CrossMark

New semantic descriptor construction for facial expression recognition based on axiomatic fuzzy set

Zedong Li[1] · Qingling Zhang[1] · Xiaodong Duan[2] ·
Cunrui Wang[2,1] · Yu Shi[2]

Abstract In this paper, we propose a new semantic descriptor based on axiomatic fuzzy set (AFS) to describe facial expressions. The new descriptor has two advantages: The first one is that it does not depend on priori-knowledge, when one uses it to construct semantic concepts. According to the distribution of feature data, one can quickly establish semantic concepts using the fuzzy membership degree. The second one is that the descriptor can describe complex features by implementing operation on semantic concepts. The developed descriptor can provide variations and relations of expression features. Finally, we implement our method on FEI and CK+ database, and make semantic interpretations for various expressions. Meanwhile, the performance is evaluated with the state-of-the-art methods such as C4.5, Bayes, Decision Table, Cart and Reduced error pruning tree.

Keywords Semantic descriptor · Axiomatic fuzzy set · Expression features · Semantic interpretation

1 Introduction

Facial expression plays an important role in our daily life and provides rich and powerful source of communicative information about human emotions. In the past two decades, it not only has attracted much attention, but also numerous solution approaches have been proposed for expression recognition.

The existing methods for facial expression recognition can be generally categorized into subspace methods [1, 41, 42], local feature methods [6, 43, 45], facial action unit methods [26,

✉ Xiaodong Duan
   minzudigital@dlnu.edu.cn

[1] Institute of Systems Science, Northeastern University, Shenyang, Liaoning, China

[2] Dalian Key Laboratory of Digital Technology for National Culture, Dalian Minzu University, Dalian, Liaoning, China

🖄 Springer

34] and geometric feature methods [3, 28]. The most popular subspace methods mainly focus on principal component analysis (PCA) [39], linear discriminant analysis (LDA) [47] and support vector machine [4]. These algorithms pay much attention on the image global properties and extract the essential features. Based on a derived projection matrix, the original expressions are represented in a new subspace. However, the facial expression images are usually non-linear, and the details are distributed in different regions. Due to emphasis on features of local region, local feature methods are usually applied to solve these troubles. For example, the local binary patterns (LBP) and Gabor [10, 45, 48] have attracted much more attention, because they can well solve expressional variations, illuminations and occlusions. However, all above methods have ignored semantic descriptions/features of facial components. When the facial action coding system (FACS) [9] emerged, it was rapidly used in expression recognition [25, 26]. According to this system, the changes of facial expression could be visually described, and clearly endowed by semantic concepts. However, the numerous landmarks in this system increase difficulties of applications because they are manually marked. Accompanied with the automatic landmark detector [5], geometric feature method has attracted much attention on facial expression again [17, 44]. Generally, the fuzzy methods are used to interpret the concepts of facial components, and related rules will be generated for the corresponding results. Specially, these landmarks can well represent the shape of facial components and bridge semantic gap between low-level visual features of image and high-level semantic concepts. Usually, the combining feature including the facial action unit and geometric feature is applied in facial expression recognition [2, 32, 38]. However, all these methods need priori-knowledge such as fuzzy models and prior-defined semantic features.

After thorough comparisons, Axiomatic fuzzy set (AFS) [18, 20] has broken the limitation of the existing fuzzy models, and it has extensive applications such as text recognition [33] and knowledge discovery [35]. Specially, it has been applied in facial characteristics description [16, 31] because of its advantage in semantic interpretability. In the AFS theory, fuzzy sets (membership functions) and their logic operations are algorithmically determined according to the distributions of original data and the semantics of the fuzzy sets. The AFS framework facilitates the studies on how to convert the information in databases into the membership functions and their fuzzy logic operations, by taking both fuzziness (subjective imprecision) and randomness (subjective uncertainty) into account. The advantage of AFS framework is that it does not be required to define membership function and initial value in advance as all these can be learned from original database.

In this paper, a new descriptor is proposed in framework of AFS. This descriptor not only interprets expression features using semantic concepts, but also serves as a classifier for expression recognition. The contributions of this paper are listed as follows:

- We propose a new descriptor extraction algorithm for facial expressions. In comparison with traditional methods, it can directly extract semantic concepts by using the data distribution.
- By considering differences among all these seven expressions, each semantic concept will be built respectively, which can be comprehended by observers intuitively.
- Moreover, in order to verify our method, we have conducted experiments for analyzing expressions on FEI [36, 37] and CK+ [13, 24]. Finally, there is a comparison about performance with some of state-of-the-art methods such as C4.5 [29], Bayes [30], Decision Table (DT) [12], Cart [15] and Reduced error pruning tree (REP) [8].

The remainder of this paper is organized as follows: in section 2, we give a brief introduction of related works. Section 3 describes facial expression features constructed by landmarks, and details the process of the proposed method. Experimental results are reported in section 4. Finally, we conclude the paper in Section 5.

2 Preliminaries

2.1 AFS theory

AFS theory was proposed by Liu [18–20], whose fuzzy sets (membership function) and logic operations are algorithmically defined following the distributions of original data and the semantic concepts of the fuzzy sets. In its framework, AFS does not demand pre-defined membership functions as those can be learned from original data, so that it is suitable for transforming facial characteristics to semantic concepts. In what follows, we offer an illustration to AFS theory with one simple example. Let us assume that there are three attributes "The distance between the points of inner canthus and the points of inner eyebrow", "The angle between the point of inner canthus and the point of inner eyebrow in right eye" and "The angle between the point of inner canthus and the point of inner eyebrow in left eye" to describe the facial characteristics. Let $M = \{m_{j,k} \; 1 \leq j \leq 3, 1 \leq k \leq 3\}$ be a set of simple semantic concepts, where $m_{j,1}$, $m_{j,2}$, $m_{j,3}$ is respectively corresponding to the semantic concept "large", "medium" and "small" associated with feature f_j. We will generate a complex semantic concepts $\prod_{m \in A} m(A \subseteq M)$ for each set of concept terms using a simple concept set belonged to M, which represents a conjunction of the concept terms in A. For example, let $A_1 = \{m_{1,3}, m_{2,1}, m_{3,2}\} \subseteq M$, a new fuzzy set "$m_{1,3}$ and $m_{2,1}$ and $m_{3,2}$" with the linguist interpretation: "The distance between the points of inner canthus and the points of inner eyebrow is smaller and The angle between the point of inner canthus and the point of inner eyebrow in right eye is larger and The angle between the point of inner canthus and the point of inner eyebrow in left eye is medium" can be represented as $\prod_{m \in A} m = m_{1,3} \; m_{2,1} \; m_{4,2}$. Similarly, Let $A_2 = \{m_{1,1}, m_{2,3}, m_{3,2}\}$, $A_3 = \{m_{1,3}, m_{3,1}\} \subseteq M$, a new concept set as the disjunction of $\prod_{m \in A1} m$, $\prod_{m \in A2} m$, $\prod_{m \in A3} m$, i.e., "$m_{1,3}m_{2,1}m_{4,2}$ or $m_{1,1}m_{2,3}m_{3,2}$ or $m_{1,3}m_{3,1}$" can be represented by $\sum 3 \; u = (\prod_{m \in Au} m) = \prod_{m \in A1} m + \prod_{m \in A2} m + \prod_{m \in A3} m$. The set EM can be defined in the following way.

$$EM = \left\{ \sum_{i \in I} \left(\prod_{m \in A_i} m \right) | A_i \subseteq M, i \in I, I \text{ is } a \text{ non-indexing set} \right\} \quad (1)$$

In addition, there is an equivalence relation [22, 23] between α and β. For example, there are two complex concepts, $\gamma_1 = m_{1,3} \; m_{2,1} \; m_{3,1} + m_{1,1} \; m_{2,2} \; m_{3,2} + m_{1,3} \; m_{3,1}$, $\gamma_2 = m_{1,3} \; m_{2,1} \; m_{3,1} + m_{1,3} \; m_{3,1}$. They are equivalent according to equivalence relation. This means that the concept "$m_{1,1}m_{2,2}m_{3,2}$" is redundant. Meanwhile, AFS demonstrates that $(EM, \; , \;)$ is a completely distributive lattice if the lattice operators " " and " " are defined as follows: for any fuzzy sets $\sum_{i \in I}(\prod_{m \in Bi} m) \in EM$, $\sum_{j \in J}(\prod_{m \in C \; j} m) \in EM$.

$$\sum_{i \in I} \left(\prod_{m \in A_i} m \right) \wedge \sum_{j \in J} \left(\prod_{m \in B_j} m \right) = \sum_{k \in I \cup J} \left(\prod_{m \in D_k} m \right) \quad (2)$$

$$\sum_{i \in I} \left(\prod_{m \in A_i} m \right) \vee \sum_{j \in J} \left(\prod_{m \in B_j} m \right) = \sum_{i \in I, j \in J} \left(\prod_{m \in A_i \cup B_j} m \right) \tag{3}$$

where for any $k \in I \cup J$ (the disjoint union of I and J, i.e., every element in I and every element in J are always regarded as different elements in $I \cup J$), $D_k = A_k$ if $k \in I$, and $D_k = B_k$ if $k \in J$.

In the AFS framework, the membership function is utilized to interpret concept of $m_{j,k}$. Its ordered relation is defined as follows. Let X be an observed data set, and M be a set of fuzzy terms on X. For $A \subseteq M, x \in X$, we define

$$A^{\succeq}(x) = \left\{ y \in X \middle| x \succeq_m y, \text{for any } m \in A \right\} \subseteq X \tag{4}$$

where a linearly ordered relation is denoted by "$\succeq$". For $m \in M$, "$x \succeq_m y$" implies that the degree of x belonging to m is larger than or equal to that of y. $A^{\succeq}(x)$ is the set of all elements in which their degrees of belonging to set $\prod_{m \in A} m$ are less than or equal to that of x. $A^{\succeq}(x)$ is determined by the semantics of fuzzy set A and the probability distribution of data.

