Journal of Visual Languages and Computing 27 (2015) 19–28

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

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

"Magic mirror in my hand, what is the sentiment in the lens?": An action unit based approach for mining sentiments from multimedia contents ☆

Luca Casaburi [a], Francesco Colace [a,*], Massimo De Santo [a], Luca Greco [b]

[a] DIEM – Department of Information Engineering, Electrical Engineering and Applied Mathematics, Università degli Studi di Salerno, Fisciano, Salerno, Italy
[b] DIIN – Department of Industrial Engineering, Università degli Studi di Salerno, Fisciano, Salerno, Italy

ARTICLE INFO

Article history:
Received 19 September 2014
Received in revised form
14 January 2015
Accepted 15 January 2015
Available online 3 February 2015

Keywords:
Affective computing
Ekman theory
Emotional intelligence
Human computer interaction

ABSTRACT

In psychology and philosophy, emotion is a subjective, conscious experience characterized primarily by psychophysiological expressions, biological reactions, and mental states. Emotion could be also considered as a "positive or negative experience" that is associated with a particular pattern of physiological activity. So, the extraction and recognition of emotions from multimedia contents is becoming one of the most challenging research topics in human–computer interaction. Facial expressions, posture, gestures, speech, emotive changes of physical parameters (e.g. body temperature, blush and changes in the tone of the voice) can reflect changes in the user's emotional state and all this kind of parameters can be detected and interpreted by a computer leading to the so-called "affective computing". In this paper an approach for the extraction of emotions from images and videos will be introduced. In particular, it involves the adoption of action units' extraction from facial expression according to the Ekman theory. The proposed approach has been tested on standard and real datasets with interesting and promising results.

1. Introduction

Affective computing is the study and development of systems and devices that can recognize, interpret, process, and simulate human affects. It is an interdisciplinary field spanning computer science, psychology, and cognitive science [2].

Affective computing gets its name from the field of psychology (where "affect" is, basically, a synonym for "emotion") and could offer benefits in an almost limitless range of applications: e-learning, e-health, e-therapy, entertainment, marketing [1,4,5,11,14,3,48].

In this scenario, the face and the facial expressions can be a powerful communication channel to convey emotions and opinions [7]. A facial expression can be defined as a visible manifestation of the emotional state, cognitive activity, intention, personality and psychology of a person [1]; it has been observed that facial expressions contribute strongly to a multimedia message, more than the vocal and the verbal components [11].

In general, the detection of affective states from facial expression in multimedia contents follows two main approaches: the recognition of discrete basic affects (by the adoption of a template matching approach); the

☆ This paper has been recommended for acceptance by Shi Kho Chang.

* Corresponding author.
E-mail addresses: lcasaburi@unisa.it (L. Casaburi), fcolace@unisa.it (F. Colace), desanto@unisa.it (M. De Santo), lgreco@unisa.it (L. Greco).

http://dx.doi.org/10.1016/j.jvlc.2015.01.001
1045-926X/

recognition of affects by the inference from movement of facial muscles according to the Facial Action Coding System (FACS) [8]. FACS classifies the facial movement as Action Units (AUs) and describes the facial expressions as a combination of AUs.

The first approach requires the execution of two main steps: the face's representation through features (landmarks or filtered images) and the classification of facial expression. Many papers deal with this approach such as [18] that shows how to represent facial expressions in a space of faces. In this case the face is encoded as a landmark (58 points) and the classification is performed through a probabilistic recognition algorithm based on the manifold subspace of aligned face appearances.

Zhang et al. [21] analyze the space of facial expressions to compare two classification systems: *geometric-based* (face is encoded by a landmark) and *Gabor-based* [9] (face is encoded by Gabor features); the classification is performed with two-layer preceptor network. They show that the best results are obtained with a network of 5–7 hidden preceptors to represent the space of expression. In this way, the facial expression analysis can be performed on static images [21,23] or video sequence [22,24,6].

Cohen et al. [22] propose a new architecture of HMM to segment and recognize facial expression and affects from video flow, while Lee et al. [24] propose a method using probabilistic manifold appearance. Wang et al. [27] describe an automatic system that performs face recognition and affect recognition in grey-scale images of face by making a classification on a space of faces and facial expressions. This system can learn and recognize if a new face is in the image and which facial expression is represented among basic affects. In [25] a methodology to choose the Gabor features with the PCA method is shown and then LDA is used to identify the basic affects. Bartlett et al. [26] propose a system for facial expressions' extraction from video that chooses Gabor features with AdaBoost algorithm and then affects are classified by a SVM. Garbas et al. [31] extract features from the face through a LBP filter and choose the most representative ones using Real-AdaBoost algorithm. Finally the faces are classified as positive or negative by a binary classifier.

Although the template matching approach typically obtains the best performance in terms of accuracy when used on standard datasets, it also shows some drawbacks on the computational side. In fact, a large number of parameters are involved when encoding face as an image: much more complex algorithms need to be used, requiring higher computation times and memory allocation. There are also some issues with the classification process: more features require a longer training phase.

A different approach aims at mining affective states from facial expressions by the use of the Ekman Model [7,8]. This model can infer six emotional states: happiness, anger, sadness, disgust, fear and surprise; recently, it has also been enriched by the introduction of states such as attention [15], fatigue [16] and pain [17].

