CASE-BASED REASONING FOR USER-PROFILED RECOGNITION OF EMOTIONS FROM FACE IMAGES

Maja Pantic and Leon Rothkrantz*
Delft University of Technology
EEMCS / Mediamatics Dept.
Delft, the Netherlands
{M.Pantic,L.J.M.Rothkrantz}@ewi.tudelft.nl

ABSTRACT

Most systems for automatic analysis of facial expressions attempt to recognize a small set of "universal" emotions such as happiness and anger. Recent psychological studies claim, however, that facial expression interpretation in terms of emotions is culture dependent and may even be person dependent. To allow for rich and sometimes subtle shadings of emotion that humans recognize in a facial expression, user-profiled recognition of emotions from images of faces is needed. In this work, we introduce a case-based reasoning system capable of classifying facial expressions (given in terms of facial muscle actions) into the emotion categories learned from the user. The utilized case base is a dynamic, incrementally self-organizing event-content-addressable memory that allows fact retrieval and evaluation of encountered events based upon the user preferences and the generalizations formed from prior input. Two versions of a prototype system are presented: one aims at recognition of six "universal" emotions and the other aims at recognition of affective states learned from the user. Validation studies suggest that in 100%, respectively in 97% of the test cases, interpretations produced by the system are consistent with those of the two users who trained the two versions of the prototype system.

1. INTRODUCTION

The ability to detect and understand affective states of a person we are communicating with is the core of emotional intelligence [1]. Emotional intelligence is a facet of human intelligence that has been argued to be indispensable and even the most important for a successful social life [2]. When it comes to computers, however, not all of them will need emotional intelligence and none will need all of the related skills that we need. Yet man-machine interactive systems capable of sensing stress, inattention, and heedfulness, and capable of adapting and responding appropriately to these affective states of the user are likely to be perceived as more natural, more efficacious and more trustworthy [3]. The research area of machine analysis and employment of human affective states to build more natural interfaces goes by a general name of affective computing.

Since facial expressions are the naturally preeminent means for humans to communicate emotions [4, 5], automatic recognition of emotions from face images has become a central topic in affective computing. Virtually all systems for automatic facial affect analysis attempt to recognize a small set of universal/basic emotions: fear, happiness, sadness, disgust, surprise, and anger [6]. This practice follows from the work of Darwin and more recently Ekman [5], who suggested that these six basic emotions have corresponding prototypic facial expressions. Yet alternative psychological studies argue for culture dependency; they claim that the comprehension of a given emotion label and the expression of the related affective state is culture dependent and may even be person dependent [4]. Also, in everyday life, prototypic expressions of emotions occur relatively infrequently. Typically shown facial expressions convey conversational signals and signs of attitudinal states such as interest and boredom that are usually displayed as one or few facial actions such as raising the eyebrows in disbelieve [4]. To allow for these rich shadings of emotion that humans detect in the face, user-defined choices must be made regarding the selection of affective states to be recognized by an automated facial affect analyzer.

This work describes an automatic facial expression recognition system that performs classification of facial actions (which produce expressions) into the emotion categories learned from the user. The proposed method is based on Facial Action Coding System (FACS) [7]. FACS is a system designed for human observers to code any anatomically possible facial expression in terms of 44 action units (AUs), each of which corresponds to a specific visually observable facial muscle action. Hence, AUs can be seen as being analogous to phonemes for facial expression. Several methods for automatic AU detection from face images were reported [8, 9]. Though our facial expression recognition system can utilize any of these methods to detect the AUs that produced the expression shown in an examined face image, it employs the AU detector proposed in [10]. We did so because the chosen method can detect 29 AUs while other existing automated AU detectors, at best, can detect 16 to 20 AUs [8, 9].

