Spatial-Temporal Model for Emotion Interpretation

Lucio C. Vázquez S.[1], Ivo H. Pineda T.[2], María J. Somodevilla[2], Concepción Pérez de Celis H.[2], and Mario Rossainz L.[2]

Benemérita Universidad Autónoma de Puebla, Fac. Ciencias de la Computación, Puebla, México.
[1]luciovs@prodigymovil.com,
[2]{ipineda,mariasg,cpelish,mrossainz}@scs.buap.mx

Abstract. Interpreting facial expressions is a task that humans perform everyday automatically. Most of gesture recognition research are focused to discriminate different facial expressions, while interpreting them is relatively a new area. This paper deals with how to recognize an emotional state, since it can be represented as states at a time t that are influenced directly by a previous state at $t-1$ therefore exists an inherent temporarily.

Keywords: Hiden Markov Models,spatial temporal variability, face tracking, emotion identification , emotion interpretation

1 Introduction

From the perspective of computer science the idea that a secuence of facial gestures might represent a complex emotional state which can be identified automatically , this is precisely the motivation of this research work. An example would be analyze a sequence of facial gestures during a job interview to determine if the interviewer is lying or even determine a psychological condition more complex. Another example is the problem of lie detection, facial gestures proven valuable information for determining whether a person is lying or not. Psychologists have concluded that no single gesture represents a lie, it is involved a sequence of gestures which would help to identify when a person lies. Human gestures are a powerful source of communication and represent an unconscious emotional response many times. The human being is capable of creating a number of gestures that often follow patterns given by culture, geographical location, etc. Although a group of people share a set of gestures to communicate subtle differences.

Interpret facial expressions is a task that humans perform every day automatically during the communication process either verbally or not, almost regardless of lighting conditions or perspective that is taken of the face. In contrast, the gesture recognition systems are sensitive to these conditions. Previous work mainly consisted on to discriminate gestures such as anger, joy, sadness, disgust,anger and surprise in a single frame. However some human emotions such as surprise

merges into fear,amusement, relief,anger , disgust depending upon what it was that surprised us. So determine it was a pleasant or an unpleasant surprise could be interesting.This research deals with how to recognize an emotional state, since it can be represented as states at a time t that are influenced directly by a previous state at $t-1$ therefore exists an inherent temporality.

2 Previous Work

Facial gestures are result of evolution and natural selection [2] in response to various situations. According to Ekman [3] the facial gestures are strongly related to emotions that are expressed unconsciously. Ekman [6] differs in functionality of gestures together with facial gestures to Frilund, he considered that gestures are oriented messages it might reveal the intention of adopting a behavior.

Perhaps its origin is not as important as their interpretation; Ekman have developed methodologies for the analysis of gestures from the observation, concluding that a person can be trained and be able to interpret them correctly. Further analysis show that some people can fake a suggested expression: happy, sad or angry, but they do not now how to emerge suddenly, how long to keep it, or make it disappear in that instant. Over a thousand different facial gestures are anatomically possible , but only a few have a real sense according to Ekman.

Ekman's work consisted on to interpret facial gesture correctly, one of his most important contributions [4] addresses the problem of detecting when a person lies. This analysis is very comprehensive because it includes changes in the face, body movements, tone and speed of speech. Ekman found several criteria for determining whether an emotion was being sincere or was real. In one experiment, Ekman found that over a period of 1/25 seconds or so people in some situations show a facial gesture so quickly that an untrained person is almost impossible to see. Ekman called these gestures *micro expressions* that provides leakage of a concealed emotion.*Squelched expressions* are much more common.Sometimes when an expression emerges it is interrupted and also covered with another expression. Smile is the most common cover or mask [4].

