Self-adaptive Expert System for Facial Expression Analysis

Maja Pantic, *Student Member, IEEE,* and Leon J.M. Rothkrantz

Abstract*--We attempt to automate facial expression recognition and introduce it into the man-machine interaction as a new modality. This will make the interaction compact and more efficient. As the first step, we developed a self-adaptive expert system that accepts the facial features contours localized in a static dual-view facial image and returns the expression interpretation label(s) used by the user. Expression identification in terms of the encountered facial actions is also displayed to the user. Reasoning with uncertainty about the extracted facial expression data is employed for facial action coding and quantification. A memory of experiences, inspired by the Shank's theory of human autobiographical memory organization and the instance-based learning, expounds the encoded facial actions in terms of the learned interpretation labels. Validation studies on the prototype suggest that the expressions' identifications and interpretations achieved are generally consistent with those defined by the users.

I. INTRODUCTION

The human face is an independent communication channel that transmits emotional and conversational signals encrypted as facial expressions. Facial displays can be viewed as communicative signals that help harmonize conversation and play a major role in human interaction and non-verbal communication [11]. If the goal is to design a human-like man-machine interaction, then human face-to-face communication provides an ideal model. To achieve this goal, automatic encoding and interpretation of facial signals should be facilitated.

Humans detect and interpret faces and their expressions in a scene with little or no effort. Still, development of an automated system that accomplishes this task is rather difficult. The main problems are identification of the encountered facial expression (i.e. facial action coding) and interpretation of the facial expression.

One of the fundamental issue about the latter is to define the set of categories we want to deal with in classifying/ interpreting expressions. Most of the existing studies on automatic facial expression analysis perform a singular classification into one of the six *basic emotion categories* as defined by Ekman [6] (anger, disgust, happiness, fear, surprise and sadness). Some recent examples of automatic singular emotional classifiers of facial expressions are proposed by Black et al. [1], Huang et al. [9], Hong et al. [8], Otsuka et al. [14], Edwards et al. [4].

Yet, it is definitely not certain that all facial expressions able to be displayed on the face can be classified under one of the six basic emotion categories. Think about so-called blended emotional expressions (e.g. raised eyebrows and smiling mouth) or an "I don't know" expression. Besides, some user might have different interpretation of a particular expression than some other user. Therefore, an advanced user-oriented facial expression analyzer should be capable of adapting its classification mechanism according to the user's subjective interpretation of facial expressions.

In order to conceive pairing of arbitrary expression (i.e. arbitrary facial actions) with a given interpretation label, automatic facial action encoding is needed. The Facial Action Coding System (FACS) [5] has been developed to facilitate objective measurement of facial activity for behavioral investigations of the face. It is a system designed for human observers to detect subtle changes in facial appearance caused by contractions of the facial muscles. In a form of rules, FACS provides a linguistic description of all possible visually detectable facial changes in terms of 44 Action Units (AUs). So far, several studies on vision-based facial gesture analysis suggested that FACS AUs could be detected from digitized facial images [3].

Essa et al. [7] use spatio-temporal templates to recognize 2 facial actions and 3 prototypic emotional expressions. Cohn et al. [2] achieved some success in automating facial action coding by feature point tracking of some points, manually located in the first frame of an examined facial image sequence. Their method can identify 8 individual AUs and 7 AUs combinations under the constraint that each image sequence starts with a neutral expression and doesn't contain more than one facial action in a row. In fact, it is not known whether any of the methods reported up-to-date is sufficient for describing full range of facial actions. None of the systems presented in the literature deals with both, quantified facial action coding and self-adaptive multiple classification of expressions in terms of user-defined labels.

This paper presents a system that can robustly perform interpretation of static dual-view facial images in terms of:

- 30 different facial actions and their intensity and
- multiple quantified user-defined interpretation labels.

The system consists of three major parts (Fig. 1): data extractor, facial action encoder, and expression classifier.

The authors are with the Department of Knowledge Based Systems, Faculty of ITS, Delft University of Technology, P.O. Box 356, 2600 AJ Delft, Netherlands. E-mail: {M.Pantic,L.J.M.Rothkrantz}@cs.tudelft.nl

The Data Extractor is a framework for "hybrid" facial expression data extraction, which for each prominent facial feature applies multiple feature detectors on the examined dual-view facial image. The Data Extractor is explained in section II. The Facial Action Encoder makes the best possible selection from the redundantly localized contours of the facial features, deals with missing data, assigns a certainty measure to the evaluated data (i.e. our confidence in data) and infers about the encountered facial actions. An expert system performs this. The Facial Action Encoder is presented elsewhere [16] and here we are providing just a short overview of it in section III. The Facial Expression Classifier expounds the identified facial actions in terms of the interpretation labels defined by the user. The dynamic memory of experiences facilitates this. The memory is dynamic since new interpretation labels can be learned with experience. The Expression Classifier, i.e. system's learning facility, is explained in section IV. The testing results are presented in section V. Section VI concludes the paper.

