Fuzzy aesthetic semantics description and extraction for art image retrieval

Qingyong Li\textsuperscript{a,}\textsuperscript{*}, Siwei Luo\textsuperscript{a}, Zhongzhi Shi\textsuperscript{b}

\textsuperscript{a} School of Computer and Information Technology, Beijing Jiaotong University, 100044, Beijing, China
\textsuperscript{b} Institute of Computing Technology, Chinese Academy of Sciences, 100080, Beijing, China

\section*{Article Info}

\textbf{Keywords:}
Art image retrieval  
Content-based image retrieval  
Fuzzy set  
Linguistic variable

\section*{Abstract}

More and more digitized art images are accumulated and expanded in our daily life and techniques are needed to be established on how to organize and retrieve them. Though content-based image retrieval (CBIR) made great progress, current low-level visual information based retrieval technology in CBIR does not allow users to search images by high-level semantics for art image retrieval. We propose a fuzzy approach to describe and to extract the fuzzy aesthetic semantic feature of art images. Aiming to deal with the subjectivity and vagueness of human aesthetic perception, we utilize the linguistic variable to describe the image aesthetic semantics, so it becomes possible to depict images in linguistic expression such as 'very action'. Furthermore, we apply neural network approach to model the process of human aesthetic perception and to extract the fuzzy aesthetic semantic feature vector. The art image retrieval system based on fuzzy aesthetic semantic feature makes users more naturally search desired images by linguistic expression. We report extensive empirical studies based on a 5000-image set, and experimental results demonstrate that the proposed approach achieves excellent performance in terms of retrieval accuracy.

\section*{1. Introduction}

Content-based image analysis and retrieval have become an active research area in the past few years. Content-based method always uses low-level visual features such as color, texture, and shape to represent image content. The advantage of this method is that it is automatic and easy to extract such features by a computer, and it is also simple to design similarity measurement for these features. However, the drawback of the content-based method is that low-level visual features could hardly be interpreted into the high-level concepts that are commonly understood by humans. On the other hand, earlier approaches for content-based image retrieval (CBIR) systems are usually based on Query-By-Example (QBE) strategy [1]. These methods are inflexible since users may have difficulties in describing the query by low-level visual features or getting suitable example images. Especially for art images, it is very difficult to devise effective features that reflect the aesthetic characteristic of images. Furthermore, the users of these systems are accustomed to browsing and searching images by aesthetic concepts described in natural language. Thus how to describe and extract the semantics of art images becomes a significant challenge.

From technical point of view, art image retrieval (AIR) encounters many challenges that are also common in CBIR, such as effective image features, and the gap between low-level features and high-level concepts. Though previous research on
CBIR has provided a good basis for our work on AIR, there are some other unique characteristics coming with AIR. In the following we will discuss some major issues.

- **Huge semantic gap**: There is a huger semantic gap for AIR between the low-level visual features (e.g. color, texture and shape) and the higher-level abstract properties, e.g. painting styles and expressed feelings or emotions. In some sense, this gap is more important for AIR since the higher-level abstract properties are always more indicative for the expression of art images. However, such higher-level abstract properties are not like the object-based concepts (e.g. flower, dog) in CBIR, and the abstract properties in AIR is somewhat subjective and vague; moreover, it cannot be directly obtained from image content. This paper is facing a challenge — how to describe and extract such higher-level abstract properties for art image retrieval.

- **Uncertainty of user perception**: It is always subjective and fuzzy for users to judge the higher-level abstract properties of art images. In our experiments we had the following observations: most of users were hesitating to mark art images and frequently utilized some adjective words, such as ‘very’, ‘somewhat’ and ‘few’ when they were asked to annotate art images with higher-level abstract properties (we focused on emotional properties in this paper, e.g. joy, relaxation and fear); moreover, when we re-present art images to some users who annotated the same images two days ago, we observed that many of them changed their opinions for some of images. This case indicates that we should find effective means, e.g. fuzzy set, to model the uncertainty of user perception.

