approach that represents emotions as coordinates in a multi-dimensional space.

An example is Russell's circumplex model (1980), which asserts that the affective state of human feeling can be considered as a point in two dimensional space.

The axis x represents the *valence* (pleasant-unpleasant continuum) and the axis y the *arousal* (activation-deactivation continuum).

- Valence represents the intrinsic attractiveness or averseness of an emotion
- Arousal represents the physiological and psychological state of being reactive to stimuli.

Russell's circumplex



Unimodal affect recognition

Visual modality

- The detection of naturalistic visual emotions, as **facial expression** or **body gestures**, has many applications as medical (such as pain detection), monitoring of depression, helping individuals on the autism spectrum, commercial uses.
- There are two types of facial expression features:
 - Permanent features remain the same through ages, which include opening and closing of lips and eyes, pupil location, eyebrows and cheek areas.
 - Transient features are observed only at the time of facial expressions, such as contraction of the corrugator muscle that produces vertical furrows between the eyebrows.

Audio modality

Vocal parameters, especially pitch, intensity, speaking rate and voice quality play an important role in recognition of emotions.

Unimodal affect recognition

Text modality

- Identify positive, negative, or neutral sentiment associated with words, phrases, sentences, and documents (sentiment analysis).
- In the last decade, researchers have been focusing on emotion extraction from texts of different genres such as news, blogs, Twitter messages, and customer reviews.
- Emotion extraction from social media content helps to predict the popularity of a product release or the results of an election poll, etc.
- Identifying emotions in text is a challenging task, because of ambiguity of words in the text, complexity of meaning and interplay of various factors such as irony, politeness, writing style, as well as variability of language from person to person and from culture to culture.

Multimodal affect recognition

- Multimodal affect recognition can be seen as the fusion of information from different channels, e.g., visual, audio, text.
- The fusion of multimodal data can provide surplus information with an increase in accuracy of the overall result or decision.
- It is an important prerequisite to the successful implementation of agent–user interaction.
- There is currently rather scarce literature on multimodal sentiment analysis. Most of the work in sentiment analysis has been carried out in the field of natural language processing (NLP).
- One of the primary obstacles is the development and specification of a methodology to integrate cognitive and affective information from different sources on different time scales and measurement values.

Multimodal affect recognition

Types of fusion

Feature-level (or early fusion)

The features extracted from each channel were combined in a "joint vector" before any classification operations are performed.

- Advantage: the correlation between various multimodal features at an early stage can potentially provide better task result.
- Disadvantage: time synchronization, as the features obtained belong to diverse modalities and can differ widely in many aspects, so before the fusion process takes place, the features must be brought into the same format.

Multimodal affect recognition

Types of fusion

Decision-level(or late fusion)

The features of each modality are examined and classified independently. The unimodal results are combined at the end of the process by choosing suitable metrics.

- Advantage: the fusion of decisions obtained from various modalities becomes easy compared to feature-level fusion, since the decisions resulting from multiple modalities usually have the same form of data. Moreover every modality can utilize its best suitable classifier or model to learn its features.
- Disadvantage: as different classifiers are used for the analysis task, the learning process of all these classifiers at the decision-level fusion stage, becomes tedious and time consuming.

Towards an intelligent framework for multimodal affective data analysis

Soujanya Poria^a, Erik Cambria^b, Amir Hussain^a, Guang-Bin Huang^c

^a*School of Natural Sciences, University of Stirling, UK*

^b*School of Computer Engineering, Nanyang Technological University, Singapore*

^c*School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore*

- They developed a big multimodal affective data analysis framework (text, audio and visual data).
- Only the six universal emotions (anger, disgust, fear, happiness, sadness, and surprise) were considered.

Training

For training they used three datasets corresponding to the three modalities:

- ISEAR dataset (2004) to build a model for emotion detection from **text**.
- CK++ dataset (2010) to construct a model for emotion detection from **facial expressions**.
- eNTERFACE dataset (2006) to build a model for emotion extraction from **audio** and to evaluate the trained models for the other two modalities.

ISEAR dataset

- The International Survey of Emotion Antecedents and Reactions (ISEAR) dataset contains 7666 such statements, which include 18,146 sentences and 449,060 running words.
- The data were collected conducting a survey in the 1900s across 37 countries and had approximately 3000 respondents.
- They have to describe a situation or event in which they felt a particular emotion, in term of a *statement* - a short text of a couple of sentences.
- Each statement is associated with the emotion felt in the situation.