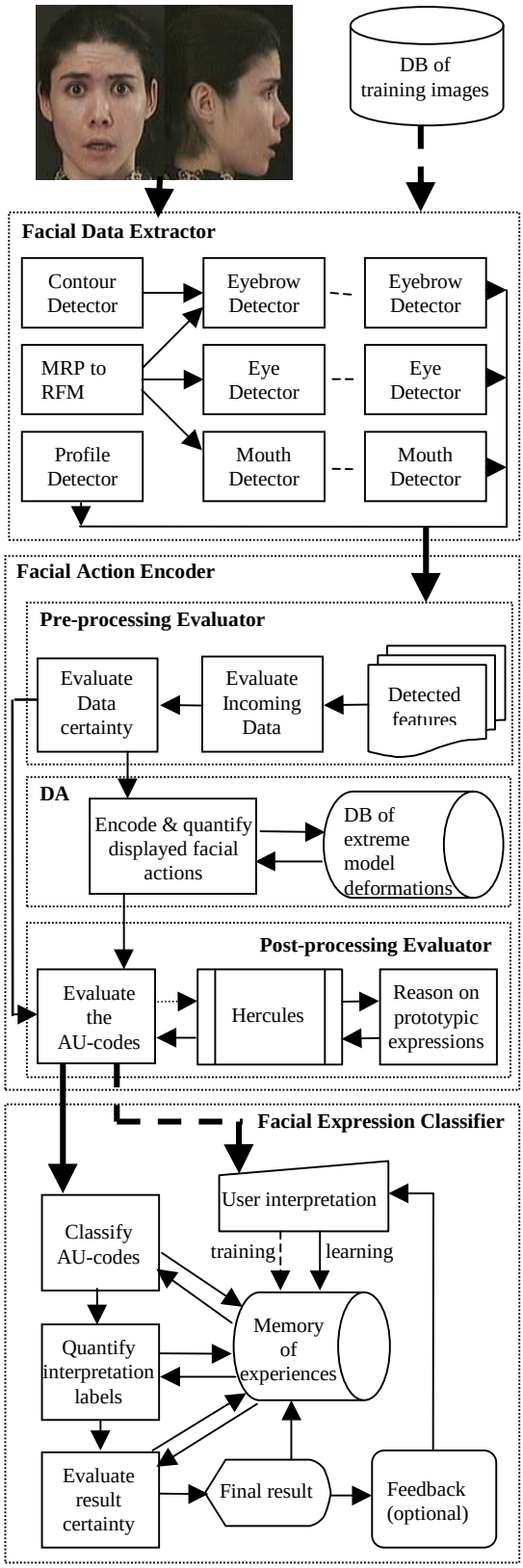

Fig. 1. System structure

II. Facial Data Extractor

Efforts have recently turned to expression classification by image processing of video sequences [1], [7], [9], [8], [14]. Our approach is more in line with FACS [5] and Ekman's work on prototypic emotional facial expressions [6] – we perform expression classification in photographs. In our system only the end-state of facial movement is measured and classified in comparison to an expressionless face of the same subject. The movement itself is not measured.

The system deals with static, dual-view facial images. The images are acquired using two digitized cameras mounted, on the head of the user, on the holders attached to a headphone-like device. One camera holder is placed in front of the face at approximately 15 centimeters from the tip of the nose (frontal view). The other camera holder is placed on the right side of the face at approximately 15 centimeters from the center of the right cheek (side view). Hence the presence of the face in the scene is ensured and some out-of-plane head motions cannot be encountered (i.e. the images are scale and orientation invariant).

The existing automated face analyzers usually utilize a single kind of facial feature detector [3]. In contrast, we are proposing a "hybrid" approach to facial expression data extraction. Per facial feature (eyebrows, eyes, nose, mouth) the Data Extractor applies multiple detectors of different kinds. For instance, a neural network originally proposed by Vincent [19] that finds the micro-features of the eyes or an active contour method proposed by Kass [10] perform currently automatic localization of the eye contour. But, any other detector picked up "off the shelves" that performs localization of the eye contour can be used instead. Rather than fine-tuning the existing- or inventing new techniques, known detectors are combined. The motivation for combining detectors is the increase in quality of the "hybrid" detector. Each detector has circumstances under which it performs extremely well. So a hybrid detector should perform better than the best single detector. In turn, introducing redundancy by applying multiple detectors per facial feature and then selecting the best of the acquired results yields more accurate and more complete set of detected facial features (i.e. less missing data) than it is the case when utilizing a single detector per facial feature.