Since AUs can occur in more than 7000 combinations, the classification of AUs in an arbitrary number of emotion categories learned from the user is an extremely complex problem. To tackle this problem, one can apply either eager or lazy learning methods. Eager learning methods such as neural networks extract as much information as possible from training data and construct a general approximation of the target function. Lazy learning methods such as case-based reasoning simply store the presented data and generalizing beyond these data is postponed until an explicit request is made. When a query instance is encountered, similar related instances are retrieved from the memory and used to classify the new instance. Hence, lazy methods have the option of selecting a different local approximation of the target function for each presented query instance. Eager methods using the same hypothesis space are more restricted because they must choose their approximation before presented queries are observed. In turn, lazy methods are usually more appropriate for complex and incomplete problem domains than eager methods, which replace the training data with abstractions obtained by generalization and which, in turn, require excessive amount of training data. Therefore, we chose to achieve classification of the AUs detected in an input face image into the emotion categories learned from the user by case-based reasoning about the content of a dynamic memory. The memory is dynamic in the sense that, besides generating facial expression interpretations by analogy to those accompanying similar expressions "experienced" in the past, it is able to learn new

* The work of Maja Pantic is supported by the Netherlands Organization for Scientific Research (NWO) Grant EW-639.021.202.

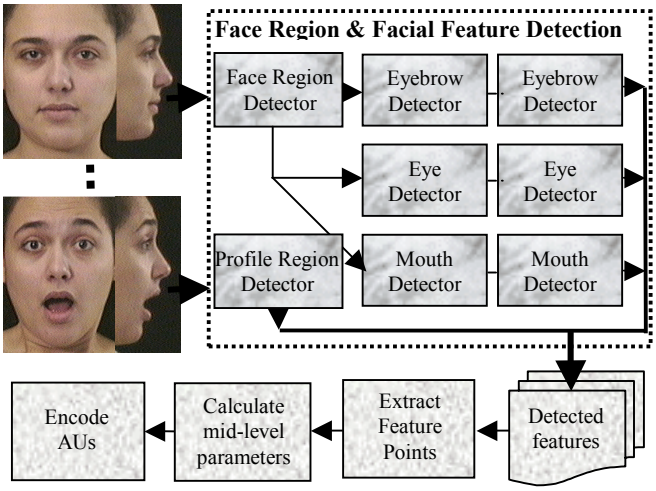

Fig. 1: Outline of the utilized automated AU-detector

emotion labels and associated AUs, thereby increasing its expertise in user-profiled expression interpretation. AU detection, dynamic memory organization, case-based reasoning, and experimental evaluation are explained in sections 2, 3, 4, and 5.

2. FACIAL ACTION DETECTION

Before any facial expression recognition can be achieved, the best features to describe the anatomical phenomena should be defined and then extracted. Facial muscle actions (i.e., AUs of the FACS system) represent the best possible choice of features since they produce changes in the appearance of the facial components (eyes, lips, etc.), which go by a general name of facial expressions. To represent the shown facial expression in terms of the facial muscle actions that produced it, we employ an automated system that we developed to recognize 29 AUs occurring alone or in combination in a static frontal- and/or profile-view color face image [10].

Fig. 1 outlines the employed AU detector. First, an image of an expressionless face of the observed subject is processed. Each subsequent image of that subject is processed under the assumption that all of the input images acquired during the same monitoring session with the pertinent subject are non-occluded, scale-, and orientation-invariant face images. If the frontal-view of the face is available, the face region is extracted from the input frontal-view face image. If the profile-view of the face is available, the face-profile region is extracted from the input profile-view face image. To do so, watershed segmentation with markers is applied on the morphological gradient of the input color image. For the frontal view, the segmented face region is subjected to a multi-detector processing which, per facial component (eyes, eyebrows, mouth), generates a spatial sample of its contour. A set of 19 frontal face feature points (Fig. 2) is then extracted from the spatially sampled contours of the facial features. For the profile view, 10 feature points (Fig. 2) are extracted from the contour of the segmented face-profile region. Subtle changes in the appearance of the facial components are measured next. Motivated by AUs of the FACS, these changes are represented as a set of mid-level parameters describing the motion of the feature points (up, down, left, right), the increase and decrease of the distances between feature points, and the shapes formed by certain feature points (linear, parabolic). Based upon these mid-level parameters, a rule-based algorithm interprets the extracted information in terms of 29 AUs occurring alone or in combination. For details about this method, see [10].