ISEAR dataset

Examples

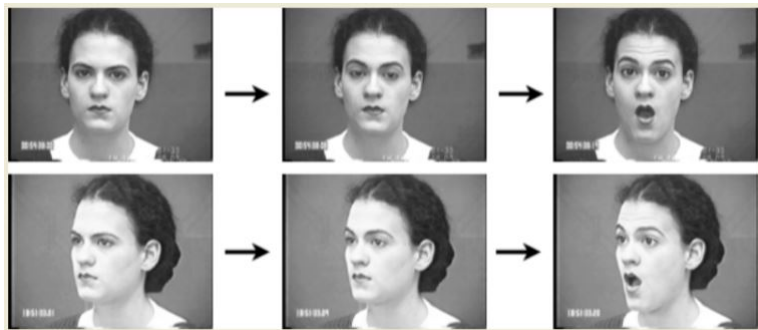
- **Joy:** "On days when I feel close to my partner and other friends. When I feel at peace with myself and also experience a close contact with people whom I regard greatly."
- **Fear:** "Every time I imagine that someone I love or I could contact a serious illness, even death."
- **Anger:** "When I had been obviously unjustly treated and had no possibility of elucidating this."
- **Sadness:** "When I think about the short time that we live and relate it to the periods of my life when I think that I did not use this short time."
- **Disgust:** I have felt this feeling when a person whom I believe and respect, lied to me.

The emotions in ISEAR dataset are little different from the six universal emotions: there are guilt and shame, that was removed and it misses surprise, that was added from another dataset.

CK++ dataset

- The Cohn-Kanade dataset consists of images of the facial behavior of 210 adults. The participants were 18–50 years old, 81% Euro-Americans, 13% Afro-Americans, and 6% from other ethnic groups; 69% were females.
- The experimenter asked the participants to perform a series of 23 facial displays. Each sequence begins with a neutral expression and proceeds to a peak expression.
- The facial expressions performed follow the FACS.
- The sequence of the facial images of each of the subjects was manually annotated with one of the six emotion categories.
- It contains 593 facial image sequences, but only 327 of them have specific emotion labels.

CK++ dataset



Each sequence begins with a neutral expression and proceeds to a target expression. The target expression is surprise, AU 1+2+5+27.

1 Inner brow raiser, 2 Outer brow raiser, 5 Upper lid raiser, 27 Mouth stretch.

eNTERFACE dataset

- It is an audio-visual emotion database that can be used as a reference database for testing and evaluating video, audio or joint audio-visual emotion recognition algorithms.
- 42 subjects of 14 nationalities were asked to listen to six short stories- each of them eliciting a particular emotion (Ekman's six basic emotions were used) - and to "immerge" himself into the situation. Then they had to read, memorize and pronounce the five proposed utterances, which constitute five different reactions to the given situation. Two experts judged whether the reaction expressed the emotion in an unambiguous way. If this was the case, the sample was added to the database.

eINTERFACE dataset

Situation to elicit anger

“You are in a foreign city. A city that contains only one bank, which is open today until 4pm. You need to get 200\$ from the bank, in order to buy a flight ticket to go home. You absolutely need your money today. There is no ATM cash machine and you don't know anyone else in the city. You arrive at the bank at 3pm and see a big queue. After 45 minutes of queuing, when you finally arrive at the counter, the employee tells you to come back the day after because he wants to have a coffee before leaving the bank. You tell him that you need the money today and that the bank should be open for 15 more minutes, but he is just repeating that he does not care about anything else than his coffee. . .”

R1: What??? No, no, no, listen! I need this money!

R2: I don't care about your coffee! Please serve me!

R3: I can have you fired you know!

R4: Is your coffee more important than my money?

R5: You're getting paid to work, not drink coffee!

eNTERFACE dataset

Situation to elicit surprise

"Your best friend invites you for a drink after your day at work. You join him on the Grand Place of Mons2, for a beer. Then, he suddenly tells you that he's actually gay! You are very surprised about it, you really didn't expect that!"

R1: You have never told me that!

R2: I didn't expect that!

R3: Wahoo, I would never have believed this!

R4: I never saw that coming!

R5: Oh my God, that's so weird!

eNTERFACE dataset

Situation to elicit fear

“You are alone in your bedroom at night, in your bed. You cannot sleep because you are nervous. Your bedroom is located on the second floor of your house. You are the only person living there. Suddenly, you start hearing some noise downstairs. You go on listening and realize that there is definitely someone in the house, probably a thief. . . or maybe even a murderer! He’s now climbing up the stairs, you are really scared.”

