Image Search by Modality Analysis: A Study of Color Semantics

Xiaohui WANG, Jia JIA, Jiaming YIN, Yongxin WANG and Lianhong CAI[*]

[*] Key Laboratory of Pervasive Computing, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology（TNList）
Department of Computer Science and Technology, Tsinghua University, Beijing 10008
E-mail: wangxh09@mails.tsinghua.edu.cn, jjia@tsinghua.edu.cn, YJM19850724@gmail.com,
wangyongxin@mails.tsinghua.edu.cn, clh-dcs@tsinghua.edu.cn

***Abstract*—In this paper, we present a novel image search system, *color-based image search by modality analysis*. This system enables users to search images with modality words as query, like "romantic", "quiet", etc. Modality here is defined as emotions and conscious subjective aspect of feeling (interests, attitudes, opinions, etc). Modality presents richer semantics than emotions on human perception and understanding of images. We first propose seven modality dimensions to quantitatively analyze modality. A Chinese corpus containing 1200 modality words is constructed, in which each word is annotated with a seven-dimensional modality vector. Then three-color combinations are extracted from images as visual signatures by utilizing K-means clustering algorithm and CIE $L^*A^*B^*$ color space. To bridge the gap between modality (high-level semantic concepts) and three-color combinations (visual signatures), the prediction model for modality vectors is built up based on *Kobayashi's Color Image Scale* by using the decision tree algorithm. Finally, we establish the image search system. Objective evaluations show the accuracy of the prediction model for modality vectors, while subjective evaluation results indicate that the extracted three-color combinations are in agreement with human perception and search results of our system are consistent with queries.**

I. Introduction

The content-based image retrieval (CBIR) [1, 2, 3] uses objects as their prime descriptors of image content [4]. A typical task is searching images containing a house, a mountain, etc. However, there are times and situations when we know semantic content we desire, but are not clear about the objects in the images. Take, for instance, a desire to find a "romantic" image. The objects to express "romantic" can be a sandy beach with the sunset, a wedding scene, etc, since the same semantic content may have different visual appearances [5]. CBIR techniques cannot meet this desire, since it can hardly establish the relationship between its visual query, e.g. an example image or a painted sketch, and high-level semantic concepts.

Many recent researches in image retrieval have focused on high-level semantic concepts, such as emotions and aesthetics [2, 4, 6, 7, 8]. Emotional Semantic Image Retrieval (ESIR) analyzes and retrieves images at the affective level [9]. But human perception and understanding of images reflects not only on emotions, but also on other aspects, such as attitudes and opinions from the perspective of cognitive psychology [10]. In this paper, we extend emotions to modality in describing human perception and understanding of images. *Modality* here is defined as emotions and conscious subjective aspect of feeling (interests, attitudes, opinions, etc). With *modality*, we can describe the high-level semantic concepts of images and focus on designing a new image search system which enables users to search images using *modality words* like "romantic", "quiet".

Modality is the extension of emotions in describing human perception and understanding of images. Emotion analysis methods are used as references to describe modality, so we would like to introduce emotion analysis first. Emotion analysis is a core task in affective computing. There are two methods to describe emotions: category based method and dimension based method. In research of emotional categories, Ekman et al. propose six basic emotions (happy, sad, surprise, fear, anger and disgust) [11]. Machajdik et al. use eight emotional categories (amusement, awe, contentment, excitement, anger, disgust, fear and sad) for image classification [6]. In research of the dimensional approach, three basic dimensions are widely adopted to describe the affective information, namely *pleasure*, *arousal* and *dominance* [12]. The dimensional approach is more suitable for computation than the category based method. In this paper, we propose seven modality dimensions, which establish a modality space, to describe modality quantitatively. To represent a specific modality, values on the seven dimensions are used to form a modality vector.