After invoking all integrated detectors, each localized contour of a prominent feature is stored in a separate file. Those files form the input to the Facial Action Encoder.

III. Facial Action Encoder

The Facial Action Encoder consists of three parts illustrated in Fig. 1: pre-processing evaluator, data analyzer (DA) and post-processing evaluator. Since the Facial Action Encoder has been presented elsewhere [16], we are providing here just a short overview of this part of the system.

A. Facial Data Evaluator

The Facial Data Evaluator operates in two stages. First it delimits the geometry of the encountered expression by choosing the "best" of the redundantly localized facial features' contours stored in the files that form the output of the Facial Data Extractor. In the second stage, the defined facial expression geometry is represented in terms of our face model. The set of the face-model points together with the assigned *certainty factors* (CFs) form the input to the Facial Data Analyzer.

The reasoning of the first stage applies the knowledge about the facial anatomy (e.g. the inner corners of the eyes are immovable points) to check the correctness of the results achieved by the facial feature detectors. For example, a file containing the result of an eye detector is discarded if the localized inner corners of the eyes deviate for more than 5 pixels from the inner corners of the eyes localized in the expressionless face of the same subject. An inter-file consistency check is also performed. If the contour stored in the tested file deviates for more than 10 pixels in any direction from relevant contours stored in the other files, the tested file is discarded. To make the best choice between the results of different detectors, which localize a contour of the same feature, we are using the priorities $\boldsymbol{m} \in \mathrm{N}$ being off-line manually assigned to the integrated detectors based on their testing results. Each facial feature is delimited by the content of a not discarded file comprising that feature's contour localized by the highest priority detector. The priority of the selected detector $\boldsymbol{m}$ (where $\boldsymbol{M}$ is the highest priority a detector can have) defines the CF assigned to the pertinent facial feature as given in (1).

$$\boldsymbol{CF} = (1/\boldsymbol{M}) * \boldsymbol{m} \tag{1}$$

If localizing a certain feature's contour fails (i.e. all of the relative files are discarded), the pertinent feature's contour localized in the expressionless face of the same subject is used to substitute missing data. The CF assigned to the feature being substituted in this way is set to $1/2\boldsymbol{M}$.

We utilize a point-based face model composed of two 2D facial views, namely the frontal and the side view (Fig. 2). Since the images are scale- and orientation invariant, extraction of the model points from the localized contours of the facial features is straightforward. To each of the model points a CF is set to the CF of the facial feature (e.g. profile, mouth) to which the point belongs.

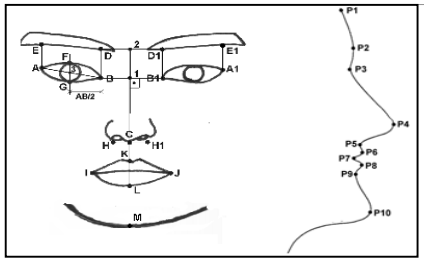

Fig. 2. Face model

B. Facial Action Coder

This part of the system performs quantified facial action coding based on the facial expression data extracted and evaluated by the parts of the system previously explained. 30 rules of the employed expert system encode the knowledge that has been acquired from FACS [5] in a straightforward manner. Each rule recognizes activation of a single AU based on the deviation of the model points from the pertinent points localized in an expressionless face of the same person. For instance, the rule for recognition of AU5 activation ("raised eyelid") has the following pseudo-code "*IF the distance 3F is increased OR the distance 4F1 is increased THEN AU5 is activated*" (see Fig. 2). The system can recognize the following AUs: 1, 2, 4, 5, 6, 7, 8, 9, 10, 12, 13, 15, 16, 17, 18, 19, 20, 23, 24, 25, 26, 27, 28, 28b, 28t, 36b, 36t, 38, 39 and 41.