- **Imprecise query intention**: One major change from CBIR to AIR is that the users' query intention is always imprecise. For example, in CBIR a user may input a query 'find all pictures depicting Madonna', while in AIR a user may just implicitly keep a vague query in mind ‘return some images which can let me feel relaxation’. In the first case ‘Madonna’ is a query concept that conveys explicit information of contents. The retrieval criterion is somewhat doubtless since all people agree what kind of images should be returned. However, in the AIR case the query has no hard requirement for contents but highly depends on the preference of the user. This also inspires our proposed methodology to deal with such imprecision.

Colombo [2] advocated a syntactic construction called compositional semantics to build the semantic representation of art images for the first time. This approach put forward some prior rules that map the perceptual and expressive representation to emotional semantics, such as joy, fear and so on. However, it did not provide explicit semantics described with natural language and its rules were experiential. Unlike Colombo, Shuqiang [3] brought forward an ontology-based approach to retrieve digitized art images. This scheme utilized certain ontology to manage the semantics of art images. Unfortunately, it could not deal with the vagueness appearing in the human process of understanding art images. Unlike these methods, in our previous work we proposed a fuzzy semantics description framework, named linguistic expression based image description framework (briefly as LEBID), and validated its feasibility in texture image retrieval [4].

In this paper, we put forward a fuzzy aesthetic semantics description (briefly as FASD) methodology based on LEBID framework. Furthermore, we utilize neural network to extract the aesthetic semantic feature vector based on fuzzy set. At last, we experiment it on Corel image set and demonstrate the FASD-based art image retrieval system. More specifically, this paper makes two contributions:

1. FASD provides a flexible aesthetic semantics description scheme to annotate art images with natural language. In addition, it can adequately deal with the problem of vagueness in human aesthetic comprehension and image retrieval, deriving the advantage of fuzzy set.

2. To some extent, FASD merges the gap between the low-level feature and higher-level feature, since we apply neural network to extract the fuzzy aesthetic semantic feature vector that can help AIR systems to effectively depict art images from human perceptual perspectives. Furthermore, our experiments demonstrate that the neural network based feature extraction method outperforms genetic programming method and typical content-based method.

The rest paper is organized as follows: the next section introduces the related work; we illustrate the FASD method in the Section 3 and FASD-based art image retrieval system in Section 4. And the experimental result is showed in Section 5. At last we make conclusion.

2. Related work

At present, more and more digitized art images are exhibited and sold through the World Wide Web. It is becoming possible to analyze and spread art works at a larger scale. So, how to organize and retrieval digitized art images eventually becomes an important research topic.

The DELOS-NSF working group discusses the problems of retrieving art images and bridging the semantic gap, and points out that this area is still in the early stages of research [5]. In our opinion, the state-of-art techniques in semantic-based art image retrieval can be roughly divided into two categories: human-based methods and machine-based methods. For the human-based methods, the system designers had to provide the system with some prior knowledge, such as certain rules to generate semantics from the low-level visual features [2], or with direct semantics description manually labeled through certain technology, for example, Shuqiang [3] brought forward an ontology-based approach to retrieve digitized art images. Obviously, it is very expensive and subjective to carry out manual annotation with large databases. On the contrary, the machine-based methods automatically extracted semantics description by machine learning algorithms. Details about machine learning in semantics-based image retrieval can be seen in [6]. Datta [7] utilized SVM and the classification and
regression trees (CART) algorithm to learn the aesthetics in photographic images. Generally speaking, these methods are crisp and deterministic, for instance, an image is classified into ‘traditional Chinese painting’ or not. However, there is much ambiguity about the content of art images [4]. Furthermore, we human beings always utilize imprecise information when we browse or retrieval images. For example, the artists are accustomed to using subjective and qualitative words or phrases to search images from large collection such as ‘action’ images, ‘very fear’ images. In our previous work, we proposed a semantics description framework, named linguistic expression based image description (LEBID) [4], which provided a method to describe the image semantics in natural language expression, for example, it utilized linguistic express ‘very coarse’ to describe a texture image, and associated the expression with a fuzzy set. Furthermore, LEBID achieved good performance in texture image retrieval. How about this methodology for art image retrieval? Therefore this paper research on fuzzy aesthetic semantics based on LEBID.