R1: Oh my god, there is someone in the house!

R2: Someone is climbing up the stairs

R3: Please don’t kill me...

R4: I’m not alone! Go away!

R5: I have nothing to give you! Please don’t hurt me!

eNTERFACE dataset

Situation to elicit disgust

"You are in a restaurant. You are already a bit sick and the restaurant looks quite dirty, but it is the only restaurant in the village, so you don't really have the choice. . . When you finally receive your plate, which is a sort of noodle soup, you take your spoon, ready to eat. Although you are very hungry, the soup does not taste very good. It seems that it is not very fresh. . . Suddenly you see a huge cockroach swimming in your plate! You're first surprised and you jump back out of your chair. Then, you look again at your plate, really disgusted. "

R1: That's horrible! I'll never eat noodles again.

R2: Something is moving inside my plate

R3: Aaaaah a cockroach!!!

R4: Eeeek, this is disgusting!!!

R5: That's gross!

eNTERFACE dataset

Situation to elicit happiness

“You learned this morning that you won the big prize of 5.000.000€ at the lottery! You're in a very happy mood of course, because you realize that some of your dreams will now become true! After the surprise to learn that you have won, comes the happy state of mind when you start dreaming about your new projects. You are in a restaurant, inviting your friends for a good meal, and telling them how happy you feel.”

R1: That's great, I'm rich now!!!

R2: I won: this is great! I'm so happy!!

R3: Wahoo... This is so great.

R4: I'm so lucky!

R5: I'm so excited!

eNTERFACE dataset

Situation to elicit sadness

"You just came back from an exhausting day at work. You are in a neutral state of mind when suddenly the telephone rings. You take the phone call and realize that it is your boy (girl) friend. He (she) announces you that he (she) doesn't want to go on the relationship with you. You first don't believe it, but after a while you start realizing what just happened. When you think about all the good moments you spent with your boy (girl) friend, and associate these memories with the fact that the relationship just finished, you start feeling really sad"

R1: Life won't be the same now

R2: Oh no, tell me this is not true, please!

R3: Everything was so perfect! I just don't understand!

R4: I still loved him (her)

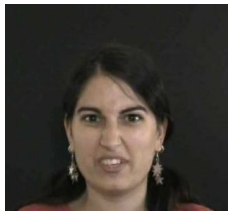
R5: He (she) was my life

eNTERFACE dataset

fear



disgust



happiness



sadness



anger



surprise



Overview of the method

They classified video clips that contained emotions in three modalities: visual information, sound track (speech), and captions (text).

Algorithm

- **Preprocessing:** Data for each modality were processed.
- **Feature extraction:** Features for building training models were extracted from the datasets for each modality. (For visual data, the feature extraction process includes a classification step.)
- **Fusion:** Feature-level fusion is used. They concatenating the feature vectors of all three modalities, to form a single long feature vector.
- **Training:** Using these features, a multimodal model was built and evaluated.

Use of visual data for emotion recognition

The method of feature extraction for visual modality of the video clips requires previous classification of still images.

Still images: data preparation

The CK++ dataset contains, for each subject, a sequence of n facial images expressing a particular emotion, from time T_0 to T_n . At time T_0 the subject starts to express the emotion in front of the camera, and expresses this emotion till time T_n . The first few images of the sequence correspond to a neutral expression, and the rest to the expression of a particular emotion.

They manually separated the images in each sequence into two categories: those expressing a neutral emotion and those expressing a given emotion.

Use of visual data for emotion recognition

These individual images, with their assigned categories (either neutral or one of the six emotions) formed the dataset. In total, the resulting dataset contained 5877 facial images corresponding to the 7 emotions (including neutral).



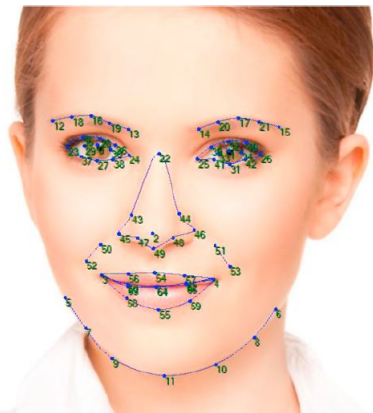
Labeling facial images in the sequence as neutral or carrying a specific emotion.

Use of visual data for emotion recognition

Still images: feature extraction

From each image they extracted 66 facial characteristic points (FCPs). The FCPs were used to construct facial features, which were defined as distances between FCPs. So there were a total of $\binom{66}{2} = 2145$ features per image.