Facial Action Coding System (FACS) [5] shows how to classify muscular activity during facial gestures. This changes are called Action Unis (AU).AUs are grouped into AUs in the Upper Face (eyebrows,forehead,eyelids) and Lower Face AUs (up/down,horizontal,oblique,orbital,miscellaneous).
There is an important difference between facial expression recognition and human emotion recognition, while facial recognition classifies into abstract classes or labels of the deformations caused by facial muscle movement.Human emotion recognition are result of a variety of factors such as voice's tone, posture, gestures or even facial expressions. Basically emotion recognition is an attempt to understand a situation including its context. This research works on digital images of a face in order to interpret facial expressions and its temporal context. Figure 1 from B. Fasel and J. Luettin work [7], shows the most widely used architecture in computer science for facial gesture recognition.

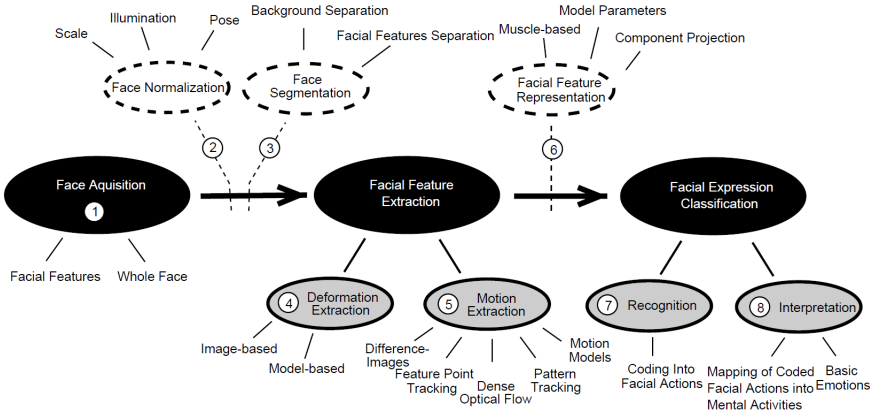

Fig. 1. Generalization of a facial gesture analysis. [7]

There is more of one approach to gesture recognition for instance using mathematical models that incorporate changes in form and lighting. This includes the features geometry of iris or nostrils, the position of these characteristics determine the location of the face. Probabilistic techniques such as the [13] are also mentioned in the literature. Other authors report the use of active contour models [8] , wavelets [10] and rule-based techniques such as FACS [5] among others .

In 2007 S.Mitra and T. Acharya [11] presented a complete revision of the state of the art concerning the recognition of gestures involving hands, arms, face, head and body in general. Based on [11] gesture recognition is divided into three categories such as:

– Hand and arm gestures.
– Head and face gestures.
– Body gestures.

Most of the tools used for gesture recognition, use statistical modeling, computer vision and/or pattern recognition. Most of the problem has been solved using statistical modeling, specially PCA, HMM, Kalman filter and even finite state machines. According to this it is concluded that there are four major approaches:

– HMM
– Filtering particles and condensation algorithm.
– Finite State Machines.
– Soft Computing and connectionist approaches.

3 Methodology

The main goal of this research is to recognize emotions through facial gestures. As it was stated in previous section there are several approaches for facial gesture recognition given a single frame.In contrast emotion recognition could be required a sequence of facial gestures. This methodology considered the following steps :

1. Image Acquisition
2. Feature Extraction
3. Gesture Recognition
4. Emotion Recognition

In general the process begins by capturing a **video sequence**, which is segmented into *frames* ,after video has been segmented most significant features on face are marked. Every frame in the sequence is labeled as one of the following gestures: neutral, anger,surprise, joy, disgust, sadness , fear combined with the intensity of the emotion (low,mid,high). Therefore exists 19 possible labels since neutral have only one intensity. After gesture recognition the labeled sequence is evaluated into a first order HMM. Figure 2 shows the proposed methodology.