The Ekman theory can be improved by the introduction of the Facial Action Coding System (FACS). As previously said, Facial Action Coding System is a system to taxonomize human facial movements by their appearance on the face. The process of categorizing physical expressions of emotions has proven useful to psychologists and to animators. Anyway, due to subjectivity and time consumption issues, FACS is actually employed in automated systems that detect faces in videos, extract the geometrical features of the faces, and then produce temporal profiles of each facial movement [47].

The recognition of affects by the inference from movement of facial muscles according to the FACS requires three steps: feature extraction, AUs recognition and basic affect classification. Parts of the face, such as eyebrows, eyes, nose and lips, are analyzed and encoded in sets of points [28],[29] or as texture features [17],[30] to detect AUs. Valstar et al. [28] introduce a method to detect the AUs starting from eyebrows, classifying their movements as spontaneous or voluntary by the use of a Relevance Vector Machine approach. After the detection of the AUs, it classifies their affective class by the adoption of a probabilistic decision function.

The FACS approach has been adopted in the automated Facial Image System (AFA) [32] which analyzes video in real-time to detect the sentiments. In this case, the face is encoded with a 2D mask which is used to interrogate a SVM to detect the associated affect. Robinson et al. [15] have developed a system that analyzes real-time video streams to detect the presence of one of the following moods: concordant, discordant, focused, interested, thinking and unsure. The face is encoded by 24 points and the distances between these points are used as features to identify different situations (open mouth, head movements, position of the eyebrows); the expressions encoded by FACS are recognized by a chain of HMM for each possible action and the computation of the probability of each state is obtained by the use of a Bayesian Network. Anyway, most Ekman based systems in literature show limited performance when compared to template matching approach.

In this paper we propose a novel approach for analyzing facial expressions and recognizing emotion from multimedia contents. This method uses the AUs approach for recognizing basic emotions [18] and implements a new technique for extracting feature points from the face and measuring emotion. The classic prototypes have been extended introducing the concept of combinations of AUs: when a combination occurs, a bonus or a penalty is assigned to the measure of emotions. In this way, a more detailed model can be obtained. Our method has proven to obtain better results than the classic Ekman based approaches and also outperforms template matching based systems in some cases. In particular, it provides better performance not only in terms of precision but also for execution time and this makes it suitable for real time video applications.

The paper is organized as follows: the proposed approach is discussed in the next section. In Section 2 results of test on CK+ dataset [35], for image analysis and video analysis, and on MMI Facial Expression [34], eNTERFACE'05 [12] and Cam3D [19] datasets for video analysis are presented. The obtained results are discussed in last section.

2. The proposed framework

As previously said, in this paper a framework for mining the emotive states of people appearing in multimedia

contents will be introduced. In particular, the proposed system is based on the Facial Expression Analysis. In particular it relies on the Facial Action Coding System (FACS) and the Ekman's theory. As depicted in Fig. 1 the proposed framework is organized in three main modules:

- *Features detection module*: a face skeleton composed by feature points is obtained from an image.
- *AUs detection module*: the probability that a specific action unit has been performed is here calculated. The action units (AUs) are obtained from the position of the feature points in the face skeleton. A vector of pairs (AU, probability) is built as result of this module.
- *Affect detection and classification*: recognition is carried out with Ekman's prototype. Detected affects are classified according to the Ekman's categories: happy, sad, angry, fear, disgust and surprise

2.1. Features detection module

According to Ekman's theory, the feature points of interest are depicted in Fig. 2. The features detection process consists of the following steps:

- Face detection: the Region Of Interest (ROI) of the image, where a face appears, is detected.
- ROI selection: the ROIs of eye, eyebrows, mouth and nose are extracted from the face's ROI.
- Eye feature detection: feature points of the eyes (points 8 and 9) are identified. These points are useful also to detect feature points of eyebrows and the orientation of the head (roll).
- Nose feature detection: the feature points of the nose (point 10) are identified. These points are used for detecting the orientation of the head (yaw and pitch).
- Eyebrow feature detection: the feature points of eyebrows (points 4, 5, 6 and 7) are identified.
- Mouth feature detection: the feature points of the mouth (points 0, 1, 2 and 3) are identified.

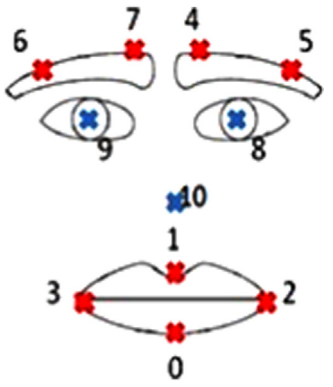

Fig. 2. Feature points for emotion extraction.