The CFs associated with the facial points $\boldsymbol{p}_1, \cdots, \boldsymbol{p}_n$ define the CF of the related distances as given in (2).

$$\boldsymbol{CF}_{\boldsymbol{feature}} = 1/2 * \min\{\boldsymbol{CF}_{\boldsymbol{p}_1}, \cdots, \boldsymbol{CF}_{\boldsymbol{p}_n}\} \tag{2}$$

The overall certainty of a premise of a fired rule is calculated as follows:

1. For the portion of the premise that contains clauses $\boldsymbol{c}1$ and $\boldsymbol{c}2$ related as $\boldsymbol{c}1 \wedge \boldsymbol{c}2$, $\boldsymbol{CF} = \min(\boldsymbol{CF}_{\boldsymbol{c}1}, \boldsymbol{CF}_{\boldsymbol{c}2})$.
2. For the portion of the premise that contains clauses $\boldsymbol{c}1$ and $\boldsymbol{c}2$ related as $\boldsymbol{c}1 \vee \boldsymbol{c}2$, $\boldsymbol{CF} = \max(\boldsymbol{CF}_{\boldsymbol{c}1}, \boldsymbol{CF}_{\boldsymbol{c}2})$.
3. If the premise contains only $\boldsymbol{c}$, $\boldsymbol{CF} = \boldsymbol{CF}_{\boldsymbol{c}}$.

Each premise clause of each rule is related with a certain Sigmoid function whose parameters are on-line defined based on the contents of the database containing the extreme model deformations. The value of this function defines so-called *membership grade* (MG) of the given rule's premise. Further, the cumulative membership grade $\boldsymbol{MG}_{\boldsymbol{p}}$ of the premise $\boldsymbol{p}$ is calculated and multiplied by 100% to obtain the intensity of the activation of the AU recognized by that rule. $\boldsymbol{MG}_{\boldsymbol{p}}$ of a rule's premise is calculated from the membership grade(s) $\boldsymbol{MG}_{\boldsymbol{c}}$ associated with the premise's clause(s) $\boldsymbol{c}$.

1. For the portion of the premise that contains $\boldsymbol{c}1 \wedge \boldsymbol{c}2$, $\boldsymbol{MG}_{\boldsymbol{p}} = \boldsymbol{avg}(\boldsymbol{MG}_{\boldsymbol{c}1}, \boldsymbol{MG}_{\boldsymbol{c}2})$.
2. For the portion of the premise that contains $\boldsymbol{c}1 \vee \boldsymbol{c}2$, $\boldsymbol{MG}_{\boldsymbol{p}} = \max(\boldsymbol{MG}_{\boldsymbol{c}1}, \boldsymbol{MG}_{\boldsymbol{c}2})$.

3. If the premise contains only c, $MG_p = MG_c$.

A processing loop ends with eventual updating of the database of extreme model deformations (in the case that an encountered deformation is greater than the pertinent stored deformation) and searching for a rule that the process will try to fire in the next loop. Fast direct chaining is applied as the inference procedure. It is a breadth-first search algorithm that starts with the first rule of the knowledge base and continues with the rule whose premise-clause forms the conclusion of the fired rule.

C. Post-Processor

In the case a certain facial feature fails to be detected, the Facial Action Encoder utilizes the pertinent feature detected in the expressionless face to substitute missing data. Hence, exact information about the examined expression is lost. To avoid this, we exploit a higher level "grammar" of basic emotional expressions as defined by Ekman [6]. We expect for instance, that there is a higher possibility that a smile is coupled with "smiling" eyes than with expressionless eyes.

TABLE 1
RULES FOR DETERMINING AU CODE OF UNDETECTED FACIAL FEATURE BY EMOTIONAL EXPRESSION CLASSIFICATION

	Eyes	Eyebrows	Mouth
Sadness	7 if 1	1	15
Fear	5+7	1 if 5	20
Happiness	6	-	12
Surprise	5	1+2	26
Disgust	9	9	9
Anger	7	4	24

The post-processor of the Facial Action Encoder utilizes an existing expert system, called HERCULES [15], to classify the encoded facial actions, say $AU_1, \ldots, AU_i$, into one of the six basic emotion categories and to set a hypothesis about the possible appearance of undetected facial features. The expression's AU-coded description is adjusted upon the returned emotion label by a kind of backward reasoning of the HERCULES' inference engine (see Table 1 for the utilized rules). The CF assigned to the newly added AU is calculated according to formula (3).

$$CF_{added_AU} = 1/2 * \min\{CF_{AU_1}, \cdots, CF_{AU_i}\} \quad (3)$$

IV. Facial Expression Classifier

Nowadays psychological theories on facial expressions of emotions are ambiguous, doubtful and most of all inadequate for interpretation of each and every facial expression able to be displayed on human face [17]. Besides, most of the explicit attempts to automate facial expression emotional interpretation use culture-specific emotion terminology such as "angry" or "sad" [1], [4], [7], [8], [9], [14], [15], instead of linguistic universals such as "he means bad" or "he feels disappointed". In order to facilitate a valuable interpretation of facial expressions in a user-defined domain, we should allow the user to define his/her own facial expression interpretation space. With this in mind we developed a learning facility, which alleviates multiple quantified classification of the shown expression into the user-defined interpretation categories.

