

# Emotional Recognition and towards Context based Decision

Ayesha Butalia  
MIT College of Engg  
Pune, India

Dr. A.K. Ramani  
Devi Ahilya University,  
Indore, India

Dr. Parag Kulkarni  
Capsilon,  
Pune, India

## ABSTRACT

Non-verbal communication may be used to enhance verbal communication or even provide developers with an alternative for communicating information.

Emotion or Gesture recognition is been highlighted in the area of Artificial Intelligence and advanced machine learning. Emotion or gesture is an important feature for an intelligent Human Computer Interaction. This paper basically is a literature survey paper which reveals with the research work already dealt with in this area. Facial expression has been concluded as the most important part involved in it. Even Facial features are also distinguished out of which eyes and mouth is probably more prominent. Neural networks are the widely used. Approaches towards Rough Fuzzy definition can be probably resolve the complexity. Context based recognition can be added so as to resolve the ambiguity involved in different scenarios.

## Keywords

Facial expression, emotion recognition, rough sets, fuzzy sets, neural networks, context.

## 1. INTRODUCTION

Emotion recognition is very important for human-computer intelligent interaction. Facial expressions, gestures, and body postures can portray emotions in a non-verbal way. These methods are frequently employed by actors in theatrical plays or in movies, or even by virtual characters such as those found in computer games, animated storybooks, and website e-assistants. Signals for emotion expressions (“cues”), such as a raised fist and narrowing of the eyes, substantially influence the viewers’ assumptions on the emotional state of the person portraying it. To enable humans to recognize emotions of virtual characters, the characters’ cues must be portrayed according to the human counterparts. Previous research has shown that emotions can be effectively portrayed through non-verbal means [Atkinson et al. 2004; Coulson 2004; Ekman 2003].

Emotion Recognition is generally performed on facial or audio information by artificial neural network, fuzzy set, support vector machine, hidden Markov model, and so forth. Although some progress has already been made in emotion recognition, several unsolved issues still exist. For example, it is still an open problem which features are the most important for emotion recognition. It is a subject that was seldom studied in computer science. However, related research works have been conducted in cognitive psychology. In recent years, there has been a growing interest in improving all aspects of the interactions between humans and computers. It is argued that to truly achieve effective human-computer intelligent interaction (HCII), there is a

requirement for computers to be able to interact naturally with users, similarly to the way human-human interaction. HCII is becoming more and more important in such applications as smart home, smart office, and virtual reality, and it will be popular in all aspects of daily life in the future. To achieve the purpose of HCII, it is essential for computers to recognize human emotion and to give a suitable feedback. Consequently, emotion recognition attracts significant attention in both industry and academia. There are several research works in this field in recent years and some successful products such as AIBO, the popular robot dog produced by Sony.

In the first section of the study, we first look at existing systems for synthesizing cues for facial expressions, gestures and body postures. This is followed by examining the emotion recognition problems that arise from utilizing the various systems. Lastly, the systems are integrated together and the implications that arise from the integration are analyzed.

The second section deals with the comparison of Roles of Postures, Facial and Gestures in Emotion Recognition. The study showed that hand gestures aided in the emotion recognition rate for postures which others [Coulson 2004] had previously assumed as unimportant. Additionally, it was discovered that the emotion recognition rate using gestures can be greatly improved when emblematic actions are combined with functional actions. This study also confirmed our assumption that an integrated system covering facial expression, gesture and body postures is capable of enhancing the emotion recognition rate beyond the rates of the single systems.

Usually, emotion recognition is studied by the methods of artificial neural network (ANN), fuzzy set, support vector machine (SVM), hidden Markov model (HMM), and based on the facial or audio features, and the recognition rate often arrives at 64% to 98% [1–3]. In the third section, the comparison of Rough Set theory, Adaptive Neuro - Fuzzy Inference Systems (ANFIS) and Rough – ANFIS approach is analyzed.

Lastly, the section describes the context approach towards facial recognition.

## 2. FUNDAMENTALS OF EMOTION EXPRESSION

This section looks at how cues are able to produce emotions. In addition, it provides an overview of how certain factors may affect emotion recognition in a 3D virtual agent.

### 2.1 Cues

Cues are non-verbal signals involving either the movement/positioning of individualized parts of the body or the

movement/positioning of a group of body parts in concert with each other [Ekman 1978]. Cues such as a raised fist and narrowing of the eyes can indicate that the individual is angry. Movement/positioning classes like facial expressions, body postures, and gestures can involve one or more of these cues, which people (subconsciously) use for interpreting the emotional state.

## 2.2 Facial Expressions:

Facial expressions involve facial cues that are displayed using body parts from the head region (e.g., eyebrows, mouth, lips). Common facial expressions such as the raising of the lips (facial cue) as part of a smile (facial expression) is interpreted by others to be a display of emotion of the actor; happiness in this example [Ekman 1978].

## 2.3 Body Posture:

Body cues involved in body postures are displayed using body parts such as the torso, arms and legs. They are another component of non-verbal emotion expression. For example, the clenching of a fist and raising it to appear like the actor is trying to attack someone is usually interpreted by others as a display of anger [Ekman 2003].

## 2.4 Gestures and Actions:

Gestures are actions/movements of body parts and they are another component of non-verbal communication of emotion. For example, a high frequency gesture such as jumping up and down very quickly can be interpreted by others to be a sign of happiness [Raouzaoui et al. 2004].

## 2.5 Factors Affecting Emotion Recognition:

Although the exact factors which can influence the interpretation of emotion have not yet been thoroughly researched upon, four factors have recently surfaced based on current experiments and research. They are gender, job status, culture, and age.

Men and women express emotions differently [Brody and Hall 1992] in terms of the frequencies of occurrence (men often experience anger more often than women). It was also proven that recognition of ambiguous facial expressions is influenced by the gender of the person performing it [Condry 1976; Devine et al. 2000] whereby “masculine” emotions (e.g., anger) are assigned to men while “feminine” emotions (e.g., happiness) are assigned to women. As such, there is a need to be mindful of these gender stereotypes

when trying to synthesize emotions.

Stereotypes of job status are known to exist too [Algoe et al. 2000]. For example, managers are often associated with “masculine” emotions and character traits while nurses are associated with “feminine” emotions and character traits. If the virtual agent is assigned human jobs (usually identified by the type of uniform they are wearing), ambiguous emotion expressions may lead others to wrongly assign “masculine” and “feminine” emotions to it. Culture can also affect the interpretation of emotions [Bianchi-Berthouze et al. 2006]. It was discovered that the Japanese are less animated in their body expressions for emotion than the Sri Lankans leading to the same emotion being read differently.

Lastly, there is neurological evidence to suggest that age can affect the interpretation of emotions. It was shown that people in

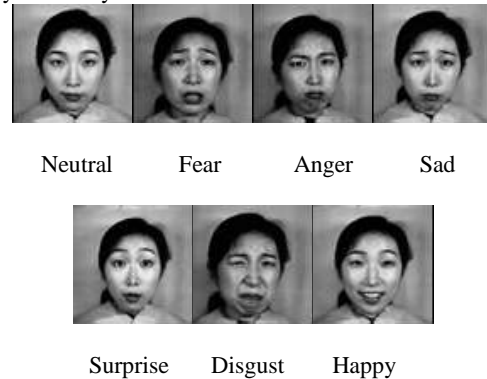
the 60-80 years old age group tend to suffer from emotion processing impairments and therefore require stronger or more cues to be displayed before being able to associate an emotion.

## 2.6 Emotion Blending and Transition:

Human beings are capable of feeling multiple emotions simultaneously [Ekman 2003]. These emotions may transition/morph in time from one state of emotion to another (e.g., a loud noise may suddenly cause a passerby to feel surprise momentarily, which might later transition into a feeling of fear if the passerby feels that his/her life is in eminent danger), or they may also be displayed at the same point in time (e.g., the loss of a loved one in a car accident may cause a person to feel both angry and upset at the same time). Emotion blending is the mechanism by which multiple emotional expressions are altered or combined simultaneously to convey more subtle information about the performer. Unfortunately, such a process is a complicated one and has not yet been well understood and researched by behavioral psychologists and animators.
