3. The methodology of fuzzy aesthetic semantics description and extraction

3.1. Linguistic variable and some basic concepts

Linguistic variable is a variable whose values are natural language expressions referring to some quantity of interest [8–10]. Just as numerical variables take numerical values, in fuzzy logic, linguistic variables take linguistic values that are words (linguistic terms) associated with degree of membership. Thus, for a variable aesthetics, instead of assuming a numerical high-dimensional vector, it is treated as a linguistic variable that may assume, for instance, linguistic values of ‘very uneasiness’ with a degree of 0.91, ‘somewhat joy’ with a degree of 0.03. This concept was introduced by Zadeh to provide a means of approximate characterization of phenomena that are too complex or too ill-defined to be described in conventional way.

**Definition 1.** Fuzzy set A in U is a set of ordered pairs \( A = \{ u, M(u) : u \in U \} \), where \( M \) maps \( u \in U \) to the real interval \([0, 1]\) and is known as the membership function.

**Definition 2.** A linguistic variable is characterized by a quintuple denoted by \(< x, T(x), U, G, M >\). Here, \( x \) is the name of the variable, and \( T(x) \) is the set of the linguistic values of \( x \), of which each element is a word or phrase describing the variable \( x \) and coupled with a fuzzy set on the universal \( U \) which represents universal of discourse. \( G \) is the syntax rule responsible for generating some elements in \( T(x) \), and \( M \) is the semantic rule governing the generation of the meaning (or degree of membership) \( M(L, u) \) for each linguistic value \( L, L \in T(x) \) and each object in universal of discourse \( u, u \in U \).

3.2. Fuzzy aesthetic semantics description

Our previous work showed that linguistic variable was a powerful tool to describe the image semantics and model the image vagueness [4]. Next, we will demonstrate the detail of fuzzy aesthetic semantics description (FASD).

3.2.1. Linguistic name and linguistic value set

Since we describe the aesthetics for art image, it is easy to determine the linguistic name \( x \) as ‘aesthetics’. To discuss the linguistic value set, we first introduce two definitions.

**Definition 3.** A basic term set is the linguistic value set in which each element cannot be separated any more in semantics.

**Definition 4.** An extended term set is the linguistic value set in which each element is created by the words in basic term set through the syntax rule \( G \).

Determine the basic term set is the key problem in LEBID. Among many authors who recently addressed the psychology of art images, Arnheim discussed the relationships between artistic form and perceptive processes [11] and Itten [12] formulated a theory about using of color in art and about the semantics it induces. Itten observed that color combinations induce effects such as harmony, disharmony, calmness, and excitement that artists consciously exploit in their paintings. Most of these effects relate to high-level chromatic patterns rather than to physical properties of single points of color.

Colombo divided the semantics of art images into two types: *expressive semantics* and *emotional semantics*. Furthermore, he firstly investigated the emotional semantics on art image retrieval [2]. The psychological analysis of effects induced by images suggests that we can use only a limited subset of primary emotions to express the great variability of human emotions (secondary emotions). So we identified five primary emotions, namely, *action*, *relaxation*, *joy*, *fear*, and *uneasiness*, as the most relevant to express the interaction of humans with images [2,12].

Furthermore, we can construct the extended term set through the syntax rule (details about syntax rule to be seen in Section 3.2.4), so extended terms will be generated, such as ‘very action’, ‘few fear’. The whole linguistic value set is shown in Table 1.
Table 1
The linguistic value set for aesthetic semantics.

<table>
<thead>
<tr>
<th>Basic terms</th>
<th>Extended terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action</td>
<td>Very action</td>
</tr>
<tr>
<td>Relaxation</td>
<td>Very relaxation</td>
</tr>
<tr>
<td>Joy</td>
<td>Very joy</td>
</tr>
<tr>
<td>Uneasiness</td>
<td>Very uneasiness</td>
</tr>
<tr>
<td>Fear</td>
<td>Very fear</td>
</tr>
</tbody>
</table>

Fig. 1. The diagram of fuzzy color semantic feature extraction.