The kernel of the learning facility is the dynamic memory of experiences. The organization of the dynamic memory and the reasoning style of the Facial Expression Classifier are explained in section A. The dynamic memory performs two functions. In interpret mode, it accepts the expression identification (i.e. facial actions) determined by the Facial Action Encoder and returns the appropriate quantified interpretation label(s). In learn mode, it accepts the identification and the attributed interpretation label and adds them to its repertoire for future use. Learn mode is evoked each time a new expression is encountered or the user is not satisfied with the given interpretation. The memory is initially endowed with 52 identification-interpretation pairs learned during system's training phase. Processing of the training phase is delimited in Fig. 1 with broken arrows and explained in section B. The interpret- and learn mode are explained in section C and section D.

A. The Reasoning Style

The organization of the memory of experiences brings together Schank's theory on functional organization of human autobiographical memory [18] and the instance-based learning as a representation of these conceptual structures in the design of self-adaptive system.

Schank's model of human memory is an attempt to explain how the memories of autobiographical social events are stored, organized and remembered. Based on Schank's model, the Facial Expression Classifier is characterized by three peculiarities. Since humans are more capable of recalling experiences than articulating internal rules, the memory of experiences is derived primarily by enumeration of specific experiences. To resemble human problem solving behavior when one is confronted with a novel situation, the Expression Classifier reasons on some general similarities to come with an answer when a problem is encountered for which no specific case or rule can match exactly. To achieve this, similar cases are represented within the same dynamic memory pool, which organizes experiences according to thematic similarities (e.g. expressions affixed with the same interpretation label belong to the same memory pool). Finally, the Expression Classifier learns automatically – the dynamic memory is augmented by each additional case that is presented.

Traditional learning methods usually construct an explicit description of the target function when training examples are provided. In contrast, instance-based methods simply store the training examples [12]. This makes the instance-based methodology very suitable for design of the memory of experiences so that it resembles human autobiographical memory. The instance-based learning is also referred to as "lazy" because generalizing beyond the training examples is postponed until a new instance must be classified. When a

new query instance is encountered, a set of similar related instances is retrieved from the memory and used to classify the new instance. The "laziness" of the instance-based methods alleviates emulating of human problem solving behavior when one is confronted with a novel experience. Moreover, the "laziness" of instance-based approaches facilitates a different approximation to the target function for each distinct query instance that must be classified. This has significant advantages in the case of our target function (i.e. multiple classification of a single expression), which is very complex but can be described by a set of less complex local approximations (i.e. expression components forming the encountered facial display can be separately classified).

B. Training

It is widely believed that there are six so-called basic emotions, which are associated with distinctive facial expressions one can universally recognize [6]. The basic emotions are believed to be a part of the biological makeup of human species and to be therefore "hardwired". In contrast to this view, Ortony and Turner [13] suggested that it is not emotions themselves but components of emotions which are universally linked with some facial displays (e.g. "squared" mouth or raised eyebrows) and which can be described by culture independent linguistic universals.

The theory of Ortony and Turner forms the fundamental assumption underlying our choice of training images. Rather than using prototypic emotional expressions we are using a set of facial expression components – individual facial actions (AUs) and some combinations of these – to compose a representative set of expressions used for an initial furnishing of the dynamic memory. The choice of facial expression components (Table 2) was influenced by both, the set of facial actions that the Facial Action Encoder actually recognizes and the facial actions that characterize the six basic expressions of emotions (for the representation of the prototypic expressions in terms of AUs see [15]).

The database of training images was created with help of eight certified FACS coders. The subjects were asked to display the expressions listed in Table 2. 720×576 pixels dual-view images were acquired under constant illumination and none of the subjects had a moustache, a beard or wear glasses. Then each subject was asked to assign an index of impression to each of 364 dual-views of the other 7 subjects, reflecting his/her opinion about the distinctiveness and clarity of the judged expression when compared to the expressionless face of the pertinent subject. The displays of the 52 expressions listed in Table 2, having the highest average index of impression, were selected to make the database of training images. The chosen images are of 6 distinct subjects of both sexes, ranged in age and ethnicity.

In the training phase, 52 dual views are retrieved one by one from the training database and shown to the user together with the expressionless face of the pertinent subject. The user assigns one and only one interpretation label to each expression. The expression identification (i.e. the activated AUs) and the attributed interpretation label are then stored in the dynamic memory.