## 3. CONCEPTS FOR REPRESENTING AND MODELING FACIAL EXPRESSIONS, POSTURES, AND GESTURES

This section looks at the approaches taken by others to represent and model emotion expression. It also describes the approach taken by this study.



**Figure 1. 6 Basic Emotions and Neutral Expression**

**Table 1a. Some AU and their associated facial change obtained from Ekman’s study [Ekman 1978].**

AU1  Inner brow raiser	AU2  Outer brow raiser	AU4  Brow Lowerer	AU5  Upper lid raiser	AU6  Cheek raiser
AU7  Lid tighten	AU9  Nose wrinkle	AU12  Lip corner puller	AU15  Lip corner depressor	AU17  Chin raiser
AU23  Lip tighten	AU24  Lip presser	AU25  Lips part	AU26  Jaw drop	AU27  Mouth stretch

**Table 1b. Table of the six basic emotions and the AUs involved.**

Basic Expressions	Involved Action Units
Surprise	AU1, 2, 5, 15, 16, 20, 26
Fear	AU1, 2, 4, 5, 15, 20, 26
Disgust	AU2, 4, 9, 15, 17
Anger	AU2, 4, 7, 9, 10, 20, 26
Happiness	AU1, 6, 12, 14
Sadness	AU1, 4, 15, 23

### 3.1 Representing and Modeling Postures

There exist a variety of sources which offer more or less detailed descriptions of emotional postures [Birdwhistell 1975; Boone and Cunningham 2001; Darwin 1872]. For instance, in the descriptions put forward by these authors, anger is described as involving a jutting chin, angular body shape, forward weight transfer, chest out and angled forwards, and a bowed head. Unfortunately, prior to Coulson's study [Coulson 2004], no formal research has been done to quantify the anatomical features which produce the emotional posture (i.e., posture was mostly descriptively documented).

Coulson found that the anatomical descriptions obtained from the studies mentioned earlier could be translated into joint rotations, which he then attempted to quantify via testing on human volunteers. The approach taken by this study to modeling postures relies on Coulson's findings for the angle representation of each emotion as it is the only study which quantifies the respective joint angles.

Gestures in this study are modeled by animating the virtual agent since gestures are essentially non-static displays of emotion. As there is a relative paucity of studies on dynamic emotion gestures [Atkinson et al. 2004], the approach taken by this study relies on actors' knowledge of gestures.

Raouzaoui et al. (2004) and Atkinson et al. (2004) formulated a short table of emotions that depict general hand and head gestures for each emotion. This study relies upon those data to provide a basic framework for generating gestures as it is also compatible with the joint-angle system used for modeling of posture.

**Table 2. Table of gestures extracted from Raouzaoui's study [Raouzaoui et al. 2004].**

Emotion	Gesture Class
Joy	Hand clapping-high frequency
Sadness	Hands over the head-posture
Anger	Lift of the hand-high speed
	Italianate gestures
Fear	Hands over the head-gesture
	Italianate gestures
Disgust	Lift of the hand-low speed
	Hand clapping-low-frequency
Surprise	Hands over the head gesture

**Table 3. Recognition rate from integrating posture and facial expression.**

Emotion(Postures)	No. of correct recognition	Emotion Recognition rate
Happy	16	100.00%

Anger	16	100.00%
Sad	16	100.00%
Surprise	9	56.25%
Fear	14	87.50%
Disgust	1	6.25%

**Table 4. Recognition rate from facial expressions alone.**

Emotion(Postures)	No. of correct recognition	Emotion Recognition rate
Happy	16	100.00%
Anger	16	100.00%
Sad	13	81.25%
Surprise	10	62.50%
Fear	11	68.75%
Disgust	2	12.50%

**Table 5: Recognition rate from integrating gesture, posture, and facial expression.**

Emotion(Postures)	No. of correct recognition	Emotion Recognition rate
Happy	16	100.00%
Anger	16	100.00%
Sad	16	100.00%
Surprise	4	25.00%
Fear	12	75.00%
Disgust	2	12.50%

*Conclusion: Infact Gesture lowered the recognition rate when added to Posture and Facial features. There was not much difference in the recognition through facial rather than from facial and posture together. Infact posture recognition would add much complexity in real life recognition. A single error to it could make lot of difference and consuming time also.*

## 4. CLASSIFICATION BASED ON RS, ANFIS, RS-ANFIS FOR FACIAL EXPRESSION:

Models and automated systems have been created to recognize the emotional states from facial expressions. The leading method Facial Action Coding System [23], for measuring facial movements in behavioral science was developed by Ekman and Friesen in 1977. Other methods such as electromyography, which directly measures the electrical signals generated by the facial muscles and deducing the facial behavior from it, are both obtrusive non-comprehensive. According to the survey [7], FACS is the leading method for measuring facial expression in behavioral science. It uses 46 defined Action Units to correspond into each independent motion of the face. However this method takes over 100 hours of training to achieve minimal competency for a human expert [8]. Faster automation approaches, such as measurement of facial motion through optic flow [9, 10] and analysis of surface textures based on principal component analysis (PCA) [11]. Newer techniques include using Gabor wavelets [12], linear discriminant analysis [13], local feature analysis [14], and independent component analysis [15]. The techniques are

benchmarked [8] and best classification accuracy of about 95% for the recognition of the twelve facial actions, was obtained using Gabor filter representation. Human experts and naïve human tester were benchmarked as well; scored about 94% and 78% respectively, and experiments were supported by Zhang et. al. [16].

Most of these systems use a set of feature vectors to represent facial images, without describing the relationship between the feature vectors. A method for emotion recognition by transforming the feature vector data into tree structure representation, which encodes the feature relationship information among the face features was then proposed [Automated knowledge engg]. Sixty Localized Gabor Features (LGF) and one Global Gabor Feature are obtained as a feature vector and transforming them into a Facial Emotion Tree Structure (FETS) representation. Tsoi [17] proposed using tree structures to preserve and make use of these relationships and processing them by specific machine learning models [1, 18-20]. Cho and Wong proposed using Gabor features in tree structure representation for face recognition with achieving high accuracy rate [1]. Gabor Feature extraction makes use of Gabor wavelets, which capture the properties of spatial localization, orientation selectivity, spatial frequency selectivity, and quadrature phase relationship, seem to be a good approximation to filter response profiles encountered experimentally in cortical neurons. A probabilistic based recursive neural network is proposed for classification of the FETS in this paper. This method is benchmarked against Support Vector Machines (SVM) [21], K nearest neighbors (KNN) [22], Naïve Bayes algorithm [23] where the flat vector representations were used in the recognition experiments. QuadTree tree structure processed using our probabilistic recursive neural network is also benchmarked. We made use of the Japanese Female Facial Expression (JAFPE) [24] database to illustrate the performance of the recognition system. Our proposed emotion recognition system is illustrated in Figure 2. This system constitutes the low-level feature extraction and the high-level tree structure representation for emotion recognition. The details of the major components in the proposed system will be described in the following sections.

Although some progress has been made in emotion recognition, several unsolved issues still exist. For example, it is still an open problem which features are the most important for emotion recognition. It is a subject that was seldom studied in computer science. However, related research works have been conducted in cognitive psychology [4–6]. Affective computing is becoming an important research area in intelligent computing technology. Furthermore, emotion recognition is one of the hot topics in affective computing. It is usually studied based on facial and audio information with technologies such as ANN, fuzzy set, SVM, HMM, etc. Many different facial and acoustic features are considered in emotion recognition by researchers.

## 4.1 Rough Sets (RS):

Rough set (RS) is a valid mathematical theory for dealing with imprecise, uncertain, and vague information; it was developed by Professor Pawlak in 1980s [28, 29]. RS has been successfully used in many domains such as machine learning, pattern recognition, intelligent data analyzing, and control algorithm acquiring [30–32]. The most advantage of RS is its

great ability of attribute reduction (knowledge reduction, feature selection).

