J Ambient Intell Human Comput (2013) 4:121–140
DOI 10.1007/s12652-012-0107-1

ORIGINAL RESEARCH

Human emotion and cognition recognition from body language of the head using soft computing techniques

Yisu Zhao · Xin Wang · Miriam Goubran · Thomas Whalen · Emil M. Petriu

Received: 15 January 2011 / Accepted: 13 February 2012 / Published online: 28 February 2012

Abstract To make a computer interface more usable, enjoyable, and effective, it should be able to recognize emotions of its human counterpart. This paper explores new ways to infer the user's emotions and cognitions from the combination of facial expression (happy, angry, or sad), eye gaze (direct or averted), and head movement (direction and frequency). All of the extracted information is taken as input data and soft computing techniques are applied to infer *emotional and cognitional* states. The fuzzy rules were defined based on the *opinion of an expert in* psychology, a pilot group and annotators. Although the creation of the fuzzy rules are specific to a given culture, the idea of integrating the different modalities of the body language of the head is generic enough to be used by any particular target user group from any culture. Experimental results show that this method can be used to successfully recognize 10 different emotions and cognitions.

Keywords Emotion recognition · Cognition recognition · Body language of the head · Fuzzy inference systems · Human–computer intelligent interaction

Y. Zhao (✉) · X. Wang · M. Goubran · T. Whalen · E. M. Petriu
University of Ottawa, Ottawa, ON, Canada
e-mail: ysusanzhao@gmail.com

X. Wang
e-mail: xinwang@discover.uottawa.ca

M. Goubran
e-mail: mgoub019@uottawa.ca

T. Whalen
e-mail: thom@thomwhalen.com

E. M. Petriu
e-mail: petriu@site.uottawa.ca

1 Introduction

When people interact with each other, they not only listen to what the other says, they react to facial expressions, gaze direction, and head movement. Human–computer interaction would be enhanced in a friendly and non-intrusive way if computers could understand and respond to users' body language in the same way.

However, current human–computer interfaces (usually involving traditional keyboard-based and mouse-based interfaces) ignore implicit information from the users and thus do not provide a natural, human-like interaction. To achieve effective human–computer intelligent interaction (HCII), a computer should be able to interact with its human counterpart in a fashion similar to human–human interaction. Specifically, HCII interfaces should detect changes of a user's head and facial movements to infer changes in emotional or cognitional state. For example, unobtrusively detecting and responding to students' level of frustration would allow a computer-based training system to take action to increase motivation and increase learning (Picard 1997).

Human–human interaction provides the groundwork for developing ways for computers to recognize emotional expression automatically. Humans interact naturally with each other to express emotions and feelings through non-verbal modalities (such as facial expressions, hand gestures, body postures, gaze direction, and so forth). They use these simultaneously to complement and enhance spoken communication. Both psychologists and engineers have tried to analyze these modalities in an attempt to understand and categorize emotions for HCII. This emerging field is also attracting increased attention from researchers in computer science, cognitive science, neuroscience, and related disciplines.

Most existing methods in machine analysis of human emotion have focused on the recognition of the six basic facial expressions (happiness, sadness, surprise, fear, anger, and disgust) (Pantic and Bartlett 2007; Gunes and Piccardi 2009; Zeng et al. 2009). However, detecting these six universal, basic emotions from facial expressions is only the tip of the iceberg of emotion recognition. There are far more secondary and tertiary emotions than the basic facial expressions. Recognizing these subtle secondary emotions is critical for HCII. In addition, researchers have shown that cognitive mental states (e.g., agreement, disagreement, concentrating, thinking, and interest) occur more often in everyday interactions than the six basic facial expressions (Baron-Cohen and Tead 2003).

Although emotion is universal, the way humans express it is culture-based (Marsh et al. 2003; Matsumoto 1989; Matsumoto et al. 1999). People from different cultures, or even different individuals within a culture, can express nonverbal communication in quite different manners. For example, nodding the head up and down is conveys different messages in different parts of the world. In North America, it means "I agree." But in some cultures, like parts of Greece, Yugoslavia, Bulgaria, and Turkey, a nodding head means "no." In a conversation among Japanese, it often simply means "I am listening." In most cases, head nodding is associated with positive emotions, whereas head shaking is associated with negative emotions. However, for some individuals, high-frequency head nodding is likely to occur with resentment, which is a negative emotion.

Humans express emotions through simultaneous combinations of nonverbal acts including facial expressions, head or eye movements, hand signals, body postures and so forth. Attempting to infer emotion from facial expression alone does not always lead to an accurate conclusion. In particular, psychological research shows that the direction of eye gaze and head movement has a strong effect on facial expression (Reginald et al. 2005, 2009; Ganel 2011). When a facial expression is accompanied by other nonverbal cues, entirely different emotions must be inferred. For example, the smiling facial expression usually means happiness. However, if the person is smiling while shaking his or her head at the same time, this means that he or she is not happy. In this case, the head movement is the more important variable in emotion recognition and negates the facial expression. Therefore, considering other modalities is crucial for analyzing emotions.

Although researchers commonly advocate combining information from different modalities, few systems have been implemented (Zeng et al. 2006, 2009; Calvo and D'Mello 2010). Recently, more effort has been made to integrate facial expression with body language to enhance emotion recognition (Balomenos et al. 2005; Gunes and Piccardi 2005a, b; Valstar et al. 2007; Baenziger et al. 2009). Scherer and Ellgring (2007) considered the possibility of combining speech and facial features with body movement (posture and gesture) to discriminate among 14 different emotions. Castellano et al. (2008) considered the possibility of detecting eight emotions (some basic emotions plus irritation, despair, etc.) by examining facial features, speech contours, and gestures. Kapoor and Picard (2005) developed a context-based system to predict a child's interest level on the basis of face, body postures, context information, skin conductance, and a pressure-sensitive mouse. Arroyo et al. (2009) took into consideration the integration of context, facial features, seat pressure, galvanic skin conductance, and pressure mouse to recognize levels of cognitions and emotions (frustration, excitement, confidence, and interest) of students in naturalistic school settings. D'Mello and Graesser (2010) considered a combination of facial features, body language, and speech cues for detecting some of the learning-centered emotional states.

Psychological researchers have found that body language of the head can provide crucial information for detecting and interpreting emotions (Reginald et al. 2005, 2009; Ganel 2011). The information we gathered from multi-modal observations of the human head has provided more information about emotional state than a single facial channel. Some studies investigated the fusion of facial expressions and head movements (Cohn et al. 2004; Ji et al. 2006; Zhang and Ji 2005, 2006; Valstar et al. 2007). El Kaliouby and Robinson (2004) looked into the use of facial expressions and head gestures to detect cognition. They described a system to inferring complex cognitive states (agreeing, concentrating, disagreeing, interested, thinking and unsure). Asteriadis et al. (2009) estimated the behavioral state of the user based on information extracted from head, eye and hand movements. Their method can detect attention states of users while they are reading. Detailed and thorough reviews of the state of the art of emotion detection through multimodal approaches can be found in (Calvo and D'Mello 2010; Zeng et al. 2009; Gunes and Pantic 2010; Sebe et al. 2005; Jaimes and Sebe 2007).

Despite these efforts, no previous research has simultaneously considered all modalities of the face and head to infer emotions and cognitions. The goal of this paper is to improve HCII by showing how to integrate information from the body language of the head to infer emotional and cognitive states. In particular, we concentrated on the combination of facial expression, eye gaze and head movement. To analyze head movements, we not only observe the direction but also the frequency of head movement. Eye gaze also plays an important role in emotion detection. An eye gaze detector was proposed to

analyze whether the subject's gaze direction was direct or averted. Since emotions have fuzzy boundaries (Calvo and D'Mello 2010), we considered all of the information extracted as input for soft computing techniques (Gegov 2007; Zadeh 1973; Jang 1993). The fuzzy rules are derived from the opinion of an expert in psychology, observations of a pilot group, and annotators. Although the creation of the fuzzy rules are specific to a given culture, the idea of integrating the different modalities of the body language of the head is generic enough to be adapted to target user groups from other cultures. Experimental results show that our method can be used to successfully recognize 10 different emotions and cognitions.

This paper is organized into five sections. Section 2 describes the detection of the head movement by observing the direction of nostril movements. Section 3 provides detailed methods of pupil detection and eye gaze analysis. Section 4 describes the general modal for emotion and cognition detection, input variables and their quantification, generation of fuzzy rules, and output variables. Section 5 gives detailed experimental procedures and results for each of the proposed methods as well as the fuzzy inference system (FIS).

2 Head movement detection

The goal of this section is to determine whether head movement is nodding, shaking, or stationary (no movement). Nodding and shaking are very common head gestures that can send and receive messages and express emotions and cognitions. In this paper, we propose a simple, fast, and effective automatic head movement detection approach by simply using a webcam. The general procedure of head movement detection is shown in Fig. 1 below.

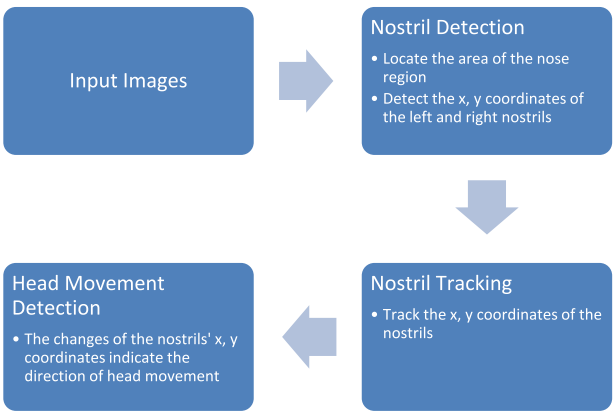


Fig. 1 General steps of proposed head movement detection

2.1 Face and nose region detection

The first step of head movement recognition is to detect the face region of the input image. In order to automatically capture the human face, we adopted a fast and robust face detection algorithm proposed by Viola and Jones (2001) using *Haar-like* feature-based AdaBoost classifiers. In the human face, nostrils are relatively static feature points that will not be significantly affected by different facial expressions. Therefore, the coordinates of these points can be used to track and predict head movements. We focused on the nose region instead of the whole face image to detect feature points. There are two advantages of focusing on the nose region: (1) It can eliminate interruption from other facial features, and (2) incorrect detection of nostrils does not affect the recognition of head movement since any facial features within the nose region are relatively static in the face region and will not be affected by facial expressions.

We obtained the nose region by geometrical proportion of the human face since the position of the nose region is always approximately 2/4–3/4 of the height and 2/7–5/7 of the width of the face. Let (x_1, y_1) be the coordinates on the upper left corner of the face region, and (x_2, y_2) be the coordinates on the bottom right corner of the face region. We have:

$$X_L = 2 * (x_2 - x_1)/7,\ Y_L = 2 * (y_2 - y_1)/4 \quad (1)$$

$$X_R = 5 * (x_2 - x_1)/7,\ Y_R = 3 * (y_2 - y_1)/4 \quad (2)$$

where X_L and Y_L are the coordinates on the upper left corner of the nose region; X_R and Y_R are the coordinates on the bottom right corner of the nose region. The following steps are all based on the located nose region.