TABLE 2
THE SET OF 52 FACIAL EXPRESSION COMPONENTS; DISPLAYS OF THESE FORM THE TRAINING EXAMPLES

AU	Description	AU	Description
1	Inner brow raised	24	Lips pressed
2	Outer brow raised	26	Mouth wide
1+2	Raised eyebrows	27	Jaw dropped
4	Frown eyebrows	28	Lips sucked in
5	Eye(s) wide open	28b	B. lip sucked
6	Smiling eyes	28t	U. lip sucked
7	Eyelid(s) tensed	41	Eyelid drop
1+4+5+7	From "fear"	20+25	From "fear"
1+4+5	From "fear"	15+20	-
1+4+7	From "sadness"	15+26	-
1+5+7	From "fear"	23+17+26	From "anger"
1+4	From "sadness"	23+17	From "anger"
1+5	From "fear"	23+26	From "anger"
1+7	From "sadness"	24+17+26	From "anger"
5+7	From "fear"	24+17	From "anger"
8	Lips tensed open	24+26	From "anger"
9	Nose wrinkled	10+16+25	From "anger"
10	Upper lip raised	10+17+25	From "disgust"
12	Lip corners raised	9+17+25	From "disgust"
13	Sharp AU12	10+17	From "disgust"
15	Lip corners down	10+25	From "disgust"
17	Chin raised	9+17	From "disgust"
18	Puckered lips	9+25	From "disgust"
19	Tongue show	12+16+25	From "happy"
20	Mouth stretched	12+25	From "happy"
23	Lips tightened	16+25	From "anger"

The training phase of the system ends with indexing the training examples, i.e. classifying the training instances according to the attributed interpretation label. This results in partitioning of the dynamic memory into the expression pools. Each expression pool is a tree with "full" expression stored at the root and the parts of that expression stored at the leaves. The expression stored at the root of a certain tree is defined by the system as a collection of all distinct components of the expressions being classified into the same interpretation category. Hence, the dynamic memory organizes the experiences (expressions) according to their thematic similarity (interpretation), where each experience is represented by a set of its components (facial actions).

C. Interpret Mode

Any recurring expression is first identified by the Facial Action Encoder and the resulting list of facial actions is channeled down the expression pools until it reaches an identical event previously encountered. This results in "reminding" the interpretation of that expression. In the case that there is no identical instance stored in the memory pockets, the encountered expression is decomposed into its components by the first-nearest neighbor algorithm. The algorithm matches the facial actions of an input event against each of the stored experiences to decide the one that includes most of it. The algorithm is iterated until the

encountered expression is fully represented with a set of events stored in the expression pools. Since the training database contains displays of each and every individual facial action that the Action Encoder is actually able to recognize, the expression pools are initially endowed with those micro-events and the termination of the algorithm is therefore ensured. The interpretations coupled with the matched memory instances are displayed to the user.

The system also assigns an intensity level to (each of) the resulting interpretation label(s). The intensity coupled with an ensuing interpretation is the output on the assumption that each AU, forming a component of the "full" expression having that interpretation, has the same influence on the intensity. Thus, the ratio of matched AUs in the input expression to the number of AUs in the expression stored at the root of the pertinent expression pool decides the issue. For example, if the expression stored at the root of the "happy" pool is composed of 5 distinct AUs and 4 AUs of the input expression match the events of the "happy" pool, the intensity of the resulting "happy" label is 80%.

A certainty factor (CF) is associated with the ensuing expression interpretation label(s) as well. For each resulting interpretation label, the CF assigned to that conclusion is determined by the CFs assigned to the AUs (see formulas given in section III) forming the matched events of the pertinent expression pool (say $\boldsymbol{event}_1,\cdots,\boldsymbol{event}_m$) and an *index of confidence* (IOC) linked automatically to each instance stored in the dynamic memory of experiences.

$$\boldsymbol{CF}_{label} = \boldsymbol{avg}\{ \boldsymbol{IOC}_{event_1} * \boldsymbol{avg}\{\boldsymbol{CF}_{AU_1_of_event_1},\cdots,\boldsymbol{CF}_{AU_n_of_event_1}\},\cdots \boldsymbol{IOC}_{event_m} * \boldsymbol{avg}\{\boldsymbol{CF}_{AU_1_of_event_m},\cdots,\boldsymbol{CF}_{AU_n_of_event_m}\}\}$$

The IOC reflects our confidence in the correctness of an expression's interpretation based on "typicality" of that expression. IOC has a value from the set $\{0.1,\cdots,1\}$. We assume that an event is atypical if it is encountered in less than 1% of all query instances ($\boldsymbol{IOC} = 0.1$) and that an event is typical if it is encountered in at least 5% of all query instances ($\boldsymbol{IOC} = 1$). For this purpose, we set a global counter, counting a total number of recurring events, and local counters for each of the stored experience.