## 4.2 The problem here is:

a) Rough Sets are not efficient in dealing with data sets made up of continuous attribute values. Various quantization or Discretization techniques exist to address this issue, but no global technique applicable to a variety of data sets exists.

b) Rough Sets are purely rule based and due to the limited number of rules that such a system has, it fails to efficiently evaluate new cases or unseen situations. A Rough system can thus be said to be very fragile at its boundaries.

c) Adaptive Neuro – Fuzzy Inference Systems inherently are disadvantaged in modeling systems which have a large number of inputs or outputs. Though time efficient, the amount of pre – processing required before an ANFIS model can be generated, can be enormous depending upon the given problem or situation.

## 4.3 Adaptive Neuro - Fuzzy Inference Systems (ANFIS):

The ANFIS proposed by Jang [24] can be described as adaptive networks which are similar to Fuzzy Inference Systems functionally. They also have outlined a methodology to help decompose the parameter set (of the adaptive network nodes), so as to help implement the Hybrid Learning algorithm within the said systems. A successful attempt to represent the Sugeno and Tsukamoto Fuzzy models through ANFIS has been undertaken and further, it is also observed that the Radial Basis Function Network when subjected to certain constraints is equivalent to the ANFIS functionality. The idea of such an approach towards modeling systems is to interpret fuzzy rules in terms of a neural network. The fuzzy sets can be representative of the weights and the input – output functions along with the rules are representative of the neurons of a network. The (hybrid) learning as well is implemented as in a connectionist system i.e. the system learns by continuously modifying and adapting the neural structure and the parameters. The advantage of such a system is that the learning can be interpreted from perspectives of both neural and fuzzy systems. And more importantly such a system enables viewing the problem solution in a linguistic fashion Architecture: In the simplest terms, the structure of an ANFIS consists of a first layer where the inputs are mapped to their respective and relevant input membership functions. These membership functions taken it on to the rules layer further and then onto the output membership functions and finally to the output characteristic function which computationally produce the final single valued output. The following synopsis about the ANFIS architecture has been adopted from [25]. Consider a set of standard Sugeno style if then rules i.e.

Rule1: if x is A and y is B, then  $F1 = px + qy + r$  and;

Rule2: if x is C and y is D, then  $F2 = sx + ty + u$ .

Modeling the above using an ANFIS would result in a 5 layer network (excluding the input and the output layers) with the following sequential operations being accomplished,

Layer 1: The objective function of the nodes (all of the nodes in the same layer are characterized by the same objective function) is basically to assign a relevant fuzzy membership function to the said inputs i.e. ( )

$$O_{i, i(1-2)} = \mu_{A, i(x)}$$

$$O_{i,i(3-4)} = \mu_{B,i-2(y)} \quad (1)$$

Moreover, initially a user chosen parameterized general membership function is also adopted and the parameters of  $\mu_{Ai}$ ,  $\mu_{Bi}$  are assigned some random values to start off with.

Layer 2: The outputs of the individual preceding layers are multiplied to retrieve the first set of real parameters which are called the premise parameters. i.e.

$$O_{2,i} = w_i = \mu_{Ai(x)} \mu_{Bi(y)} \quad (2)$$

Layer 3: The outputs of this layer are normalized or weighted weights i.e.

$$O_{3,i} = \bar{w}_i = \frac{w_i(1-2)}{w_1 + w_2} \quad (3)$$

Layer 4: Unlike the nodes from layers 2 and 3, layer 4 nodes are adaptive with the following output, (the final output individual vector components are computed). The parameters of p, q, r, s, t, u, are all called Consequent parameters.

$$O_{4,i} = \bar{w}_i F_i \quad (4)$$

Layer 5: Usually composed of a single static node, the operation of summing up the incoming individual vector components of the final single valued output.

$$O_{5,i} = \sum_i \bar{w}_i F_i \quad (5)$$

As described earlier, the ANFIS network shall comprise of a minimum of 5 layers with at least two of them being adaptive in nature. This structure generates two sets of varied parameters to solve for, namely the premise and the consequent parameters. These parameters and their values are generated appropriately during the learning process (as described in the following section). Hybrid Learning: Conventionally the Gradient optimization methods or back propagation techniques have been used to identify and optimize the various nodal parameters associated with adaptive networks. But Jang [24] proposed a hybrid learning rule approach to identify the parameters. This method involves combining the back propagation steepest descent and least squares method for a faster identification. The process of continuous parameter update can be achieved by two methods i.e. off – line or batch learning which involves updating the parameters iteratively at the end of all training data pair runs and on- line learning wherein the parameters are updated posthumously after every layered data run.

Batch Learning: Consider the following output function for an adaptive network

$$O = f(i, S) \quad (6)$$

where ‘O’ is the output function, ‘i’ inputs set and ‘S’ is the parameter set.

Now if there exists a function ‘h’ such that applying it to Eq

(1) would generate a new function linear in certain parameters of the set ‘S’ i.e.

$$S = S1 \oplus S2 \quad (7)$$

therefore applying ‘h’ to the objective function,

$$h(O) = h \circ f(i, S) \quad (8)$$

where ‘h’ is linear in  $S2$  parameters.

Assuming a set of values for the parameter set  $S2$ . We can generate  $P$  number of linear equations in  $S2$  parameters.

Representing the same in a matrix we get,

$$A\theta = y \quad (9)$$

and by the Least Squares Estimator the solution to the above equation can be given as,

$$\theta = (A^T A)^{-1} A^T y \quad (10)$$

The above described procedure is implemented within forward pass of the network. Once all of the  $S2$  parameters are identified, then the backward pass is initiated. The errors are calculated and the gradient vector is determined. At the end of all training pairs, in the backward pass, the parameters in  $S1$  are updated by steepest descent method.

$$f = (\bar{w}_1 x)p + (\bar{w}_1 y)q + (\bar{w}_1 r) + (\bar{w}_2 x)s + (\bar{w}_2 y)t + (\bar{w}_2)u \quad (11)$$

$\bar{w}_1$  and  $\bar{w}_2$  are the premise parameters and p, q, r, s, t, and u are the consequents.

Data sets used: Following is a brief outline of the data sets which have been used, and a little bit more about them individually.

a) Breast Cancer:

Source: Orange Data Mining [26]

Attributes (9 in all): recurrence, age, menopause, Tumor

Size, inv nodes, node caps, Deg

Malig, breast and breast quad.

Missing Values: No

Data Objects: 190 (one per patient)

Decision Classes: 2

b) Lung Cancer:

Source: Orange Data Mining [26]

Attributes (56 in all): Attributes 1 to 56.

Missing Values: Yes

Data Objects: 32 (one per patient)

Decision Classes: 2

Data Reliability: Though emphasis was laid on the conceptual implementation in this work, an effort has been made to use reliable data. The data is a common testing data available for cost free downloads from various servers. The UCI data repository and the Orange Data Mining Repository are just a couple examples. This data is widely used by AI researchers in order to establish a performance scale for their respective techniques.

**Table 6a. Breast Cancer**

Attribute	Rough	ANFIS	Rough-ANFIS
Number of Rules	152	512	64
Error Rate	0	0.175	0.08
Attributes in use	6 out of 9	9 out of 9	6 out of 9
Reduct Cardinality	1(100%)	1(100%)	-NA-

**Table 6b. Lung Cancer**

Attribute	Rough	Rough-ANFIS
Number of Rules	32	27
Error Rate	0	0.1 (per unity)
Attributes in use	3 out of 56	3 out of 56
Reduct Cardinality	1(100%)	-NA-

Neuro Fuzzy systems: Neural networks exhibit lack of interpretability and Fuzzy systems lack the capability of effective learning. Thus Neuro Fuzzy systems present the learning capability of the neural networks and the fuzzy interpretation skills both in one tool. Moreover with respect to dynamic systems, neural networks provide the requisite skills for knowledge acquisition through learning and the fuzzy systems top it up with their ability to automatically approximate the knowledge bases for non – deterministic events. Due to the presence of these characteristics belonging to both the techniques, Neuro Fuzzy systems are widely used in machine learning applications. ANFIS is one of the kinds.