2.2 Nostril detection and tracking

Our nostril detection method was inspired by the methods described by Vukadinovic and Pantic (2005). We have extended their methods by first applying the Harris corner detection algorithm (Harris and Stephens 1988) to automatically detect feature point candidates instead of manually labeling them in the training stage. The Harris corner detection algorithm not only can detect corners but also any number of isolated feature points in the region of interest (ROI) simultaneously. The number of feature points is flexible and can be predefined in the program. Since nostrils are obvious feature points in the nose region, 10 interest points will be enough for our recognition. The order of the selection of the interest points begins with the one that has the maximum eigenvalue and proceeds consecutively until it reaches the number we defined. Apparently, only two of these 10 feature points are the nostrils. If

we define only two interest points in the ROI, in most cases, they will be the nostrils. However, in some cases, such as when a person has a mole in the nose area, there is a high possibility that the mole will be misinterpreted as feature points. Therefore, a nostril classifier needs to be trained in order to accurately detect nostrils. An automatic nostril detection method using the Gabor feature-based boosted classifier (Friedman et al. 2000) was applied.

In the training stage of nostril detection, the 10 interest points that were detected in the previous step were used both as positive samples and negative samples. We manually identified the two out of 10 interest points that were positive feature points (nostrils). The rest of the eight detected interest points were used for the negative ones. Gabor filters (Jones and Palmer 1978) with eight orientations and six spatial frequencies were applied for feature extraction. The feature vector for each sample point was extracted from a 5×5-pixel bounding box centered at each positive and negative sample point. The 5×5-pixel bounding box was extracted from both the grayscale image and the Gabor filter bank (consisting of $8 \times 6 = 48$ representations). Therefore, $49 \times 5 \times 5$ features are used to represent one feature point. Since there are four negative points and one positive point for each nostril, the size of the training data matrix would be $5 \times 49 \times 5 \times 5$, which is computationally expensive. To solve this problem and avoid redundancy, the GentleBoost algorithm (Friedman et al. 2000) was used to reduce the dimensionality.

In the testing stage, a 5×5-pixel sliding bounding box was used to slide across the ROI pixel by pixel. The GentleBoost classifier outputs a response depicting the similarity between the trained feature point model and the current sliding bounding box. When the entire ROI region has been scanned, the position with the highest response shows the location of the feature point (nostril). This method is very effective for feature point detection, especially in the case of person-specific applications since the moles are already trained as negative samples and will not be misidentified as true feature points.

We applied the iterative Lucas-Kanade (LK) method with pyramids (Shi and Tomasi 1994) for real-time tracking of the nostrils. This method implements the sparse iterative version of the LK optical flow in pyramids and calculates the coordinates of the nostrils on the current video frame given their coordinates on the previous frame. Since the head may move back and forth in front of the camera, the size of the face segment changes. Therefore, multi-level matching is desired. The window size for computing the local coherent motion was set as 25×25, and the depth of the pyramid level was three. In extreme cases when the speed at which the object is moving is too fast, the feature points may be lost. For the purpose of detecting the natural rhythm of the shaking and nodding of the head, we found that this tracking algorithm was fairly reliable.

2.3 Head movement detection

After the nostrils had been detected and tracked, the coordinates were used to determine the head movements. We adopted a statistical pattern matching approach, which was trained and tested on real data to detect head movement. To improve recognition accuracy and to achieve real-time performance, a boosted-based pattern analyzer (Friedman et al. 2000) was used to adaptively select the best features by using a linear combination of individually weak classifiers and combining them into a strong classifier. Initially, the boosted analyzer assigned equal weight to each training sample. For the next stage, more weight was added to the training samples that were missed in the previous stage. The iteration went on by adding new classifiers until the overall accuracy met the desired requirement. There was a tradeoff between the overall recognition accuracy and the speed. For each type of head movement (head nods, head shakes and stationary), 20 weak classifiers in total were selected.

To determine head movements, a feature extraction module was used to locate the nostrils and map out the displacement onto the x, y coordinates. When the head movement is vertical, the shift of the nostrils in the y-coordinates is greater than that of the x-coordinates. On the other hand, when the head movement is horizontal, the shift of the nostrils in the x-coordinates is greater than that of the y-coordinates. If the shift in the displacement of the nostril is within a user-defined limit on the x and y axes, then the head movement is considered to be still. Suppose the coordinate of a nostril is (x_{n-1}, y_{n-1}) and (x_n, y_n) in two consecutive frames respectively, then $|y_n - y_{n-1}| \gg |x_n - x_{n-1}|$ indicates head nods, and $|y_n - y_{n-1}| \ll |x_n - x_{n-1}|$ indicates head shakes. If both $|y_n - y_{n-1}|$ and $|x_n - x_{n-1}|$ are lower than a certain threshold, then stationary status has occurred. For each frame, an observation sequence O consisting of 10 consecutive frames is generated:

$$O = \{[(x_n - x_{n-1}), \ldots, (x_{n+9} - x_{n+8})], [(y_n - y_{n-1}), \ldots, (y_{n+9} - y_{n+8})]\} \quad (3)$$

The coordinate differences between each of these frames were used as inputs for the boosted classifier to recognize head nodding and shaking for each frame. The nodding and shaking of the head can have different frequencies and amplitudes depending on the context. Ten is considered to be a reasonable number of an observation sequence, since it can prevent misclassification of a random head movement as a nod or shake, but is enough to retain recognition of subtle head movements. Only when all 10 consecutive frames respond with the same output, is the final interpretation being considered accurate.

3 Eye gaze analysis

3.1 Eye gaze evaluation

The purpose of this section is to determine whether the subject's gaze direction was direct or averted. A geometric relationship of human face organs was used to analyze eye gaze status. Since there were only two statuses that we considered, direct gaze and averted gaze, we compared the face image to a known direct gaze image as the template in order to evaluate the possible gaze status. The positions of nostrils and pupil locations were analyzed as evaluators. We defined the geometry of the eyes and nostrils as shown in Fig. 2.

We computed the four parameters r_R, r_L, α, and β where r_R, r_L are defined as follows:

$$r_R = \frac{|AC|}{|CD|}, \quad r_L = \frac{|BD|}{|CD|} \tag{4}$$

α and β are angles in radians. We denoted the values of the four parameters of the template in which the subject's gaze status was a direct gaze as r_{R0}, r_{L0}, α_0, and β_0. For each face image, we defined an evaluation parameter S, computed as follows:

$$S = |r_R - r_{R0}| + |r_L - r_{L0}| + |\alpha - \alpha_0| + |\beta - \beta_0| \tag{5}$$

The nostrils are relatively static points in frontal face images. The pupils move when eye gaze status changes. Assuming the face images are all the same size, the length and orientation of CD are stable. When the subject is in averted gaze status (i.e., he or she looks left, right, up, or down), |AC| and |BD| change, and the angles α and β change as well. To eliminate the effects of image size, we used the fractions of lengths r_R and r_L instead of absolute values |AC| and |BD|. In the case of direct gaze status, $|r_R - r_{R0}|$, $|r_L - r_{L0}|$, $|\alpha - \alpha_0|$, and $|\beta - \beta_0|$ are approaching zero, respectively. On the contrary, these values are greater than zero in the averted gaze status. By adding these values together, we can distinguish the direct gaze and averted gaze.

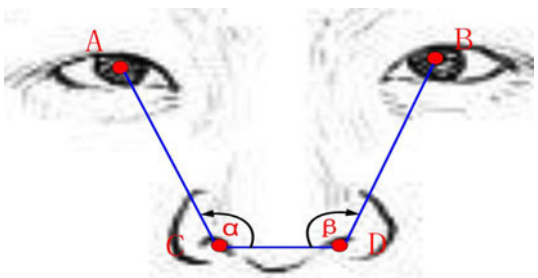


Fig. 2 The geometrical eye and nostril model

The average S value was computed for each test video and normalized into the range of 0–1. The final eye gaze input value of the fuzzy inference system was $1 - S$.

3.2 Pupil detection

We used a knowledge-based method to detect the location of the pupil. To ensure accurate pupil detection, our approach was built on the following conditions:

1. The frontal view face images, which were picked out from the sequence of video frames, were used.
2. Only the image segment of the human face, which is detected from the frontal view face image, was analysed.
3. The environment light uniformly illuminated the whole face.

3.2.1 Detection of frontal view face image

In order to keep the consistency of the geometric relationship of face organs such as eyebrows, eyes, nose and mouth, we considered only frontal view face images in which the subjects were facing the camera directly. A frontal view image was used as a reference. As stated in the previous section, the coordinates of nostrils are determined for each frame. There are two bounding boxes that are defined according to the location of the reference. Let x, y be the reference nostril coordinates. Two bounding boxes corresponding to the left and right nostrils are defined as follows:

$$B_{nostril} = \{(x, y) | x \in [x^0 - \varepsilon, x^0 + \varepsilon],\ y \in [y^0 - \varepsilon, y^0 + \varepsilon]\} \tag{6}$$

where ε is the small number that we define as a threshold. Let (x, y) be the detected location of the nostrils. If the two nostrils are both inside the bounding boxes, then this frame is a frontal view image.

3.2.2 Generation of the face mask

A face mask is used for removing the information that does not belong to the face region. The process for generating a face mask is shown in Fig. 3. In the first step, the color image is converted into grayscale. To unify the brightness intensity distribution of the image, the brightness of the image is normalized using the following equation:

$$F_n(x, y) = Mean_0 + sign \times \sqrt{Var_0 (F(x, y) - Mean)^2 / Var} \tag{7}$$

where $F(x, y)$ and $F_n(x, y)$ are the input face image and normalized image, respectively. $Mean_0$ and Var_0 are the

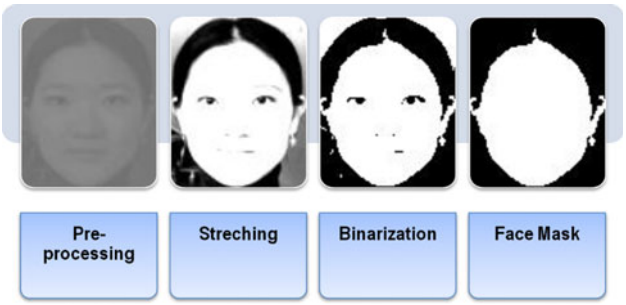


Fig. 3 Generation of the face mask

mean and variance of the destination image. *Mean* and *Var* are the mean and variance of input image; *sign* is equal to 1 when $F(x, y) > Mean$, and -1 when $F(x, y) \leq Mean$.

Secondly, enhancing the contrast of the image improves the performance of binarization. Thus, we applied gray-scale stretching (Al-amri et al. 2010) to increase the contrast as shown in the following equations:

$$F_{str}(x,y) = \begin{cases} 255 \times \frac{F_n(x,y)-low}{(high-low)}, & low \leq F_n(x,y) \leq high \\ 255, & F_n(x,y) > high \\ 0, & F_n(x,y) < low \end{cases} \quad (8)$$

where *low* and *high* are set to $Mean_0 \cdot P_0$ and $Mean_0 \cdot P_1\%$, respectively.