D. Learn Mode

When the system's final result is displayed to the user, i.e. the expression identification and its interpretation, the user may trigger system's learn mode if he/she is not satisfied with the given interpretation. Still, irrespectively of user's action, learn mode is triggered automatically each time the current expression or a part of it has not been encountered before. The AUs that led to a certain interpretation label are matched to the events of the pertinent expression pool and if there is no identical event, new experience is added.

If the user triggers system's learn mode, the system-user interaction starts with an explanation of the achieved result. Per interpretation label, the description of the facial actions that led to that interpretation is displayed. Then the dynamic memory is reconstructed to reflect desired modifications. Two cases can be distinguished.

1. The encountered expression or its component part has an interpretation label that the user wants to change and there is no identical event stored in the dynamic memory. The pertinent facial-action list and the newly defined label are stored in that case at an appropriate leaf of the pertinent tree. A new leaf is created "above" all leaves containing the events that form the components of the expression to be stored in the new leaf. The expression stored at the root of the pertinent tree is redefined to form a collection of all distinct components of expressions being classified into that interpretation category.
2. The encountered expression or its component part has an interpretation label that the user wants to change and there is an identical event already stored in the dynamic memory. In that case, the identical event is removed from the dynamic memory, the expression stored at the root of the pertinent tree is redefined and the processing follows the procedure defined for the first case.

The user may also combine several components of the expression having different interpretation labels and assign a single interpretation label. Here, the procedure defined for the first case is applied. If the user introduces a new category, a new expression pool is created and the expression to be learned is stored at the root of the new tree.

V. System Evaluation

Validation studies addressed the question whether the facial action encoding and its emotional interpretation acquired by the system are acceptable to human experts judging the same facial photographs. The performance of the system has been evaluated on a database containing 968 dual views. Utilized testing images have been obtained in the same manner and under same conditions the training images have been acquired. The testing images are of eight certified FACS coders displaying 2x52 expressions listed in Table 2, maximal intensity displays of AU8, AU18, AU39, AU41, 2x6 basic emotional expressions and a neutral expression.

The AU correct recognition rate achieved was 89.6% - 90% for the upper face AUs, 85% for the lower face AUs and 94% for the AUs combinations - when compared to human coding of all images in the testing database.

The average disagreement between the AU activation intensity level assigned by the system and the pertinent average of the *indexes of intensity impression* assigned by human experts was 0.08 (i.e. 8%) in the case of the correctly recognized AU with a $CF \geq 0.3$.

To validate the classification (i.e. "reminding") function of the dynamic memory, that is, whether the "correct" expression pool (i.e. interpretation category) is selected, we asked one subject to "train" the system using the database of training images. Eleven categories have been defined: disappointed, thinking (problem), surprised, happy, "please don't", "what a slimy thing", ironic, "I don't know", funny, angry and sleepy. Thereafter we asked her to label 520 selected testing images with the defined interpretation labels

(per subject we used 52 displays of the expressions listed in Table 2, 2x6 prototypic emotional expressions and the neutral expression). For 23 testing images the labeling was not consistent with that of the training images and system's classification performance has been evaluated on the rest of 497 images. In 94% of the correctly identified expressions the user approved the given interpretations. Discrepancies were encountered in classification of basic expressions of anger and sadness. Namely, the system classified those into two categories (e.g. "angry" and "thinking (problem)"), while the user used just one of these labels.

Learning function of the dynamic memory was tackled next. A set of 144 dual views different from those used in previous tests, showing arbitrary expressions displayed by our eight subjects, was presented to two subjects. Each subject trained the system first and then, for the first 72 images, triggered learn mode whenever the achieved identification was not satisfactory. The question addressed at that point was: How acceptable is the interpretation given by the system while running just in the interpret mode for the other 72 testing images? The analysis showed that 92% of the interpretations of the correctly identified expressions (i.e. twice 65 dual views in total) were approved in some measure: good 104 (80%), fair (i.e. approbation of at least half of the given labels) 16 (12.3%), poor 10 (7.7%). This result suggests that system's learning capability is rather high – after 52 training and 72 learning loops the interpretations achieved were in 92% approved by the user.

Nevertheless a deeper scrutiny, involving more testing images and a larger group of users, is necessary for establishing a confident measure of system's learning and interpretative performance. Also, how acceptable is the intensity of an interpretation label is still to be examined.