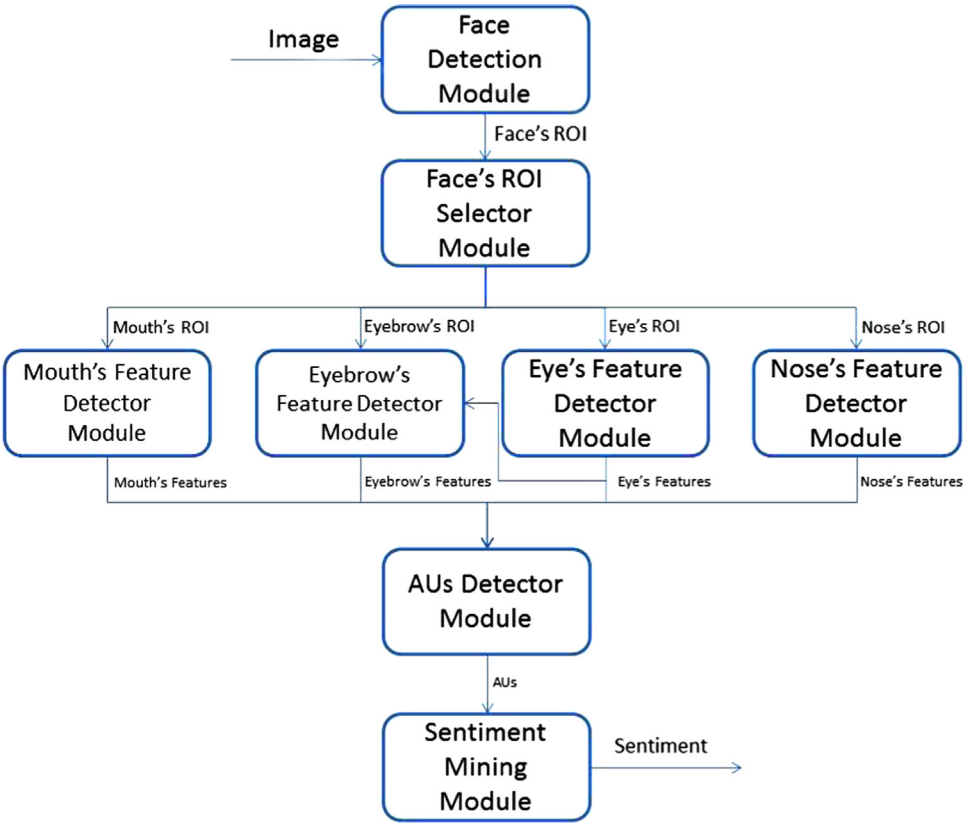

Fig. 1. General architecture of the proposed framework

2.2. Face detection

The problem of face detection refers to a specific case of object detection and in literature various solutions have been proposed: the SIFT (Scale Invariant Feature Transform) algorithm [36], SURF (Speeded Up Robust Features) algorithm [37], Haar Cascade, also known as the algorithm of Viola-Jones [38], SEMB-LBP Cascade (Statistically Effective Multi Block Local Binary Pattern Cascade) [39] and SURF Cascade (Speeded Up Robust Features Cascade) [46]. The solutions based on the SIFT and SURF algorithms include a first phase of feature extraction and then a second phase of classification that is typically performed using SVM. They have the following properties:

- Scale invariant.
- Rotation invariant.
- Symmetric invariant.
- Partially invariant to brightness changes.
- Highly repeatable.

Haar Cascade, SEMB-LBP Cascade and SURF Cascade are based on the idea of the Cascade classifier [40] and, in particular, on the AdaBoost algorithm [41].

Face detection based on SIFT or SURF typically obtains high performance in terms of accuracy of detection, but has the drawback of being too complex for real-time applications [42–46]. For our aims we selected four classifiers based on the OpenCV framework [13]: Stump-based Haar Cascade, Tree-based Haar Cascade, SEMB-LBP classifier and SURF Cascade.

We compared the four classifiers using the following datasets: IMM Face DB [47], CMU-MIT Face Test Set [44], Caltech Faces 1999 [33], Caltech 10,000 Web Faces [20], Cam3D [19]. The results show that the SURF Cascade has a false positive rate close to zero and has an excellent hit rate for high-quality images; it shows also the best performance on video with moving subjects. The Haar Cascade stump-based has the highest hit rate for images with both low and high resolutions.

In addition to precision evaluation, for each algorithm processing time has been calculated using the Caltech dataset. The tests were performed on a personal computer with the following features: CPU: Intel I3-2328M 2.20 GHz, RAM: 4 GB DDR3, Video Card: NVIDiA GeForce GT 635M, HD: 400 GB SATA2, SO: Windows 8×64. The results are shown in Table 1.

In conclusion, for the proposed system we chose to adopt the SURF Cascade because it shows good performance for face detection on video streams: it is very fast and has a false positive rate close to zero.

Table 1
Test results for the face detection module.

Caltech with 450 images		
Classifier	**Total time (s)**	**Time on 1 image (s)**
Haar Cascade 1	537,17	1.1937
Haar Cascade 2	334,10	0.7424
LBP Cascade	38,43	0.0854
SURF Cascade	36,04	0.0809

2.3. ROI selection

In this module the ROI of eyes, eyebrows, nose and mouth are extracted. The image of the face is splitted using a 24×24 grid, where each grid cell outlines a part of the face. The set of certain cells defines a search area where the parts of the face may be present. The search areas are the ROIs used in the feature point detection.

2.3.1. Eye feature detection

Eye feature detector receives the ROIs of eyes as input and returns the feature points as output. This module carries out the following phases:

- Eye detection: the position of the eyes is detected using a Haar Cascade classifier. The output is a rectangle that circumscribes the eye by defining its location and size.
- Feature point detection: the center of the rectangle is located and corresponds to the center of the pupil.

2.3.2. Nose feature detection

The module of nose feature detection is similar to eye feature detection, but returns the position of the nose as a feature point. The performed steps are

- Nose detection. The nose is detected by a Haar Cascade classifier: a rectangle is identified and it surrounds the tip of the nose.
- Feature point detection. The center of the rectangle is identified as feature point of the nose.