Use of visual data for emotion recognition



Facial characteristic points of a facial image.

Facial point	Description
0	Left eye
1	Right eye
24	Left eye inner corner
23	Left eye outer corner
38	Left eye lower line
35	Left eye upper line
29	Left eye left iris corner
30	Left eye right iris corner
25	Right eye inner corner
26	Right eye outer corner
41	Right eye lower line
40	Right eye upper line
33	Right eye left iris corner
34	Right eye right iris corner
13	Left eyebrow inner corner
16	Left eyebrow middle
12	Left eyebrow outer corner
14	Right eyebrow inner corner
17	Right eyebrow middle
54	Mouth top
55	Mouth bottom

Some relevant facial characteristic points.

Use of visual data for emotion recognition

Unimodal classification of still facial images

To classify facial images by emotion, they designed a two-step classifier:

- 1) A two-way classifier was used to decide whether the image expressed no emotion (neutral) or some emotion.
- 2) A 6-way classification was carried out to identify the specific emotion category of the image.

Use of visual data for emotion recognition

Video clips (visual modality): feature extraction for multimodal fusion

To build a feature vector of a video clip showing the human face, they first divided the clip into a set of individual frames. Next, they extracted the features from these individual frames, and subsequently classified these images as described before. Finally, they built the feature vector for the video clip using coordinate-wise averaging of the feature vectors of individual frames:

$$x_i = \frac{1}{N} \sum_{j=1}^N x_{ij},$$

where x_i is the i th coordinate of the video clip's feature vector, x_{ij} is the i th coordinate of its j th frame's vector, and N is the number of frames in the video clip.

Use of audio (speech) for emotion recognition

First, the audio signal was extracted from video files in the eNTERFACE dataset. Then relevant features were extracted from the audio signal. To extract all audio features, they used the JAudio toolkit, which is a music feature extraction toolkit written in Java. There are two broad kinds of audio features:

- *Short time-based features* are mainly used to distinguish the timbral characteristics of the signal and are usually extracted from every short-time window (or frame), during which the audio signal is assumed to be stationary. (*Mel-frequency cepstral coefficients, Spectral centroid, Spectral rolloff, Spectral flux, Root mean square, Compactness, Time domain zero crossing*)
- *Long time-based features* can be generated by aggregating the short-term features extracted from several consecutive frames within a time window. (*Beat histogram, Beat sum, Strongest beat*)

Text-based emotion recognition

They considered the text as expressing both semantics and sentics (i.e., the conceptual and affective information associated with natural language).

- **Bag of concepts:** For each concept in the text, they obtained a 100-dimensional feature vector from their resource, EmoSenticSpace. Then they aggregated the individual concept vectors into one document vector through coordinate-wise summation:

$$x_i = \sum_{j=1}^N x_{ij},$$

where x_i is the i th coordinate of the document's feature vector, x_{ij} is the i th coordinate of its j th concept vector, and N is the number of concepts in the document.

Text-based emotion recognition

- **Sentic feature:** The polarity scores of each concept extracted from the text were obtained from SenticNet (a lexical resource that contains 30,000 concepts along with their polarity scores in the range from -1.0 to +1.0) and summed to produce one scalar feature.
- **Negation:** Negations can change the meaning of a statement. They followed the approach of Lapponi, Read, and Ovreid (2012) to identify the negation and reverse the polarity of the sentic feature corresponding to the concept that followed the negation marker.

After extracting the features, they built the text analysis by training model on the ISEAR dataset and evaluated this model on the transcriptions of the video files in the eNTERFACE dataset.

Results

- As testing data for all three modalities, they used the eNTERFACE dataset.
- They evaluated various supervised classifiers for each modality: for textual and speech modality, the best accuracy was achieved by using SVM, and for visual modality, by means of the Extreme Learning Machine (ELM).

Classifiers	Modalities			Fusion
	Visual	Speech	Text	
KNN	57.90%	57.25%	49.12%	59.45%
ANN	65.45%	67.28%	61.20%	68.25%
ELM	81.21%	72.17%	73.17%	84.45%
SVM	81.20%	78.57%	78.70%	87.95%

Results

- The current multimodal system outperformed the best state-of-the-art system by more than 10%.
- With feature-level fusion better accuracy was achieved compared with unimodal classifiers.
- The use of text-based features enhanced the accuracy of our system by 2,72% as compared with using only audio–visual fusion.