For binarization, we assumed that the hair is darker than the face skin. Consequently, we simply took the mean value of the stretched image as the threshold to binarize the image. The threshold can be adjusted for opposite cases when the skin is darker than the hair. As shown in step three of Fig. 3, the face skin area turns white.

In the last step, we want to further remove the face organs such as the eyes, lips, and nostrils to turn the entire face region white. For most cases, the face skin is brighter than the face organs; the eyes, nostrils and lips appear to be black. To take out these small black areas, we find all connected components in which all pixel values are the same and connected to each other from the binary face image. Later on, the pixels of these components are counted respectively, and then the small black areas, whose pixel number was less than a certain percentage of the total pixels in the image, are turned white. Using the same technology, we remove the small white areas in a large black region for the purpose of turning the white pixels in the background to black. The resulting image is used as the face mask. The face mask is a binary matrix that has the same size as the original face image.

3.2.3 Determination of eye location

The procedure for locating eyes is illustrated in Fig. 4. First, we apply histogram equalization to the original gray face image (Fig. 4a). The resulting image of histogram equalization (Fig. 4b) is denoted by $F_{hsq}(x, y)$. We multiply the binary face mask (Fig. 4f) obtained from Sect. 3.2.2 with $F_{hsq}(x, y)$ for the purpose of removing the confusing objects not located in the face area, such as the hair region. The image obtained after multiplying the face mask is shown in Fig. 4c.

The following binarization equations are applied to the masked face image $F_{mf}(x, y)$ to extract the dark components on the face:

$$F_b(x,y) = \begin{cases} 1, & F_{mf}(x,y) < 95 \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

Based on our test, we found the pixel number of the eye region to be less than 3% and greater than 0.1% of the total pixel number. We removed the connected white component that was larger than 3% and smaller than 0.1% of the total image area. The resulting image, in which the white connected pixels are extracted face organs, is shown in Fig. 4e.

Among the extracted face organs, eyebrows, lips, nostrils, and other noise regions are the false candidates. The knowledge of the face's geometry and spatial relationship is applied to locate the eyes. The proposed algorithm is described as follows: Given the binary image $F_{bfo}(x, y)$ resulting from the last step, horizontal integral $H(y)$ is computed. Local peaks of $H(y)$ are detected. The first peak

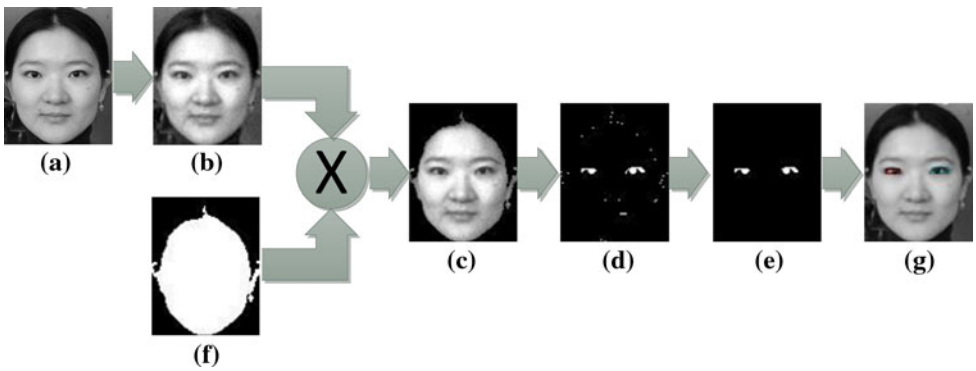


Fig. 4 Determination of eye location

y_{eye} above the 40% of the image height from the bottom is the y-coordinate of eyes. The objects along the horizontal line $y = y_{\text{eye}}$ are candidates of eyes. The two objects close to the central vertical line are the eyes. The two rectangular regions, which exactly bound two detected eyes, are defined as eye regions R_{eye}.

3.2.4 Determination of pupil location

The detection of pupil locations is broken into two steps: (1) estimating the y-coordinates of two eyes and (2) estimating the x-coordinates of two eyes. The local horizontal integral projections of two grayscale eye regions are computed separately. The horizontal integral is defined by:

$$H(y) = \sum_{x} F_{hsq}(x, y), \quad (x, y) \in R_{eye} \tag{10}$$

where R_{eye} represents the region of the eye determined in the Sect. 3.2.3.

Because the darkest area is most likely the center of the iris, the y-value corresponding to the minimal H(y) value is selected as the y-coordinate of the pupil. The vertical integral projections are computed as follows:

$$V(x) = \sum_{y=y_p-W_p}^{y_p+W_p} F_{hsq}(x, y), \quad (x, y) \in R_{eye} \tag{11}$$

where y_p denotes the detected y-coordinate of the pupil, and W_p is the estimated pupil diameter. The x-value corresponding to the minimal $V(x)$ value is selected as x-coordinate of the pupil.

4 Emotion and cognition detection

4.1 Introduction

Effective design of emotion recognition systems is a very challenging task since it relies on multidisciplinary collaboration among areas such as computer science, electrical engineering, and psychology for the sharing of knowledge. Moreover, emotion detection is difficult since the boundaries among different emotions are not crisp (Calvo and D'Mello 2010); for example, sometimes different emotions have the same facial expression. In this paper, we applied fuzzy logic techniques (Zadeh 1973; Gegov 2007) for emotion and cognition detection. Fuzzy logic techniques are widely used in the area of pattern recognition; for example, in (Taur and Tao 2000) fuzzy classifier was applied for face recognition. In (Khezri et al. 2007; Engin 2004; Khezir and Jahed 2007), fuzzy techniques were used in biometric applications.

With the development of affective computing (Picard 1997) and HCII, the fuzzy logic technique has been used for emotion recognition. Developing linguistic rules is the most important step in developing a fuzzy logic system. These rules can be knowledge-based, which are provided by experts (Feng 2006; Gegov 2007). When expert knowledge is not sufficient, the rules can also be extracted from a given training data set (data-driven-based) by using computer learning algorithms (Feng 2006; Gegov 2007). Chakraborty and Konar (2009) used a knowledge-based fuzzy system to recognize emotion from facial expressions. By dividing face images into localized regions, facial features including eye opening, mouth opening, and the length of eyebrow constriction were extracted, fuzzified, and mapped into facial expressions. Contreras et al. (2010) presented a knowledge-based fuzzy reasoning system that could recognize the intensity of facial expressions by using the facial action units and facial animation parameters. Esau et al. (2007) proposed a fuzzy emotion model that could analyze facial expressions in video sequences. Mandryk and Atkins (2007) presented a knowledge-based fuzzy logic model using physiological data to recognize emotion. Arousal and valence values were generated by the physiological signals and were then used as inputs to a fuzzy logic model to detect emotional states including boredom, challenge, excitement, frustration, and fun.

Chatterjee and Hao (2010) used a data-driven method (one successful technique is neuro fuzzy (Jang 1993)) to model face emotion by identifying the spots, edges and corners of a face and training the neuro fuzzy model. Ioannou et al. (2007) have successfully proposed an emotion recognition system that analyzes and evaluates facial expressions incorporating psychological knowledge about emotion. A neuro fuzzy rule-based system has been created to classify facial expressions by analyzing facial animation parameter variations from the discrete emotional space and the continuous 2D emotion space. Katsis et al. (2008) presented a methodology of assessing the emotional state of car-racing drivers using bio-signals. Neuro fuzzy and a support vector machine were used as the classification techniques. Lee and Narayanan (2003) and Giripunje and Bawane (2007) used a data-driven based fuzzy inference system for emotion recognition from human speech.

It is worth mentioning that neural fuzzy has not only been applied in emotion detection but also in human behavior recognition. Asteriadis et al. (2009) have recently presented a system that detects and tracks movements of the head, eye, and hand in order to estimate the level of interest by applying a neural-fuzzy technique. Their method has been successfully used to detect attention states of users.

4.2 General fuzzy model

The goal of this paper is to explore new ways of HCII by integrating information from the body language of the head

to infer emotions and cognitions. In particular, we concentrated on the combination of facial expression, eye gaze, and head movement. This is a difficult task since it requires the incorporating of knowledge from various disciplines, especially psychology. The general model of emotion detection in this paper is shown in Fig. 5. First of all, we extract information from the body language of the head. This extracted information is used as input for the fuzzification interface, which defines a mapping from a crisp value to a fuzzy number. These fuzzy numbers are then given to the knowledge-based inference rules, which give conditions to derive reasonable actions. The defuzzification interface then maps the fuzzy value into a crisp emotional or cognitive state.

4.3 Input variables

In this section, the input variables used in the fuzzy system are discussed. Figure 6 shows the taxonomy structure of the body language of the head. For the fuzzy inference systems, we defined five input linguistic variables: Happiness, Angry, Sadness, Head Movement, and Eye Gaze. The input variable "Happiness" was mapped into membership functions as "Maybe Happy", "Little Happy", "Happy", and "Very Happy" according to the degree of happiness. The input variable "Angry" was mapped into membership functions as "Maybe Angry", "Angry", and "Very Angry" based on the degree of angriness. The input variable "Sadness" was mapped into membership functions as "Maybe Sad", "Sad", and "Very Sad" based on the degree of sadness. The input variable "Head Movement" was mapped into membership function as "High Frequency Shaking", "Low Frequency Shaking", "Stationary", "Low Frequency Nodding", and "High Frequency Nodding" based on the direction and frequency of head movement. The input variable "Eye Gaze" was mapped into membership function as "Avert Gaze" and "Direct Gaze" based on the scale of eye gaze direction. The mapping strategy from input variables to membership function boundary parameters will be explained in Sect. 5.3.

These linguistic variables need to be further quantified into fuzzy input values. The variables can be quantified into input values ranging in degree between 0 and 1 using the following proposed method respectively:

1. We quantified the linguistic variable "Head Movement" into values using the following formulas:

$$\text{Head-movement} = 0.5 + (\text{sgn}) \times f_{normal}$$
$$\text{sgn} = \begin{cases} 1, & nodding \\ -1, & shaking \end{cases}$$
$$f_{normal} = \frac{f_{hm}}{2 \times MaxFrequency}, \quad hm = nod\ or\ shake \tag{12}$$

where head-movement is the quantified input value, f_{hm} is the frequency of a certain head nod or head shake movement, and *MaxFrequency* is the maximal frequency of head nods and shakes. After repeated testing, we found that the maximum frequency human beings can achieve for both head nods and shakes is 3 cycle/sec. Therefore, 3 would be a reasonable value for *Maxfrequency*.

2. We computed the value of S for the scale of eye gaze direction analysis, which is defined in formula (5). The input value of the "Eye Gaze" variable can be defined using the following formula:

$$EyeGaze = 1 - \bar{S}_{normal} \tag{13}$$

where $\bar{S}_{normal}$ is the normalization value S. The detailed procedures will be explained in the experiment section.