2.3.3. Eyebrow feature detection

Eyebrow feature detection involves the segmentation process of the image to obtain a binarized image of eyebrows. The segmentation algorithm performs these steps:

- *Extraction of the red channel.*
- *Image equalization.* The equalization technique allows obtaining a uniform histogram by redistributing grey levels.
- *Thresholding.* The binarized image B is derived from the equalized image Ceq in the following way:

$$B(x,y) = \begin{cases} 1 & \text{if } C_{eq}(x,y) > \theta \\ 0 & \text{otherwise} \end{cases}$$

where

$$\theta = \overline{C} + m\alpha\sigma$$

$$m = \pm 1$$

$$\overline{C} = \frac{1}{N}\sum_{x,y} C_{\exp}(x,y)$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{x,y}(C_{\exp}(x,y))^2 - \overline{C}^2} \quad \alpha \in [0,1]$$

x and y are the coordinates of pixels of the image.

The binarized image is enhanced through dilation with an elliptical kernel. Then the feature points are obtained by taking the projections of the ends of the eyes to the top limit of the eyebrows.

2.3.4. Mouth feature detection

Mouth feature detection is similar to the eyebrow feature detection. A segmentation algorithm is used to obtain the binarized image and includes the following steps:

- *Mouth detection*. Haar Cascade classifier is used to locate the precise position of the mouth in the ROI.
- *Extraction of green channel of the image*.
- *Calculation of the cumulative probability histogram*. The cumulative histogram (CH) is obtained from the histogram of the image (H) in the following way:

$$CH(i) = CH(i-1) + H(i) \quad \text{with} \quad i = 1, \ldots, 255$$

$$CH(0) = H(0)$$

- *Tresholding*. The binarized image B is obtained from the equalized image C in the following way:

$$B(x,y) = \begin{cases} 1 & \text{if } CH(C(x,y) = i) < \theta \\ 0 & \text{otherwise} \end{cases}$$

where

$$i = 0, \ldots, 255; \quad x = 0, \ldots, \dim_x(C); \quad y = 0, \ldots, \dim_y(C)$$

and x and y are the coordinates of pixels of the image.

The binarized image is improved through dilation with elliptical kernel. Then the Canny algorithm is applied to detect the contours of the mouth. The right, left, top and bottom extremes of these contours are the feature points.

2.4. AU detection

The feature points represent the face skeleton and are used to recognize a particular facial expression. A facial expression is described by 8 feature points; in more detail, 3 feature points (for eyes and nose) describe the rotation of the head.

In a neutral expression feature points of eyebrows and mouth are in a well-defined region; they move out if an AU is performed. AU detector calculates the distance of the points from the neutral region and in this way recognizes the performed AU. The distance is defined as a normalized distance with respect to the distance of the pupils, the feature points from eye line and the normal line. The eye line is the segment connecting the feature points of eyes: the normal line is the segment perpendicular to the eye line and passing through the center of the eye line (Fig. 3).

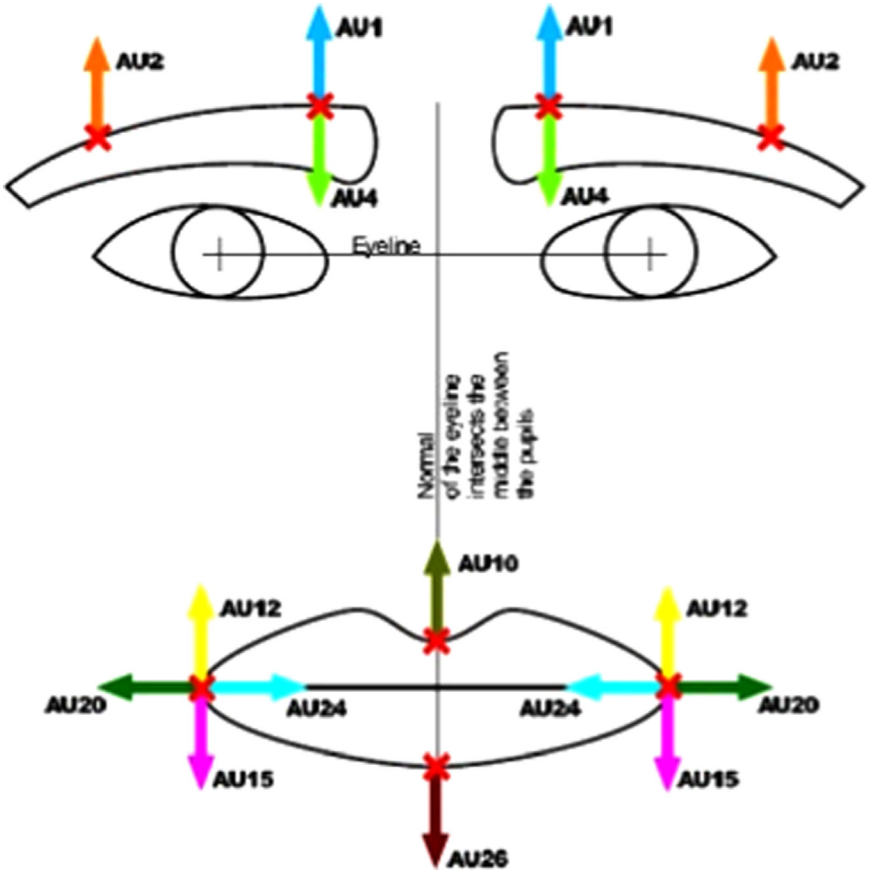

Fig. 3. AU definitions.