3. DYNAMIC MEMORY OF EXPERIENCES

The utilized dynamic memory of experiences is based on Schank's theory of functional organization of human memory of experiences [11]. According to this theory, for a certain event to remind one spontaneously of another, both events must be represented within the same dynamic chunking memory structure, which organizes the experienced events according to their thematic similarities. Both events must be indexed further by a similar explanatory theme that has sufficient salience in the person's experience to have merited such indexing in the past. Indexing, in fact, defines the scheme for retrieval of events from the memory. The best indexing is the one that will return events most relevant for the event just encountered.

In the case of the utilized memory of experiences, each event is one or more micro-events, each of which is a set of AUs displayed with the goal of communicating the affective state of the person. Micro-events related by the goal of communicating one specific affective state are grouped within the same dynamic memory chunk. In other words, each memory chunk represents a specific emotion category and contains all micro-events to which the user assigned the emotion label in question. The indexes associated with each dynamic memory chunk comprise individual AUs and AU combinations that are most characteristic for the emotion category in question. Finally, the micro-events of each dynamic memory chunk are hierarchically ordered according to their typicality: the larger the number of times a given micro-event occurred, the higher its hierarchical position within the given chunk.

Before any case-based reasoning about the content of a dynamic memory can be achieved, the cases that will constitute the dynamic memory (i.e., the case base) should be obtained. One approach of achieving this is to generate the case base from scratch through on-line interaction with the user. The other approach is to start with an initial furnishing of the case base, by utilizing the available data on the subject problem, and then incrementally enhance the case base by learning from the user about newly encountered cases. Complex adaptation procedures make case-based reasoning systems more difficult to build and to maintain and significantly reduce user's confidence in the system since faulty adaptations are encountered more often due to incompleteness of knowledge. Therefore, we favor the second approach and for each novel user we perform an initial endowment of the memory using 40 cases. The choice of the related 40 micro-events (Table 1) has been influenced by both the 29 AUs that the utilized AU detector can encode from a frontal- and a profile-view of the face and the components of expressions (i.e., micro-events) that might be hardwired to emotions [4].

The initial endowment of the dynamic memory is accomplished further by asking the user to associate an interpretation (emotion) label to each of the 40 facial expressions picturing the 40 micro-events listed in Table 1 (see Fig. 3 for examples of this stimulus material). This process is repeated until the consistency of the

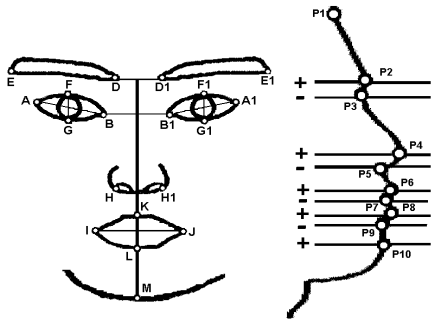

Fig. 2: Facial feature points

Table 1: 40 AU combinations used for initial endowment of the dynamic memory of experiences

AUs		*AUs*	
1	raised inner eyebrow	6+13	from "happiness"
2	raised outer eyebrow	15	depressed lip corners
1+2	from "surprise"	15+17	from "sadness"
4	furrowed eyebrows	16+25	from "anger"
5	raised upper eyelid	17	raised chin
7	raised lower eyelid	18	puckered lips
1+4+5+7	from "fear"	19+26	showed tongues
1+4+5	from "fear"	20	horiz. stretched mouth
1+4+7	from "sadness"	23	tightened lips
1+5+7	from "fear"	24	pressed lips
1+4	from "sadness"	24+17	from "anger"
1+5	from "fear"	27	vert. stretched mouth
1+7	from "sadness"	28+26	sucked lips
5+7	from "fear"	28t+26	sucked upper lip
8+25	lips tensed open	28b+26	sucked lower lip
9	wrinkled nose	29	jaw forward
9+17	from "disgust"	35+26	sucked cheeks
10	raised upper lip	36t+26	tongue under upper lip
10+17	from "disgust"	36b+26	tongue under lower lip
6+12	from "happiness"	41	lowered upper eyelid