Theorem 1 ([20]) *Let v be a fuzzy term on X. $v : X \Rightarrow R+ = [0,\infty)$. ρ_v is called a weight function of the simple concept v if ρ_v satisfies the following conditions:*

1. *For $x \in X, \rho_v(x) = 0 \Leftrightarrow x \npreceq_m x$;*
2. *For $x, y \in X, \rho_v(x) \ge \rho_v(y) \Leftrightarrow x \npreceq_m y$.*

The coherence membership function is defined as follows:

$$\mu_\xi(x) = \sup_{i \in I} \inf_{\gamma \in A_i} \frac{\sum_{u \in A_i^{\succeq}(x)} \rho_\gamma(u) N_u}{\sum_{u \in X} \rho_\gamma(u) N_u}, \forall x \in X \tag{5}$$

where, Nu is the number of observed times of sample u. Some other properties and its applications of AFS are detailed in the literature [20–23].

2.2 The max-dependency, max-relevance and min redundancy algorithm (mRMR)

In this section, a feature selection algorithm will be briefly introduced, which is established according to mutual information theory [7, 27]. Its aim is to find a feature set M with n features $\{m_i\}$, which jointly exhibit the largest dependency on the target class C. Usually, it searches the subset $M_n - 1$ with $n - 1$ features according to the contribution to the largest increase of entropy $I(M_n; C)$, which is presented as follows:

$$
\begin{aligned}
I\ (M_n; C) &= \int \int p\ (M_n, C)\ \log \frac{p(M_n, C)}{p(M_n)p(C)} dM_n dC \\
&= \int \int p\ (M_{n-1}, m_n, C) \log \frac{p(M_{n-1}, m_n, C)}{p(M_{n-1}, m_n)p(C)} dM_{n-1} dm_n dC \\
&= \int \cdots \int p(m_1, \cdots, m_n, C) \log \frac{p(m_1, \cdots, m_n, C)}{p(m_1, \cdots, m_n)p(C)} dm_1 \cdots dm_n dC
\end{aligned}
\tag{6}
$$

where $M_n = \{M_{n-1}, m_n\}$ can be treated as a multivariate variable. Thus, by the definition of mutual information, it can be represented in the form:

$$I(M_n; C) = H(C) + H(M_n) - H(M_n, C) = H(C) + H(M_{n-1}, m_n) - H(M_{n-1}, m_n, C) \quad (7)$$

where $H(\cdot)$ represents the entropy of the respective multivariate (or univariate) variables. Similarly, the quantity $J(M_n) = J(m_1, m_2, \ldots, m_n)$ is defined for scalar variables $m_1, m_2, \ldots, m_n$

$$J(M_n) = J(m_1, \cdots, m_n) = \int \cdots \int p(m_1, \cdots, m_n) \log \frac{p(m_1, \cdots, m_n)}{p(m_1) \cdots p(m_n)} dm_1 \cdots dm_n \quad (8)$$

$$J(M_n, C) = J(m_1, \cdots, m_n, C)$$

$$= \int \cdots \int p(m_1, \cdots, m_n, C) \log \frac{p(m_1, \cdots, m_n, C)}{p(m_1) \cdots p(m_n) p(C)} dm_1 \cdots dm_n dC \quad (9)$$

Then, the formula (11) and (10) can be easily derived as follows:

$$H(M_{n-1}, m_n) = H(M_n) = \sum_{i=1}^{n} H(m_i) - J(M_n) \quad (10)$$

$$H(M_{n-1}, m_n, C) = H(M_n, C) = H(C) + \sum_{i=1}^{n} H(m_i) - J(M_n, C) \quad (11)$$

By substituting them in the corresponding term in formula (6), we can rewrite the form of entropy $I(M_n; C)$ as follows:

$$I(M_n; C) = J(M_n, C) - J(M_n) = J(M_{n-1}, m_n, C) - J(M_{n-1}, m_n) \quad (12)$$

However, due to insufficient samples, high-dimensional and intensive computing required in above estimation, it is hard to arrive at an accurate estimation. Therefore, the maximal-relevance-minimal-redundancy (mRMR) algorithm [27], is proposed to select a subset of M, by choosing mutually exclusive features via optimizing the maximal relevance criterion $D(M,C)$ and minimal redundancy condition $R(M)$. In summary, by combining the above constraints, the operator $\Phi(D, R)$ can be defined. Here, the terms D and R need to be optimized simultaneously. The simplest form is in the following form:

$$\max \Phi \quad (D, R) \quad , \quad \Phi \quad = \quad D \quad - \quad R$$
$$s.t. \begin{cases} \max D(M, C), D = \frac{1}{|M|} \sum_{m_n \in M} I(m_n; C) \, \min R(M), R = \frac{1}{|M|^2} \sum_{m_i, m_j \in M} I(m_i, m_j) \end{cases} \quad (13)$$

where the mutually exclusive (independent) features are extracted through mining $J(Mn)$ and the max-relevance features will be chosen by maxing the $J(Mn, C)$. Finally, mRMR can select a compact set of superior features. The details are illustrated in literature [7, 27].

3 SDAFS: a semantic descriptor based on AFS

In this section, we will create a semantic descriptor based on AFS (SDAFS). The algorithmic process mainly contains three stages. The first one is to select features. The second stage is to establish semantic description for each individual. The third one is to optimize concept set for obtaining the best interpretation for each expression. We first interpret the key symbols used in this paper in Table 1.

3.1 Construction of expression features construction

When expression is changed, the geometrical shape of facial components will be varied intuitively (see Fig. 1).

According to literature [10, 14], the characteristics are main concentrated on facial principal regions such as eyes, mouth and nose. Therefore, we construct 30 expression features by landmarks [5] for describing expression variations. Let $L_K = \{l_1, l_2, \ldots, l_k\}$ be a set of landmarks extracted from each expression image. $l_i = (r_i, c_i) \in L_K$ presents the points of facial component. r_i and c_i are i-th coordinates of point l_i. In this case, the $d(l_h, l_k)(l_h, l_k \in L_K)$ represents the distance between two points. The $\angle(l_h, l_k, l_p)(l_h, l_k, l_p \in L_K)$ indicates the angle among three points. Their calculations are illustrated as follows:

$$d(l_h, l_k) = \sqrt{\left(l_h^r - l_k^r\right)^2 + \left(l_h^c - l_k^c\right)^2} \tag{14}$$

$$\cos\angle\left(l_h, l_k, l_p\right) = \frac{d^2(l_h, l_k) + d^2\left(l_p, l_k\right) - d^2\left(l_h, l_p\right)}{2 \times d(l_h, l_k) \times d\left(l_p, l_k\right)} \tag{15}$$

For all the measurements including $d(l_h, l_k)$ and $cos\angle(l_h, l_k, l_p)$, there are 30 expression features $F_{total} = \{f_1, f_2, \cdots, f_{30}\}$ to be established, which are listed illustrated in Table 2.

3.2 Construction of compact feature subset by using mRMR

The constructed features can well describe the characteristics of facial components. However, the effectiveness is distinction among various features for different expressions. In this study, mRMR is used to select the compact feature subset for differentiating expressions. The aim of mRMR is to search a subset $F \in F_{total}$, which could satisfy the minimal redundancy and maximal relevance simultaneously. This strategy consists of two steps.

- First, according to formula (13), the score $S\ score\ j$ is built to evaluate the performance for corresponding $f_j \in F_{total}$. Then, the score $S^{score} = \{S\ score\ 1, S\ score\ 2, \cdots, S\ score\ 30\}$ will be obtained.
- Second, the element $S\ score\ j \in S^{score}$ will be sorted in a descent order. Then, according to the score $S\ score\ j$, the feature subset F for each expression can be determined. The f_j, whose score is bigger, will be remained.

Table 1 The key symbols in this paper

| Key symbol	Description of symbol
F	A set of expression features
x_i	The individual sample
X	The set of all samples
X_{Ci}	The i-th subset in X
$m_{j,k}$	The k-th semantic concept associated with the j-th expression feature
M^{xi}	A set of semantic concepts for xi
EM^{Ci}	The conjunction semantic set for C_i class
ξ_{Ci}	The semantic description for cluster C_i
δ_1	A threshold value for selecting semantic concept of $m_{j,k}$
δ_2	A threshold value for selecting conjunction semantic
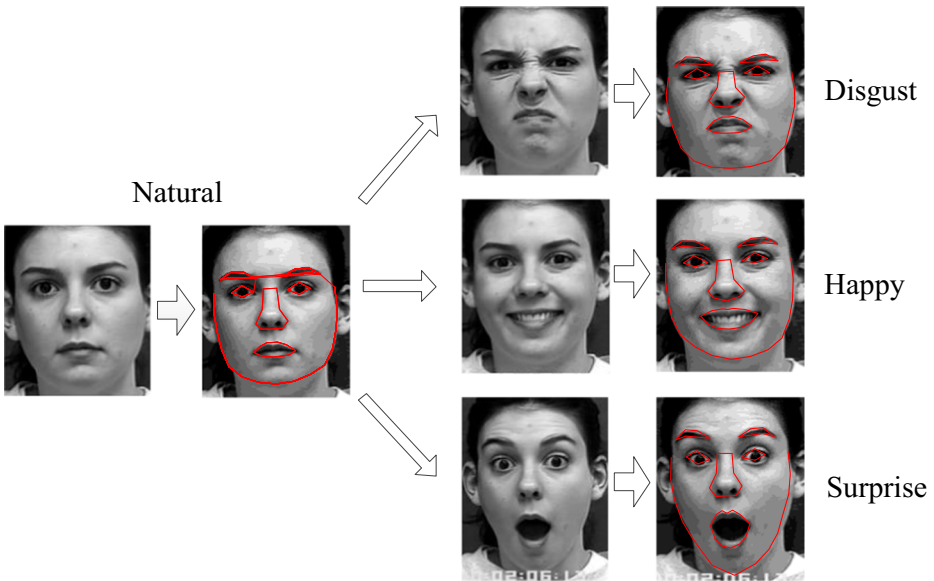
Fig. 1 The geometrical shape of facial components in various expressions