3. For the facial expression analysis, we applied our previous work in paper (Zhao 2009). Instead of classifying the six basic facial expressions, we trained a new classifier that divided expressions by intensity. The results were then manually marked with values. In the future, we plan to extract facial features by applying facial action units (AUs) and recognize facial expression using fuzzy logic. This could provide more accurate quantified value for facial expression results.

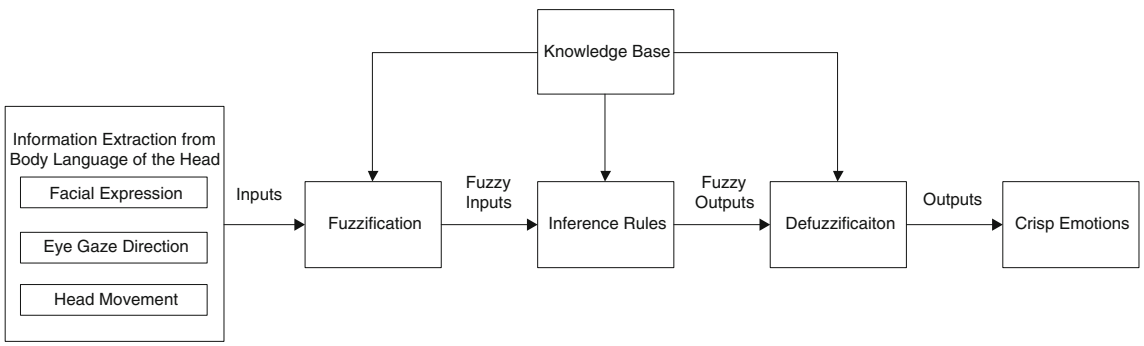


Fig. 5 General model of emotion detection

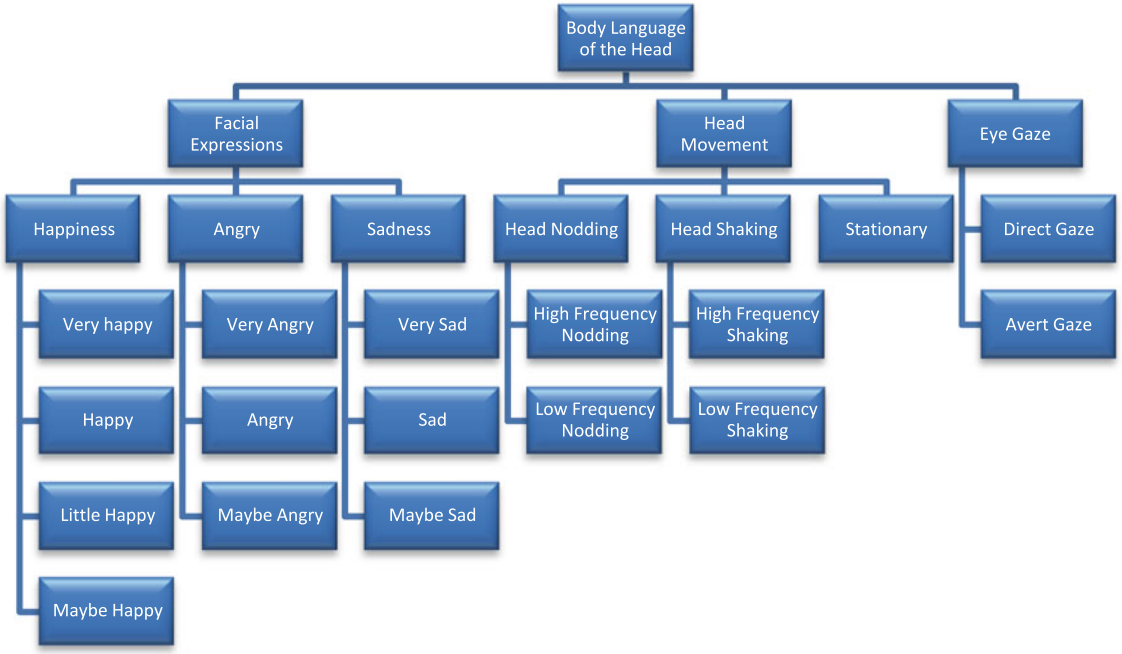


Fig. 6 Taxonomy structure of body language of the head

4.4 Generation of fuzzy rules

In order to obtain valid emotions to generate fuzzy rules, the ideal conditions for the experiment should be that:

1. The subject feels the emotion internally.
2. The subject should be in a real-world environment instead of a lab environment, and emotions should occur spontaneously.
3. The subject should not be aware that he or she is being recorded.
4. The subject should not know that he or she is part of an experiment.

The ideal experiment cannot be conducted due to privacy and ethical concerns. Therefore, we used an alternative method for obtaining ground truth and fuzzy rules from the knowledge and opinions of an expert in psychology, a pilot group, and annotators (as shown in Fig. 7).

Step 1: A psychology expert gave advice on high-level instructions (for example, emotions can be inferred by a combination of facial expressions, eye gaze direction, and head movements). These instructions were used for the pilot group, which consisted of two Ph.D. students (one majoring in Computer Science and the other in Electrical Engineering) and one undergraduate student (majoring in Psychology). The pilot group made decisions about candidate rules. After discussion, 15 candidates for fuzzy rules (each corresponding to an emotion or cognition) were defined.

Step 2: Fifteen video clips representing these 15 emotions and cognitions were recorded by one member of the pilot group. We established a questionnaire to evaluate the proposed rules. In the questionnaire, the following question was asked: "After watching this video clip, what emotion or cognition have you experienced?" for each video clip.

Step 3: Ten annotators, who were not members of the pilot group, were asked to annotate these video clips by answering the questionnaire. When answering the questionnaire, some annotators used alternative words to depict emotion. For example, for "resentment," some annotators used "hate." For the "decline" response, some annotators used "reject politely" instead. In such cases, we considered that both terms represented the same emotion. In order to minimize the influence of out-group culture differences (in-group members have an advantage in recognizing emotions (Schalk et al. 2011)), we asked only Chinese students to be annotators in this experiment. It is worth mentioning that although emotions are universal, the ways humans express emotions vary between cultures (Marsh et al. 2003; Matsumoto 1989, 2001). To the best of our knowledge, only the six basic facial expressions (happiness, anger, fear, surprise, sadness, and disgust) have been proven by psychologists to be universal. This is the reason why we asked students with the same cultural backgrounds to participate in this experiment. Although the fuzzy rules were drawn by in-group people, the idea of integrating the different modalities of the body language of the head is generic enough to be used by any particular target user group from any culture.

Step 4: The pilot group analyzed the answers in the questionnaire. A fuzzy rule would be generated if the

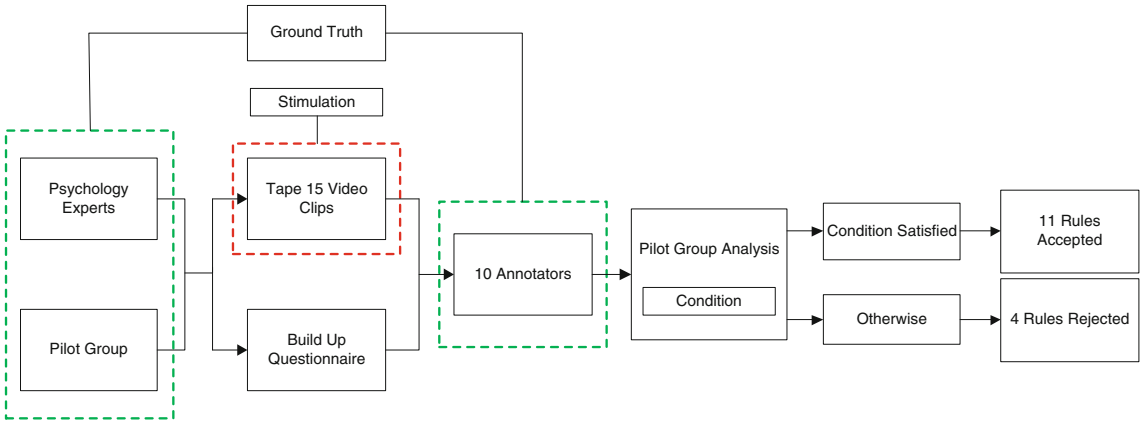


Fig. 7 Procedure of fuzzy rules creation

following condition is satisfied: at least 8 out of 10 annotators have experienced the same emotion from watching the clip. After analysis, 11 out of 15 candidate fuzzy rules were retained. The other four were rejected.

It is true that some facial emotion researchers use movie clips to stimulate emotions. However in our case, emotions were complex and many (such as guilt) could not be easily evoked by movie clips or other methods. Therefore, video clips were taped and shown to the annotators in an effort to provoke emotion and cognition. These video clips play the same role as movie clips to provide the stimulation. The 11 generated rules based on the above procedures are listed below:

Rule 1: IF (*Happiness* is *Very-Happy*) AND (*Head-Movement* is *High-Frequency-Nodding*) AND (*Eye-Gaze* is *Direct-Gaze*) THEN (*Emotion-Set-A* is *Agree*)

Rule 2: IF (*Happiness* is *Happy*) AND (*Head-Movement* is *Low-Frequency-Nodding*) AND (*Eye-Gaze* is *Direct-Gaze*) THEN (*Emotion-Set-A* is *Admire*)

Rule 3: IF (*Happiness* is *Little-Happy*) AND (*Head-Movement* is *Low-Frequency-Shaking*) AND (*Eye-Gaze* is *Direct-Gaze*) THEN (*Emotion-Set-A* is *Decline*)

Rule 4: IF (*Happiness* is *Maybe-Happy*) AND (*Head-Movement* is *Low-Frequency-Nodding*) AND (*Eye-Gaze* is *Avert-Gaze*) THEN (*Emotion-Set-A* is *Thinking*)

Rule 5: IF (*Happiness* is *Maybe-Happy*) AND (*Head-Movement* is *Stationary*) AND (*Eye-Gaze* is *Avert-Gaze*) THEN (*Emotion-Set-A* is *Thinking*)

Rule 6: IF (*Angry* is *Very-Angry*) AND (*Head-Movement* is *Low-Frequency-Shaking*) AND (*Eye-Gaze* is *Direct-Gaze*) THEN (*Emotion-Set-B* is *Resentment*)

Rule 7: IF (*Angry* is *Maybe-Angry*) AND (*Head-Movement* is *High-Frequency-Shaking*) AND (*Eye-Gaze* is *Direct-Gaze*) THEN (*Emotion-Set-B* is *Disagree*)

Rule 8: IF (*Sadness* is *Very-Sad*) AND (*Head-Movement* is *Low-Frequency-Shaking*) AND (*Eye-Gaze* is *Avert-Gaze*) THEN (*Emotion-Set-C* is *Distressed*)

Rule 9: IF (*Sadness* is *Sad*) AND (*Head-Movement* is *High-Frequency-Nodding*) AND (*Eye-Gaze* is *Avert-Gaze*) THEN (*Emotion-Set-C* is *Guilty*)