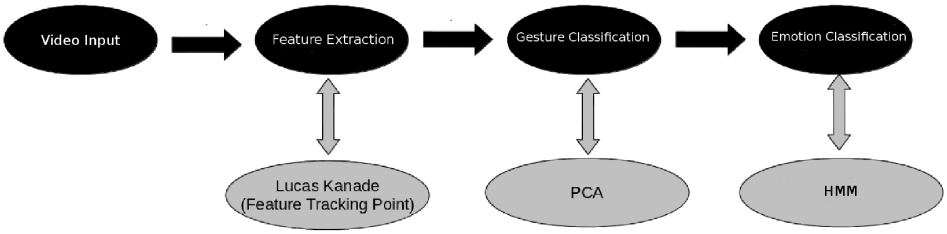

Fig. 2. Proposed Methodology.

Feature Extraction. Let I, J be gray scaled images ,$\boldsymbol{u} = [u_x, u_y]$ where $\boldsymbol{u}$ is an image point on the first image I and (u_x, u_y) are the two pixel coordinates. The goal of feature tracking is to find the location $\boldsymbol{v} = \boldsymbol{u} + \boldsymbol{d} = [u_x + d_x, u_y + dy]$. Lucas-Kanade's algorithm [9] with Bouguet's improvement [1] was used to track and mark must significant features on face , even in areas of low contrast. Since this algorithm is sensitive to changes in lighting face normalization results a reasonable approach for reducing variations.There are twelve facial features without contrast or texture problems as it is showed in figure 3. These points are tracked and marked in every frame when a human face is detected. Since the background is removed in every frame, face detection is used to crop and re size

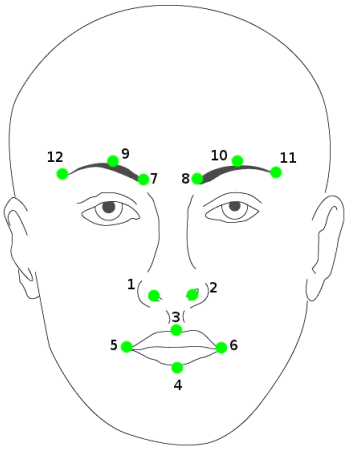

Fig. 3. Feature points. *1.-Left nostril. 2.-Right nostril. 3.-Upper lip. 4.-Lower Lip. 5.-Left corner. 6.-Right corner. 7.-Outer of left eyebrow. 8.-Outer of right eyebrow. 9.-Center of left eyebrow . 10.-Center of right eyebrow. 11.-Inner of left eyebrow. 12.-Inner of right eyebrow.*

the region of interest (ROI). As result of this process every frame contains just a marked face and have the same size.

Gesture Recognition. Let be $A = \{x|x_i$ is a frame of the video sequence to analyze$\}$, and $B = \{$*neutral,low anger,mid anger,high anger, low disgust,mid disgust,high disgust,low fear,mid fear,high fear,low joy,mid joy,high joy,low sadness,mid sadness,high sadness,low surprise,mid surprise,high surprise*$\}$ a set of emotions respectively. Every frame is subject to Principal Component Analysis (PCA) in order to reduce dimensionality of the feature space by considering the first 30 components. Given a training set of N images $C = [I_1, I_2, ..., I_N]$ where each I_J is an image representing one of the six basic emotions, PCA deals with the weights of image's training set then Euclidean and Mahalanobis distances are obtained to determined the best match . Mahalanobis distance is given by $((\boldsymbol{x} - \boldsymbol{y_i})'C^{-1}(\boldsymbol{x} - \boldsymbol{y_i}))^{1/2}$ where $\boldsymbol{x}$ and $\boldsymbol{y_i}$ are elements from the testing and training set respectively. The covariance matrix C is calculated between the tested frame and every frame in the training set.