3.2.2. Universe of discourse \(U\)

The universe of discourse \(U\) is the whole input range allowed for a given fuzzy linguistic variable. In the domain of image retrieval, \(U\) always refers to the image visual feature since it is easier to manipulate than the raw pixel matrix of image. An image’s aesthetic semantics relates to its color content and the presence of elements such as lines that induce dynamism and action [2].

Colors. For representing color, we used HSV (Hue, Saturation, and Value) color model because this model is closely related to human visual perception. Color quantization is useful for reducing the calculation cost. Furthermore, it provides better performance for semantic rule because it can eliminate the detailed color components that can be considered as noises. The human visual system is more sensitive to hue than saturation and value so that hue should be quantized finer than saturation and value. In the experiments, we non-uniformly quantized HSV space into ten bins for hue, three bins for saturation and three bins for value for lower resolution. So each image has \(10 \times 3 \times 3 = 90\) dimensional color vector.

Lines. Detecting significant line slopes in an image can be accomplished by using the Hough transform to generate a line slope histogram. For the computation complexity, we just estimate a four-dimensional histogram, in which four bins are assumed as \((0, 90], (90, 180], (180, 270], (270, 360]\), respectively.

So a 94-dimensional vector, combining the color feature and line feature, represents universe of discourse \(U\).

3.2.3. Semantic rule \(M\)

\(M\) is the semantic rule mapping from the low-level visual feature to the high-level fuzzy semantics of \(T\), that is to say, for \(u \in U, t \in T(x)\), it assumes a value in \([0, 1]\) called degree of membership to \(u\) according to the linguistic value \(t\).

We can formally describe the task as following: through sample set \(
\{(V_1, y_1), \ldots, (V_n, y_n)\}
\), in which \(V_i \in U\) \((i = 1, \ldots, n)\) represents the 94-dimensional vector and \(y_i\) \((i = 1, \ldots, n)\) is the degree of membership in \([0, 1]\) regarding to certain linguistic value \(t\), we learn to construct a map:

\[
M_t: V \rightarrow y
\]

where \(t\) is the linguistic value.

It is very important to learn the semantic rule. There are many machine learning approaches that can work for it. For example, Datta [7] utilized SVM and the classification and regression trees (CART) algorithm to learn aesthetics, and we utilized genetic programming to construct the semantic rule in our previous work [4]. However, artificial neural network has shown more advantages in the human perception modeling, and it is also convenient in dealing with fuzzy set. So the neural network based methodology is proposed to learn the semantic rule. The learning diagram is shown in Fig. 1.

The architecture of the neural network is a standard three-layer perceptron with a slight modification as described below. The 94-dimensional visual feature vector is used as input to the neural network and the degrees of membership for five basic linguistic terms are used as the output. The number of hidden units is taken as twenty for our experiment. The activation function of hidden layer is a hyperbolic tangent function as following:

\[
\Phi(x) = \tanh(\beta x)
\]
where $\beta$ is a constant ($\beta = 0.20$ is used in our experiment). Because the output of sigmoid function is between 0 and 1, similar with the definition of degree of membership, the activation function of output unit choose sigmoid function. Learning rate ($\eta$) and momentum ($\alpha$) are kept as 0.02 and 0.25, respectively.

For the output representation, every unit of output layer models the semantic rule of certain basic linguistic term, so its output represents the degree of membership for the input art image obeying to the basic linguistic term. For example, if the first unit models the basic term 'action', and its output value is 0.9, then we can say that the input art image is regarded as 'action' with degree of 0.9.

3.2.4. Syntax rule $G$

Semantic rule $G$ is the part of syntax in FASD, which extends the linguistic values and enriches the expression ability of FASD. For the aesthetic semantics, we construct the syntax rule as in Fig. 2.