Color has been shown to play an important role in image analysis [13]. Arnheim claims that all visual appearance owes its existence to color and the boundaries determining the shape of objects derive from the eyes' capacity to distinguish between areas of different colors [14]. Color, one of the main features affecting image semantics, is effectively used by artists to induce modality effects [14]. For example, red makes people feel excited and blue makes people calm. So we choose color features as image signatures. The modality of images is influenced by many other factors, such as shape, texture, etc. Possible relationships between modality and other types of features are beyond the scope of this paper. The method presented is a first step to a broader use of high-level semantic concepts in image search.

Most related works use common or generic features [6, 15, 16, 17] in image processing, such as color histogram, Itten contrasts [18] and the summarization of kinds of color features [2, 6]. Few researchers extract image features from a semantic perspective [8]. However, Kobayashi [20] proposes color semantics of three-color combinations, called the *Color Image Scale*, which is a system that maps 1,170 three-color combinations to 180 high-level semantic concepts, like “romantic”, “elegant”, etc. Within Graphic Arts, the *Color Image Scale* has already been successfully used in the selection of colors and color combinations. Three-color combination of an image is three prominent color centers. Dominant color extraction seeks for the one prominent color center that represents color information [21]. But using only one color to describe the color feature of an image is not enough. The three-color combination is able to not only capture the overall color feature of an image simply and easily, but also reflect the difference of high-level semantic concepts between images [4, 20]. So in this paper, we use three-color combinations as visual signatures of images. We also adopt the *Color Image Scale* to establish the relationship between modality (high-level semantic concepts) and three-color combinations (visual signatures).

The query words of ESIR are specific in most cases. In [4], 180 high-level semantic concepts, like “elegant”, “romantic”, are used for image indexing and search. Only 24 emotional words can be used as query in the image retrieval system of [8]. In our system, we establish a Chinese modality word corpus, which contains 1,200 modality words. These words can retrieve the images directly from our system. To further expand the set of acceptable query words, we find the most similar word in the corpus with the help of Hownet if the query is not in the corpus. Hownet [22] is a knowledge database, which contains more than 100,000 Chinese words, and it provides an algorithm to calculate the similarity between any two words in it.

In this paper, we propose a novel image search system, which enables users to intuitively indicate the search goal by modality words, like “romantic”, “quiet”, etc, rather than to think of the object in desired images. Compared with the traditional textual query, the texts describe not the objects in images, but emotions, interests, attitudes, opinions, etc. First, we construct a modality space with the proposed seven dimensions inspired by cognitive psychology. A Chinese corpus containing 1200 modality words is built, in which each word is annotated with a seven-dimensional modality vector. Then, we extract three-color combinations from images as visual signatures using K-means clustering algorithm and CIE $L^*A^*B^*$ color space. To bridge the gap between modality and three-color combinations, we set up a prediction model, which builds a mapping from three-color combinations to modality vectors, based on Kobayashi’s *Color Image Scale* by using decision tree algorithm. In image search process, for each query, we first project the query onto the modality space to get the modality vector. Then all the images are ranked by the similarities between the modality vector of query and those of images in target database. The search results are the top ranking images.

In summary, the key contributions of this paper are as follows:

1. We extend emotions to modality in the high-level semantic concepts to describe human perception and understanding of images, and propose seven modality dimensions to quantitatively analyze modality.
2. We establish the prediction model for modality vectors based on *Kobayashi’s Color Image Scale* to bridge the gap between high-level semantic concepts and visual signatures of images.
3. We set up an image search system, which allows modality words as query, and images with their corresponding visual signatures (three-color combinations) as output. Furthermore, with the help of HowNet， the high-level semantic concepts (queries) can be any modality word, which makes users’ search intention to be better represented.

The rest of this paper is organized as follows. Section II gives an overview of related works. Section III presents the framework of our image search system. The proposed dimensional description of modality is introduced in Section IV. Section V presents the extraction algorithm of three-color combinations and the color-based prediction model for modality vectors. The image search system is introduced in Section VI. Experimental results are shown in Section VII. Section VIII draws the conclusion.