Method	Algorithms and modalities used	Accuracy
Datcu and Rothkrantz (2009)	HMM, audio and video	56.27%
Paleari and Huet (2008)	SAMMI framework, audio and video	67.00%
Mansoorizadeh and Charkari (2010)	Async. feature fusion, audio and video	71.00%
Dobrišek et al. (2013)	GMM, audio and video	77.50%
Proposed uni-modal method	SVM, audio	78.57%
Proposed uni-modal method	SVM, text	78.70%
Proposed uni-modal method	ELM, video	81.21%
Proposed bi-modal method	SVM, audio and video	85.23%
Proposed multi-method	SVM, audio, video, and text	87.95%

Results

The two-stage classifier for facial expression outperforms a simple seven-way classifier.

Confusion matrix for the CK++ facial expression dataset using a **one-stage** emotion classifier (ELM classifier, tenfold cross-validation).

Actual classification	Predicted classification							Precision
	Surprise	Joy	Sadness	Anger	Fear	Disgust	Neutral	
Surprise	1142	57	19	43	26	11	31	85.92%
Joy	65	1121	27	45	25	19	29	84.22%
Sadness	13	23	461	19	13	15	4	84.12%
Anger	29	21	3	770	65	77	57	75.34%
Fear	11	9	3	47	396	42	38	72.52%
Disgust	20	13	24	38	45	639	89	73.61%
Neutral	3	6	9	5	7	2	201	86.26%

Confusion matrix for the CK++ facial expression dataset using a **two-stage** emotion classifier (ELM classifier, tenfold cross-validation).

Actual classification	Predicted classification							Precision
	Surprise	Joy	Sadness	Anger	Fear	Disgust	Neutral	
Surprise	1170	49	25	43	15	6	21	88.03%
Joy	41	1191	21	37	17	6	18	89.48%
Sadness	7	12	492	17	9	4	5	89.78%
Anger	22	19	31	832	47	53	18	81.40%
Fear	9	7	14	32	445	27	12	81.50%
Disgust	14	10	12	34	37	732	29	84.33%
Neutral	3	7	3	0	0	0	220	94.42%

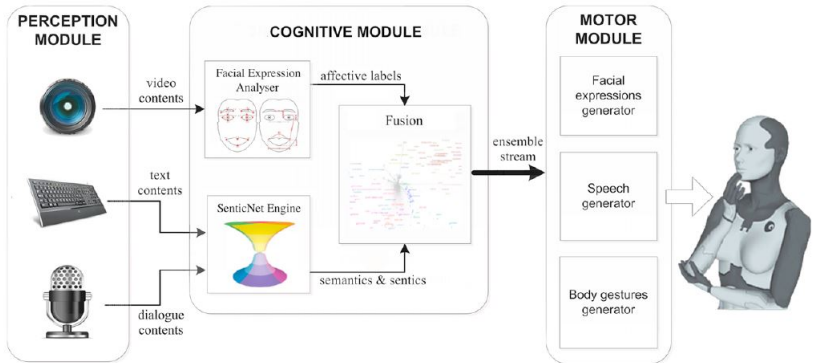
Results

Confusion matrix for the feature-level fusion (SVM classifier).

Actual classification	Predicted classification						Precision
	Surprise	Joy	Sadness	Anger	Fear	Disgust	
Surprise	195	10	3	2	3	7	88.63%
Joy	7	203	19	0	0	0	92.27%
Sadness	5	3	199	7	5	1	90.45%
Anger	15	2	3	196	2	2	89.09%
Fear	10	3	8	7	183	9	83.18%
Disgust	3	2	9	7	14	185	84.09%

Real-time emotion analysis system




Finally they have developed a real-time multimodal emotion recognition system. The system allows the users to upload emotional videos and it then shows the emotion expressed by the speaker of each video.



Conclusions

- Multimodal classifiers outperform unimodal classifiers. Furthermore, text modality plays an important role in supporting the performance of an audio-visual affect detector.
- There are still many challenges as estimating noise in unimodal channels, synchronization of frames, voice and utterance, time complexity, etc.
- We are still far from producing a real-time multimodal affect detector which can effectively and affectively communicate with humans, and feel our emotions.

References

-  *S.Poria, E.Cambria, A.Hussain, G.-B.Huang. Towards an intelligent framework for multimodal affective data analysis, Neural Netw. 63 (2015) 104–116.*
-  *S.Poria, E.Cambria, R.Bajpai, A.Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion. 37 (2017) 98-125.*
-  *O.Martin, I.Kotsia, B.Macq, I.Pitas. The eNTERFACE'05 audio-visual emotion database. In Proceedings of the first IEEE workshop on multimedia database management (2006).*