Then 3 feature points are used to define the eye line and normal-line and to calculate the distance of the points from these lines.

The regions of neutral state are identified by an upper and lower limit, which are calculated in the following way:

$$position_{new} = position_{neutral} + \gamma\left(dif_{UpDown}, dif_{EyesNose}\right)$$

$$\gamma\left(dif_{UpDown}, dif_{EyesNose}\right) = \alpha\left(dif_{EyesNose}\right) + \beta(dif_{UpDown})$$

The "neutral position" distance can be considered constant and it is determined empirically by analyzing different subjects in neutral poses. The values are reported in Table 2. α and β are two variables and depend on the rotation of the head (yaw and pitch). The following parameters are used:

$$dif_{EyesNose} = \begin{cases} > 0 & head\ turned\ to\ the\ right \\ < 0 & head\ turned\ to\ the\ left \end{cases}$$

$$dif_{UpDown} = \begin{cases} > 0 & head\ turned\ downward \\ < 0 & head\ turned\ upward \end{cases}$$

where

$$dif_{EyesNose} = (x_{nose} - x_{left\_eye}) - (x_{right\_eye} - x_{nose})$$

$$dif_{UpDown} = \left(y_{eye\_\ mod\ el\_position} - y_{eye}\right) - \left(x_{right_{eye}} - x_{nose}\right)$$

α and β are functions that vary according to the AU. These functions have been obtained by inferring mathematical models empirically through a dataset of images designed specifically for this purpose.

2.5. Affect detection and classification

For the affect detection, Ekman's prototypes have been modified. These variants calculate a measure, between 0 and 1, that identifies if a particular affect is detected (Table 3). If particular combinations of AU arise, the result of these adapted prototypes is amended by adding or subtracting a score.

Working on the CK+ dataset [35], we found that different AUs can occur at the same time when a particular affective status is considered. For example, if a person smiles, AU12 is detected with high intensity while AU10 and AU20 can be also detected with low intensity: the subject could be happy, scared and disgusted at the same time.

Table 2
Thresholds for the neutral regions.

AU	Lower limit	Upper limit	Reference line
1	0.238	0.392	Eyeline
2	0.278	0.472	Eyeline
4	0.228	0.188	Eyeline
10	0.902	0.825	Eyeline
12	1.086	0.822	Eyeline
15	1.056	1.111	Eyeline
20	0.556	0.583	Normal line
24	0.415	0.276	Normal line
26	1.203	1.284	Eyeline

Table 3
Original definition and adapted definition of the prototypes.

Affect	Original definition	Adapted definition
Fear	1+2+4+5+20+25	(1L+1R+2L+2R+20L+20R)/6
Surprise	1+2+5+26	(1L+1R+2L+2R+26)/5
Anger	4+5+7+24	(4L+4R+24L+24R)/4
Sad	1+4+15	(4L+4R+15L+15R)/4
Disgust	4+9+10+17	(4L+4R+10)/3
Happy	6+12+25	(12L+12R)/2

With the combinations of AUs, a bonus is given to the emotion of "happiness" and penalties are given to "fear" and "disgust" highlighting the difference. A bonus or a penalty is added to calculated measure with adapted prototypes. This bonus or penalty is obtained according to the combinations of AUs that occur on the face (Table 4).

2.6. Video analysis

The proposed architecture (Fig. 1) could be easily extended to the problem of affect detection in video. In this case an emotional state *tracking* problem has also to be considered. The modified architecture is shown in Fig. 4.

Such an architecture aims at analyzing videos both in real time and off-line operation. It consists of three modules:

- *User tracking*. This module receives as input a new frame and the history of a user present in the scene. This information is stored in a vector of users: for each user the position at the last frame and the *affect measures* are stored. This module checks if there are any other people in the scene, if they are new or were already present. The output is the position of the user within the scene. The problem of user tracking is resolved by evaluating the minimum Euclidean distance from the last position and current position of the users' face. Each detected user's face is compared to the vector of users' faces of previous frame to infer if such a face appeared in the previous frame. We make the assumption that the face can be subjected to minimal movements between two consecutive frames, and then two faces having a slightly different position in two consecutive frames can be the same. The faces without a previous position are considered as new users in the scene and their location and size are stored as a new entry in the vector of users. Faces undetected in the current frame, but present in the previous frame, can be false positive or users that leave the scene. So, for each detected face in a frame, the tracking is obtained by determining the minimum Euclidean Distance with respect to all the other faces detected in the previous frame.
- *Face emotion analysis*. This module considers the faces of all users and calculates the probabilities of the six basic affects. In this step emotion detection process (Fig. 2) is applied to all users' faces.
- *Affect tracking*. This module calculates the measures of all users' affects by considering mood changes over time. The problem of the affect tracking is resolved by

Table 4
Bonus and penalty definition.