Rough Sets framework was chosen predominantly for its complimentary behavior when amalgamated with other approaches. Rough Sets Theory (RST) is well equipped to deal effectively with *imprecise, noisy and missing information* [27]. Unlike in other approaches, RST effectively sets the accuracy and precision value as per the requirement of the user for various classificatory processes. Further the concept of indiscernibility relations coupled with the concept of Reducts provides *discernibility – preserving elimination of irrelevant information* [27]. Also the issues arising due to multi and partial memberships of the objects in various sets have been reasonably addressed using the RST. And similar to the ES answering module, Rough Set models can be made capable of providing a description of analysis which led to the final decision. Fuzzy set theory and Rough Set theory are complimentary and not competitive. Amalgamation of the said techniques or theories would result in constructive determination since each of them refers to different aspects of imprecision i.e. Fuzzy set theory represents imprecision in the form of a partial membership whereas Rough sets accommodate the imprecision in the form of indiscernibility relations and the set upper and lower approximations.

#### 4.4 Facial Feature Recognition: Parametric Feature Representation:

The contours of the facial features, generated by the facial feature detection method (Fig. 1), are utilized for further analysis of shown facial gestures. First, we carry out feature points' extraction under two assumptions: (1) the face images are non-

occluded and in frontal view, and (2) the first frame is in a neutral expression. We extract 22 fiducial points: 19 are extracted as vertices or apices of the contours of the facial features (Fig. 2), 2 represent the centers of the eyes (points X and Y), and 1 represents the middle point between the nostrils (point C). We assign a certainty factor to each of the extracted points, based on an "intra-solution consistency check". For example, the fiducial points of the right eye are assigned a certainty factor  $CF \in [0, 1]$  based upon the calculated deviation of the actually detected inner corner  $B_{current}$  from the pertinent point  $B_{neutral}$  localized in the first frame of the input sequence. The functional form of this mapping is:

$CF = \text{sigm}(d(B_{current}, B_{neutral}); 1, 4, 10)$  where  $d(p_1, p_2)$  is the block distance between points  $p_1$  and  $p_2$  (i.e., maximal difference in x and y direction) while  $\text{sigm}(x; \alpha, \beta, \gamma)$  is a Sigmoid function. The major impulse for the usage of the inner corners of the eyes as the

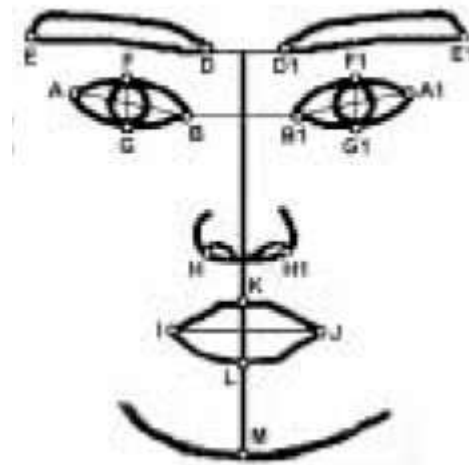


Figure 2: Feature points (fiducials of the features' contours)

E, E1: outer corner of the eyebrow  
D, D1: inner corner of the eyebrow  
A, A1: outer corner of the eye  
B, B1: inner corner of the eye  
F, F1: top of the eye  
G, G1: bottom of the eye  
H, H1: Outer corner of the nostril  
K: top of the upper lip  
L: bottom of the lower lip  
I, J: mouth corner  
N: tip of the chin

referential points for calculating CFs of the fiducial points of the eyes comes from the stability of these points with respect to non-rigid facial movements: facial muscles' contractions do not cause physical displacements of these points. For the same reason, the referential features used for calculating CFs of the fiducial points of the eyebrows, nose/chin and mouth are the size of the relevant eyebrow area, the inner corners of the nostrils and the medial point of the mouth respectively. Eventually, in order to select the best of sometimes redundantly available solutions (e.g., for the fiducial points belonging to the mouth), an intersolution

consistency check is performed by comparing the CFs of the points extracted by different detectors of the same facial feature. AUs of the FACS system are anatomically related to contractions of facial muscles [4]. Contractions of facial muscles produce motion on the skin surface and changes in the shape and location of the prominent facial features. Some of these changes are observable

from changes in the position of the fiducial points. To classify detected changes in the position of the fiducial points in terms of AUs, these changes should be represented first as a set of suitable feature parameters. Motivated by the FACS system, we represent these

changes as a set of mid-level feature parameters describing the state and motion of the fiducial points. We defined a single mid-level feature parameter, which describes the state of the fiducials. This parameter, which is calculated for each frame for various fiducial points by comparing the currently extracted fiducial points with the relevant fiducial points extracted from the neutral frame, is defined as:

$$\text{inc/dec}(AB) = AB_{\text{neutral}} - AB_{\text{current}}, \text{ where } AB = \sqrt{\{(xA - xB)^2 + (yA - yB)^2\}}$$

If  $\text{inc/dec}(AB) < 0$ , distances increases.

#### 4.4.1 Action Unit Recognition:

The last step in automatic facial gesture analysis is to translate the extracted facial information (i.e., the calculated feature parameters) into a description of shown facial changes, e.g., into the AU codes.

**Table 7. The description of 22 AUs to be recognized and the related rules for AU recognition**

AU	AU description & the related rule
1	Raised inner portion of the eyebrow(s) IF $\text{inc/dec}(\text{BD}) < 0$ OR $\text{inc/dec}(\text{B1D1}) < 0$ THEN AU1
2	Raised outer portion of the eyebrow(s) IF $\text{inc/dec}(\text{AE}) < 0$ OR $\text{inc/dec}(\text{A1E1}) < 0$ THEN AU2
3	Eyebrows pulled closer together (frown) IF $\text{inc/dec}(\text{DD1}) > 0$ THEN AU4
4	Raised upper eyelid(s) IF $\text{inc/dec}(\text{FG}) < 0$ OR $\text{inc/dec}(\text{F1G1}) < 0$ THEN AU5
5	Raised cheeks (smile); IF AU12 OR AU13 THEN AU6
6	Raised lower eyelid(s) IF $\text{not}(\text{AU12})$ AND $((\text{FG} > 0 \text{ AND } \text{inc/dec}(\text{GX}) > 0)$ OR $(\text{F1G1} > 0 \text{ AND } \text{inc/dec}(\text{G1Y}) > 0))$ THEN AU7
7	Lips pulled towards each other IF $\text{not}(\text{AU12 OR AU13 OR AU15 OR AU18 OR AU20 OR AU23 OR AU24 OR AU35})$ AND $\text{KL} > 0$ AND $\text{inc/dec}(\text{CK}) < 0$ THEN AU8