labeling is ensured. Namely, if a certain facial expression has been labeled in the second round differently than in the first, the user is asked to label anew the facial expression in question. A number of *n*-tuple vectors, one for each emotion label defined by the user, are generated eventually. Except of the emotion label, each of those vectors contains all the relevant cases, that is, all the micro-events that the user interpreted with the emotion label in question (e.g., *["surprise", AU1+AU2, AU27]*). From each n-tuple vector and, hence, for each engendered interpretation category, 3 vectors are defined: ***index***, ***label***, and ***cases***. The ***label*** vector holds the label associated with the interpretation category in question (e.g., ***label****["surprise"]*). The ***cases*** vector contains all the relevant cases, each of which is initially assigned the typicality equal to zero (e.g., ***cases****[(AU1+AU2, 0), (AU27, 0)]*). For the ***index*** vector to contain only the AUs and AU combinations that characterize the given interpretation category, it is derived from ***cases*** by excluding each AU combination whose component AUs are also cases in their own right. For example, if the ***cases*** vector is ***cases****[(AU1+AU2, 0), (AU1, 0), (AU27, 0)]* the ***index*** vector will be ***index****[AU1, AU27]*.

4. CASE-BASED REASONING

The classification of the AUs detected in an input face image into the emotion categories learned from the user is achieved by case-based reasoning about the content of the dynamic memory. To solve a new problem of classifying a set of input AUs into the user-defined interpretation categories, the following steps are taken:

1. Search the dynamic memory for similar cases, retrieve them, and interpret the input set of AUs using the interpretation labels suggested by the retrieved cases.

Fig. 3: Sample stimulus images used for initial endowment of the case base. Left to right: AU1, AU5, AU6+AU13, AU9+AU17

Consider a list of input AUs, the ***AUs-list***. To classify this input into the user-defined emotion categories, i.e., to create a list of solutions, the ***solutions*** list, do:

1. Create the following empty lists: ***clusters***, ***cases-list***, ***best-cases***, and ***solutions***. Go to 2.

2. Match the AUs of ***AUs-list*** with the AUs of ***index*** vectors of the emotion chunks constituting the case base. Each time a match is established, exclude the matching AUs from ***AUs-list***, add ***label*** of the chunk in question to ***clusters***, add all ***cases*** of that chunk to ***cases-list***. If ***AU-list*** is empty, go to 3.

3. Re-establish ***AUs-list***. Examine the cases of ***cases-list***. If the current case is composed of AUs that belong to ***AUs-list***, add it to ***best-cases***. Go to 4.

4. Find the longest AU-combination, and then with the highest typicality, in ***best-cases***. Match the found case with AUs of ***AUs-list***. Each time a match is established, exclude the matching AUs from ***AUs-list***, find ***label*** related to the matched case in ***clusters***, add that ***label*** and the case to ***solutions***. If ***AU-list*** is empty, redefine ***solutions*** so that the AUs related to a specific label are grouped. E.g., if ***solutions*** is *["surprise", (AU1+AU2, 21), "surprise", (AU27, 18), "happy", (AU6+AU12, 35)]*, redefined ***solutions*** will be *[("surprise", AU1, AU2, AU27), ("happy", AU6, AU12)]*.

Fig. 4: Retrieval algorithm

2. If the user is satisfied with the generated interpretation, store the case in the dynamic memory. Otherwise, adapt the memory according to user-provided feedback on the interpretation he associates with the input facial expression.

The simplest form of retrieval is to apply the first nearest neighbor algorithm, that is, to match all cases of the case base and return a single best match. This method is usually to slow. A pre-selection of cases is therefore usually made based on the indexing structure of the utilized case base. Our retrieval algorithm employs a pre-selection of cases that is based upon the clustered organization of the dynamic memory, the indexing structure of the memory, and the hierarchical organization of cases within the clusters/ chunks according to their typicality. The algorithm is given in Fig. 4.

A successful termination of the retrieval algorithm, resulting in the classification of input expression into the user-defined classes, is ensured. This is because the case base is initially endowed with the 40 cases listed in Table 1, which comprehend each and every AU that the utilized AU detector is able to encode from an input face image. In other words, the dynamic memory is initialized with each and every micro-event that can possible be encountered.