Combination	Bonus	Penalty
AU4L–AU4R	–	Surprise – 0.1 Fear – 0.1 Happy – 0.3
AU1L–AU1R–AU2L–AU2R	Surprise – 0.2	Disgust – 0.3 Anger – 0.3 Sadness – 0.2
AU24L–AU24R (low probability)	Disgust – 0.2	–
AU24L–AU24R (high probability)	Anger – 0.2	–
AU12L–AU12R–AU10 (low probability)	Happy – 0.2	Disgust – 0.1

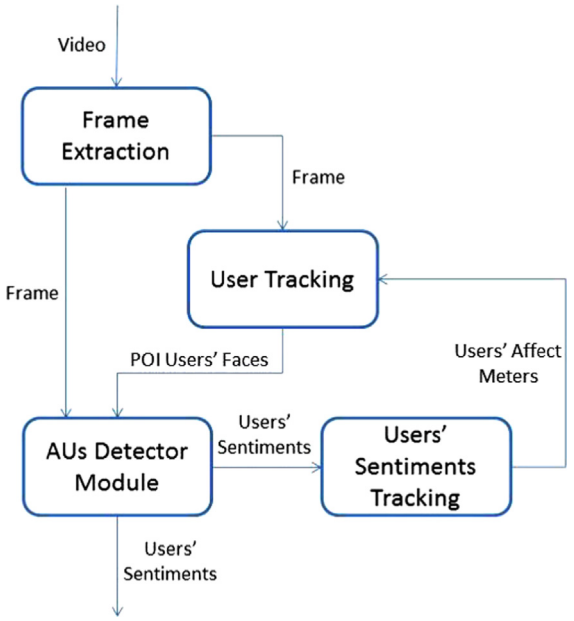

Fig. 4. Architecture of the video analysis system.

Table 5
Test results on AUs.

AU	Precision (%)	Recall (%)
1	81.81	96.11
2	92.94	88.76
4	82.65	84.37
10	90.00	69.23
12	69.33	66.67
15	73.08	46.34
20	75.00	64.28
24	75.61	59.61
26	81.18	66.35

Table 6
Test results of image analysis on base sentiment (dataset CK+).

Affect	Test results obtained	
	Precision (%)	Recall (%)
Anger	51	48
Disgust	55	56
Fear	67	95
Happy	79	75
Sadness	59	48
Surprise	80	87

Table 7
Test results on the MMI, eNterface 05′ and Cam3D datasets on base sentiments.

	MMI		eNTERFACE 05′		Cam3D	
	Pre (%)	Rec (%)	Pre (%)	Rec (%)	Pre (%)	Rec (%)
Happy	80.55	60.42	67.44	47.54	67.86	100.00
Fear	62.50	37.50	45.58	57.30	–	–
Anger	46.15	40.00	39.06	48.27	–	–
Disgust	40.00	53.33	43.25	84.54	75.00	60.00
Sadness	41.93	66.67	67.90	83.90	100.00	100.00
Surprise	57.14	95.23	42.32	55.48	60.00	50.00
Neutral	–	–	–	–	50.00	50.00
Average	54.71	58.86	50.93	62.84	70.57	72.00

Fig. 5. Frames extracted from the dataset UniSA.

Table 8
Performance comparison with Cohn [35] on the CK+ dataset for base sentiments. Results shown in bold text are obtained by the proposed system.

Emotion	Neutral	Happy	Anger	Disgust	Fear	Sad	Surprise
Neutral	*78.59*	1.03	3.51	8.18	1.85	5.78	1.03
	82.33	**1.5**	**0.74**	**4.6**	**1.02**	**7.8**	**2.01**
Happy	0	*86.22*	4.91	5,0.65	3.19	0	0
	1.2	***84.45***	**6.14**	**3.11**	**0**	**5.1**	**0**
Anger	2.04	4.76	*66.46*	14.28	5.21	6.09	1.14
	2.61	**3.2**	***73.51***	**9.7**	**2.01**	**8.22**	**0.75**
Disgust	3.4	1.13	10.9	*62.27*	10.9	9.09	2.27
	3.11	**2.32**	**14.12**	***65.23***	**4.39**	**5.57**	**5.26**
Fear	1.19	13.57	7.38	7.61	*63.8*	3.8	1.9
	0	**13.19**	**2.14**	**1.6**	***71.23***	**0.87**	**10.97**
Sad	5.55	1.58	13.25	11,19	3.96	*61.26*	3.17
	8.14	**2.33**	**8.7**	**10.3**	**8.14**	***58.9***	**3.49**
Surprise	0	0	0	0	2.02	4.04	*93.93*
	3.2	**0**	**0**	**0**	**1.08**	**0**	***95.72***

Table 9
UniSA dataset base sentiments retrieval: test results.

UniSA dataset			
Emotion	**No. video**	**Pre (%)**	**Rec (%)**
Happy/surprise	21	66.67	63.63
Sadness/anger	19	68.42	54.17
Fear	19	42.10	61.54
Average	59	59.06	59.78

using an "affect meter". For each basic emotion there is an affect meter that shows the measure of affect recognized within the previous frame. This measure is represented as a real value $A \in [0; 10]$, which is increased if the calculated measure for the current single frame is greater than the average of the measures of the last five frames.

3. Experimental results

The image analysis experimental stage was carried out by implementing a module that receives an input image and returns as output the position of the faces within the frame and the associated affective states. Image analysis was tested using the CK+ dataset. Two tests were performed. The first one tests the capabilities of the system in identifying the AUs on the eyebrows and the mouth; the second one detects the basic sentiments. The results are shown respectively in Tables 5 and 6.

In the case of video analysis, performance on Ekman's sentiments' detection (happy, surprise, anger, disgust, sadness, fear) has been tested on CK+, MMI, Cam3D and eNTERFACE'05 standard datasets, as shown in Table 7.

We also made a performance comparison between our system and the method proposed by Cohn [35]; Cohn's system codes the face through a wireframe model and the extracted features are classified by means of a Gaussian TAN Classifier. For the comparison, standard CK+ dataset was used; results are reported in Table 8.