8	Mouth corner(s) pulled up IF $(\text{inc/dec}(\text{IB}) > 0 \text{ AND } \text{inc/dec}(\text{CI}) < 0)$ OR $(\text{inc/dec}(\text{JB1}) > 0 \text{ AND } \text{inc/dec}(\text{CJ}) < 0)$ THEN AU12
9	Mouth corner(s) pulled sharply up IF $(\text{inc/dec}(\text{IB}) > 0 \text{ AND } \text{inc/dec}(\text{CI}) > 0)$ OR $(\text{inc/dec}(\text{JB1}) > 0 \text{ AND } \text{inc/dec}(\text{CJ}) > 0)$ THEN AU13
10	Mouth corner(s) pulled down IF $\text{inc/dec}(\text{IB}) < 0$ OR $\text{inc/dec}(\text{JB1}) < 0$ THEN AU15
11	Mouth pushed medially forward (as when saying "fool") IF $\text{not}(\text{AU28})$ AND $\text{IJ} \geq t1$ AND $\text{inc/dec}(\text{IJ}) > 0$ AND $\text{inc/dec}(\text{KL}) \leq 0$ THEN AU18
12	Mouth stretched horizontally IF $\text{inc/dec}(\text{IJ}) < 0$ AND $\text{inc/dec}(\text{IB}) = 0$ AND $\text{inc/dec}(\text{JB1}) = 0$ THEN AU20
13	Tightened lips IF $\text{KL} > 0$ AND $\text{inc/dec}(\text{KL}) > 0$ AND $\text{inc/dec}(\text{IJ}) \leq 0$ AND $\text{inc/dec}(\text{JB1}) \geq 0$ AND $\text{inc/dec}(\text{IB}) \geq 0$ THEN AU23
14	Lips pressed together IF $\text{not}(\text{AU12 OR AU13 OR AU15})$ AND $\text{KL} > 0$ AND $\text{inc/dec}(\text{KL}) > 0$ AND $\text{IJ} > t1$ AND $\text{inc/dec}(\text{IJ}) > 0$ THEN AU24
15	Parted lips IF $\text{inc/dec}(\text{KL}) < 0$ AND $\text{inc/dec}(\text{CM}) \geq 0$ THEN AU25
16	Parted jaws IF $\text{inc/dec}(\text{CM}) < 0$ AND $\text{CM} \leq t2$ THEN AU26
17	Mouth stretched vertically; IF $\text{CM} > t2$ THEN AU27
18	Lips sucked into the mouth; IF $\text{KL} = 0$ THEN AU28
19	Cheeks sucked into the mouth; IF $\text{IJ} < t1$ THEN AU35
20	Widened nostrils IF $\text{not}(\text{AU8 OR AU12 OR AU13 OR AU18 OR AU24})$ AND $\text{inc/dec}(\text{HH1}) < 0$ THEN AU38
21	Compressed nostrils IF $\text{not}(\text{AU8 OR AU15 OR AU18 OR AU24 OR AU28})$ AND $\text{inc/dec}(\text{HH1}) > 0$ THEN AU39
22	Dropped upper eyelid(s) IF $\text{not}(\text{AU7})$ AND $((\text{FG} > 0 \text{ AND } \text{inc/dec}(\text{FG}) > 0)$ AND $\text{inc/dec}(\text{FX}) > 0)$ OR $(\text{F1G1} > 0 \text{ AND } \text{inc/dec}(\text{F1G1}) > 0)$ AND $\text{inc/dec}(\text{F1Y}) > 0))$ THEN AU41

#### 4.4.2 Feature transformation and Retention:

**Definition 1.1.:** A decision information system is a continuous value information system, and it is defined as a quadruple  $s = (U, C \cup D, V, f)$ , where  $U$  is a finite set of objects,  $C$  is the condition attribute set, and  $D = \{d\}$  is the decision attribute set. For all  $c \in C$ ,  $c$  is continuous value attribute.

A facial expression information system is a continuous value information system according to the above Definition. If a condition attribute value is a continuous value, indiscernibility relation cannot be used directly since it requires that the condition attribute values of two different samples are equal, which is difficult to satisfy. Consequently, a process of discretization must be taken, in which information may be lost or changed. The result of attribute reduction would be affected. Since all measured facial attributes are continuous value and imprecise to some extent, the process of discretization may affect the result of emotion recognition. We argue that it is suitable for the continuous value information systems that the attribute values are taken as equal if they are similar in some range. Based on this idea, a method based on tolerance relation that avoids the process of discretization is proposed.

**Definition 1.2.:** A binary relation  $R(x, y)$  defined on an attribute set  $B$  is called a tolerance relation if it satisfies

$$(1) \text{ symmetrical: } \forall_{x, y \in U} (R(x, y) = R(y, x)); \quad (12)$$

$$(2) \text{ reflexive: } \forall_{x \in U} (R(x, x) = 1). \quad (13)$$

From the standpoint of a continuous value information system, a relation could be set up for a continuous value information system as follows.

**Definition 1.3.:** Let an information system  $S = (U, C \cup D, V, f)$  be a continuous value information system; a relation  $R(x, y)$  is defined as

$$R(x, y) = \{(x, y) \mid x \in U \wedge y \in U \wedge \forall_{a \in c} (|a_x - a_y| \leq \varepsilon, 0 \leq \varepsilon \leq 1)\}. \quad (14)$$

Apparently,  $R(x, y)$  is a tolerance relation according to Definition 2.4 since  $R(x, y)$  is symmetrical and reflexive. In classical rough set theory, an equivalence relation constitutes a partition of  $U$ , but a tolerance relation constitutes a cover of  $U$ , and equivalence relation is a particular type of tolerance relation.

**Definition 1.4.:** Let  $R(x, y)$  be a tolerance relation based on (12) and (13),

$$n_R(x_i) = \{x_j \mid x_j \in U \wedge \forall_{a \in c} (|a_x - a_y| \leq \varepsilon)\} \quad (15)$$

is called a tolerance class of  $x_i$ , and  $|n_R(x_i)| = |\{x_j \mid x_j \in n_R(x_i), 1 \leq j \leq U\}|$  is the cardinal number of the tolerance class of  $x_i$ .

According to above Definition, for all  $x \in U$ , the bigger the tolerance class of  $x$  is, the more uncertainty it will be and the less knowledge it will contain. On the contrary, the smaller the tolerance class of  $x$  is, the less uncertainty it will be and the more knowledge it will contain. Accordingly, the concept of knowledge entropy and conditional entropy could be defined as follows.

**Definition 1.5.:** Let  $U = \{x_1, x_2, \dots, x_{|U|}\}$ ,  $R(x, y)$  be a tolerance relation; the knowledge entropy  $E_R$  of relation  $R$  is defined as

$$E(R) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{n_R(x_i)}{|U|}. \quad (15)$$

**Definition 1.6.:** Let  $R$  and  $Q$  be tolerance relations defined on  $U$ , a relation satisfying  $R$  and  $Q$  simultaneous can be taken as  $R \cup Q$ , and it is a tolerance relation too. For all  $x_i \in U$ ,  $|n_{R \cup Q}(x_i)| = |n_R(x_i) \cap n_Q(x_i)|$ ; therefore, the knowledge entropy of  $R \cup Q$  can be defined as

$$E(R \cup Q) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{n_{R \cup Q}(x_i)}{|U|}. \quad (16)$$

**Definition 1.7.:** Let  $R$  and  $Q$  be tolerance relations defined on  $U$ ; the conditional entropy of  $R$  with respect to  $Q$  is defined as

$$E(Q | R) = E(R \cup Q) - E(R).$$

Let  $S = (U, C \cup D, V, f)$  be a continuous value information system, let relation  $K$  be a tolerance relation defined on its condition attribute set  $C$ , and let relation  $L$  be an equivalence relation \_a special tolerance relation\_ defined on its decision attribute set  $D$ . According to Definitions 2.7, 2.8, and 2.9, we can get

$$\begin{aligned} E(D | C) &= E(L | K) \\ &= E(K \cup L) - E(K) \\ &= -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{n_{K \cup L}(x_i)}{|U|} - \left( -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{n_K(x_i)}{|U|} \right) \\ &= -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{n_{K \cup L}(x_i)}{|U|} \end{aligned} \quad (17)$$

where the conditional entropy  $E(D | C)$  has a clear meaning; that is, it is a ratio between the knowledge of all attributes (condition attribute set plus decision attribute set) and the knowledge of the condition attribute set.

A novel attribute reduction algorithm is proposed based on rough set theory and domain-oriented data-driven data mining (3DM).