Rule 10: IF (*Sadness* is *Maybe-Sad*) AND (*Head-Movement* is *Low-Frequency-Shaking*) AND (*Eye-Gaze* is *Direct-Gaze*) THEN (*Emotion-Set-C* is *Disappointed*)

Rule 11: IF (*Angry* is *Angry*) AND (*Head-Movement* is *High-Frequency-Shaking*) AND (*Eye-Gaze* is *Avert-Gaze*) THEN (*Emotion-Set-B* is Annoyed)

4.5 Output variables

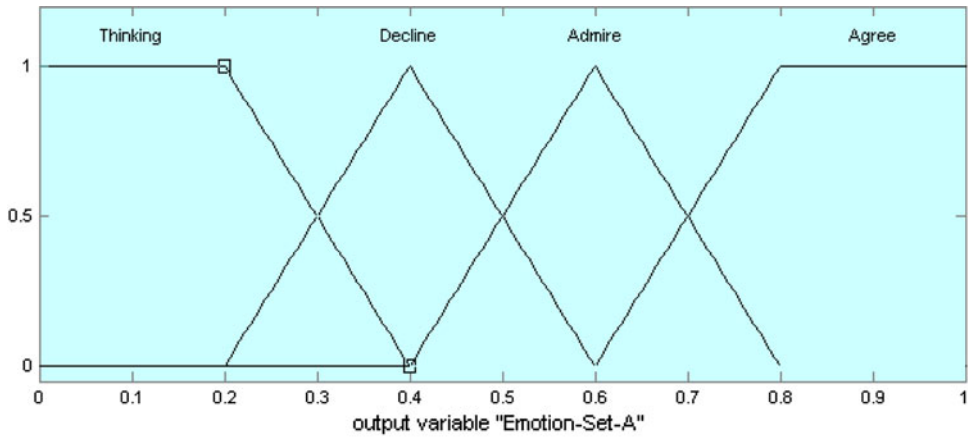
The output variables from our emotion detection system are Emotion set-A, Emotion set-B, and Emotion set-C. From the rules described above, we came up with 10 emotions and cognitions. These emotions and cognitions can be divided into three categories, which are derived from the three facial expressions: happiness, angriness, and sadness respectively. Emotion set-A includes "Thinking", "Decline", "Admire", and "Agree", which are co-related with the happy expression. Emotion set-B includes "Disagree", "Annoyed", and "Resentment", which are co-related with the angry expression. Emotion set-C includes "Disappointed", "Guilty", and "Distressed", which are co-related with the sad expression. Therefore, we set up three output variables as shown in Fig. 8.

5 Experimental results and discussion

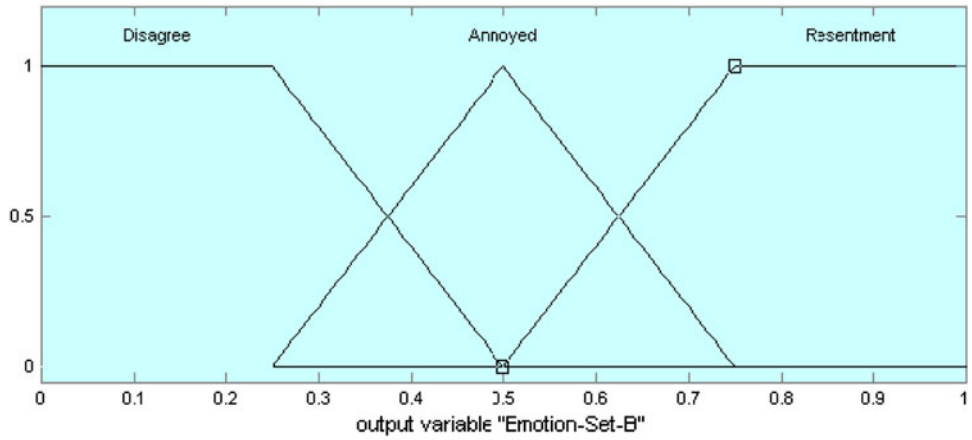
5.1 Head movement evaluation

5.1.1 Nostril detection evaluation

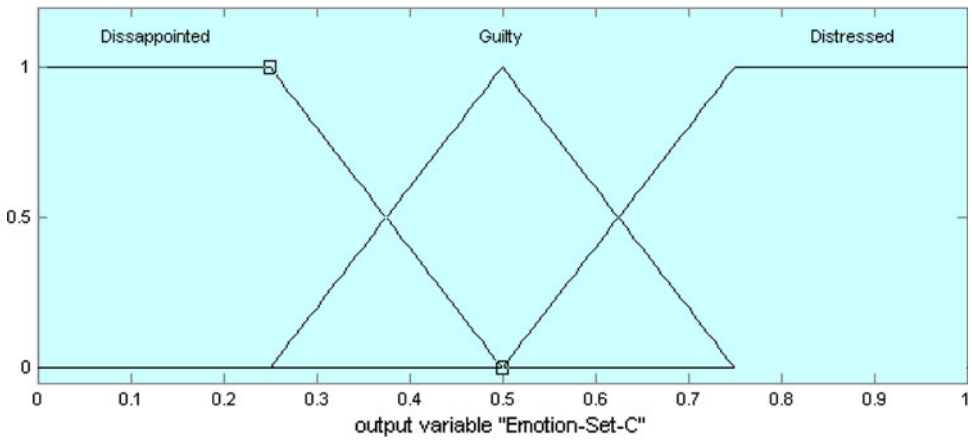
To evaluate the performance of the proposed method, we tested our nostril detection method on both on the JAFFE database (Lyons et al. 1999) and images collected



(a) Output variable Emotion set-A



(b) Output variable Emotion set-B



(c) Output variable Emotion set-C

Fig. 8 Output variables and their respective membership functions

from a webcam. The JAFFE database contains 213 images of 10 Japanese female frontal face images expressing seven different facial expressions (see Fig. 9 for examples). The images collected from the webcam were taken by two volunteer colleagues, one male and one female. A total of 200 image samples were collected using the Microsoft LifeCam 1.4 with a resolution of 640 × 480. In order to complicate the data, the participants were asked to perform different facial expressions. Some of the collected images have in-plane and out-plane head rotations. For both of these two experiments, each automatically detected nostril point was compared to the manually annotated nostril point (considered as true point). If the automatically detected feature point was displaced within 3 pixels distance from

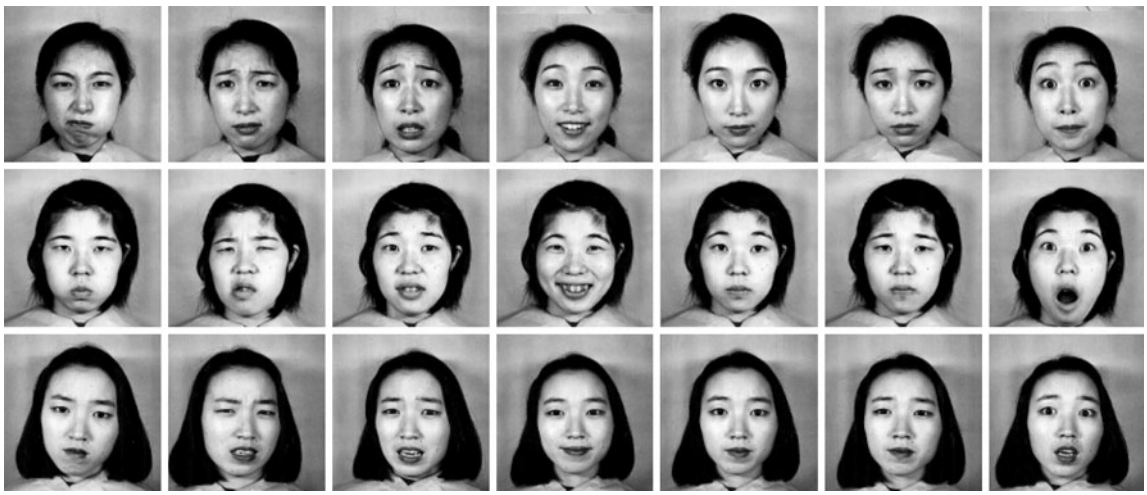


Fig. 9 Examples of the JAFFE face images

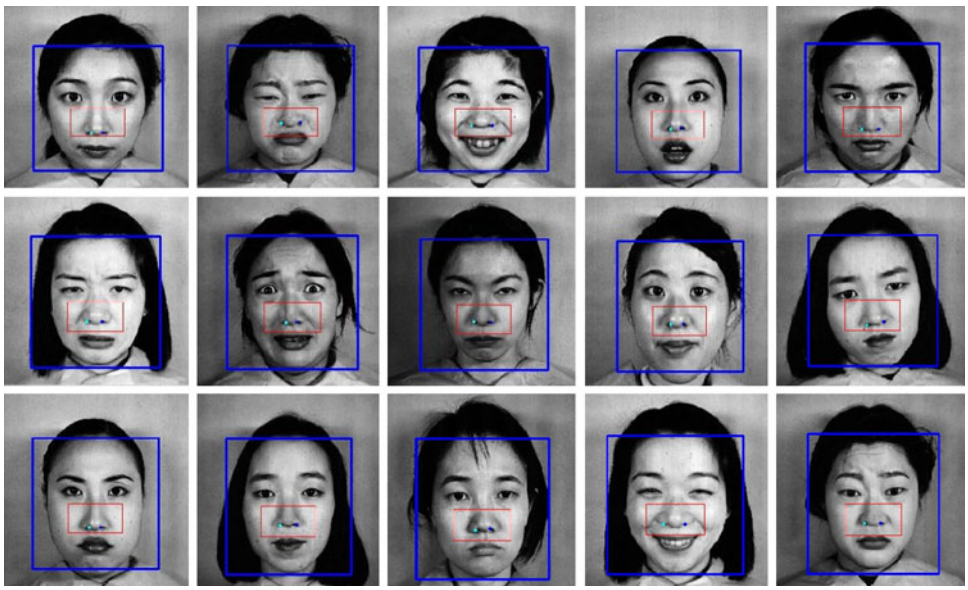


Fig. 10 Examples of nostril detection results using the JAFFE database

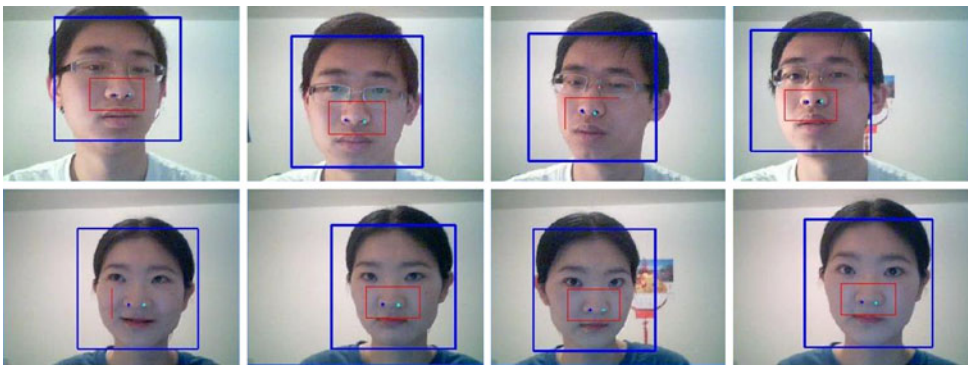


Fig. 11 Examples of nostril detection results using webcam images

the manually annotated true point, then this point is regarded as successful detection. Figures 10 and 11 show some examples of nostril detection results using the JAFFE database and webcam respectively. The detection accuracy of JAFFE database is 93%. No images from the database were missed. The measurements of the mean and variance of the distance between a detected point and true point of the seven facial expressions from the JAFFE database are

Table 1 Measurements of the distance between detected point and true point of the seven facial expressions from the JAFFE database

Emotion	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Overall
Mean	1.048016	1.452378	1.626345	1.712985	1.464439	1.47756	1.406209	1.45388
Variance	1.01864	1.450082	0.868573	1.214302	0.802119	1.190863	0.672096	1.05854

Table 2 Example questions

	Questions
1	Are you a university student?
2	Are you male?
3	Are you female?
4	Do you like basketball game?
5	Are you Canadian?
6	Do you like travelling?
7	Do you like apples or oranges?