Note that the semantic rules of extended terms are not modeled by neural network as the basic terms, but rather are modeled as operators on the fuzzy set of the basic terms. There are two reasons for this choice: first, the extended terms are derived from basic terms, and their membership values also have inherent correlation, so it is reasonable to define the semantic rules of extended terms based on those of corresponding basic terms by operators; on the other hand, it can save the computation resource when we extract the semantic features.

Derived from our previous work [4], we define three operators according to 'very', 'somewhat', and 'few' as follows,

$$V_t(x) = \{ M_t^2(x) \mid x \in U \}$$
$$S_t(x) = \{ \text{Sin}(M_t(x) \ast \pi) \mid x \in U \}$$
$$F_t(x) = \{ 1 - M_t(x) \mid x \in U \} \quad (3)$$

where $t$ refers to certain basic term, and $M_t(x)$ represents the semantic rule of $t$, which is learned by neural network.

3.3. FASD feature representation

As discussed above, the linguistic term set includes basic term set and extended term set, each element of extended term set is formed based on basic term by syntax rule, furthermore, its semantic rule is also derived from that of corresponding basic term. Obviously, if we store all linguistic terms in feature database for an art image, it will augment the volume of database. Fortunately, it is unnecessary to conserve all linguistic terms since the degree of membership for each extended term can be easily computed from that of corresponding basic term. In conclusion, we only keep the basic term's fuzzy value in the feature database, and the extended term's value will be calculated through the syntax rule $G$ and the operators as discussed in Section 3.2.4.

At last, the FASD feature for art image is formed by a 5 dimensional vector:

$$F^1 = (y_1, y_2, \ldots, y_5) \quad (4)$$

where $y_i$, $i = 1, \ldots, 5$ refers to the fuzzy value for certain basic linguistic term. In this paper the corresponding aesthetic concepts in $F^1$ are (action, relaxation, joy, uneasiness, fear).

The fuzzy value for certain extended term can be computed based on $F^1$ through corresponding operator. For example, if we get the semantic feature vector of an art image, assumed as $F^1 = (0.32, 0.83, 0.92, 0.21, 0.48)$, its degree of membership for extended term 'very joy' can be calculated as $V_{joy}(0.92) = 0.92^2 = 0.85$.

4. Art image retrieval based on FASD

Query-by-example (QBE) strategy is popular utilized in typical content-based image retrieval system, however, users are accustomed to browsing and searching by semantic concepts. In this section, we demonstrate the query and retrieval strategy based on FASD feature in our approach.

The AIR system proposed in this paper is based on semantics, which is evidently different from typical content-based method. It has two distinguished characteristics: (1) the query is natural language-like expression based on FASD, rather than an example image; (2) the similarity between query description and FASD feature vector replaces the similarity between feature vectors when AIR searches images in database. Next, we will demonstrate the details about query expression and similarity measurement.
**Query expression**: In our AIR system, users can input natural language-like query description based on FASD to search art images. The query description can be expressed as a logic composition of linguistic terms. The logic operators include AND (denoted by $\land$), OR ($\lor$). Note that we do not define operator NOT, because we have designed the modifier 'few' and operator $F_t(\cdot)$ in the syntax rule $G$, which is similar with operator 'not' in function. Therefore, similarity between the query description and an art image can be interpreted as the similarity between the query expression and the image FASD feature vector.

**Similarity measurement**: First, we have to define the similarity function between single linguistic term $t$ and FASD feature vector. Since the degree of membership $y_t$ indicates the extent that an art image satisfies the linguistic term $t$ from aesthetic viewpoint. It is natural to define the degree of membership $y_t$ as the similarity between the query linguistic term $t$ and the art image. If the linguistic term $t$ belongs to basic term set we can directly get the degree of membership $y_t$ from FASD feature vector $F^t$ that is stored in feature database; otherwise, for an extended term $t' = G(t)$, which is generated ground on basic term $t$ by syntax rule $G$, its degree of membership is calculated based on $y_t$ by corresponding operator, just as discussed in Section 3.2.4.