#### 4.4.3 Handling Facial Dynamics:

Yacoub and Davis [35] proposed an approach for analyzing and representing the dynamics of facial expressions from image sequences. This approach is divided into three stages: locating and tracking prominent facial features (i.e., mouth, nose, eyes, and brows), using optical flow at these features to construct a mid-level representation that describes spatio-temporal actions, and applying rules for classification of mid-level representation of actions into one of the six universal facial expressions. Matsuno *et al.* [37] proposed an approach for recognizing facial expressions from static images based on precomputed parameterization of



facial expressions. Their approach lays a grid over the face and warps it based on the gradient magnitude using a physical model. The amount of warping is represented in a multivariate vector that is compared to learned vectors of four facial expressions (happiness, sadness, anger, and surprise). Mase [36] used optical flow computation for recognizing and analyzing facial expressions in both a top-down and bottom-up approach. In both cases, the focus was on computing the motion of facial *muscles* rather than of facial *features*. Four facial expressions were studied: surprise, anger, happiness, and disgust. The top-down approach assumed that the face's image can be divided into muscle units that correspond to the action units (AU's) suggested by Ekman and Friesen [38]. Optical flow is computed within rectangles that include these muscle units, which in turn can be related to facial expressions. However, Mase did not report any results on mapping the optical motion results into facial expressions. This approach relies heavily on locating rectangles containing the appropriate muscles—a difficult image analysis problem since the muscle units correspond to smooth, featureless surfaces of the face. Furthermore, it builds upon a model that is suitable for synthesizing facial expressions but remains untested in analysis of facial expressions (for more details see [39]). The bottom-up approach covered the area of the face with evenly divided rectangular regions over which feature vectors derived from an optical flow computation are computed. The feature vectors are defined over a 15-dimensional space that is computed based on the means and variances of the optical flow. The recognition of expressions is based on a k-nearest-neighbor voting rule. The optical flow calculation was averaged within each window to smooth the results over edges. Furthermore, the optical flow is treated on a per-frame basis without considering the time-sequence of frames. The experiments considered the expressions of just one face and the results were compared with the performance of human subjects that were asked to classify the displayed emotions. Terzopoulos and Waters [40] proposed an approach for synthesis and analysis of facial expressions based on physical modeling of the muscles of the face. They devised a six level representation of the face that consists of expression level (which includes the six primary expressions); control level (implements a subset of the FACS-facial action coding system for controlling muscles), muscle level (models the muscles' contraction and expansion as springs), physics level (models the facial tissue's deformations), geometry level (provides a geometric representation of the face as a mesh of polyhedral elements that depend on the curvature of surface), and the image level (visualizes the data). The analysis part assumes that 11 principle contours are initially located manually on the face. These contours are tracked throughout the sequence by applying an image force field that is computed from the gradient of the intensity image. The estimation of the muscle contractions takes place after the contours' reference points have been determined. In addition to assuming a frontal view, it was assumed that the projection is orthographic. Once the muscle contractions have been estimated, they were resynthesized onto the 3-D range data model of the subject to recreate the muscle contractions. It remains to be determined whether the computation of muscle mapped onto a 3-D physical model of the human face to activate muscles that create the same facial expressions on this model.

Cottrell and Metcalfe [41] proposed a back propagation neural network that recognizes facial expression (along with gender and identity) from static images. Their network compresses the input images in a "holistic" manner into 40 hidden units that were fed

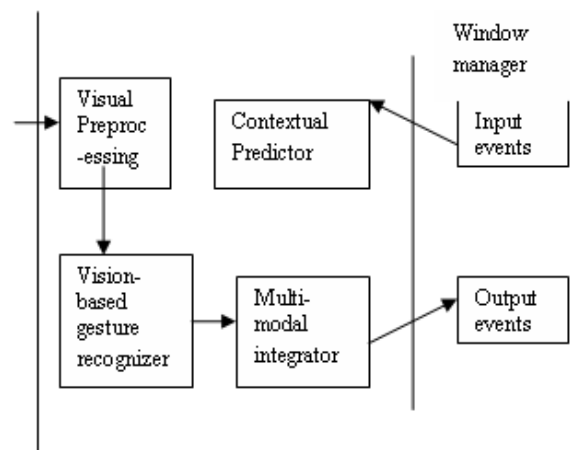
into a three-layer classification network. The expression network was able to distinguish some of the positive emotions but was less able to handle negative expressions. The ability of the network to generalize to new faces was poor. Essa [42] recently proposed a physically based approach for modeling and analyzing facial expressions. This approach extends the FACS model to the temporal dimension (thus calling it FACS+) to support combined spatial and temporal modeling of facial expressions. It is assumed that a mesh is originally overlaid on the face; then its vertices are tracked based on the optical flow field throughout the sequence. The emphasis is on accuracy of capturing facial changes, which is important for synthesis. Recognition results were reported in [42] on six subjects displaying four expressions and eyebrow raising.

## 5. TOWARDS CONTEXT BASED APPROACH TOWARDS FACIAL RECOGNITION

### 5.1 Contextual features

There are several sources of potential errors with a gesture based recognition system. For example when people search for their cursor on the screen, they perform fast short movements similar to head nods or head shakes, and when people switch attention between the screen and keyboard to place their fingers on the right keys, the resulting motion can appear like a head nod. These types of false positives can cause trouble, especially for users who are not aware of the tracking system.

In research on interactions with ECAs, it has been shown that contextual information about dialog state is a productive way to reduce false positives [33]. A context-based recognition framework can exploit several cues to determine whether a particular gesture is more or less likely in a given situation. To apply the idea of context-based recognition to non-embodied interfaces, i.e. windows-based interfaces, here we define a new set of contextual features based on window manager state.



**Figure 3: Framework for context-based gesture recognition.** The contextual predictor translates contextual features into a likelihood measure, similar to the visual recognizer output. The multi-modal integrator fuses these visual and contextual likelihood measures.

We want to find contextual features that will reduce false positives that happen during interaction with conventional input devices, and contextual features that can be easily computed using

pre-existing information. For our initial prototype we selected two contextual features:  $fd$  and  $fm$ , defined as the time since a dialog box appeared and time since the last mouse event and respectively. These features can be easily computed by listening to the input and display events sent inside the message dispatching loop of the application or operating system. We compute the dialog box feature  $fd$  as

$$f_d(t) = C_d \text{ if no dialog box was shown} \\ t - t_d \text{ otherwise}$$

where  $td$  is the time-stamp of the last dialog box appearance and  $Cd$  is default value if no dialog box was previously shown. The same way, we compute the mouse feature  $fm$  as

$$f_m(t) = C_m \text{ if no mouse event happened} \\ t - t_m \text{ otherwise}$$

where  $tm$  is the time-stamp of the last mouse event and  $Cm$  is default value if no mouse event happened recently. In our experiments,  $Cd$  and  $Cm$  were set to 20. The contextual features are evaluated at the same rate as the vision-based gesture recognizer (about 18Hz).

We wish to learn a measure of likelihood for a gesture given only the contextual features described in the previous section. This measure will later be integrated with the measure from our vision-based head gesture recognizer to produce the final decision of our context-based gesture recognizer (see Figure 3). The measure of likelihood is taken to be the distance to a separating surface of a multi-class Support Vector Machine (SVM) classifier that predicts the gesture based on contextual features only. The SVM classifier learns a separating function whose distance  $m(x)$  to training labels is maximized. The margin  $m(x)$  of the feature vector  $x$ , created from the concatenation of the contextual features, can easily be computed given the learned set of support vectors  $x_i$ , the associated set of labels  $y_i$  and weights  $w_i$ , and the bias  $b$ :

$$m(x) = \sum_{i=1}^l y_i w_i K(x_i, x) + b \quad (18)$$

where  $l$  is the number of support vectors and  $K(x_i, x)$  is the kernel function. In our experiments, we used a radial basis function (RBF) kernel:

$$K(x_i, x) = e^{-\gamma \|x_i - x\|^2} \quad (19)$$

where  $\gamma$  is the kernel smoothing parameter learned automatically using cross-validation on our training set. After training the multi-class SVM, we can easily compute a margin for each class and use this scalar value as a prediction for each visual gesture. The contextual predictor was trained using a subset of twelve participants. Positive and negative samples were selected from this data set based on manual transcription of head nods and head shakes.