For a better characterization of the proposed method an experimental campaign on a real dataset has also been conducted. In particular, a UniSA[1] dataset (Fig. 5) has been built: it contains 59 videos obtained by shooting 21 different users, aged among 21 and 45 years, of both genders.

Each user watched three different videos:

- the first video shows funny sketches about animals and should inspire happiness and surprise sentiments (positive emotional states) to the users;
- the second video is a public service announcement on road safety and should inspire sadness and anger sentiments (negative emotional states); and
- the third video is a scene from a horror movie and should inspire fear sentiments (negative emotional states).

At the end of the shooting, users filled a questionnaire declaring the intensity of their sentiments during the views. The intensity has been expressed as a number among 1 (very negative) and 5 (very positive). The results obtained from the analysis of the UniSA dataset are shown in Table 9.

Also in this case our method obtained encouraging results. Anyway, the proposed approach shows troubles in the detection of "fear" sentiment. In particular, the system is not able to distinguish "fear" from "sadness/anger" properly. This is probably due to the common impulsive nature of these sentiments.

4. Conclusions

In this paper a novel approach to the detection and classification of sentiments within multimedia contents has been introduced. This technique is based on the definition of a head tracking strategy. Face detection and the extraction of points of interest involve image processing techniques, properly improved or adapted to our aims. The recognition of sentiment relies basically on the Ekman's theory. The proposed approach has been tested

[1] The dataset can be downloaded at the following link: http://193.205.190.208/unisadataset/

on standard and custom made datasets; the results are very encouraging. The future works aim at applying the proposed approach to real time video streams and collecting emotional states from various users.

References

[1] L. Shen, M. Wang, R. Shen, Affective e-learning: using emotional data to improve learning in pervasive learning environment, Educ. Technol. Soc. 12 (2) (2009) 176–189.

[2] Tao Jianhua, Tieniu Tan, Affective computing: a review, in: Gerhard Goos, Juris Hartmanis, Leeuwen Jan van (Eds.), Affective Computing and Intelligent Interaction, 3784, LNCS, Springer, 2005, pp. 981–995.

[3] Francesco Colace, Massimo De Santo, Luca Greco, A probabilistic approach to Tweets' sentiment classification, in: Proceedings of 2013 IEEE Humaine Association Conference on Affective Computing and Intelligent Interaction, Ginevra 3–6 Settembre, vol. 1, 2013, pp. 37–42.

[4] D. McDuff, R. Kaliouby, E. Kodra, R.W. Picard, Measuring voter's candidate preference based on affective responses to election debates, in: Proceedings of the the 5th Biannual Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII 2013), pp. 2–5.

[5] A. Luneski, E. Konstantinidis, P.D. Bamidis, Affective medicine: a review of affective computing efforts in medical informatics, Methods Inf. Med. 49 (3) (2010) 207–218.

[6] F. Colace, P. Foggia, G. Percannella, A probabilistic framework for TV-news stories detection and classification, in: Proceedings of the IEEE International Conference on Multimedia and Expo, 2005, pp. 1350–1353.

[7] P. Ekman, W. Friesen, Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues, Prentice-Hall, Upper Saddle River, New Jersey, 1975.

[8] P. Ekman, W. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, Palo Alto, 1978.

[9] Michel F. Valstar Timur R. Almaev, Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition, in: Proceedings of the Humane Association Conference on Affective Computing and Intelligent Interaction, pp. 356–361.

[11] Martin Wollmer, Felix Weninger, Tobias Knaup, Bjorn Schuller, Congkai Sun, Kenji Sagae, Louis-Philippe Morency, Youtube movie reviews: sentiment analysis in an audio-visual context, IEEE Intell. Syst. 28 (3) (2013) 46–53.

[12] 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, pp. 8–15.

[13] OpenCV Documentation: ⟨http://docs.opencv.org/⟩.

[14] Zhihong Zeng, Maja Pantic, Glenn I. Roisman, Thomas S. Huang, A survey of affect recognition methods: audio, visual, and spontaneous expressions, IEEE Trans Pattern Anal Mach Intell 31 (1) (2009) 39–58.

[15] R. El Kaliouby, P. Robinson, Real-time inference of complex mental states from facial expressions and head gestures, in: Proceedings of the International Conference on Computer Vision & Pattern Recognition, 2004, pp. 181–200.

[16] H. Gu, Q. Ji, An automated face reader for fatigue detection, in: Proceedings of the International Conference on Face & Gesture Recognition, 2004, pp. 111–116.

[17] M.S. Bartlett, G. Littlewort, M.G. Frank, C. Lainscsek, I. Fasel, J. Movellan, Fully automatic facial action recognition in spontaneous behavior, in: Proceedings of the International Conference on Automatic Face & Gesture Recognition, 2006, pp. 223–230.

[18] Ya Chang, Changbo Hu, Matthew Turk, Probabilistic expression analysis on manifolds, in: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'04, 2004, pp. 520–527.

[19] Marwa Mahmoud, Tadas Baltrušaitis, Peter Robinson, Laurel Riek, 3D corpus of spontaneous complex mental states, in: Proceedings of Affective Computing and Intelligent Interaction, 2011, pp. 205–214.

[20] Caltech 10, 000 Web Faces, ⟨http://www.vision.caltech.edu/Image_Datasets/Caltech_10K_WebFaces/⟩.