Table 2 The expression features and their semantic interpretation

Feature	Semantic interpretation
f_1	The distance between the points of inner canthus and the points of inner eyebrow
f_2	The angle between the point of inner canthus and the point of inner eyebrow in right eye
f_3	The angle between the point of inner canthus and the point of inner eyebrow in left eye
f_4	The angle between the point of outer canthus and the point of both ends of eyebrow in right eye
f_5	The angle between the point of outer canthus and the point of both ends of eyebrow in left eye
f_6	The radian of right eyebrow
f_7	The radian of left eyebrow
f_8	The height of eyes
f_9	The width of eyes
f_{10}	The angle between points of top and down of eye and inner canthus in right eye
f_{11}	The angle between points of top and down of eye and inner canthus in left eye
f_{12}	The area of right eye
f_{13}	The area of left eye
f_{14}	The angle between points of inner canthus of eyes and points of center of inner canthus eyebrows
f_{15}	The height of mouth
f_{16}	The width of mouth
f_{17}	The perimeter of mouth
f_{18}	The area of mouth
f_{19}	The angle of right corner of the mouth
f_{20}	The angle of left corner of the mouth
f_{21}	The angle between corners of mouth and nasal peak
f_{22}	The angle between corners of mouth and the top point of mouth
f_{23}	The angle between corners of mouth and the down point of mouth
f_{24}	The angle between points of outer canthus and the down point of mouth
f_{25}	The distance between the point of nasal peak and the point of under-jaw tip
f_{26}	The width of nose
f_{27}	The length of nose
f_{28}	The area of nose
f_{29}	The angle between points of inner canthus and nasal peak
f_{30}	The angle of under-jaw

3.3 Semantic description for each expression

By using the mRMR, the compact feature subset F can be obtained. However, any expression characteristics are different from others. Therefore, a new approach is proposed in this section to extract the representative features, and establish semantic concepts for describing various expressions. The process is divided into three steps, and the flow chart is illustrated in Fig. 2.

3.3.1 Step A: features selected from each individual

According to the sample label, it is very important to extract salient features, which can describe $x_i \in X$ effectively. By comparing with literature [31], we propose a new approach here to select the valuable features according to discriminative power of features. Technically, let $F = \{f_1, f_2, \cdots, f_S\}$ be a set of features, $X = \{x_1, x_2, \cdots, x_n\}$ be a set of samples. The $\mathscr{R}_{k/x_i}^{f_s}$ $(1 \leq s \leq S)$ denotes the k nearest neighbors around x_i with $f_j \in F$, $\mathscr{R}_{k/x_i}^{F}$ presents the k nearest neighbors around x_i with $F = \{f_1, f_2, \cdots, f_S\}$. Here, the distance $Dis^{(fs)}(x_i - x_j)$ between x_i and x_j
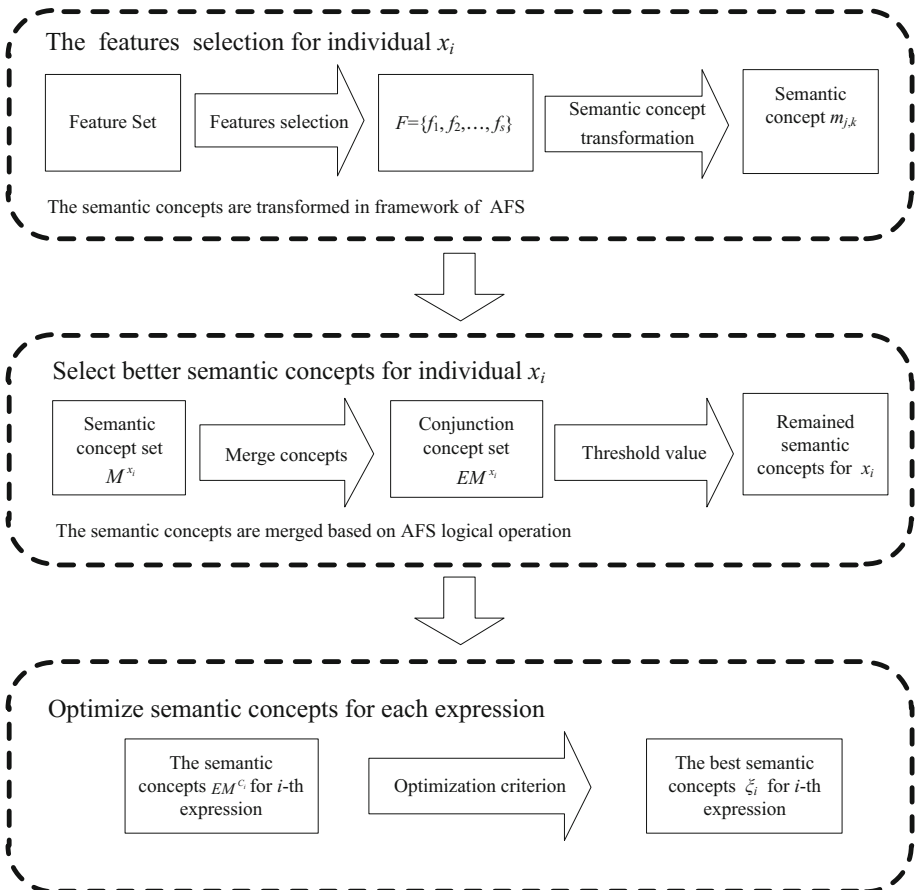

Fig. 2 The flow chart of the proposed method

$(x_i, x_j \in X_{Ci}, x_i \neq x_j)$ is estimated by the Euclidean Distance. Similarly, the distance $Dis^{(F)}(x_i\text{-}x_j)$ between x_i and x_j is also estimated by the Euclidean Distance.

For each x_i, we utilize the mutual neighbors to estimate the influence of f_s between $\mathscr{R}^{f_s}_{k/x_i}$ and $\mathscr{R}^{F}_{k/x_i}$. If the neighbors of x_i in f_s are similar to that of x_i in F, we consider f_s could represent x_i in feature set F.

$$\mathscr{R}^{f_s}_N = \mathscr{R}^{f_s}_{k/x_i} \cap \mathscr{R}^{F}_{k/x_i}, (f_s \in F) \tag{16}$$

where $\mathscr{R}^{f_s}_N$ is the number of neighbors between f_s and F, which can be used to reflect the performance of feature f_s. It means that the $\mathscr{R}^{f_s}_N$ contains more discriminative information than F on data X. Its extraction process is described in **Algorithm 1**.

Algorithm 1 Features Selection for individual x_i

Input: $F = \{f_1, f_2, \cdots, f_s\}$ is a set of expression features for individual x_i; $K_{neighbor}$ is the number of neighbors; $K_{f/F}$ is the mutual number of neighbors;

Output: $F'(x_i)$ is a subset of F, in which features are more effective for describing x_i.

1:   **for** each $x_i \in X_{Ci}$ **do**
2:        Compute Euclidean distance $Dis^{(F)}(x_i\text{-}x_j)$, $x_j \in X_{Ci}$, $x_i \neq x_j$;
3:        The elements of $Dis^{(F)}$ are sorted in descent order;
4:        Compute $\mathcal{R}^{F}_{k/x_i} = Dis^{(F)}(1:K_{neighbor})$;
5:        **for** each $f_s \in F$ **do**
6:             Compute Euclidean distance $Dis^{(fs)}(x_i\text{-}x_j)$, $x_j \in X_{Ci}$, $x_i \neq x_j$;
7:             The elements of $Dis^{(fs)}$ are sorted in descent order;
8:             Compute $\mathcal{R}^{f_s}_{k/x_i} = Dis^{(f_s)}(1:K_{neighbor})$;
9:             The elements of $\mathcal{R}^{f_s}_{k/x_i}$ are sorted in descent order;
10:            $\mathcal{R}^{f_s}_N = \mathcal{R}^{f_s}_{k/x_i} \cap \mathcal{R}^{F}_{k/x_i}$;
11:            **if** $\mathcal{R}^{f_s}_N \geq K_{f/F}$ **then**
12:                 The f_s is remained;
13:            **end if**
14:        **end for**
15:   **end for**
16:   $F'(x_i)= \{\Sigma f_s \mid f_s \in F\}$;

3.3.2 Step B: semantic concepts for each individual

In section 3.2, we obtained a set of features F' for each x_i. And each f_j is divided into three semantic concepts "$m_{j,1}$", "$m_{j,2}$" and "$m_{j,3}$", which corresponds to "large",

"medium" and "small" respectively. Then, we can obtain the semantic concept set for every x_i as follows.

$$M^{x_i} = \left\{ \sum m_{j,k} \Big| \mu_{m_{j,k}}(x_i) \geq \max\left\{ \mu_{m_{j,k}}(x_i) \right\} - \delta_1 \right\} \tag{17}$$

where δ_1 is a threshold for selecting $m_{j,k}$, which can describe x_i well. However, when the expression happens, several facial components will change their shape simultaneously. Therefore, the conjunction semantic concept set A^{xi} is selected for describing x_i by:

$$A^{x_i} = \left\{ \sum A_{\text{remain}}^{x_i} \Big| \mu_{A_{\text{remain}}^{x_i}}(x_i) - \delta_2 \geq 0 \right\} \tag{18}$$

where δ_2 is a threshold for selecting the semantic concept A^{xi}. Then, $A_{\text{remain}}^{x_i} = \prod m_{j,k}$ $\left(m_{j,k} \subset M^{x_i}\right)$ is a set of conjunction concepts set. However, we expect to obtain the best description $A\ x_i\ select \in A^{xi}$ for describing x_i. Assume that $\mathscr{R}_{k/x_i}^{A_s^{x_i}}$ is the k neighbors around x_i with semantic concept $A_s^{x_i} \in A^{x_i}$ and for each x_i, we can extract the best semantic concept $A_{\text{select}}^{x_i}$ according the number of mutual neighbors.

$$\mathscr{R}_N^{A_s^{x_i}} = \mathscr{R}_{k/x_i}^{A_s^{x_i}} \cap \mathscr{R}_{k/x_i}^{A^{x_i}}, \left(A_s^{x_i} \in A^{x_i}\right) \tag{19}$$

where $\mathscr{R}_N^{A_s^{x_i}}$ is the number of neighbors between $A_s^{x_i}$ and A^{x_i}. Then the $A_{\text{select}}^{x_i}$ corresponding to maximum $\mathscr{R}_N^{A_s^{x_i}}$ will be selected for x_i. Similarly as in section 3.2, the distance $D^{\left(A_s^{x_i}\right)}(x_i - x_j)$ between x_i and x_j ($x_i, x_j \in X_{Ci}, x_i \neq x_j$) is measured by the Euclidean Distance. However, the feature value, which describes x_i under $A_s^{x_i}$, is replaced by membership value calculated by formula (5).