VI. Conclusion

In comparison to the existing approaches to automatic facial action coding [2], [1], [7], and facial expression emotional interpretation [1], [4], [8], [9], [14], [15], the system presented in this paper is fundamentally different. First it brings together diverse theories and technologies - FACS [5], studies on facial expression of emotion [6], [13], image analysis and various AI methodologies [18], [12]. Second, it is the only system we are aware of that can robustly perform both: encode and quantify 30 different facial actions in a dual view facial image and interpret encountered expression in terms of multiple quantified user-defined labels.

By a high number of experiments, we obtained confident measurements that indicate rather accurate quantified expression identification accomplished by the system. We established as well that the system performs a generally approved expression classification in multiple interpretation categories defined by the user. Validation of the recall and learning functions of the dynamic memory suggests that system's learning and interpretative capability is rather high. Additional testing procedures are currently performed.

An obvious limitation of the system is its incapability to identify all visually distinguishable facial action. As a result, the expressions unlike in terms of the displayed facial actions might have an identical interpretation. Modeling the facial motion and adding new features' detectors could improve the system performance in this sense.

Other limitations of the system are time-consuming image processing (processing of single dual view takes in average 3 minutes) and incapability to process images of faces with facial hair or glasses. The system should be enhanced in these terms if it is to be used as a part of a realistic man-machine interface.

VII. References

[1] M.J. Black and Y. Yacoob, "Recognising Facial Expressions in Image Sequences using Local Parameterised Models of Image Motion", *International Journal on Computer Vision*, vol. 25(1), pp. 23-48, 1998.

[2] J.F. Cohn, A.J. Zlochower, J.J. Lien and T. Kanade, "Feature-Point Tracking by Optical Flow Discriminates Subtle Differences in Facial Expression", *Proc. FG*, pp. 396-401, 1998.

[3] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman and T.J. Sejnowski, "Classifying Facial Actions", *IEEE Trans. PAMI*, vol. 21(10), pp. 974-989, 1999.

[4] G.J. Edwards, T.F. Cootes and C.J. Taylor, "Face Recognition Using Active Appearance Models", *Proc. European Conf. Computer Vision* 2, pp. 581-695, 1998.

[5] P. Ekman and W.V. Friesen, *Facial Action Coding System (FACS)*. Palo Alto: Consulting Psychologists Press, 1978.

[6] P. Ekman, *Emotion in the Human Face*. Cambridge University Press, 1982.

[7] I.A. Essa and A.P. Pentland, "Coding Analysis Interpretation and Recognition of Facial Expressions", *IEEE Trans. PAMI*, vol 19(7), pp. 757-763, 1997.

[8] H. Hong, H. Neven and C. von der Malsburg, "Online Facial Expression Recognition based on Personalised Galleries", *Proc. FG'98*, pp. 354-359, 1998.

[9] C.L. Huang and Y.M. Huang, "Facial Expression Recognition Using Model-Based feature Extraction and Action Parameters Classification", *Journal of Visual Communication and Image Representation*, vol. 8(3), pp. 278-290, 1997.

[10] M. Kass, A. Witkin and Terzopoulos, "Snake: Active Contour Model", *Proc. 1st ICCV*, pp. 259-269, 1987.

[11] A. Mehrabian, "Communication without words", *Psychology Today*, vol. 2(4), pp. 53-56, 1968.

[12] T.M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

[13] A. Ortony and T.J. Turner "What's basic about basic emotions ?", *Psychological Review*, vol. 97, pp. 315-331.

[14] T. Otsuka and J. Ohya, "Spotting Segments Displaying Facial Expression from Image Sequences Using HMM", *Proc. FG'98*, pp. 442-447, 1998.

[15] M. Pantic and L.J.M. Rothkrantz, "Expert system for automatic analysis of facial expressions", *Image and Vision Computing Journal*, vol. 18(11), pp. 881-905, 2000.

[16] M. Pantic and L.J.M. Rothkrantz, "An Expert System for Recognition of Facial Actions and Their Intensity", *Proc. Int'l Conf. Innovative Applications of AI*, Austin, August 1st –3rd,2000.

[17] J.A. Russell, "Facial Expression of Emotion: What Lies Beyond Minimal Universality?", *Psychological Bulletin*, vol. 118(3), pp. 379-391, 1995.

[18] R.C. Schank, "Memory based expert systems". Internal report, AFOSR.TR.84-0814, Yale University, Computer Science Department, New Haven, CT, 1984.

[19] J.M. Vincent, D.J. Myers and R.A. Hutchinson, “Image feature location in multi-resolution images”, *Neural Networks for Speech, Vision and Natural Language*, pp. 13-29, 1992.