After we measure the similarity between a single linguistic term and FASD feature vector, the final similarity between the whole query expression and FASD feature vector can be aggregated. For general, the logic composition of linguistic terms can be represented as a tree, named 'query-tree', in which each leaf corresponds to a linguistic term and each internal node corresponds to a logic operator. So the final similarity can be computed by aggregating the single similarity through min–max composition rules. Let consider a query expression 'fear $\lor$ (very uneasiness $\land$ few relaxation)'. We firstly transform the expression to a query-tree, as shown in Fig. 3(a). After that, the similarity value for each leaf is firstly achieved. For a basic term, its similarity value is directly derived from FASD feature vector $F^t$; on the other hand, for an extended term, its similarity value is calculated based on certain dimension of $F^t$ by specific operator defined in syntax rule $G$. In our example, the similarity value for leaf 'very uneasiness' is $V(y_4) = (y_4)^2$. At last, the final similarity can be computed by aggregating the three similarity functions through min–max composition rules, as shown in Fig. 3(b). In this example, the final similarity is formalized as $\text{Max}(y_5, \text{Min}(V(y_4), F(y_2)))$.

**5. Experiment results**

We have presented the methodology about fuzzy aesthetic semantics description and extraction for art image retrieval. In this section, we will demonstrate the details about AIR system implemented by the methodology proposed in this paper and show some experimental results.

Our experiments are divided into three parts:

*Fuzzy aesthetic semantics description* demonstrates the semantic features based on linguistic variable and shows some examples of FASD feature vector.

*FASD-based AIR* reports the details of an AIR system implemented by FASD.

*Retrieval performance* tests the performance of FASD feature extraction scheme, i.e. neural network approach, compared with other competitive algorithms.

For art image retrieval, we used a test set of 5000 images, including Renaissance, contemporary painters and traditional Chinese painting. The test set also includes images from a large online photo sharing community, http://www.photo.net, in which there are more than one million photographs uploaded by many amateur and professional photographers. Specially, many of these photographs are peer-rated in terms of two qualities, namely aesthetics and originality.

**5.1. Fuzzy aesthetic semantics description**

How to describe content of art images is very difficult, because it is not explicit about how the low-level features (such as color, shape) act on the aesthetic quality and how to measure the similarity between feature vectors even we toughly
define the visual feature for art images. On the other hand, the aesthetic perception of human being for art images is full of vagueness and subjectiveness.

However, in our approach, we describe the image semantics with an abundant and flexible linguistic expresses set, which is generated by basic term set and syntax rule $G$ (as discussed in Section 3). Benefitting from syntax rule $G$, which is a type of phrase-structure grammar, the linguistic express has powerful description ability. Opposite to utilizing the crisp probability in image classification systems, FASD utilizes fuzzy value (degree of membership) to label the image’s semantics according to certain linguistic value, and such fuzzy value is calculated by the semantic rule $M$. The rationale is that for every art image, FASD assigns a fuzzy value to it according to each linguistic value in linguistic term set. After an image is annotated, it is associated with a $K$-dimension linguistic-vector, where $K$ refers to the number of linguistic value. Each element in the linguistic-vector includes a keyword that represents the semantics description, and a value that represents the degree of membership. A typical vector may look like $(\text{action}; 0.35), (\text{few fear}; 0.83) \ldots (\text{very joy}; 0.79))$.

Just as discussed in Section 3.3, every image aesthetic semantic feature comprises of a 5-dimensional vector $(y_1, y_2, y_3, y_4, y_5)$, and the corresponding semantic concept is $(\text{action}, \text{relaxation}, \text{joy}, \text{uneasiness}, \text{fear})$. The three images in Fig. 4 are randomly selected from our semantic feature database. Generally speaking, most of the aesthetic semantics of art image obey the human aesthetic perception. For instance, Orange, the warmest color, resembles the color of fire and thus communicates action, and green that is always regarded as a cold color, communicates relaxation. Correspondingly, the degree of membership for linguistic value of ‘action’ in last image is 0.83, which is greater than the values of former two images. As well as, the degree of membership for ‘relaxation’ in middle image is obviously greater than other two images.