Emotion Recognition. Let be a set of T emotional hidden states $\omega = \{\omega_1, \omega_2, \omega_3, ..., \omega_T\}$. Given a visible state sequence $V = \{x|x \in B\}$ we determine the probability that this particular sequence was generated by a particular model θ. Determine the most likely sequence of hidden states ω led to those observations that defines which emotional state is. Given the coarse structure of a model *(the number of states and the number of visible states)* but not the probabilities a_{ij} and b_{jk} and a set of training observations of visible symbols *(video examples)*, we determine a_{ij} and b_{jk} probabilities using the Baum-Welch algorithm. For

instance surprise is the briefest emotion and most of the times merges to pleasant or unpleasant gestures depending upon what it was that surprised us. Such an ergodic HMM for the emotion surprise have the states *neutral, pleasant surprise, unpleasant surprise* as shown in Figure 4. The probability that the model θ

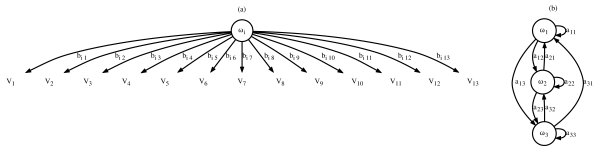

Fig. 4. HMM for the surprise emotion.(a) Every hidden state ω_i emits surprise,joy,disgust or anger with three different intensities.Neutral emission has only one intensity **(b)** Since surprise is never expected, *neutral* as hidden state is included in this model besides pleasant and unpleasant surprise.

produces this sequence V^T of visible states is the probability that this video contains a surprised expression regardless it was pleasant or unpleasant. Given this sequence of visible states V^T , the problem is to find the most probable sequence of hidden states. Therefore it is concluded whether it was a pleasant or unpleasant surprise. Table 1 shows a simple example for classifying an emotional state using the model shown in Figure 4.

States / Emissions	Neutral	Pleasant Surprise	Unpleasant Surprise
Neutral	0.4	0.025	0.025
High Surprise	0.00671875	0.02203125	0.02203125
High Joy	5.1806640625E-4	0.003206542	0.0

Table 1. Forward algorithm and its probabilities for $V^3 = \{V_1, V_4, V_7\}$.This table shows in the last iteration that pleasure surprise is the must probable state for the sequence V^3.

4 Results

The training set for the gesture recognition step has 900 faces depicting a neutral expression and the six basic emotions: anger, disgust, fear, joy, sadness, surprise. Some images are from the University of Stirling face database [14] and the Japanese Female Facial Expression **(JAFFE)** database [12]. Gender and race distribution is shown in Table 2. A classifier for each of the six fundamental facial gestures (joy, sad, angry, fear, surprise, disgust) and one for neutral gesture were

Ethnicity	Men	Women
American (U.S. citizen)	10	13
Scottish	32	34
Japanese	0	10
Iranian	0	35
Mexican	25	15

Table 2. Gender and race distribution. Training set for gesture recognition step

trained to build a facial gesture classifier. The normalized distance is used as a deciding factor.

4.1 Detection of pleasant/unpleasant surprise gestures

A HMM was trained with 20 videos depicting a pleasant or unpleasant surprise using the Baum-Welch algorithm. In order to induce surprise at this experiment several videos showing suddenly a scary images were used. Reaction differs for each person depending its emotional state. The entire set of videos were recorded with a frame rate of 30 frames per second but for this analysis 15 frames per second are only used in order to reduce the length of the emission sequences V^T without losing capability for capture even micro-expressions. It were calculated the most probable sequence of hidden states for 15 testing videos. Table 3 shows the most probable state reached in last observation which is used as a deciding factor where *P=Pleasant,U=Unpleasant,O=Other*. Results are acceptable since the classificator fails only in video 12 and 13. Video 12 shows a surprise expression but it not merges into another emotion, therefore this test shows a desirable response. Otherwise the video 13 starts with an unpleasant surprise expression omitting a neutral expression at the begining.