## 5.2 Context-Based Scene Recognition Using Bayesian Networks

Scene understanding is an important problem in intelligent robotics. Since visual information is uncertain due to several

reasons, we need a novel method that has robustness to the uncertainty. Bayesian probabilistic approach is robust to manage the uncertainty, and powerful to model high-level contexts like the relationship between places and objects. At first, image pre-processing extracts features from vision information and objects existence information is extracted by SIFT that is rotation and scale invariant. This information is provided to Bayesian networks for robust inference in scene understanding.

### 5.2.1 Visual Context-Based Low-Level Feature Extraction:

It would be better to use features that are related to functional constraints, which suggests

to examine the textural properties of the image and their spatial layout [43]. To compute texture feature, a steerable pyramid is used with 6 orientations and 4 scales applied to the gray-scale image. The local representation of an image at time  $t$  is as follows:

$$v_x^L(x) = \{v_t, k(x)\}_{k=1, N}, \quad \text{where } N=24 \quad (20)$$

It is desirable to capture global image properties, while keeping some spatial information. Therefore, we take the mean value of the magnitude of the local features averaged over large spatial regions:

$$m_t(x) = \sum_{x'} |v_t^L x'| w(x'-x), \quad \text{where } w(x) \text{ is averaging window} \quad (21)$$

The resulting representation is down-sampled to have a spatial resolution of 4x4 pixels, leading to the size of  $m_t$  as 384(4 x 4 x 24), whose dimension is reduced by PCA (80 PCs). Then, we have to compute the most likely location of the visual features acquired at time  $t$ . Let the place be denoted as  $Q_t \in \{1, \dots, N_p\}$  where  $N_p = 5$ . Hidden Markov model (HMM) is used to get place probability as follows:

$$P(Q_t = q | v_{1:t}^G) \propto p(v_{1:t}^G | Q_t = q) P(Q_t = q | v_{1:t-1}^G)$$

$$p(v_{1:t}^G | Q_t = q) \sum_{q'} A(q', q) P(Q_{t-1} = q' | v_{1:t-1}^G) \quad (22)$$

where  $A(q', q)$  is the topological transition matrix. The transition matrix is simply learned from labeled sequence data by counting the number of transitions from location  $i$  to location  $j$ . We use a simple layered approach with HMM and Bayesian networks. This presents several advantages that are relevant to modeling high dimensional visual information: learning each level independently with less computation, and although environment changes, only first layer requires new learning with the remaining unchanged [45]. The HMM is for extracting place recognition and BNs are for high-level inference.

### 5.2.2 High-Level Context Extraction with SIFT

Scale-Invariant Feature Transform (SIFT) is used to compute high-level object existence information. Since visual information is uncertain, we need a method that has robustness to scale or camera angle change. It was shown that under a variety of

reasonable assumptions the only possible scale-space kernel was the Gaussian function [44]. Therefore, the scale space of an image is defined as a function,  $L(x, y, \sigma)$  that is produced by the convolution of a variable-scale Gaussian,  $G(x, y, \sigma)$ , with an input image,  $I(x, y)$ :

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y), \quad (23)$$

where  $*$  is the convolution operation in  $x$  and  $y$ , and

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2 + y^2) / 2\sigma^2} \quad (24)$$

To efficiently detect stable key-point locations in scale space, scale-space extrema in the difference-of-Gaussian function are convolved with the image,  $D(x, y, \sigma)$ , which can be computed from the difference of two nearby scales separated by a constant multiplicative factor  $k$ :

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned} \quad (25)$$

Extracted key-points are examined in each scene image, and the algorithm decides that the object exists if match score is larger than a threshold.

### 5.2.3 Context-Based Bayesian Network Inference

A Bayesian network is a graphical structure that allows us to represent and reason in an uncertain domain. The nodes in a Bayesian network represent a set of random variables from the domain. A set of directed arcs connect pairs of nodes, representing the direct dependencies between variables. Assuming discrete variables, the strength of the relationship between variables is quantified by conditional probability distributions associated with each node [46].

Consider a BN containing  $n$  nodes,  $Y_1$  to  $Y_n$ , taken in that order. The joint probability for any desired assignment of values  $\langle y_1, \dots, y_n \rangle$  to the tuple of network variables  $\langle Y_1, \dots, Y_n \rangle$  can be computed by the following equation:

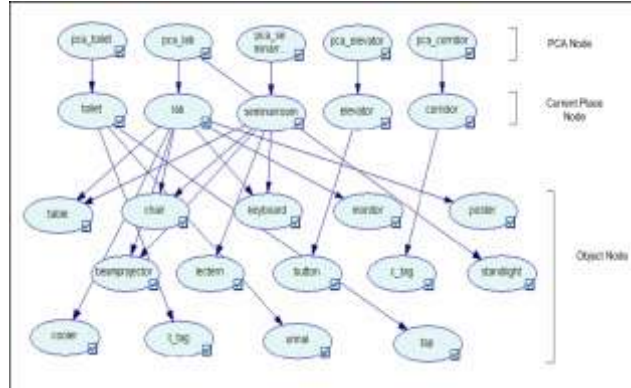
$$p(y_1, y_2, \dots, y_n) = \prod_i P(y_i | \text{Parents}(Y_i)) \quad (26)$$

where  $\text{Parents}(Y_i)$  denotes the set of immediate predecessors of  $Y_i$  in the network.

BN consists of 4 types of nodes: (1) 'PCA Node' for inserting global feature information of current place, (2) 'Object Node' representing object existence and correlation between object and place, and (3) 'Current Place Node' representing the probability of each place. Let the place be denoted  $Q_{t,i} \in \{1, \dots, N_p\}$  where  $N_p = 5$  and object existence is denoted by  $O_{t,i} \in \{1, \dots, N_{object}\}$  where  $N_{object} = 14$ . Place recognition can be computed by the following equation:

$$\text{CurrentPlace} = \arg \max_{1:t, O_{t,i}, \dots, O_{t,N_{object}}} P(Q_t = q | v_{1:t}^G, O_{t,i}, \dots, O_{t,N_{object}}) \quad (27)$$

The BNs are manually constructed by expert, and nodes that have low dependency are not connected to reduce computational complexity. Fig. 2 shows a BN that is actually used in experiments.



**Figure 4. A BN manually constructed for place and object recognition**

The work has been verified[34] that the context-based Bayesian inference for scene recognition shows good performance in the complex real domains. Even though the global feature information extracted is the same, the proposed method could produce correct result using contextual information: relationship between object and place. But SIFT algorithm showed low performance when objects had insufficient textual features, and this lack of the information caused to the low performance of scene understanding. To overcome it, we need a method that disjoints objects with ontology concept, and extracts SIFT key-points in each component. Besides, we could easily adopt more robust object recognition algorithm to our method.

## 6. CONCLUSION AND FUTURE WORKS

Theoretically all the aspects including cues, facial and gesture are important for non verbal communication. Adding facial features to gestures don't cause much difference to the results, rather can include complexity. Adding gesture again to it causes ambiguity, and probably the results go down. Hence facial features are the most important aspect for emotion recognition. Further, ANN, Fuzzy are extensively used for this purpose. Adding Rough set approach gives better results, and rough set is easy to implement too. Some Context based research in these areas have been done using support vector machine and Bayesian networks. Furthermore, context based approach can be further extended with the help of Rough-fuzzy approach towards refining artificial intelligence.

## 7. ACKNOWLEDGE

This is the research work associated to the Ph.D work going on Under Devi Ahilya University, Indore, and College of Engg, Pune. My sincere thanks the centre and their faculty for their extensive support.