5.2. FASD-based art image retrieval

Traditional CBIR systems do not have the facility of querying in the form of natural language. For instance, if a painter wants images that look like ‘action’, she (or he) may feel inconvenient to find an example image of ‘action’ firstly and then to submit it to the system to search the desired images. Rather she (or he) prefers to utilize natural language, such as ‘action’, ‘few action’.

We implements the FASD-based art image retrieval system here, that is, when the user enters a linguistic express, constructed by a linguistic value or logic combination of linguistic values, the system returns the images which are relevant to the description.

In FASD-based art image retrieval system, users can retrieval the art image database through the linguistic expression, such as ‘very action’, ‘few joy’ or their logic combination. Fig. 5 illustrates the FASD-based art image retrieval interface and the retrieval result of the query ‘very action and few uneasiness’.

5.3. Retrieval performance

In FASD-based AIR system, how to extract the FASD feature vector (i.e. the semantic rule $M$) is one of the most important parts, and it is also the key factor impacting the retrieval performance since the search procedure is based on similarity between query expression and FASD feature vectors. In order to test the performance of extraction approach ground on neural network (briefly as NN) in this paper, we compare NN approach with two competitive algorithms:

- Genetic programming (GP) based method trains the semantic rule $M$ utilizing the GP algorithm, and then applies the rule to extract the fuzzy aesthetics semantics description of art image. This algorithm was utilized for texture image retrieval in our previous work [4].
- Content-based (CB) method directly applies the 94-dimension visual feature vector to search images just as typical content-based image retrieval.

To our best knowledge, there is not any common ground benchmark for semantic-based art image retrieval. Thus to enable objective evaluation, we have to first produce a ‘standard’ relevant image collection for each query description. Five judges who have no image retrieval related background and have normal visual perception are asked to produce the standard relevant image collections. According to an aesthetics description (noted as $q$), every judge marks each art image
with a confidence value $y$ ranged in $[0, 1]$. The confidence value represents the extent an art image satisfying the aesthetics description $q$. 0 means an art image is totally not suited for $q$; on the contrary, 1 means an art image is very likely for $q$. Then, the average confidence value $\bar{y}_I^q$ is calculated according to five judges’ confidence value, where $I$ represents certain art image. At last, all art images in collection are ranked according to confidence value $\bar{y}_I^q$, and the top $N$ art images compose standard relevant collection for aesthetics description $q$, noted as $S^q_N$. In our experiment, we produce 60 standard relevant collections, and the 60 aesthetics descriptions comprise 20 single linguistic terms defined in linguistic value set (see Section 2), 20 two-terms composition description which is a logic composition of two linguistic values (e.g. very fear and somewhat uneasiness) and 20 multiple-terms composition description which is a logic composition of more than two linguistic values.

We use top-$N$ accuracy to measure the retrieval performance in our experiments. Since they are both semantic-based image retrieval schemes in NN and GP, we measure the top-$N$ accuracy as following:

$$Ac^q_N = \frac{\text{Number of relevant images}}{N}$$

(5)

where $q$ refers to a query which is a pre-defined aesthetics description in standard collection $S$. If a returned art image $I$ is included in $S^q_N$, we take $I$ as a relevant image for query $q$. However, since CB is a content-based method and the linguistic query does not work any longer. Alternatively, the first art image in $S^q_N$ is taken as the query example, because it is the most satisfying art image for aesthetics description $q$. And the similarity between the 94-dimensional visual vector of query example and those of images in collection are calculated, and then the most $N$ similar art images are returned as the search result. The top-$N$ accuracy is counted as Eq. (5).