Table 3 Head movement test results

Head movements	Hits	Missed	False	Recognition rate (%)
Nods	40	0	2	95
Shakes	30	0	2	93.3
Stationary	30	0	1	96.7

listed in Table 1. The nostril detection accuracy using webcam data is 95% and 11 images were missed. By analyzing the results, we found that most of the missed images were due to excessive head in-plane and out-of-plane rotation invariance over $\pm 15°$.

5.1.2 Head movement detection evaluation

In order to evaluate the head movement detection performance, four subjects (two male and two female) participated in the experiment. The Microsoft LifeCam 1.4 at a resolution of 320×240 and 15 frames/s was used for the head movement analysis experiment. A database was collected of natural head nods and head shakes by asking a number of questions that the subjects were instructed to answer with a head nod, a head shake or keeping his or her head stationary. Examples of such questions are listed in Table 2. For the questions that required the participants to make a choice (for example, question 7 in Table 2), they were asked to answer the question by keeping their head stationary. A total of 100 samples were collected by asking 25 questions from each of the participants. The participants were asked to sit with frontal face and perform facial expressions while answering the questions. Data was collected during different times of the day near the window in order to vary the lighting conditions. We randomly selected 40% of each of the head nods, shakes and stationary samples for training. To train a head nod classifier, head nod samples are used as positive samples, and the rest of the head shake and stationary samples are used as negative samples. The same strategy was applied for head shakes and stationary classifiers, in turn. In the recognition stage, the classification decision was determined by the classifier with the largest value. The recognition results are shown in Table 3. The errors could be caused by in-plane and out-plane rotations of the head movement.

The goal of head movement analysis is not only to observe the direction of head movement but also to calculate the frequency of head movement. Head movement frequencies were quantified into values for the further fuzzy inference system inputs. To achieve this, tracking results of the nostrils were used.

Figure 12 shows examples of nostril tracking results under five different conditions: high frequency nodding, low frequency nodding, high frequency shaking, low frequency shaking and stationary. The blue plots indicate the x-coordinates of the nostril points and the pink plots indicate the y-coordinates of the nostril points. The x-axes represent the number of detected nostril points and the y-axes represent the corresponding pixel value of these points in the 320×240 images.

To determine the frequency of head movement, we applied the Fast Fourier Transform (FFT) (Brigham 2002) to transform the tracking results from the time domain to the frequency domain. The highest peak after FFT in the spectrum shows the main frequency for the input signals. Figure 13 shows an example of the result after FFT for the status of Low-Frequency-Nodding. For each video sample, we recorded the coordinate of the nostril tracking result and calculated the frequency value using FFT. These frequency values were used to quantify the input variables based on the formula (12) we defined in Sect. 4.3.

5.2 Gaze status analysis

5.2.1 Pupil detection evaluation

The JAFFE database of frontal face images was chosen for evaluation of the pupil detection procedure. The resolution of each image is about 256×256 pixels. We chose $Mean_0 = 127$, $Var_0 = 85$, $P_0 = 80$, $P_1 = 100$. The detected pupils are shown in Fig. 14. Since the face detection

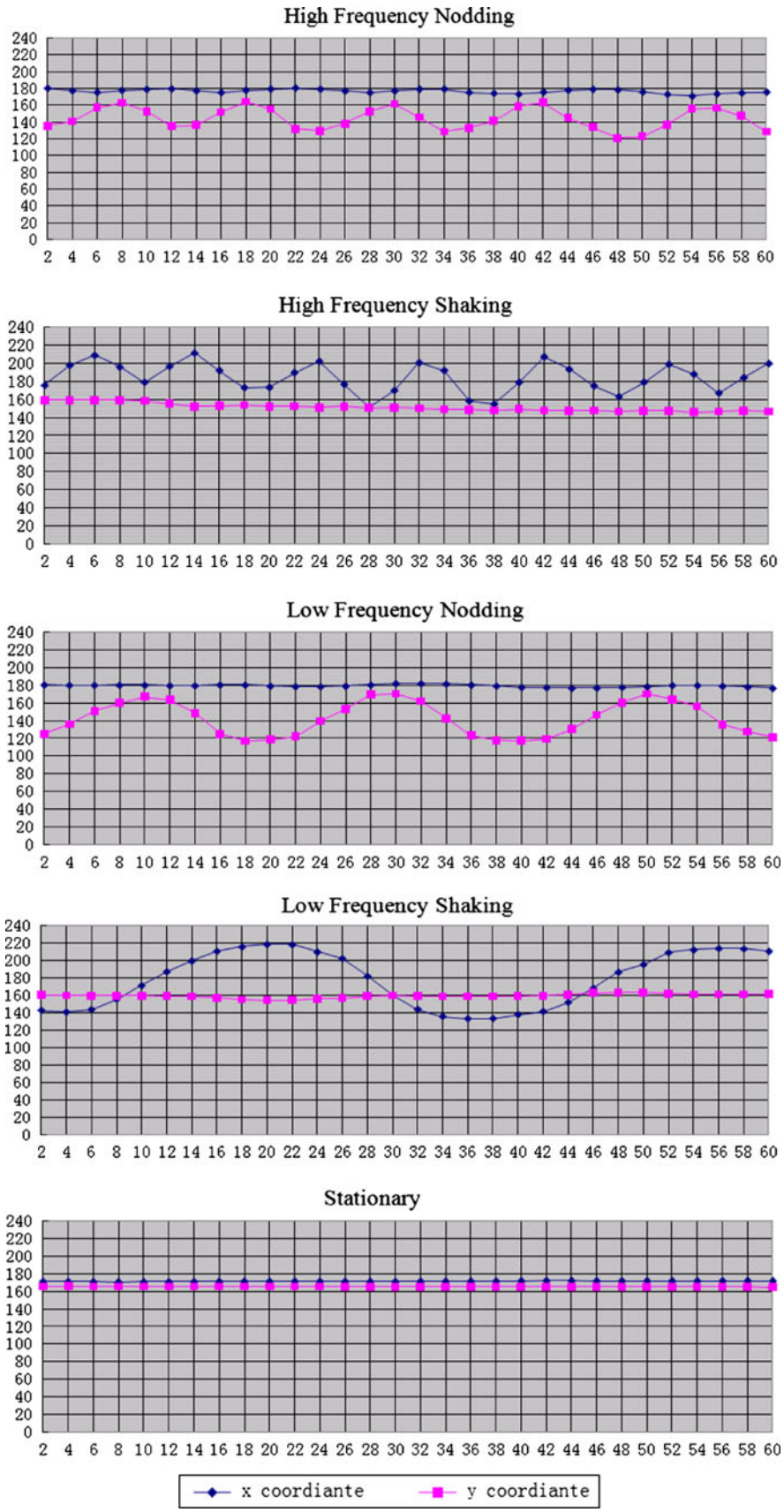


Fig. 12 Example of nostril tracking results

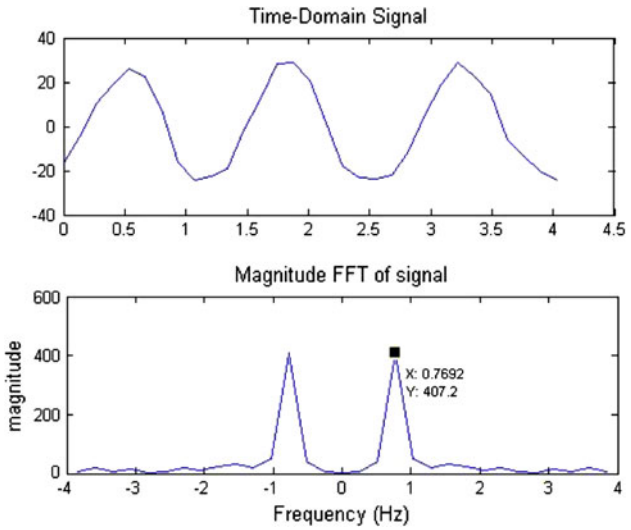


Fig. 13 Example of FFT for low frequency nodding

algorithm is a scale invariant (Viola and Jones 2001), the size of the detected face region will not be a constant. The eye regions were detected from these face regions with various sizes based on geometrical characteristics of the human face. Therefore, the size of the face image does not affect eye detection. In order to further test the performance of the method, we took the images collected from the webcam (named WebCam-1) in Sect. 5.1.1 for nostril detection. We decreased the size of the same set of images to 320 × 240 and named them WebCam-2. The size of the detected face region varies and ranges from 274 × 274 to 194 × 194. The comparison of the pupil detection results is shown in Table 4. The results show that this algorithm is not sensitive to the scale of the image. However, when the image quality is poor and the eyes are no longer noticeable, this algorithm will fail. The reasons for the failure of detection are as follows: (1) The environmental light does not uniformly illuminate the face; for example, one side of the face is too dark to detect the eye location. (2) The head does not face the frontal plane; thus one eye is incorrectly taken as part of the face mask. (3) The head is tilted too much and two eyes are not aligned horizontally, causing the algorithm to fail. In further tests, the Gaussian noise was added into the image with SNR around 20 dB, and there was no significant change of detection rate.