## 8. REFERENCES

- [1] R.W. Picard, *Affective Computing*, MIT Press, Cambridge, UK, 1997.
- [2] R. W. Picard, "Affective computing: challenges," *International Journal of Human Computer Studies*, vol. 59, no. 1-2, pp. 55–64, 2003.
- [3] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: analysis of affective physiological state," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.
- [4] X. Sui and Y. T. Ren, "Online processing of facial expression recognition," *Acta Psychologica Sinica*, vol. 39, no. 1, pp. 64–70, 2007 \_Chinese\_.
- [5] Y. M. Wang and X. L. Fu, "Recognizing facial expression and facial identity: parallel processing or interactive processing," *Advances in Psychological Science*, vol. 13, no. 4, pp. 497–500, 2005 \_Chinese\_.
- [6] Paul Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Palo Alto, CA: Consulting Psychologist Press, 1978.
- [7] Paul Ekman and E.L. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System (FACS)*, New York: Oxford University Press, 1997.
- [8] Gianluca Donato, Marian Steward Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski, *Classifying Facial Actions*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21(10), pp. 974-989, 1999.
- [9] K. Mase, *Recognition of facial expression from optical flow*. *IEICE Transactions E*, vol. 74(10), pp. 3474-3483, 1991.
- [10] M. Rosenblum, Y. Yacoob, and L. Davis, *Human expression recognition from motion using a radial basis function network architecture*. *IEEE Trans. Neural Networks*, vol. 7(5), pp. 1121-1138, 1996.
- [11] A. Lanitis, C. Taylor, and T. Cootes, *Automatic interpretation and coding of face images using flexible models*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19(7), pp. 743-756, 1997.
- [12] J.G. Daugman, *Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression*. *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 36, pp. 1169-1179, 1988.
- [13] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, *EigenFaces vs. FisherFaces: Recognition Using Class Specific Linear Projection*. *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19(7), pp. 711-720, 1996.
- [14] P.S. Penev and J.J. Atick, *Local feature analysis: a general statistical theory for object representation* *Network: Computation in Neural Systems*, vol. 7(3), pp. 477-500, 1996.
- [15] M.S. Bartlett and T. Sejnowski, *Viewpoint invariant face recognition using independent component analysis and attractor networks*, in *Advances in Neural Information Processing Systems*, M. Mozer, M. Jordan, and T. Petsche, Editors, MIT Press: Cambridge, MA, 1997.
- [16] J. Zhang, Y. Yan, and M. Lades. *Face recognition: Eigenface, elastic matching, and neural nets*, in *proceedings of IEEE*, vol. 85(9), pp. 1423-1435, 1997.
- [17] A. C. Tsoi, *Adaptive Processing of Data Structure : AnExpository Overview and Comments*, Faculty Informatics, Univ. Wollongong, Wollongong, Australia, 1998.
- [18] A. Sperduti and A. Starita, *Supervised neural networks for classification of structures*. *IEEE Trans. Neural Networks*, vol. 8, pp. 714-735, 1997.
- [19] P. Frasconi, M. Gori, and A. Sperduti, *A General Framework for Adaptive Processing of Data Structures*. *IEEE Trans. Neural Networks*, vol. 9, pp. 768-785, 1998.
- [20] Siu-Yeung Cho, Zheru Chi, Wan-Chi Siu, and Ah Chung Tsoi, *An Improved Algorithm for learning longterm dependency problems in adaptive processing of data structures*. *IEEE Transactions on Neural Networks*, vol. 14(4), pp. 781-793, 2003.
- [21] J. Platt, *Fast Training of Support Vector Machines using Sequential Minimal Optimization*, in *Advances in Kernel Methods - Support Vector Learning*, B. Scholkopf, C. Burges, and A. Smola, Editors, MIT Press, pp. 185-208, 1998.
- [22] D. Aha and D. Kibler, *Instance based learning algorithms*. *Machine Learning*, vol. 6, pp. 37-66, 1991.
- [23] M. White, "Effect of photographic negation on matching the expressions and identities of faces," *Perception*, vol. 30, no. 8, pp. 969–981, 2001.
- [24] Jang. J. S. R; "ANFIS: Adaptive Network Based Fuzzy Inference Systems"; *IEEE Transactions on Systems, Man and Cybernetics*; May 1993.
- [25] Jang. J. S. R, Sun. C. T, Mizutani. E; *Neuro – Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*; Prentice Hall; 1997.
- [26] Demsar J, Zupan B; "Orange: From Experimental Machine Learning to Interactive Data Mining"; White Paper ([www.ailab.si/orange](http://www.ailab.si/orange)), Faculty of Computer and Information Science, University of Ljubljana; 2004.
- [27] Kudo. Y, Murai. T; "Missing Value Semantics and Absent Value Semantics for Incomplete Information in Object-Oriented Rough Set Models"; In Bello et. al. [13]; 2008.
- [28] Z. Pawlak, "Rough sets," *International Journal of Computer & Information Sciences*, vol. 11, no. 5, pp. 341–356, 1982.
- [29] Z. Pawlak, "Rough classification," *International Journal of Man-Machine Studies*, vol. 20, no. 5, pp. 469– 483, 1984.
- [30] Z. Pawlak, "Rough set theory and its applications to data analysis," *Cybernetics and Systems*, vol. 29, no. 7, pp. 661–688, 1998.
- [31] R. W. Swiniarski and A. Skowron, "Rough set methods in feature selection and recognition," *Pattern Recognition Letters*, vol. 24, no. 6, pp. 833–849, 2003.
- [32] N. Zhong and A. Skowron, "A rough set-based knowledge discovery process," *International Journal of Applied*

*Mathematics and Computer Science*, vol. 11, no. 3, pp. 603–619, 2001.

- [33] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell. Contextual recognition of head gestures. In *Proceedings of the International Conference on Multi-modal Interfaces*, October 2005.
- [34] Seung-Bin Im and Sung-Bae Cho, Context-Based Scene Recognition Using Bayesian Networks with Scale-Invariant Feature Transform, *ACIVS 2006*, LNCS 4179, pp. 1080 – 1087, 2006. © Springer-Verlag Berlin Heidelberg 2006.
- [35] “Computing spatio-temporal representations of human faces,” in *IEEE Conf. Comput. Vision and Pattern Recognition*, 1994, pp. 70-75.
- [36] K. Mase, “Recognition of facial expression from optical flow,” *IEICE Trans.*, vol. E74, no. 10, pp. 3474-3483, Oct. 1991.
- [37] K. Matsuno, C. Lee, and S. Tsuji, “Recognition of human facial expressions without feature extraction,” *ECCV*, pp. 513-520, 1994.
- [38] *The Facial Action Coding System*. San Francisco, CA: Consulting Psychologists Press, 1978.
- [39] V. Bruce, *Recognizing Face*, London: Lawrence Erlbaum, 1988.
- [40] D. Terzopoulos and K. Waters, “Analysis and synthesis of facial image sequences using physical and anatomical models,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, pp. 569-579, June 1993.
- [41] G. W. Cottrell and J. Metcalfe, “EMPATH: Face, gender, and emotion recognition using holons,” in *Advances in Neural Information Processing Systems*, R. P. Lippman, J. Moody, and D. S. Touretzky, Eds. San Mateo, CA: 1991, vol. 3, pp. 564-571.
- [42] I. A. Essa and A. Pentland, “A vision system for observing and extracting facial action parameters,” in *Proc. Computer Vision and Pattern Recognition, CVPR-94*, Seattle, WA, June 1994, pp. 76-83.
- [43] A. Torralba, K.P. Murphy, W. T. Freeman and M. A. Rubin, “Context-based vision system for place and object recognition,” *IEEE Int. Conf. Computer Vision*, vol. 1, no. 1, pp. 273-280, 2003.
- [44] D.G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Intl. J. Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.
- [45] N. Oliver, A. Garg and E. Horvitz, “Layered representations for learning and inferring office activity from multiple sensory channels,” *Computer Vision and Image Understanding*, vol. 96, no. 2, pp. 163-180, 2004.
- [46] R.E. Neapolitan, *Learning Bayesian Network*, Prentice hall series in Artificial Intelligence, 2003.