Repeated states appears in every video sequence since human face holds and expression at least for $\frac{1}{2}$ seconds (micro-expressions) some post-processing may be applied and just get the sequence somewhat independent of variations in rate. Convert the sequence $\{\omega_1, \omega_1, \omega_2, \omega_3, \omega_3, \omega_1\}$ to $\{\omega_1, \omega_2, \omega_3, \omega_1\}$, seems to be appropriate for emotion recognition in order to reduce the length of this sequence that can reach 30 symbols per second.

4.2 Detection of lie gestures

Many factors such a facial expression, tone of voice, slip of the tongue, or certain gestures could leak our true feelings. Exclusively facial expressions will be used for other reasons which are beyond the scope of the present research. Micro and squelched expressions are hints to discover if someone is lying but they are not conclusive. The first issue to consider in estimating whether or not there will be any clues to deceit is whether or not the lie involves sense of emotions when is happening a lie. Based on we have proposed a method for emotion recognition

States / Video	$P(Neutral\|T)$	$P(Pleasant\|T)$	$P(Unpleasant\|T)$	**Test**
1	5.180664E-4	**0.003206**	0.0	P
2	5.180664E-4	0.0	**0.002137**	U
3	5.180664E-4	0.0	**0.001068**	U
4	**0.002781**	4.199218E-4	0.001705	O
5	**0.003882**	4.199218E-4	0.002990	O
6	**0.004984**	0.004275	4.199218E-4	O
7	3.115234E-4	**0.001760**	0.0	P
8	3.115234E-4	0.0	**0.001173**	U
9	3.115234E-4	0.0	**5.869140E-4**	U
10	**0.005011**	0.002767	8.398437E-4	O
11	8.197021E-5	**5.139770E-4**	0.0	P
12	**0.008289**	0.004275	0.004275	P/U
13	**7.189025E-5**	5.123138E-6	6.508712E-5	U
14	2.426757E-4	0.0	**7.475585E-4**	O
15	7.421875E-4	**0.003632**	0.0	P

Table 3. Reached states after decoding algorithm. The decoding algorithm finds at each time step t the state that has the highest probability of having come from the previous step and generated the observed visible state V_k. (P=*Pleasant*, U=*Unpleasant*, O=*Other*)

into a temporal context our approach for evaluate lies is based on micro and squelched expressions. Figure 5 shows a model for representing a moment of *lying* and its transitions between hidden states before any training. Micro expressions provide a full expression of the concealed emotion, but so quickly that they are usually missed. The presence of any micro expression depicting any emotion is a lying sign further a smile is the most common cover or mask.

The experiment for gathering videos is based on Ekman's *nurse experiment*, that offers the chance to test out and practice the ability to control expression of your feelings. It is basically consisted on watch an unpleasant film meanwhile a person describes the film as pleasant. The same video was used to collect testing videos. Finally a HMM was trained with 15 videos using the Baum-Welch algorithm. Reaction differs for each person depending its emotional state. Figure 6 shows the difference between each hidden state probability at the last emission for some video samples. Microexpressions were the most difficult emotion to recognize.Since microexpressions are the most unusual gestures only 5 samples were obtained.

5 Conclusions

This paper describes and tests a probabilistic model for automatic emotion interpretation by using a set of observations obtained from real time situations. The use of a first order HMM is justified by the inherent temporality in human emotions. In this paper was introduced an approach for analyze human emotions

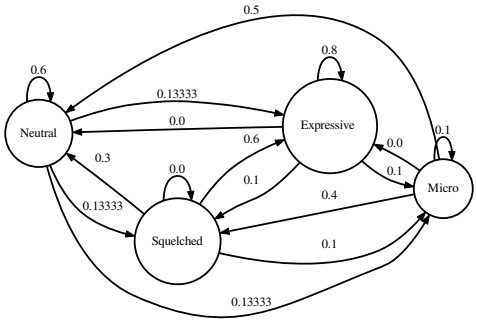

Fig. 5. HMM for lie detection. Each hidden state ω_i can emit one of the 19 available visible states from neutral to any of the six basic emotions and its intensities.