[21] Z. Zhang, M. Lyons, M. Schuster, S. Akamatsu, Comparison between geometry-based and gabor-waveletsbased facial expression recognition using multi-layer perceptron, in: Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 454–459.

[22] I. Cohen, N. Sebe, A. Garg, L.S. Chen, T.S. Huang, Facial expression recognition from video sequences: temporal and static modeling, Comput. Vis. Image Underst. (2003) 160–187.

[23] A. Elad, R. Kimmel, On bending invariant signatures for surfaces, IEEE Trans. Pattern Anal. Mach. Intell. 25 (10) (2003) 1285–1295.

[24] K. Lee, J. Ho, M.H. Yang, D. Kriegman, Video based face recognition using probabilistic appearance manifolds, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, 2003, pp. 313–320.

[25] H. Deng, L. Jin, L. Zhen, J. Huang, A new facial expression recognition method based on local gabor filter bank and PCA plus LDA, Int. J. Inf. Technol. 11 (11) (2005) 86–96.

[26] G. Littlewort, M. Stewart Bartlett, I. Fasel, J. Susskind, J. Movellan, Dynamics of facial expression extracted automatically from video, in: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'04), 2004.

[27] H. Wang, N. Ahuja, Facial expression decomposition, in: Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV'03), vol.2, 2003, pp. 958–965.

[28] M.F. Valstar, M. Pantic, Z. Ambadar, J.F. Cohn, Spontaneous vs. posed facial behavior: automatic analysis of brow actions, in: Proceedings of ICMI'06, 2006, pp. 162–170.

[29] J.F. Cohn, L.I. Reed, Z. Ambadar, J. Xiao, T. Moriyama, Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior, in: Proceedings of 2004 IEEE International Conference on Systems, Man and Cybernetics, 2004, pp. 610–616.

[30] J. Whitehill, C.W. Omlin, Haar features for FACS AU recognition, in: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR'06), 2006, pp. 96–101.

[31] J. Garbas, T. Ruf, M. Unfried, A. Dieckmann, Towards robust real-time valence recognition from facial expressions for market research applications, in: Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction, 2013, pp. 570–575.

[32] J.F. Cohn, S. Lucey, J. Saragih, P. Lucey, F. De la Torre, Automated facial expression recognition system, in: Proceedings of the IEEE International Carnahan Conference on Security Technology, 2009, pp. 172–177.

[33] Caltech Archive: ⟨http://www.vision.caltech.edu/html-files/archive.html⟩.

[34] M.F. Valstar, M. Pantic, Induced disgust, happiness and surprise: an addition to the MMI facial expression database, in: Proceedings of the International Conference on Language Resources and Evaluation, Workshop on EMOTION, 2010, pp. 65–70.

[35] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended Cohn–Kanade dataset (CK+): a complete dataset for action unit and emotion-speci_ed expression, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010, pp. 94–101.

[36] David G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, 1999, pp. 1150–1157.

[37] Herbert Bay, Tinne Tuytelaars, Luc Van Gool, SURF: Speeded Up Robust Features, in: Aleš (Ed.), Computer Vision – ECCV 2006, 3951, Lecture Notes in Computer Science, 2006, pp. 404–417.

[38] Paul Viola, Michael Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2001, pp. 511–518.

[39] Shengcai Liao, Xiangxin Zhu, Zhen Lei, Lun Zhang, Stan Z. Li, Learning multi-scale block local binary patterns for face recognition, in: Proceedings of the International Conference on Biometrics (ICB), 2007, pp. 828–837.

[40] J. Gama, P. Brazdil, Cascade generalization, Mach. Learn. 41 (3) (2000) 315–343.

[41] Robert E. Schapire, Yoram Singer, Improved boosting algorithms using confidence-rated predictions, Mach. Learn. 37 (3) (1999) 297–336.

[42] M.N. Dailey, S. Tongphu, N. Thongsak, Rapid detection of many object instances, in: Jacques Blanc-Talon, Wilfried Philips, Dan Popescu, Paul Scheunders (Eds.), Advanced Concepts for Intelligent Vision Systems, 5807, Lecture Notes in Computer Science, 2009, pp. 434–444.

[43] A. Schmidt, A. Kasinski, The performance of the Haar cascade classifers applied to the face and eyes detection, in: Marek Kurzynski, Edward Puchala, Michal Wozniak, Andrzej Zolnierek (Eds.), Computer Recognition Systems 2, 45, Advances in Soft Computing, 2007, pp. 816–823.

[44] CMU-MIT Frontal Face Dataset: ⟨http://vasc.ri.cmu.edu/idb/html/face/frontal_images/⟩.
[45] R. Rani, S.K. Grewal, K. Panwar, Object recognition: performance evaluation using SIFT and SURF, Int. J. Comput. Appl. 75 (3) (2013) 39–47.
[46] N. Younus Khan, B. McCane, G. Wyvill, SIFT and SURF performance evaluation against various image deformations on benchmark dataset, in: Proceedings of the International Conference Digital Image Computing Techniques and Applications (DICTA), 2011, pp. 501–506.
[47] J. Hamm, C.G. Kohler, R.C. Gur, R. Verma, Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders, J. Neurosci. Methods 200 (2) (2011) 237–256.
[48] F. Colace, M. De Santo, L. Greco, SAFE: a sentiment analysis framework for E-learning, Int. J. Emerg. Technol. Learn. 9 (6) (2014) 37–41.