Each time the user is not satisfied with the interpretation produced by the retrieval algorithm and renders his feedback on the issue, the case base is reconstructed according to the wishes of the user. The adaptation algorithm given in Fig. 5 does this.

1. If the set of AUs ***a1 + ... + aN*** for which a novel interpretation label "***x***" has been introduced (this can be only a part of the original input expression) matches exactly a specific case stored in the case base, the old case is removed and the case base is reconstructed to reflect these changes. Go to 2. Otherwise, go to 2 as well.

2. Match label "***x***" with ***label*** of each interpretation category chunk constituting the dynamic memory. If chunk "***x***" already exists, go to 3. Otherwise, generate a new chunk "***x***" using the following vectors: ***label****["x"]*, ***cases****[(a1 + ... + aN, 1)]*, and ***index****[(a1 + ... + aN)]*. Terminate the execution of this procedure.

3. Reconstruct the dynamic memory of experiences by adding the new case ***a1 + ... + aN*** to chunk "***x***". Add *(a1 + ... + aN, 1)* to ***cases***. Redefine ***index*** vector to contain only the AUs and AU combinations that characterize the interpretation category "***x***" (i.e., derive it from ***cases*** by excluding each AU combination whose component AUs are also cases in their own right).

Fig. 5: Adaptation algorithm

5. EXPERIMENTAL EVALUATION

Validation studies on a prototype system addressed the question of whether the facial expression interpretations generated by the system were acceptable to human observers judging the same face images. They were carried out using a face-image database containing 560 dual-views (combined frontal- and profile-views) of faces, acquired by 2 head-mounted digital PAL cameras (e.g., Fig. 1). This camera setting ascertained the assumption, adopted by the AU detector [10], that all of the input images acquired during the same monitoring session with one subject are non-occluded, scale-, and orientation-invariant face images. The utilized images were of 8 young subjects of both sexes and of European, Asian, or South American ethnicity. The subjects were asked to display series of facial expressions that included individual AUs and AU combination. Metadata were associated with the acquired images given in terms of AUs scored by two FACS coders. As the actual test data set, we used 454 images for which the coders agreed about the displayed AUs.

The aim of the 1st validation study was to measure the agreement between the human judgments of these 454 test images and those generated by the utilized AU detector. The result of the comparison is given in Table 2. For further details about this validation study, see [10]. The objective of the 2nd and 3rd validation study was to evaluate the performance of the case-based reasoning utilized by the system. The question addressed by the 2nd validation study was: How acceptable are the interpretations given by the system, after it is trained to recognize 6 basic emotions? The question addressed by the 3rd validation study was: How acceptable are the interpretations given by the system, after it is trained to recognize arbitrary number of user-defined interpretation categories? In the first case, a human FACS coder was asked to train the system. In the second case, a lay expert, without formal training in emotion signals recognition, was asked to train the system. For each case, the interpretation categories defined during the initial training of the system are given in Table 3. The same expert used to train the system was used to evaluate its performance. 392 images, which were correctly AU-coded by the AU detector (Table 2), were used for this purpose. Of those, 196 were used for further training (i.e., whenever the interpretation given by the system was not satisfactory, the expert was asked to provide a novel interpretation). In the case of the user-defined interpretation categories, this subsequent training resulted in addition of another 3 interpretation categories: bored, monkey face, and delighted. The experts judged finally the acceptability of interpretations returned by the system over the set of 196 face images that were not previously used to train the system. For basic emotions, in 100% of 196 test cases the expert approved of the interpretations generated by the system. For user-defined interpretation categories, in 83% of test cases the lay expert approved entirely of the interpretations and in 14% of test cases the expert approved of most but not of all the interpretation labels generated by the system for the pertinent cases.

Table 2: AU recognition results. Upper face AUs: AU1, AU2, AU4-AU7, AU41. AUs affecting the nose: AU9, AU38, AU39. AUs affecting the mouth: AU8, AU10, AU12, AU13, AU15, AU16, AU18-AU20, AU23-AU25, AU28, AU35, AU36. AUs affecting the jaw: AU17, AU26, AU27, AU29. # denotes the number of images. C denotes correctly recognized images. MA denotes the number of images in which some AUs were missed or they were scored in addition to those depicted by human experts. IC denotes incorrectly recognized images.