3.3.3 Step C: optimizing the semantic concepts for each group

In this section, the main aim is to search the best semantic concepts for each expression. From section 3.3.2, we have obtained the salient concept for each $x_i \in X_{Ci}$. Then, we can summarize a set of concepts EM^{Ci} for C_i class as follows.

$$EM^{C_i} = \left\{ \sum_{x=x_i, x_i \in X_{C_i}} A_{\text{select}}^{x_i} \right\} \tag{20}$$

However, there are redundant concepts in EM^{Ci}, which may generate weak influence for C_i class. As well known, the compactness within class and separability between classes are very

important for discriminant feature selection. However, it is difficult to accomplish compactness and separability simultaneously. Hence, we design a new optimization criterion for selecting the best semantic concepts for each group according to compactness, separability and weighting.

$$\max \xi_{C_i}(\omega_\eta) = R^{C_i}(\omega_\eta) \cdot W^{C_i}(\omega_\eta) \cdot G^{C_i}(\omega_\eta), \omega_\eta \in EM^{C_i}$$
$$s.t. \begin{cases} R^{C_i}(\omega_\eta) = \max\{\theta(\omega_\eta)\}, 0 < \theta(\omega_\eta) < 1 \\ W^{C_i}(\omega_\eta) = \max\{\varphi(\omega_\eta)\}, 0 < \varphi(\omega_\eta) < 1 \\ G^{C_i}(\omega_\eta) = \max\{\psi(\omega_\eta)\}, 0 < \psi(\omega_\eta) < 1 \end{cases} \tag{21}$$

where $R^{C_i}(\omega_\eta)$ is considered as a quotient for satisfying greater separability and smaller compactness simultaneously. Here $Class_{sep}$ and $Class_{com}$ signify the separability and compactness respectively. And the quotient is larger, the result is better.

$$R(\omega_\eta) = \frac{Class_{sep}}{Class_{com}} = \frac{\sum\limits_{C_i=C_1}^{C} \sum\limits_{i=1}^{n_i} \left| \mu_{\omega_\eta}(x_i) - \mu_{\omega_\eta}\left(\overline{X_{C_i}}\right) \right|}{\sum\limits_{C_i=C_1}^{C} n_i \left| \mu_{\omega_\eta}\left(\overline{x_{C_i}}\right) - \mu_{\omega_\eta}(X) \right|} = \frac{\sum\limits_{C_i=C_1}^{C} \sum\limits_{i=1}^{n_i} \left| \mu_{\omega_\eta}(x_i) - \frac{1}{n_i} \sum\limits_{i=1}^{n_i} \mu_{\omega_\eta}(x_i) \right|}{\sum\limits_{C_i=C_1}^{C} n_i \left| \frac{1}{n_i} \sum\limits_{i=1}^{n_i} \mu_{\omega_\eta}(x_i) - \frac{1}{N} \sum\limits_{j=1}^{N} \mu_{\omega_\eta}(x_j) \right|}$$

$$= \frac{\sum\limits_{C_i=C_1, C_i \neq C_j}^{C} \sum\limits_{i=1}^{n_i} \mu_{\omega_\eta}(x_i)}{\sum\limits_{C_i=C_1, C_i \neq C_j}^{C} \left[\sum\limits_{i=1}^{n_i} \mu_{\omega_\eta}(x_i) - \frac{n_i}{N} \sum\limits_{j=1}^{N} \mu_{\omega_\eta}(x_j) \right]}$$

$$\tag{22}$$

where $\mu_{\omega_\eta}(x_i)$ represents the membership degree of x_i under semantic concept ω_η. n_i is the number of samples in i-th class. N is the total number of all samples. $\mu_{\omega_\eta}\left(\overline{X_{C_i}}\right)$ is the mean value of all samples in X_{Ci}. Then, we can introduce $R^{Ci}(\omega_\eta)$, $W^{Ci}(\omega_\eta)$ and $G^{Ci}(\omega_\eta)$ as follows:

$$R^{C_i}(\omega_\eta) = \max\{\theta(\omega_\eta)\}, \theta(\omega_\eta) = \frac{R(\omega_\eta)}{\sum_{\omega_\eta \in EM^{C_i}} R(\omega_\eta)} \tag{23}$$

$$W^{C_i}(\omega_\eta) = \max\{\varphi(\omega_\eta)\}, \varphi(\omega_\eta) = \frac{\sum_{i=1}^{n_i} \mu_{\omega_\eta}(x_i)}{\sum_{\omega_\eta \in EM^{C_i}} \sum_{i=1}^{n_i} \mu_{\omega_\eta}(x_i)} \tag{24}$$

$$G^{C_i}(\omega_\eta) = \max\{\psi(\omega_\eta)\}, \psi(\omega_\eta) = \frac{\sum_{i=1}^{n_i} \left[\mu_{\omega_\eta}(x_i) - \max\{x_t\} \right]}{\sum_{\omega_\eta \in EM^{C_i}} \sum_{i=1}^{n_i} \left[\mu_{\omega_\eta}(x_i) - \max\{x_t\} \right]} \tag{25}$$

where, $x_i \in X_{Ci}$, $\omega_\eta \in EM^{Ci}$, $x_t \in X$, $x_t \notin X_{Ci}$. X_{Ci} represents the i-th class. Then, the process of optimizing semantic concepts for each group is described in detail as **Algorithm 2**.

Algorithm 2 Optimizing the semantic concepts for each group

Input: EM^{Ci} is a set of semantic concepts i-th class;

Output: ξ_{Ci} is the best description for i-th class

1: **for** each $C_i(i=1, 2, \ldots, I)$ **do**

2:      **for** each $\omega_\eta \in EM^{Ci}$ **do**

3:          Compute $\mu_{\omega_\eta}(x_i)$;

4:      **end for**

5:      Compute $\sum_{i=1}^{n_i} \mu_{\omega_\eta}(x_i)$

6:      Compute $\max\{\mu_{\omega_\eta}(x_t)\}, x_t \in X, x_t \notin X_{C_i}$;

7: **end for**

8: Compute $R^{Ci}(\omega_\eta)$, $W^{Ci}(\omega_\eta)$ and $G^{Ci}(\omega_\eta)$, $\omega_\eta \in EM^{Ci}$;

9: Compute $\max\{\xi_{C_i}(\omega_\eta)\}$

3.4 Inference

In order to predict the class label for new samples which are not included in X, we need the semantic concept set ξ expressed over the entire input space $U_1 \times U_2 \times \cdots \times U_s \subseteq R^s (X \subseteq U_1 \times U_2 \times \cdots \times Us)$, where U_i is the feature associated with the semantic concept term $m_{j,k}$ in M. If the distribution of data set is given, the membership function of semantic concept set γ can be derived according to the observed data. For any semantic concept $\xi \in EM$, we expand its universe of discourse X to $U_1 \times U_2 \times \cdots \times U_s$. For each $x = (u_1, u_2, \cdots, u_s) \in U_1 \times U_2 \times \cdots \times U_s$, $\xi = \sum_{i \in I}(\prod_{m \in A_i} m) \in EM$, the lower bound of membership function of semantic concept set γ is defined over $U_1 \times U_2 \times \cdots \times U_s$ as follows:

$$\mu_\xi^L(x) = \sup_{i \in I} \inf_{\gamma \in A_i} \frac{\sum_{u \in L_{A_i}^x} \rho_\gamma(u) N_u}{\sum_{u \in X} \rho_\gamma(u) N_u} \tag{26}$$

where $L_{A_i}^x \subseteq X$, $i \in I$ are defined as $L_{A_i}^x = \{y \in X | x \succeq_m y, \forall m \in A_i\}$. We call $\mu_\xi^L(x)$ the lower bound of membership function of ξ. In virtue of formula (26), we can expand the fuzzy set $\xi = \sum_{i \in I} (\prod_{m \in A_i} m) \in EM$, from the universe of discourse X to the universe of discourse $U_1 \times U_2 \times \cdots \times U_s$. Therefore, by the fuzzy rule-base, we can establish fuzzy-inference systems whose input space is $U_1 \times U_2 \times \cdots \times U_s$. The membership function $\mu_\xi^L(x)$ is dependent on the distribution of training examples and the AFS fuzzy logic.

When, new pattern $x \in U_1 \times U_2 \times \cdots \times U_s$ is provided with unknown class label, we can calculate the membership degree $\mu_\xi^L(x)$ with (26), where ξ_{C_k} is the fuzzy description of the class C_i. Finally, x belongs to the class C_q, if $q = \text{argmax}_{1 \leq k \leq C}\{\mu_\xi^L(x)\}$.

4 Experimental results

In all the experiments, the membership functions of semantic concepts are determined by formula (5) in Theorem 1; for any $m_{j,k} \in M_j$, $\rho_{m_{j,k}}(x_i) = 1$; $N_{x_i} = 1$ denotes x_i is observed once. Three semantic concepts are defined on each feature f_j. $m_{j,1}$, $m_{j,2}$, $m_{j,3}$ are semantic concept terms, "large", "medium", "small" associated with the feature f_j in F. $m_{j,1}$ with the semantic meaning "the value on f_j is large", "$m_{j,2}$ with the semantic meaning "the value on f_j is medium", and $m_{j,3}$ with the semantic meaning "the value is closer to the small on f_j".