Table 4 Comparison of pupil detection results

	Original image size	Detected face region (approximately)	Recognition rate (%)
JAFFE	256 × 256	173 × 173	96.7
WebCam-1	640 × 480	274 × 274	97
WebCam-2	320 × 240	132 × 132	96.5

5.2.2 *Gaze direction evaluation*

To validate the proposed eye gaze detection method, we analyzed images captured by the webcam at a resolution of

Fig. 14 Example of pupil detection results

Table 5 Example of eye gaze analysis

	r_R	r_L	α	β	S	Result	Normalization	Fuzzy input
Reference	2.2209	2.2224	2.0700	2.0081	0.0000	1	0.0000	1.0000
Direct	2.2288	2.2566	2.0670	1.9952	0.0579	1	0.0986	0.9014
Up	2.3707	2.5228	1.9745	2.0149	0.5524	0	0.9415	0.0585
Down	1.9444	2.0142	2.0543	2.0348	0.5273	0	0.8986	0.1014
Right	2.0587	2.3331	1.9270	2.0798	0.4877	0	0.8311	0.1689
Left	2.2306	2.0130	2.1663	1.8588	0.4647	0	0.7920	0.2080

640 × 480. Five colleagues volunteered to participate in the experiment. Each participant was asked to repeat all the possible combinations of seven facial expressions with five different eye gaze directions (direct, up, down, left and right) five times. Therefore, in total, $5 \times 25 \times 7 = 875$ images (175 for direct gaze and 700 for avert gaze) were collected. Table 5 shows an example of eye gaze evaluation data by comparing five images with the reference image. In this table, columns "r_R", "r_L", "α", "β", and "S" are the five parameters defined in formula (5). The column "*Result*" represents the detection result for eye gaze. We set a threshold of 0.25 for "S". When the values of "S" are less than the threshold, we categorized the image into direct gaze using 1 to represent it; it was otherwise categorized as an averted gaze using 0 to represent it. The value of S was then normalized into the range of 0–1. For each subject, we calculated his or her maximal S value based on these collected images with respect to the reference image of direct gaze. The average value of these maximal S values was used for normalization. The column "*Normalization*" shows the result after normalization and column the "*Fuzzy Input*" column shows the result of 1 minus the normalized value. The final value will be the quantified eye gaze value of fuzzy input for emotion detection. Table 6 shows the measurements of the mean and variance value of the parameter S from the collected images under different gaze directions. Since direct gaze images should compute the smallest S values when compared to the reference image, the mean value of "*Direct*" is the smallest among all gaze directions.

5.3 Emotion model test

A group of five Chinese university students (aged from 20 to 35), including three females and two males, volunteered to participate in the experiment. The experiment was carried out in a laboratory environment with uniform illumination. A commercial-grade 640 × 480 pixel Microsoft LifeCam was used to record the videos at 15 frames/s for approximately 5 s each. Each participant performed the emotions and cognitions twice according to the 11 established rules. Thus, a total of 110 video clips were recorded. Then, each video clip was divided into image sequences with time intervals of 100 ms. We extracted 30 consecutive frames from each image sequence, manually choosing the first frame that showed a frontal face. These 110 image sequences were used as a database for input value extractions for fuzzy inference systems. In the experiment, both conventional fuzzy logic (a knowledge-based technique that manually annotates the mapping strategy from input values to membership function boundaries) and neural fuzzy techniques (a data-driven-based technique that obtains membership function boundaries from training data) were utilized for emotional and cognitional detection. A conventional Mamdani-type (Mamdani and Assilian 1975) fuzzy system is an intuitive technique that has the main advantage of being capable of formalizing inputs and fuzzy rules from expert knowledge, especially in cases in which there is no sufficient information available, where using expert knowledge is the only way to collect information (Gegov 2007).

Table 6 Measurements of different gaze direction

Gaze direction	Direct	Up	Down	Left	Right
Mean	0.09231	0.52048	0.4821	0.4759	0.49456
Variance	0.004012	0.004217	0.010959	0.00266	0.002337

For the Mamdani-type fuzzy inference system, the two typical triangular and trapezoidal membership functions were applied. We manually labelled membership function boundary parameters uniformly based on the assumption that the possibility of each case is uniformly distributed. The steps for obtaining fuzzy input values for eye gaze were as follows: We extracted the frontal face images from the 30 images using formula (6). For each frontal face image, the S value was calculated using formula (5), and then the average S value was calculated for all frontal images. By applying formula (13), the fuzzy input could be obtained for the image sequence. For all image sequences, the above steps were repeated. In order to obtain the fuzzy input values for head movement, the nostril points for each image were extracted, tracked, and recorded for an image sequence. Afterwards, head movement direction and frequency were analyzed based on the provided algorithms.

Fig. 15 An example of an image sequence for admire emotion

Table 7 An example of input and output values of admire emotion

The results of each stage	Head movement		Eye gaze	Facial expression	Emotion	
	Status	Frequency	Direct	Happy	Output	Emotion
	Nodding	0.8 cycle/s				
Value of input fuzzy variable	0.633		0.887	0.6	A = 0.6 B = N/A C = N/A	Admire

Using formula (12), we obtained the fuzzy input for the image sequence. For all image sequences, the above steps were repeated. Figure 15 shows an image sequence performed by a female subject expressing the "Admire" emotion. Table 7 lists the example of the input and output values according to this image sequence. The quantified input values can be obtained from the formulas defined in Sect. 4.3. For example, after calculating the frequency value using FFT (0.8 cycle/s) and recognizing the head movement direction (nodding), formula (12) can be applied to determine the input value of head movement, which is 0.633. Following the same strategy, we obtained the input value for eye gaze and facial expression. By applying the knowledge-based Mamdani-type fuzzy inference system, which manually labels membership function boundary parameters uniformly, the output value is 0.6, which is the "Admire" emotion.

A hazard of conventional fuzzy systems (Mamdani-type) is that it may be based on qualitative and imprecise knowledge from an operator's or expert's perspective. It requires a thorough understanding of the fuzzy variables, their mapping strategy, and their membership functions, as well as the selection of fuzzy rules. In contrast, neural fuzzy systems can enhance fuzzy reasoning by acquiring knowledge from training input data. This has the advantage that it lets the data "speak" for itself. To take advantage of that, we also applied a Sugeno-type (Takagi and Sugeno 1985) neuro fuzzy system that can generate membership function and membership function boundaries automatically from training data.

For training and testing procedures, we followed the leave-one-out protocol. Hybrid (Jang 1993) provided the optimal method for defining the parameters of the input and output membership functions. After training, we calculated the average testing error between the trained FIS output and the testing data. Figure 16 shows an example that compares the outputs between the training data for participants 1, 2, 3, and 5 (red asteroids) with the testing data for participant 4 (blue dots) of Emotion set-A. The average testing error was 0.054581. The same procedure was followed for each participant. Table 8 lists the average testing errors of the neural fuzzy classifier. In order to evaluate both Mamdani-type FIS and Sugeno-type FIS, we tested the recognition rate of the 110 image sequences for each system.

Table 9 shows the recognition results of each method under different emotional sets. Clearly both performed well. In use, the choice between these two approaches will depend on practical considerations. The membership function boundaries for knowledge-based Mamdani-type FIS are manually labelled, assuming that the cases are uniformly distributed. However, there is a chance that these boundaries are not perfectly uniformly distributed for a

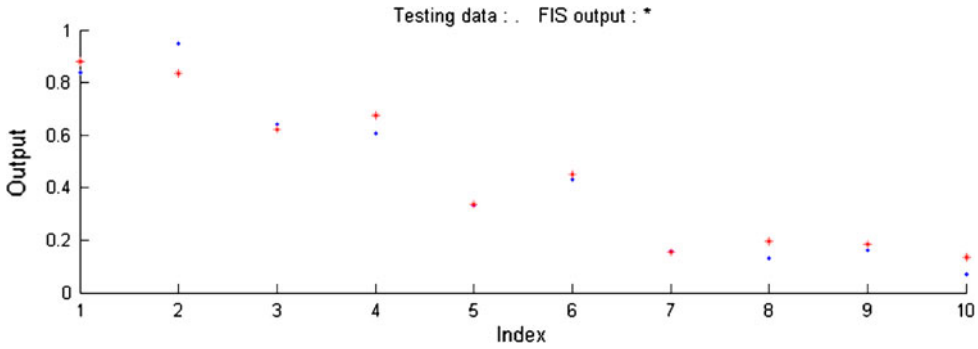


Fig. 16 Example of comparison of testing data and FIS output

Table 8 Average testing error

Participants	Average testing error		
	Emotion set A	Emotion set B	Emotion set C
1	0.071485	0.080826	0.04722
2	0.059401	0.22968	0.053492
3	0.05659	0.06852	0.025724
4	0.054581	0.076242	0.04675
5	0.078943	0.089649	0.052105
Average	0.0642	0.1089834	0.0450582

Table 9 Recognition rate

	Recognition rate		
	Emotion set A (%)	Emotion set B (%)	Emotion set C (%)
Sugeno-type	96	93.32	96.66
Mamdani-type	96	96.66	96.66

particular target user group. Therefore, we further applied the Sugeno-type neural FIS which could automatically generate membership functions boundaries based on the training data sets. When the training data is insufficient, the disadvantage of a Sugeno-type fuzzy system is that the training data set might not fully represent the target group and thus leads to incorrect recognition. On the other hand, the Mamdani-type system is easy and intuitive, but less adaptive to changes. The neural fuzzy system can be improved by training with more input and output data, while the Mamdani-type system cannot. Thus, if more input and output data are collected from experiments in the future, the recognition rate of the neural fuzzy system can be improved.

6 Conclusion

The automatic recognition and analysis of human emotions is one of the most challenging tasks in the field of HCII, which has attracted much interest in the past two decades. However, despite the fact that humans have different ways of expressing emotions, most of the past research in machine analysis of human emotion has focused on the recognition of the prototypical six basic facial expressions.

The originality of this paper is that it explored new ways of HCII to distinguish emotions and cognitions by integrating information from the body language of the head. In particular, we concentrated on integrating facial expressions, eye gaze direction and head movements. These three modalities were detected and mapped into emotional and cognitive states by applying fuzzy inference systems. We defined the fuzzy rules based on a questionnaire answered by non-expert participants. Although the creation of the fuzzy rules is culture-based, the idea of integrating the different modalities of the body language of the head is generic enough to be used by different target user groups in different cultures. Experimental results show that our method can be used to successfully recognize ten different emotions and cognitions. In the future, other modalities such as hand gestures and body postures will be combined with head information to detect a broader range of emotions and cognitions more precisely.