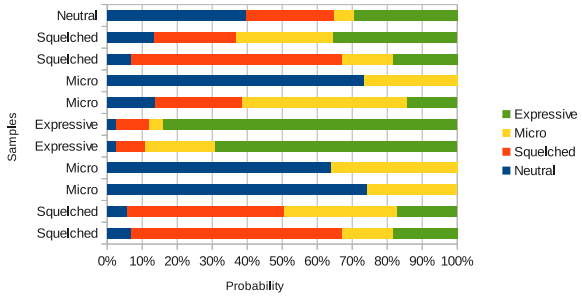

Fig. 6. Recognition results of the proposed lie detector model.

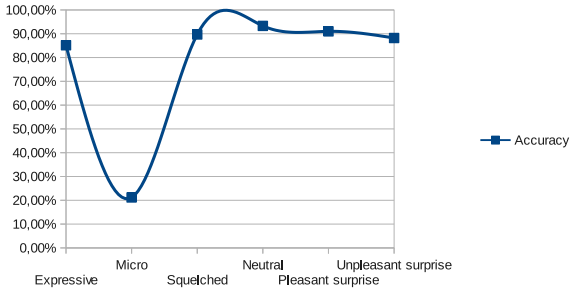

Fig. 7. Average accuracy for the proposed methodology.

in a temporal context. Preliminary results shows that this approach allows to analyze several human emotions in a temporal context, not only the two models proposed. Future work will improve the methodology and its implementation, introducing other variables such as voice tone, arm gestures, heart rate fluctuations and other sources of information like activity in social networks which are used to show several emotions.

I't is considered that this approach can be applied in differents escenarios, such as airport security where a shorts ans specific questions can trigger small change in people's face; other area of interest could be for human resources departments in order to detect when a person is lying , based on the fact that some changes develop slowly, because of these phenomena are rarely perceptible over a short space of time.

References

1. Bouguet, J.Y.: Pyramidal implementation of the lucas kanade feature tracker description of the algorithm. `http://robots.stanford.edu/cs223b04/algo_tracking.pdf` (2000)
2. Darwin, C., Ekman, P., Prodger, P.: The expression of the emotions in man and animals. Oxford University Press, USA (2002)
3. Ekman, P.: Cross-cultural studies of facial expression. Darwin and facial expression: A century of research in review pp. 169–222 (1973)
4. Ekman, P.: Telling lies: Clues to deceit in the marketplace, politics, and marriage. WW Norton & Company (2009)
5. Ekman, P., Friesen, W.: Facial action coding system, chap. IV. Consulting Psychologists Press, Stanford University, Palo Alto (1977)
6. Ekman, P., Friesen, W.: Unmasking the face: A guide to recognizing emotions from facial clues. Ishk (2003)
7. Fasel, B., Luettin, J.: Automatic facial expression analysis: a survey. Pattern Recognition 36(1), 259–275 (2003)
8. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. International journal of computer vision 1(4), 321–331 (1988)
9. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision (ijcai). In: Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI '81). pp. 674–679 (April 1981)
10. Lyons, M., Budynek, J., Akamatsu, S.: Automatic classification of single facial images. Pattern Analysis and Machine Intelligence, IEEE Transactions on 21(12), 1357–1362 (1999)
11. Mitra, S., Acharya, T.: Gesture recognition: A survey. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 37(3), 311–324 (2007)
12. Miyuki Kamachi, M.L., Gyoba, J.: The japanese female facial expression database. `http://www.kasrl.org/jaffe.html` (2012), [Online; accessed 22-November-2012]
13. Samaria, F., Young, S.: Hmm-based architecture for face identification. Image and vision computing 12(8), 537–543 (1994)
14. University, S.: Psychological image collection at stirling. `http://pics.psych.stir.ac.uk/` (2012), [Online; accessed 22-November-2012]