We conduct our experiments on database FEI [36, 37] and CK+ [13, 24]. The FEI contains 2800 images coming from 200 different subjects (100 male, 100 female), which are gathered by college students of Brazil. Each frontal face includes two expressions (natural and happy), which are used in our experiment. CK+ [13, 24] is a sequence of images including 123 subjects. In which, the expression is changed from the neutral face to the corresponding peak. In order to compare the differences among expressions, we only use the peak expression.

In addition, we extract the semantic concepts for interpreting the differences of expressions. Meanwhile, we consider SDAFS as a classifier to have a comparison of performance with some state-of-the-art classifiers such as C4.5 [29], Bayes [30], Decision Table (DT) [12], Cart [15] and Reduced error pruning tree (REP) [8].

4.1 The semantic description for expression on FEI

In this section, we select natural and smile images from FEI (NH-FEI), which include 400 samples (natural 200, happy 200), to verify our method. In order to distinguish each expression, the symbol ξ_{C1} and ξ_{C2} represent a set of semantic concepts of natural and happy respectively. The result is as follows:

$\xi_{C1} = m_{17,3}$. Its semantic interpretation is as follows: "The perimeter of mouth is smaller";
$\xi_{C2} = m_{17,1}$. Its mean is that: "The perimeter of mouth is smaller is larger".

According to the semantic concepts, we can easily observe that the mouth can obviously differentiate natural and happy (see Fig. 3). The result denotes that the mouth is best important feature for distinguishing natural and happy. In addition, the result also implies the change extent of facial component according to the semantic concept of mouth. For example, the perimeter shows the open extent in happy is larger than that of natural.

In order to observe the effect of SDAFS, we make 10-fold cross-validation on NH-FEI. And the experiments are executed 5 times. Here, "A" represents average accuracy in one experiment; "C" denotes the average number of correct testing samples in one experiment; "N" is the number of testing samples; "std" presents standard deviation of average accuracy. The result is illustrated in Table 3.

4.2 The semantic description for expression on CK+

According to the results in [48], samples are often misclassified because of the similar appearance variations between anger and sadness. In addition, due to the sample number of fear expression is very small, we only select four expressions such as natural, disgust, happy and surprise in our experiment, and the peak expressions of disgust, happy and surprise will be selected. And we execute 10-fold cross-validation on each database to estimate the performance of classifiers.
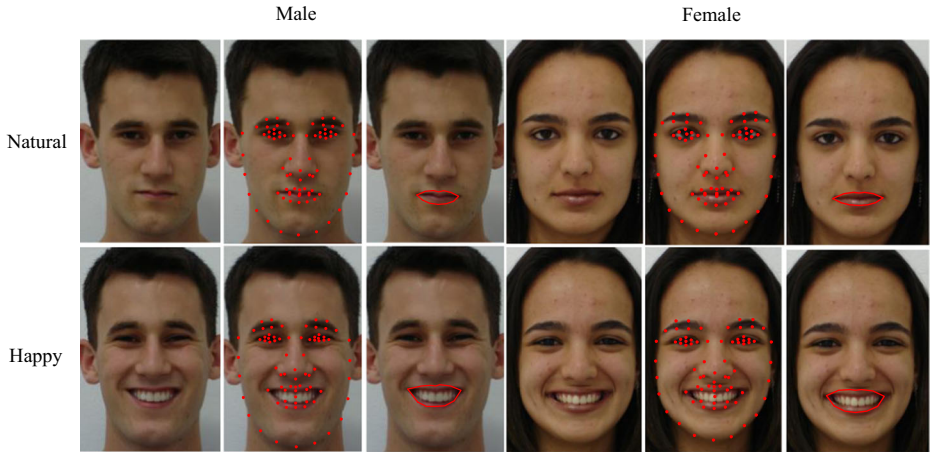
Male   Female

Fig. 3 The comparison between natural and happy on FEI

4.2.1 The difference description among natural, disgust and surprise

In this section, we extract natural, disgust and surprise images from CK+ as a new expression database (NDS-CK+). It contains 225 samples (natural 83, disgust 59, surprise 83). The symbol ξ_{C1}, ξ_{C2}, and ξ_{C3} denote a set of semantic concepts of natural, disgust and surprise respectively. The results are as follows:

> $\xi_{C1} = m_{23,1}$ with the semantic concept is that: "The angle between corners of mouth and the down point of mouth is larger";
>
> $\xi_{C2} = m_{1,3} + m_{8,3}$ with the semantic concept is that: "The distance between the point of inner canthus point and the point of inner eyebrow" or "The height of eyes is smaller";
>
> $\xi_{C3} = m_{18,1}\, m_{19,1} + m_{15,1}\, m_{19,1}$ with the semantic concept is that: "The area of mouth is larger and the angle of right corner of the mouth is larger" or "the height of mouth is larger and the angle of right corner of the mouth is larger".

According to the above results, we can obtain the differences of various expressions with semantic concepts. For example, ξ_{C1} denotes three movements of facial components such as eyebrows downed and eyes squinted. ξ_{C3} presents the mouth is opened. Moreover, the result illustrates the characteristics of disgust expression is concentrated on eyes region and that of surprise is mouth region. The differences are described in Fig. 4.

Then, a 10-fold cross-validation is carried out 5 times for comparing with other classifiers on NDS-CK+. The results are illustrated in Table 4. Here "A" represents average accuracy in one experiment; "C" denotes the average number of correct testing samples in one experiment; "N" is the number of testing samples; "std" presents standard deviation of average accuracy.

4.2.2 The difference description among natural, happy and surprise

In this section, we make a combining database, which includes natural, happy and surprise (NHS-CK+). The database contains 235 samples (natural 83, happy 69, surprise 83). In order

Table 3 The accuracy rate on NH-FEI (%)

Method	1 time		2 time		3 time		4 time		5 time		Mean value		
	A	C(N)	A	C(N)	A	C(N)	A	C(N)	A	C(N)	A	C(N)	std
C4.5	90.00	36.00(40)	88.50	35.34(40)	88.75	35.50(40)	88.00	35.20(40)	89.75	35.90(40)	89.00	35.60(40)	0.85
Bayes	88.75	35.50(40)	88.75	35.50(40)	90.00	36.00(40)	89.25	35.70(40)	90.00	36.00(40)	89.35	35.74(40)	0.63
DT	87.50	35.00(40)	88.50	35.40(40)	88.25	35.30(40)	88.50	35.40(40)	88.75	35.50(40)	88.30	35.32(40)	0.48
Cart	88.25	35.30(40)	87.50	35.00(40)	88.00	35.20(40)	89.00	35.60(40)	88.25	35.30(40)	88.20	35.28(40)	0.54
REP	87.75	35.10(40)	86.00	34.40(40)	88.75	35.50(40)	90.50	36.20(40)	89.00	35.60(40)	88.40	35.36(40)	1.66
SDAFS	90.25	36.10(40)	90.25	36.10(40)	90.25	36.10(40)	90.25	36.10(40)	90.25	36.40(40)	90.25	36.10(40)	0.00

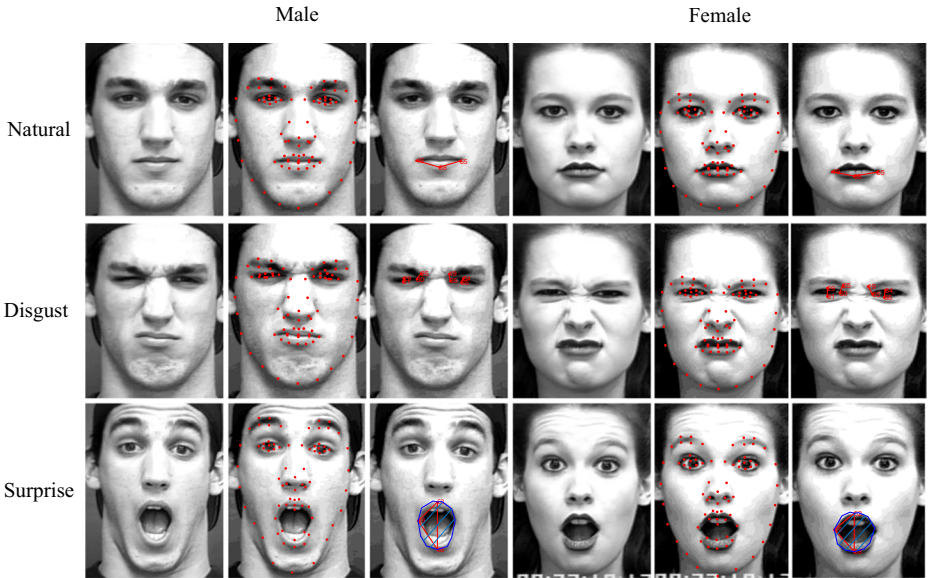Male   Female

Fig. 4 The comparison among natural, disgust and surprise on CK+

to distinguish each expression, the ξ_{C1}, ξ_{C2} and ξ_{C3} indicate natural, happy and surprise respectively. The results are as follows.

$\xi_{C1} = m_{23,1}$ with the semantic concept is that: "The angle between corners of mouth and the down point of mouth is larger";

$\xi_{C2} = m_{21,1}$ with the semantic concept is that: "The angle between corners of mouth and nasal peak is larger";

$\xi_{C3} = m_{15,1} + m_{15,1}\ m_{19,1}$ with the semantic concept is that: "The height of mouth is larger" or "The height of mouth is larger and the angle of right corner of the mouth is larger".

According to these results, we can obtain the differences among natural, happy and surprise, which are focused on mouth region (see Fig. 5.). For example, ξ_{C2} denotes the mouth corners will be stretched into sides, when happy happened. It leads to increasing the angle (f_{21}). However, ξ_{C3} denotes that the surprise expression is also emphasized on mouth region.

In this case, we execute 5 times 10-fold cross-validation on NHS-CK+. Here, "A" represents average accuracy in once experiment; "C" denotes the average number of correct testing samples in once experiment; "N" is the number of testing samples; "std" presents standard deviation of average accuracy. The result is illustrated in Table 5.