We change the number of returned images $N$ for each query $q$, i.e. 15, 30, 45, 60 and 90, to study the accuracy curve of the three compared methods. For one accuracy curve, we calculate the top-$N$ accuracy for every aesthetics description $q$ in $S$, and the average value is regarded as the final top-$N$ accuracy. The obtained final results have been shown in Fig. 6. The NN and GP significantly outperform the CB methodology, which indicates that the higher-level aesthetic semantic feature is more effective than the low-level visual features. The results confirm our analysis that although CBIR demonstrate excellent performance in query-by-example based image retrieval, it suffers the problems of semantic gap between low-level visual feature and high-level semantic feature. As for NN and GP, the accuracy of NN is better than that of GP. It is more difficult
for GP to construct the map between the low-level visual feature and the aesthetics description due to the deficiency of prior knowledge about how human understand art images. Furthermore, the semantic rule $M$ learned by GP is not a normal membership function since its output cannot be guaranteed to range in $[0,1]$ and certain post-processing is necessary. While our proposed NN methodology generally overcomes the weaknesses of GP and thus achieves the best accuracy.

The complexity of the aesthetic description also impacts the retrieval performance. During the experiments, we had the following observation: it is difficult and hesitating for judges to mark an art image according to an aesthetic description which is a logic composition of multiple linguistic terms, however, it is somewhat easy for them to mark an art image according to an aesthetic description which is a single linguistic terms. In the following, we test the accuracy for three categories of aesthetic description: single-term description, two-term description and multiple-term description. In this experiment, every category of description has 20 different aesthetic descriptions, and the average top-$N$ accuracy is calculated for every kind of description. Finally the accuracy curve is demonstrated in Fig. 7. It is easy to get that single-term description achieves the best retrieval performance, and the two-term description outperforms the multiple-term description. There are two reasons for this phenomenon: (1) single-term description always means simple aesthetic concept, and it is feasible and reliable to model the semantic rule $M$ for it. However, multiple-term description represents complex aesthetic concepts and it is a very complicated procedure even for human being to extract such concept. Obviously, logic combination of semantic rule $M$ of the component of multiple-term description is too simple to extract the complex concept. We will research powerful methodology to model the semantic rule of multiple-term description; (2) similarity measurement (discussed in Section 4), i.e. min–max operation rules, also impacts the retrieval results for the query with two linguistic terms or multiple linguistic terms. After all, a min–max operation rule is a simple numeric approximation for semantic similarity. We will probe into the similarity measure in the future. This observation indicates that simple aesthetic semantic concepts are feasible and effective for art image retrieval.
6. Conclusions and future work

Based on the analysis of properties of AIR and the existing possible solution, i.e. CBIR, we present a fuzzy semantic approach, named fuzzy aesthetic semantics description (FASD) based art image retrieval, to meet the challenges of AIR task. The approach depicts the higher-level emotional semantics of art image based on linguistic variable and extracts the fuzzy aesthetic semantic feature applying neural network ground on fuzzy set. Our experiments demonstrate that the described approach is significantly superior over typical CBIR method in terms of retrieval accuracy. Of course, the FASD-based AIR system is more convenient for users than typical CBIR scheme since users are accustomed to searching images by high-level aesthetic concepts, rather than by low-level visual feature vector or example image.

The major advantages of our work are (1) a principled framework integrating fuzzy set and image retrieval has been introduced to solve the problems in AIR, i.e. semantic gap and uncertainty of human visual perception; (2) Theoretically, every part of FASD is flexible, for instance, the linguistic term set can be redesigned according to different tasks, and semantic rule may apply any appropriate machine learning algorithms. In a word, the framework of FASD can be tailored to various tasks, like expressive aesthetic semantics.

However, there are also some limitations in this paper, which will motivate our future work. (1) The similarity measurement between query expression and FASD feature vector needs to be further probed. It is one of the key factors impacting the retrieval accuracy of AIR. (2) We only consider the emotional aesthetic semantics in this paper. Obviously, it will be more powerful to integrate other levels of aesthetic semantics, such as expressive semantics and painting style.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 60805041, and No. 60773016), 863 National High-Tech Program (No. 2007AA01Z168), National Basic Research Priorities Programme (No. 2007CB311004), and Doctoral Foundations of China (No.2005004001).

References