	#	C	MA	IC	Rate
upper face	454	422	32	0	93.0%
nose	454	443	10	1	97.6%
mouth	454	423	28	3	93.2%
jaw	454	436	17	1	96.0%
all 29 AUs	454	392	58	4	**86.3%**

Table 3: Interpretation categories defined by two experts during the initial endowment of the dynamic memory

AUs	*Expert 1*	*Expert 2*	*AUs*	*Expert 1*	*Expert 2*
1	sadness	disappoint	6+13	happiness	ironic
2	anger	angry	15	sadness	"don't know"
1+2	surprise	surprised	15+17	sadness	"don't know"
4	anger	angry	16+25	anger	angry
5	fear	"don't!"	17	sadness	"don't know"
7	anger	thinking	18	no basic	thinking
1+4+5+7	fear	"don't!"	19+26	no basic	funny
1+4+5	fear	"don't!"	20	fear	"don't know"
1+4+7	sadness	disappoint	23	anger	thinking
1+5+7	fear	"don't!"	24	anger	angry
1+4	sadness	disappoint	24+17	anger	angry
1+5	fear	"don't!"	27	surprise	surprised
1+7	sadness	disappoint	28+26	no basic	thinking
5+7	fear	"don't!"	28t+26	no basic	thinking
8+25	anger	angry	28b+26	no basic	thinking
9	disgust	"yak!"	29	no basic	funny
9+17	disgust	"yak!"	35+26	no basic	thinking
10	disgust	"yak!"	36t+26	no basic	funny
10+17	disgust	"yak!"	36b+26	no basic	thinking
6+12	happiness	glad	41	no basic	sleepy

6. CONCLUSIONS

In this paper we presented a new facial expression recognition system that performs classification of facial muscle actions (i.e., AUs that produce facial expressions) into the emotion categories learned from the user. Given that the previously reported facial expression analyzers are able to classify facial expressions only in one of the 6 basic emotion categories, the method proposed here extends the state of the art in the field by enabling facial expression interpretation in a user-adaptive manner. By a number of experimental studies, we demonstrated that the facial expression interpretation achieved by the system is rather accurate. However, additional field trials (i.e., more lay experts) and more elaborate quantitative validation studies are necessary to confirm this finding.

REFERENCES

[1] D. Goleman, *Emotional Intelligence*, Bantam Books, 1995.

[2] P. Salovey and J.D. Mayer, "Emotional intelligence", *Imaginat. Cogn. Personality*, vol. 9, no. 3, pp. 185-211, 1990.

[3] J. Cassell, T. Bickmore, "External manifestations of trust-worthiness in interface", *Communications of the ACM*, vol. 43, no. 12, pp. 50-56, 2000.

[4] J. Russell and J. Fernandez-Dols, *The psychology of facial expression*, Cambridge University Press, 1997.

[5] D. Keltner and P. Ekman, "Facial expression of emotion", *Handbook of Emotions*, Guilford Press, pp. 236-249, 2000.

[6] M. Pantic, L.J.M. Rothkrantz, "Toward an affect-sensitive multimodal HCI", *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370-1390, 2003.

[7] P. Ekman and W. Friesen, *Facial Action Coding System*, Consulting Psychologist Press, 1978.

[8] Y. Tian, et al., "Recognizing action units for facial expression analysis", *IEEE TPAMI*, vol. 23, no. 2, pp. 97-115, 2001.

[9] M. Pantic et. al., "Facial action recognition in face profile image sequences", *IEEE ICME*, pp. 37-40, 2002.

[10] M. Pantic, L.J.M. Rothkrantz, "Facial Action Recognition for Facial Expression Analysis from Static Face Images", *IEEE TSMC-B*, to appear.

[11] R.C. Schank, "Memory based expert systems", AFOSR.TR. 84-0814, Comp. Science Dept., Yale University, 1984.