4.2.3 The difference description among natural, disgust, happy and surprise

In this section, we make a combining database, which includes natural, disgust, happy and surprise (NDHS-CK+). The database contains 294 samples (natural 83, disgust 59, happy 69, surprise 83). In order to distinguish each expression, ξ_{C1}, ξ_{C2}, ξ_{C3} and ξ_{C4} indicate natural, disgust, happy and surprise respectively. The results are as follows.

Table 4 The accuracy rate on NDS-CK+(%)

| Method	1 time		2 time		3 time		4 time		5 time		Mean value		
	A	C(N)	A	C(N)	A	C(N)	A	C(N)	A	C(N)	A	C(N)	std
C4.5	86.34	18.99(22)	82.73	18.20(22)	86.82	18.97(22)	88.45	19.46(22)	86.82	18.97(22)	86.23	18.97(22)	2.11
Bayes	85.18	18.74(22)	85.55	18.82(22)	86.55	19.04(22)	87.45	19.24(22)	86.54	19.04(22)	86.25	18.98(22)	0.90
DT	77.73	17.10(22)	81.82	18.00(22)	83.64	18.40(22)	83.64	18.40(22)	82.27	18.10(22)	81.82	18.00(22)	2.65
Cart	85.00	18.70(22)	85.46	18.80(22)	85.45	18.14(22)	86.36	19.00(22)	85.00	18.70(22)	85.45	18.80(22)	0.56
REP	84.09	18.50(22)	88.64	19.50(22)	83.18	18.30(22)	84.09	18.50(22)	83.64	18.40(22)	84.73	18.64(22)	2.22
SDAFS	87.27	19.20(22)	85.00	18.70(22)	86.82	19.10(22)	85.91	18.90(22)	86.82	19.10(22)	86.36	19.00(22)	0.91
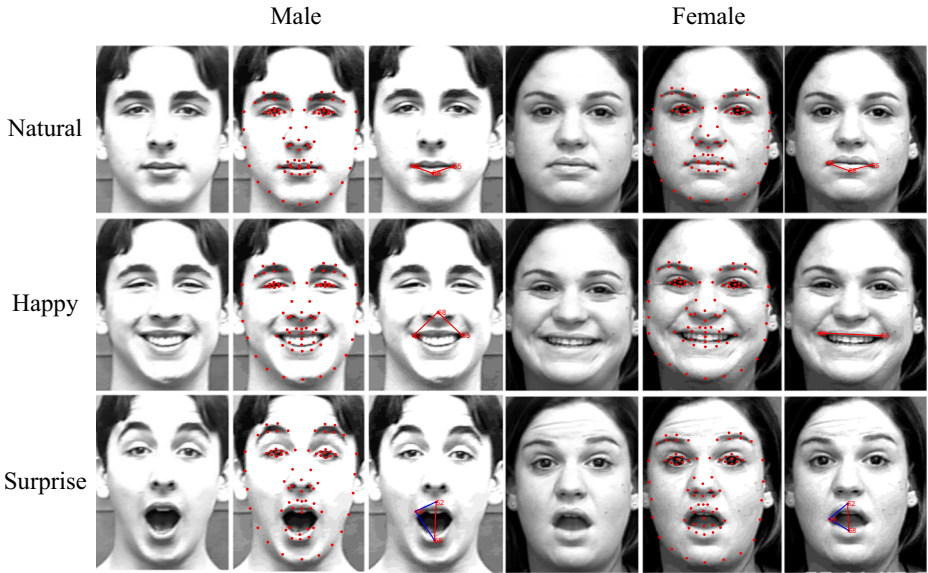
Fig. 5 The comparison among natural, happy and surprise on CK+

$\xi_{C1} = m_{23,1} + m_{16,2}\, m_{23,1}$ with the semantic concept is that: "The angle between corners of mouth and the down point of mouth is larger" or "The width of mouth is medium and the angle between corners of mouth and the down point of mouth is larger";

$\xi_{C2} = m_{8,3} + m_{8,3}\, m_{11,3} + m_{10,3}\, m_{11,3}$ with the semantic concept is that: "The height of eyes is smaller" or "The height of eyes is smaller and the angle between points of top and down of eye and inner canthus in left eye is smaller" or "The angle between points of top and down of eye and inner canthus in right eye is smaller";

$\xi_{C3} = m_{21,1} + m_{16,1}\, m_{21,1}$ with the semantic concept is that: "The angle between corners of mouth and nasal peak is larger" or "The width of mouth is larger and the angle between corners of mouth and nasal peak is larger";

$\xi_{C4} = m_{1,1}\, m_{15,1} + m_{15,1}\, m_{19,1} + m_{15,1}\, m_{20,1} + m_{15,1}\, m_{23,3} + m_{19,1}\, m_{23,3}$ with the semantic concept is that: "The distance between the points of inner canthus and the points of inner eyebrows is larger and the height of mouth is larger" or "The height of mouth is larger and the angle of right corner of the mouth is larger" or "The height of mouth is larger and the angle of left corner of the mouth is larger" or "The height of mouth is larger and the angle between corners of mouth and the down point of mouth is smaller" or "The angle of right corner of the mouth is larger and the angle between corners of mouth and the down point of mouth is smaller".

According to these semantic results, we can obtain two important observations of facial expression. First, the differences of facial regions can be extracted in various expressions such as natural, disgust, happy and surprise (see Fig. 6). Second, the movements of facial components can be represented using semantic concepts clearly. For example, eyebrows frowned and eyes squinted are two representative action in disgust expression. Grin is a typical action in happy expression. Mouth opened and eyebrows rise are also two facial movements, when surprise happens.

Table 5 The accuracy rate on NHS-CK+(%)

Method	1 time		2 time		3 time		4 time		5 time		Mean value		std
	A	C(N)	A	C(N)	A	C(N)	A	C(N)	A	C(N)	A	C(N)	
C4.5	92.17	20.44(23)	92.17	20.44(23)	94.78	21.80(23)	91.74	21.10(23)	93.47	21.50(23)	92.87	21.36(23)	1.25
Bayes	96.96	22.30(23)	96.09	22.10(23)	95.65	22.00(23)	96.18	22.12(23)	94.88	21.82(23)	95.95	22.07(23)	0.76
DT	91.74	21.10(23)	89.13	20.50(23)	87.39	20.10(23)	90.87	20.90(23)	87.07	20.03(23)	89.24	20.53(23)	2.06
Cart	91.31	21.00(23)	89.56	20.60(23)	90.00	20.70(23)	93.04	21.40(23)	91.30	21.00(23)	91.04	20.94(23)	1.36
REP	90.00	20.70(23)	90.43	20.80(23)	89.56	20.60(23)	88.26	20.30(23)	91.74	21.10(23)	90.00	20.70(23)	1.27
SDAFS	97.39	22.40(23)	96.52	22.20(23)	96.96	22.30(23)	96/09	22.10(23)	95.65	22.00(23)	96.52	22.20(23)	0.69

Fig. 6 The comparison among natural, disgust, happy and surprise on CK+

Similarly, we can still execute 5 times 10-fold cross-validation on NDHS-CK+. Here, "A" represents average accuracy in once experiment; "C" denotes the average number of correct testing samples in once experiment; "N" is the number of testing samples; "std" presents standard deviation of average accuracy. The result is illustrated in Table 6.

4.3 Performance analysis

According to these experimental results, we can obtain some insights in the performance of the classifier. However, those results do not provide enough support for drawing a strong conclusion in favor or against any of the studied methods. In order to achieve a convincing conclusion, we resort ourselves to statistical testing of the results. The Friedman Test [11, 40] is a nonparametric test that is based on the relative performance of classification method in terms of their ranks: for each dataset, the methods to be compared are sorted according to their performance, i.e., each method is assigned a rank (in case of ties, average ranks are assigned). Therefore, under the null hypothesis, the Friedman statistic is as follows:

$$\tau_{\chi^2} = \frac{12N}{k(k+1)} \left[\sum_{j=1}^{k} (R_j)^2 - \frac{k(k+1)^2}{4} \right] \tag{27}$$

where k is the number of classifiers and N is the number of data sets. r_i^j is the rank of classifier j on the data set i and $R_j = \left(\sum_{i=1}^{N} r_i^j \right) \Big/ N$ is the average rank of classifier j. The Friedman statistic is asymptotically χ^2-distributed with $k-1$ degrees of freedom.

Table 6 The accuracy rate on NDHS-CK+(%)

Method	1 time		2 time		3 time		4 time		5 time		Mean value		
	A	C(N)	A	C(N)	A	C(N)	A	C(N)	A	C(N)	A	C(N)	std
C4.5	84.14	24.40(29)	81.73	23.70(29)	81.38	23.60(29)	79.66	23.10(29)	86.21	25.00(29)	82.62	23.96(29)	2.56
Bayes	83.62	24.25(29)	83.62	24.25(29)	82.59	23.95(29)	82.24	23.85(29)	82.93	24.05(29)	83.00	24.07(29)	0.62
DT	75.86	22.00(29)	77.59	22.50(29)	76.21	22.10(29)	77.24	22.40(29)	76.48	76.48(29)	76.68	22.24(29)	0.72
Cart	80.69	23.40(29)	82.07	23.80(29)	79.66	23.10(29)	80.34	23.30(29)	82.07	23.80(29)	80.97	23.48(29)	1.07
REP	83.10	24.10(29)	81.72	23.70(29)	81.38	23.60(29)	83.79	24.30(29)	79.66	23.10(29)	81.93	23.76(29)	1.61
SDAFS	83.79	24.30(29)	81.72	23.70(29)	82.76	23.60(29)		23.10(29)	80.34	23.30(29)	81.65	23.68(29)	1.70

If N and k are not large enough, it is recommended to use the following correction that is F-distributed with $(k − 1)$ and $(k − 1)(N − 1)$ degrees of freedom:

$$\tau_F = \frac{(N-1)\tau_{\chi^2}}{N(k-1)-\tau_{\chi^2}} \qquad (28)$$

We now evaluate the performance using the Friedman test. The values of N and k are set to 3 and 7 respectively. First the average rank R_j is determined (see Table 7).