References

Al-amri SS, Kalyankar NV, Khamitkar SD (2010) Linear and non-linear contrast enhancement image. Int J Comput Sci Netw Secur 10(2):139–143

Arroyo I, Cooper DG, Burleson W, Woolf BP, Muldner K, Christopherson R (2009) Emotion sensors go to school. In: Proceedings of 14th conference on artificial intelligence in education, pp 17–24

Asteriadis S, Tzouveli P, Karpouzis K, Kollias S (2009) Estimation of behavioral user state based on eye gaze and head pose—application in an e-learning environment. J Multimedia Tools Appl 41(3). doi:10.1007/s11042-008-0240-1

Baenziger T, Grandjean D, Scherer KR (2009) Emotion recognition from expressions in face, voice, and body: the multimodal emotion recognition test (MERT). Emotion 9:691–704

Balomenos T, Raouzaiou A, Ioannou S, Drosopoulos A, Karpouzis K, Kollias S (2005) Emotion analysis in man–machine interaction system. LNCS 3361 3361:318–328

Baron-Cohen S, Tead THE (2003) Mind reading: the interactive guide to emotion. Jessica Kingsley Publishers, London

Brigham EO (2002) The fast Fourier transform. Prentice-Hall, New York

Calvo RA, D'Mello S (2010) Affect detection: an interdisciplinary review of models, methods, and their applications. IEEE Trans Affect Comput 1:18–37. doi:10.1109/T-AFFC.2010.1

Castellano G, Kessous L, Caridakis G (2008) Emotion recognition through multiple modalities: face, body gesture, speech. Affect and emotion in human-computer interaction. Springer, Berlin, pp 92–103. doi:10.1007/978-3-540-85099-1_8

Chakraborty A, Konar A (2009) Emotion recognition from facial expressions and its control using fuzzy logic. IEEE Transact Syst Man Cybernet 39(4). doi:10.1109/TSMCA.2009.2014645

Chatterjee S, Hao S (2010) A novel neuro fuzzy approach to human emotion determination. Int Conf Digital Image Comput Techniq Appl 283–287. doi:10.1109/DICTA.2010.114

Cohn JF, Read LI, Ambadar Z, Xiao J, Moriyama T (2004) Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior. In: Proceedings of IEEE international conference systems, man, and cybernetics (SMC '04), vol 1, pp 610–616

Contreras R, Starostenko O, Alarcon-Aquino V, Flores-Pulido L (2010) Facial feature model for emotion recognition using fuzzy reasoning. Adv Pattern Recogn 6256:11–21. doi:10.1007/978-3-642-15992-3_2

D'Mello S, Graesser A (2010) Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. User Model User-Adap Inter 10:147–187. doi:10.1007/s11257-010-9074-4

El Kaliouby R, Robinson P (2004) Real-time inference of complex mental states from facial expressions and head gestures. In: Proceedings international conference computer vision and pattern recognition 3:154. doi:10.1109/CVPR.2004.153

Engin M (2004) ECG beat classification using neuro-fuzzy network. Pattern Recogn Lett 25(15):1715–1722. doi:10.1016/j.patrec.2004.06.014

Esau N, Wetzel E, Kleinjohann L, Kleinjohann B (2007) Real-time facial expression recognition using a fuzzy emotion model. In: IEEE international conference fuzzy systems, 351–356. doi:10.1109/FUZZY.2007.4295451

Feng G (2006) A survey on analysis and design on model-based fuzzy control system. In: IEEE transaction on fuzzy systems, vol 14, issue 5. doi:10.1109/TFUZZ.2006.883415

Friedman J, Hastie T, Tibshirani R (2000) Additive logistic regression: a statistical view of boosting. Ann Stat 28(2):337–374

Ganel T (2011) Revisiting the relationship between the processing of gaze direction and the processing of facial expression. J Exp Psychol Hum Percept Perform 37(1):48–57

Gegov A (2007) Complexity management in fuzzy systems. Studies in fuzziness and soft computing, vol 211. Springer, Berlin

Giripunje S, Bawane N (2007) ANFIS based emotions recognition in speech. Knowledge-based Intelligent Information and Engineering Systems, vol 4692. Springer-Verlag, Berlin, Heidelberg, pp 77–84

Gunes H, Pantic M (2010) Automatic, dimensional and continuous emotion recognition. Int J Synth Emot 1:68–99. doi:10.4018/jse.2010101605

Gunes H, Piccardi M (2005a) Affect recognition from face and body: early fusion versus late fusion. In: Proceedings of IEEE international conference systems, man, and cybernetics (SMC'05), pp 3437–3443. doi:10.1109/ICSMC

Gunes H, Piccardi M (2005b) Fusing face and body display for bi-modal emotion recognition: single frame analysis and multi-frame post-integration. In: Proceedings of first international conference affective computing and intelligent interaction, pp 102–111. doi:10.1007/11573548_14

Gunes H, Piccardi M (2009) From monomodal to multi-modal: affect recognition using visual modalities. Ambient intelligence techniques and applications. Springer-Verlag, Berlin, pp 154–182. doi:10.1007/978-1-84800-346-0_10

Harris C, Stephens M (1988) A combined corner and edge detector. In: Proceedings of the 4th Alvey Vision Conference, vol 15, pp 146–151

Ioannou S, Caridakis G, Karpouzis K, Kollias S (2007) Robust feature detection for facial expression recognition. J Image Video Process 2007(2). doi:10.1155/2007/29081

Jaimes A, Sebe N (2007) Multimodal human-computer interaction: a survey. Comput Vis Image Underst 108:116–134

Jang JSR (1993) ANFIS: adaptive-network-based fuzzy inference systems. IEEE Transact Syst Man Cybernet 23(3):665–685. doi:10.1109/21.256541

Ji Q, Lan P, Looney C (2006) A probabilistic framework for modeling and real-time monitoring human fatigue. IEEE Syst Man Cybernet Part A 36(5):862–875. doi:10.1109/TSMCA.2005.855922

Jones J, Palmer L (1978) An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. J Neurophysiol 58(6):1233–1258

Kapoor A, Picard RW (2005) Multimodal affect recognition in learning environment. In: Proceedings of 13th annual ACM international conference multimedia, pp 677-682. doi:10.1145/1101149.1101300

Katsis CD, Katertsidis N, Ganiatsas G, Fotiadis DI (2008) Toward emotion recognition in car-racing drivers: a biosignal processing approach. IEEE Transact Syst Man Cybernet Part A Syst Humans 38(3), 502–512. doi:10.1109/TSMCA.2008.918624

Khezir M, Jahed M (2007) Real-time intelligent pattern recognition algorithm for surface EMG signals. BioMedical Engineering Online 6(1). doi:10.1186/1475-925X-6-45

Khezri M, Jahed M, Sadati N (2007) Neuro-fuzzy surface EMG pattern recognition for multifunctional hand prosthesis control. In: IEEE International symposium on industrial electronics, pp 269–274. doi:10.1109/ISIE.2007.4374610

Lee C, Narayanan S (2003) Emotion recognition using a data-driven fuzzy inference system. In: Proceedings of Eurospeech, Geneva, pp 157–160

Lyons M, Budynek J, Akamatsu S (1999) Automatic classification of single facial images. IEEE Trans Pattern Anal Mach Intell 21:1357–1362. doi:10.1109/34.817413

Mamdani EH, Assilian S (1975) An experiment in linguistic synthesis with a fuzzy logic controller. Int J Man Mach Stud 7:1–13. doi:10.1006/ijhc.1973.0303

Mandryk RL, Atkins MS (2007) A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. Int J Hum Comput Stud 65(4):329–347. doi:10.1016/j.ijhcs.2006.11.011

Marsh AA, Elfenbein HA, Ambady N (2003) Nonverbal "accents": cultural differences in facial expressions of emotion. Psychol Sci 14(4):373–376

Matsumoto D (1989) Cultural influences on the perception of emotion. J Cross Cult Psychol 20(1):92–105. doi:10.1177/0022022189201006

Matsumoto D (2001) The handbook of culture and psychology. Oxford University Press, USA

Matsumoto D, Kasri F, Kooken K (1999) American-Japanese cultural differences in judgments of expression intensity and subjective experience. Cogn Emot 13(2):201–218. doi:10.1080/026999399379339

Pantic M, Bartlett MS (2007) Machine analysis of facial expressions: face recognition. I-Tech Education and Publishing, Vienna, pp 377–416

Picard R (1997) Affective computing. MIT Press, Cambridge

Reginald B, Jr Adams, Robert EK (2005) Effects of direct and averted gaze on the perception of facially communicated emotion. Emotion 5(1):3–11. doi:10.1037/1528-3542.5.1.3

Reginald B, Jr Adams, Robert G, Jr Franklin (2009) Influence of emotional expression on the processing of gaze direction. Motivat Emot 33(2):106–112. doi:10.1007/s11031-009-9121-9

Schalk J, Hawk ST, Fishcher AH, Doosje B (2011) Moving faces, looking places: validation of the Amsterdam dynamic facial expression set (ADFES). Emotion 11(4):907–920. doi:10.1037/a0023853

Scherer K, Ellgring H (2007) Multimodal expression of emotion: affect programs or componential appraisal pattern? Emotion 7:158–171. doi:10.1037/1528-3542.7.1.158

Sebe N, Cohen I, Gevers T, Huang TS (2005) Multimodal approaches for emotion recognition: a survey. Proc SPIE-IS&T Electronic Imag SPIE 5670:56–67. doi:10.1117/12.600746

Shi J, Tomasi C (1994) Good features to track. In: Proceedings of IEEE Computing Society Conference on Computing Vision and Pattern Recognition, pp 593–600. doi: 10.1109/CVPR.1994.323794

Takagi T, Sugeno M (1985) Fuzzy identification of systems and its applications to modeling and control. IEEE Trans Syst Man Cybern 15(1):116–132

Taur JS, Tao CW (2000) A new neuro-fuzzy classifier with application to on-line face detection and recognition. J VLSI Sig Proc 26(3):397–409. doi:10.1023/A:1026515819538

Valstar MF, Gunes H, Pantic M (2007) How to distinguish posed from spontaneous smiles using geometric features. In: Proceedings of ninth ACM international conference on multimodal interfaces (ICMI '07), pp 38–45. doi:10.1145/1322192.1322202

Viola P, Jones M (2001) Robust real-time object detection. Cambrige Research Laboratory Technical Report Series CRL2001/01, pp 1–24

Vukadinovic D, Pantic M (2005) Fully automatic facial feature point detection using Gabor feature based boosted classifiers. IEEE Int Conf Syst Man Cybernet 2:1692–1698. doi:10.1109/ICSMC.2005.1571392

Zadeh L (1973) Outline of a new approach to the analysis of complex systems and decision processess. IEEE Trans Syst Man Cybern SMC-3(1):28–44. doi:10.1109/TSMC.1973.5408575

Zeng Z, Hu Y, Roisman G, Wen Z, Fu Y, Huang T (2006) Audio-visual emotion recognition in adult attachment interview. In: Quek JYFKH, Massaro DW, Alwan AA, Hazen TJ (eds) Proceedings of international conference multimodal interfaces, pp 139–145. doi:10.1145/1180995.1181028

Zeng Z, Pantic M, Roisman GI, Huang TS (2009) A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell 31(1):39–58. doi:10.1109/TPAMI.2008.52

Zhang Y, Ji Q (2005) Active and dynamic information fusion for facial expression understanding from image sequences. IEEE Trans Pattern Anal Mach Intell 27(5):699–714. doi:10.1109/TPAMI.2005.93

Zhao Y (2009) Facial expression recognition by applying multi-step integral projection and SVMs. In: Proceedings of IEEE instrumentation and measurement technology conference, pp 686–691. doi:10.1109/IMTC.2009.5168537

Zhu Z, Ji Q (2006) Robust real-time face pose and facial expression recovery. In: Proceedings of IEEE international conference computer vision and pattern recognition (CVPR '06) 1:681–688. doi:10.1109/CVPR.2006.259