According to formula (28), the value of _F is 11.483, while the critical value for the significance level α = 0.05 is 2.901. Thus, the null-hypothesis can quite safely be rejected, which means that there are significant differences in the classifiers performance. Given the result of the Friedman Test, we conducted the Nemenyi Test [11] as a post-hoc test to compare the classifiers in a pairwise manner. According to this test, the performance of two classifiers is significantly different if the distance of the average ranks exceeds the critical distance CD.

$$CD = q_\alpha \sqrt{\frac{k(k + 1)}{6N}} \qquad (29)$$

where the q_α is taken from the table of normal distribution, the value CD = 3.770. According to R_j in Table 7, the performance among classifiers can be observed in detail in Fig. 7. From Fig. 7, we evaluate the performance of various classifiers. Here, the performance of SDAFS and Bayes is similarity; the performance of C4.5 and Cart is similarity; the performance of Cart, REP and DT is similarity. And, the performance of SDAFS and Bayes is better than that of REP and DT.

However, considering the semantic interpretation of expression, the proposed approach is quite convincing. Then, in order to compare the effect, the pixel features will be addressed in expression recognition.

## 4.4 Expression recognition with holistic features

In last section, we will execute two experiments to show the holistic pixel features are not as effective as the features extracted in this paper. For this purpose, PCA [39], LDA [47] and Local Binary Patterns (LBP) [46] are used to recognize expressions using pixel features. Meanwhile, we carry out 10-fold cross-validation in experiments. In order to ensure the same resolution, all images are cropped into a new size 100 ∗ 100 (see Fig. 8).

Here, "A" represents average accuracy in one experiment; "C" denotes the average number of correct testing samples in one experiment; "N" is the number of testing samples. The result is illustrated in Tables 8, 9, and 10.

Table 7 The rank of various classifiers

Data set	Classifiers					
	C4.5	Bayes	DT	Cart	REP	SDAFS
NH-FEI	3	2	5	6	4	1
NDS-CK+	3	2	6	4	5	1
NHS-CK+	3	2	6	4	5	1
NHDS-CK	2	1	6	5	3	4
Average R_j	2.75	1.75	5.75	4.75	4.25	1.75

Fig. 7 The performance comparison among classifiers based on Friedman Test
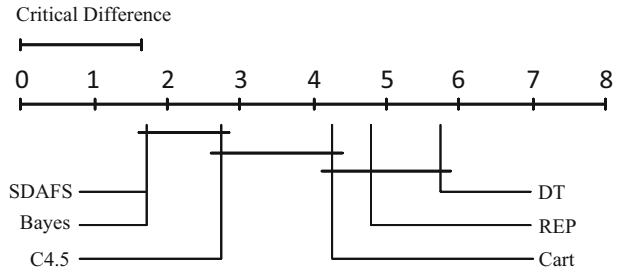
By comparing with the performance results by using the geometry features and pixel features on NDS-CK+, NHS-CK+ and NDHS-CK+. The pixel features are much worse than that of geometry features. The results indicate that the illumination intensity can influence pixel values, which could lead to error in recognition process. In addition, another advantage of geometry features is that it breaks the limitation of image resolution and illuminations.
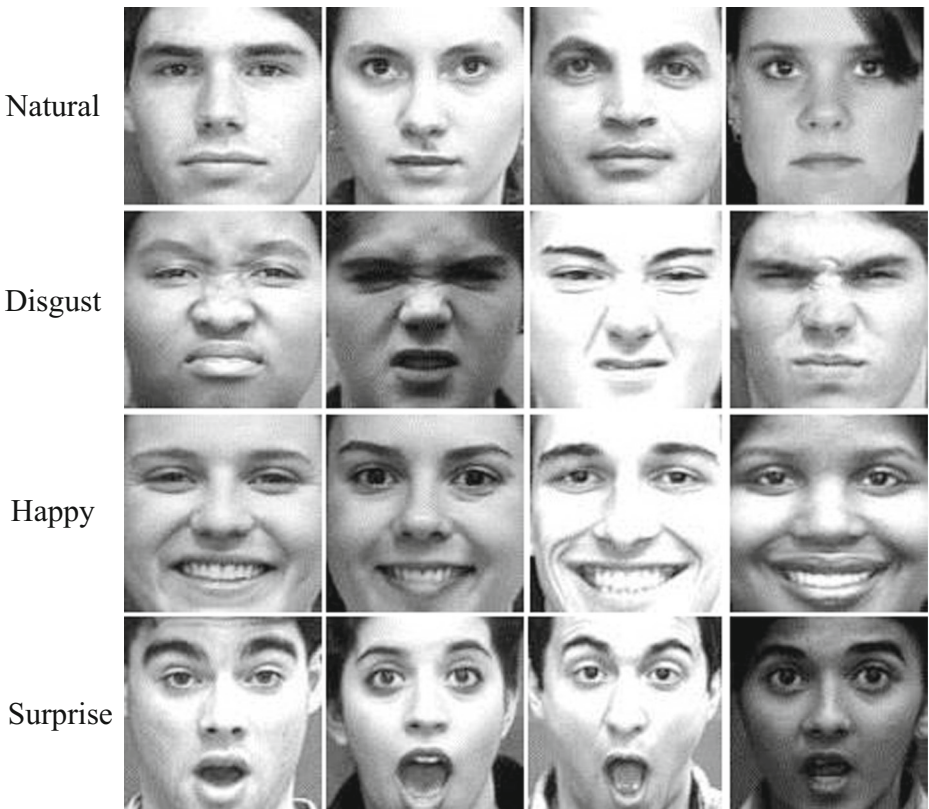

Fig. 8 A segmental facial expression images on CK+

Table 8 The result of holistic features using PCA(%)

Dataset	1 time		2 time		3 time		4 time		5 time		Mean value	
	A	C(N)	A	C(N)	A	C(N)	A	C(N)	A	C(N)	A	C(N)
NDS-CK+	35.91	8.26(23)	35.91	8.26(23)	35.45	8.15(23)	35.45	8.15(23)	36.36	8.36(23)	35.82	8.24(23)
NHS-CK+	34.34	7.55(22)	34.34	7.55(22)	34.78	7.65(22)	34.34	7.55(22)	34.36	7.56(22)	34.43	7.57(22)
NDHS-CK+	27.59	8.00(29)	27.59	8.00(29)	26.90	7.80(29)	27.59	8.00(29)	27.59	8.00(29)	27.45	7.96(29)

Table 9 The result of holistic features using LDA(%)

Data set	1 time		2 time		3 time		4 time		5 time		Mean value	
	A	C(N)	A	C(N)	A	C(N)	A	C(N)	A	C(N)	A	C(N)
NDS-CK+	35.91	8.26(23)	36.36	8.36(23)	35.91	8.26(23)	36.36	8.36(23)	36.36	8.36(23)	36.18	8.32(23)
NHS-CK+	34.78	7.65(22)	34.78	7.65(22)	34.34	7.55(22)	34.78	7.65(22)	34.34	7.56(22)	34.60	7.61(22)
NDHS-CK+	27.59	8.00(29)	27.59	8.00(29)	26.90	7.80(29)	27.59	8.00(29)	27.59	8.00(29)	27.45	7.96(29)

Table 10 The result of holistic features using LBP (%)

Data set	1 time		2 time		3 time		4 time		5 time		Mean value	
	A	C(N)	A	C(N)	A	C(N)	A	C(N)	A	C(N)	A	C(N)
NDS-CK+	53.33	12.27(23)	55.00	12.65(23)	51.67	11.88(23)	52.92	12.17(23)	53.75	12.36(23)	53.33	12.27(23)
NHS-CK+	69.60	15.31(22)	70.00	15.40(22)	69.60	15.31(22)	68.00	14.96(22)	68.00	14.96(22)	69.04	15.19(22)
NDHS-CK+	53.55	15.53(29)	54.84	15.90(29)	52.58	15.25(29)	56.45	16.37(29)	54.84	15.90(29)	54.45	15.79(29)

5 Conclusions

In this paper, we propose a new approach for facial expression based on the framework of AFS theory. The developed descriptor not only describes the characteristics of various expressions, but also bridges the gap between primitive features and semantic concepts. When certain expression happened, its shape could directly reflect the differences among expressions. Firstly, we establish expression features using landmarks, and then extract the semantic concepts with its distribution. Secondly, we design a method to select the effective features. In the same class, we estimate representative effectiveness of features for each sample. We finally select the better features as its representation. Eventually, an optimization criterion is built to extract semantic concept sets for each expression. Then, we estimate four combining data sets such as NH-FEI, NDS-CK+, NHS-CK+ and NDHS-CK+, and conduct the experiments for analyzing expression characteristics respectively. The results obtained on NH-FEI show that if one only differentiates the natural and happy, the feature of mouth is obvious ($m_{17,3}$ and $m_{17,1}$). Then, we have a study on disgust and surprise. For example, the result on NDS-CK+ represents that the differences among natural, disgust and surprise focuses on eyes ($m_{8,3}$), and eyebrows ($m_{1,3}$) and mouth ($m_{15,1}$, $m_{18,1}$ and $m_{19,1}$). Similarly, the results on NHS-CK+ indicate the main differences among natural, happy and surprise still concentrate on mouth ($m_{23,1}$, $m_{21,1}$, $m_{15,1}$ and $m_{19,1}$). Finally, the experiment on NDHS-CK+ denotes the regions of eyes and mouth are important to analyze facial expressions. In addition, the results demonstrate SDAFS can describe the various expressions using semantic concepts, which conforms to real variation of facial components in various expressions.

Furthermore, we have an accuracy rate test between SDAFS and some state-of-the-art methods such as C4.5, Bayes, DT, Cart and REP. In order to compare their performance, the Friedman Test is applied in examination. The results show that the performance of SDAFS and Bayes is similar, and better than that of REP and DT in analyzing facial expressions.

In the future, the expression images will be collected continuously. Meanwhile, more advanced optimization schemes will be researched. Specially, we will have a deep research about facial expression in different cultural backgrounds.

References

1. An G, Liu S, Jin Y, Ruanc Q, Lu S (2014) Facial expression recognition based on discriminant neighborhood preserving nonnegative tensor factorization and ELM. Mathematical Problems in Engineering 2014: 10. doi:10.1155/2014/390328
2. Chakraborty A, Konar A, Chakraborty UK, Chatterjee A (2009) Emotion recognition from facial expressions and its control using fuzzy logic. IEEE Trans Syst Man Cybern Part A Syst Hum 39:726–743
3. Cheng S-C, Chen M-Y, Chang H-Y, Chou T-C (2007) Semantic-based facial expression recognition using analytical hierarchy process. Expert Syst Appl 33(1):86–95
4. Chew SW, Lucey S, Lucey P, Sridharan S, Conn JF (2012) Improved facial expression recognition via uni-hyperplane classification. 2012 I.E. Conference on Computer Vision and Pattern Recognition (CVPR), pp 2554–2561
5. Cootes TF, Taylor CJ, Cooper DH, Graham J (1995) Active shape models-their training and application. Comput Vis Image Underst 61:38–59
6. De Marsico M, Nappi M, Riccio D (2010) FARO: face recognition against occlusions and expression variations. IEEE Trans Syst Man Cybern Part A Syst Hum 40:121–132

7. Ding C, Peng H (2005) Minimum redundancy feature selection from microarray gene expression data. J Bioinforma Comput Biol 3:185–205
8. Elomaa T, Kaariainen M (2001) An analysis of reduced error pruning. J Artif Intell Res 15:163–187
9. Ekman P, Friesen WV (1978) Facial action coding system: a technique for the measurement of facial movement. Consulting Psychologists Press, Palo Alto, CA
10. Hernandez-Matamoros A, Bonarini A, Escamilla-Hernandez E, Nakano-Miyatake M, Perez-Meana H (2016) Facial expression recognition with automatic segmentation of face regions using a fuzzy based classification approach. Knowl-Based Syst 110:1–14
11. Huhn JC, Hullermeier E (2009) Fr3: a fuzzy rule learner for inducing reliable classifiers. IEEE Trans Fuzzy Syst 17:138–149
12. Huysmans J, Dejaeger K, Mues C, Vanthienen J, Baesens B (2011) An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. Decis Support Syst 51:141–154
13. Kanade T, Cohn JF, Tian Y (2000) Comprehensive database for facial expression analysis. IEEE International Conference on Automatic Face and Gesture Recognition, pp 46–53
14. Lee SH, Ro YM (2016) Partial matching of facial expression sequence using over-complete transition dictionary for emotion recognition. IEEE Trans Affect Comput 7:389–408
15. Lee TS, Chiu C-C, Chou Y-C, Lu C-J (2006) Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. Comput Stat Data Anal 50:1113–1130
16. Li Q, Ren Y, Li L, Liu W (2016) Fuzzy based affinity learning for spectral clustering. Pattern Recogn 60:531–542
17. Liang H, Liang R, Song M, He X (2016) Coupled dictionary learning for the detail-enhanced synthesis of 3-d facial expressions. IEEE Trans Cybern 46:890–901
18. Liu X (1998) The fuzzy sets and systems based on AFS structure, EI algebra and EII algebra. Fuzzy Sets Syst 95:179–188
19. Liu X (1998) The fuzzy theory based on afs algebras and afs structure. J Math Anal Appl 217:459–478
20. Liu X, Pedrycz W (2009) Axiomatic fuzzy set theory and its applications. Studies in Fuzziness and Soft Computing 244. doi:10.1007/978-3-642-00402-5
21. Liu X, Wang W, Chai T (2005) The fuzzy clustering analysis based on AFS theory. IEEE Trans Syst Man Cybern Part B Cybern 35:1013–1027
22. Liu X, Chai T, Wang W, Liu W (2007) Approaches to the representations and logic operations of fuzzy concepts in the framework of axiomatic fuzzy set theory I. Inf Sci 177:1007–1026
23. Liu X, Wang W, Chai T, Liu W (2007) Approaches to the representations and logic operations of fuzzy concepts in the framework of axiomatic fuzzy set theory II. Inf Sci 177:1027–1045
24. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 94–101
25. Mohammadi MR, Fatemizadeh E, Mahoor MH (2016) Intensity estimation of spontaneous facial action units based on their sparsity properties. IEEE Trans Cybern 46:817–826
26. Pantic M, Rothkrantz LJ (2004) Facial action recognition for facial expression analysis from static face images. IEEE Trans Syst Man Cybern Part B Cybern 34:1449–1461
27. Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27:1226–1238
28. Pu X, Fan K, Chen X, Ji L, Zhou Z (2015) Facial expression recognition from image sequences using twofold random forest classifier. Neurocomputing 168:1173–1180
29. Quinlan JR (1993) Quinlan J R. C4. 5: Programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco
30. Ramoni M, Sebastiani P (2001) Robust bayes classifiers. Artif Intell 125:209–226
31. Ren Y, Li Q, Liu W, Li L (2016) Semantic facial descriptor extraction via axiomatic fuzzy set. Neurocomputing 171:1462–1474
32. RodrìGuez RM, MartìNez L, Herrera F (2013) A group decision making model dealing with comparative linguistic expressions based on hesitant fuzzy linguistic term sets. Inf Sci 241:28–42
33. Sarkhel R, Das N, Saha AK, Nasipuri M (2016) A multi-objective approach towards cost effective isolated handwritten bangla character and digit recognition. Pattern Recogn 58:172–189
34. Siebers M, Schmid U, Seuß D, Kunz M, Lautenbacher S (2016) Characterizing facial expressions by grammars of action unit sequences–a first investigation using ABL. Inf Sci 329:866–875
35. Silva C, Vieira SM, Sousa JMC (2015) Fuzzy decision tree to predict readmissions in intensive care unit. Lect Notes Electr Eng 321:365–373
36. Tenorio EZ, Thomaz CE (2011) Anàlise multilinear discriminante de formas frontais de imagens 2d de face. Proceedings of the X Simposio Brasileiro de Automacao Inteligente SBAI, pp 266–271
37. Thomaz CE, Giraldi GA (2010) A new ranking method for principal components analysis and its application to face image analysis. Image Vis Comput 28:902–913

38. Tseng JL (2016) An improved surface simplification method for facial expression animation based on homogeneous coordinate transformation matrix and maximum shape operator. Math Probl Eng 2016:14. doi:10.1155/2016/2370919
39. Turk MA, Pentland AP (1991) Face recognition using eigenfaces. 1991 I.E. Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Maui, p 586–591
40. Wang X, Liu X, Zhang L (2014) A rapid fuzzy rule clustering method based on granular computing. Appl Soft Comput 24:534–542
41. Wang Z, Ruan Q, An G (2016) Facial expression recognition using sparse local fisher discriminant analysis. Neurocomputing 174:756–766
42. Wang H, Gao J, Tong L, Yu L (2016) Facial expression recognition based on PHOG feature and sparse representation. 2016 I.E. Conference on Chinese control Conference (CCC), pp 3869–3874
43. Yang S, Bhanu B (2012) Understanding discrete facial expressions in video using an emotion avatar image. IEEE Trans Syst Man Cybern Part B Cybern 42:980–992
44. Zhang Z, Wang L, Zhu Q, Chen S-K, Chen Y (2015) Pose-invariant face recognition using facial landmarks and weber local descriptor. Knowl-Based Syst 84:78–88
45. Zhang K, Mistry S, Neoh C, Lim CP (2016) Intelligent facial emotion recognition using moth-firefly optimization. Knowl-Based Syst 11:248–267
46. Zhao G, Pietikäinen M (2007) Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans Pattern Anal Mach Intell 29(6):915–928
47. Zhao H, Yuen PC (2008) Incremental linear discriminant analysis for face recognition. IEEE Trans Syst Man Cybern Part B Cybern 38:210–221
48. Zhong L, Liu Q, Yang P, Huang J, Metaxas DN (2015) Learning multiscale active facial patches for expression analysis. IEEE Trans Cybern 45:1499–1510

Zedong Li received the B.S.c. degree in Automation form Chongqing University of Technology, Chongqing, China in 2006, the M.S.c. degree in Automation from Shenyang Ligong University, Shenyang, China, in 2011. Currently, he is a PhD candidate in the Institute of System Science, Northeastern University. His research interests include fuzzy system, knowledge discovery, and data process. E-mail: lizedong2008@163.com

Qingling Zhang received the B.Sc. and M.Sc. degrees from the Mathematics Department and the Ph.D. degree from the Automatic Control Department of Northeastern University, China, in 1982, 1986 and 1995, respectively. Since1997, he has been a Professor with Northeastern University. He is also a member of the University Teaching Advisory Committee of National Ministry of Education. He has published six books and more than 230 papers about control theory and applications. Dr. Zhang received 14 prizes from central and local governments for his research. He has also received the Golden Scholarship from Australia in 2000. His current research interests cover fuzzy system, singular systems, data-driven control, and robust control.

Xiaodong Duan is the director of Dalian Key Lab of Digital Technology for National Culture, Dalian Nationalities University, Dalian, PR. China. He received the the B.S. degrees in computer science and technology from Nankai University, Tianjin, P. R. China in 1985, and received the M. S. and Ph.D. degrees in applied mathematics and computer software and theory from Northeastern University, Shenyang, P. R. China in 1988 and 2001, respectively. His research interests include pattern recognition and data mining. E-mail: minzudigital@dlnu.edu.cn

Cunrui Wang received the B.S. degrees in computer science and technology from Dalian Nationalities University, Dalian, P. R. China, in 2002, and received the M.S. degrees in computer software and theory from Northeastern University, Shenyang, P. R. China, in 2005. Currently, he is a PhD candidate in the Institute of System Science, Northeastern University. His research interests include biometrics identification and data mining.

Yu Shi received the B.S. degrees in computer science and technology from Huihua College of Hebei Normal University, Shijiazhuang, P. R. China, in 2016. Currently, she is a M.S. candidate in college of computer science and engineering, Dalian Minzu University. Her research interests include fuzzy system, pattern recognition and data processing.