II. Related Works

As an important semantic scope in semantics, modality refers to the opinion or attitude to the propositions or situations [24]. From the perspective of “attitudes and opinions”, linguists investigate the expression of modality in natural languages, and try to analyze modality in a systematic way [25]. The cognitive linguistic recognizes the modality as an important mechanism in human cognitive activities [26]. In linguistics, modality is the linguistic phenomenon whereby grammar allows one to say things about situation which need not be real [19]. More generally, some linguists propose that the modality includes all linguistic expression outside the proposition, and the sentence can be considered as the composition of proposition and modality [23]. From the perspective of human perception and understanding of images and reference to the definitions in semantics and linguistics, modality in this paper is defined as emotions and conscious subjective aspect of feeling (interests, attitudes, opinions, etc).

The *Color Image Scale* [20] is a book, or collection, developed by Shigenobu Kobayashi at the Nippon Color & Design Research Institute (Japan). Early thoughts about the system are presented in [27]. It maps 130 basic colors and 1,170 three-color combinations to 180 keywords, like “sweet”, “romantic”, based on the psychophysical investigations. For each of the 130 basic colors, nine three-color combinations with other basic colors are created. The 180 keywords are high-level semantic concepts related to the ways in which people perceive colors. All three-color combinations and the keywords are located in a two-dimensional semantic space, where the axes correspond to the scales *hard-soft* and *cool-warm*. Fig. 1 illustrates a few examples of three-color combinations and their corresponding keywords located in the two-dimensional semantic space.

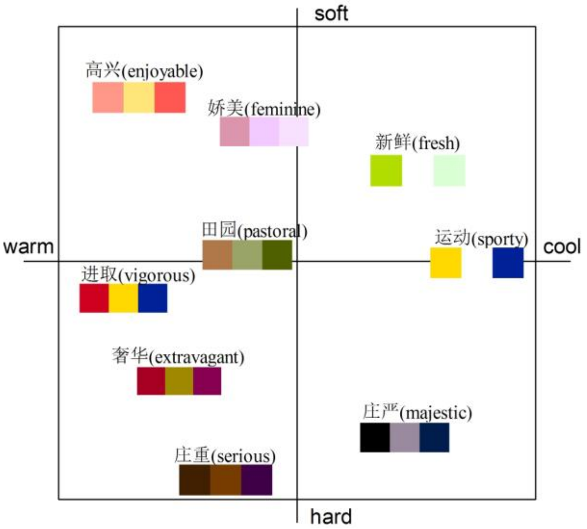

Fig. 1 Examples of three-color combinations and their corresponding keywords, located in the two-dimensional semantic space in *Color Image Scale*.

Solli et al. use *Kobayashi's Color Image Scale* for image indexing based on high-level color semantics [4]. They transform ordinary RGB-histograms of images to one of the basic colors and one of the three-color combinations in the *Color Image Scale*, and then find their corresponding keywords as the image descriptor. While in our approach, the mapping between the 1,170 three-color combinations and the 180 keywords is used to train the prediction model for modality vectors. Our approach has at least two advantages compared with [4]: 1) almost any modality word can be used as query in our system, making the users' search intention to be better represented, while the queries in Solli's system are only the 180 keywords; 2) in our system, colors in three-color combinations are in a continuous color space. In this way the image has a more precise description of color features, while only 130 basic colors are used in Solli's system.

III. Framework

The framework of our image search system is described as follows and illustrated in Fig. 2.

Before image search process, some preparation and preprocessing is completed by the following three steps.

Step 1: Construct the Chinese modality word corpus and annotate all the words in the corpus with the proposed seven modality dimensions.

Step 2: Train the color-based prediction model for modality vectors based on *Kobayashi's Color Image Scale*.

Step 3: Construct an image database, and then extract three-color combinations as the visual signatures of images. Label each image with a modality vector automatically using the prediction model.

For each query, the image search process has the following two steps.

Step 1: Project query onto the modality space and get the modality vector.

Step 2: Calculate the modality similarities between the modality vector of the query and those of all the images in the database, and rank the images in descending order of the similarities. The top ranking images are search results.

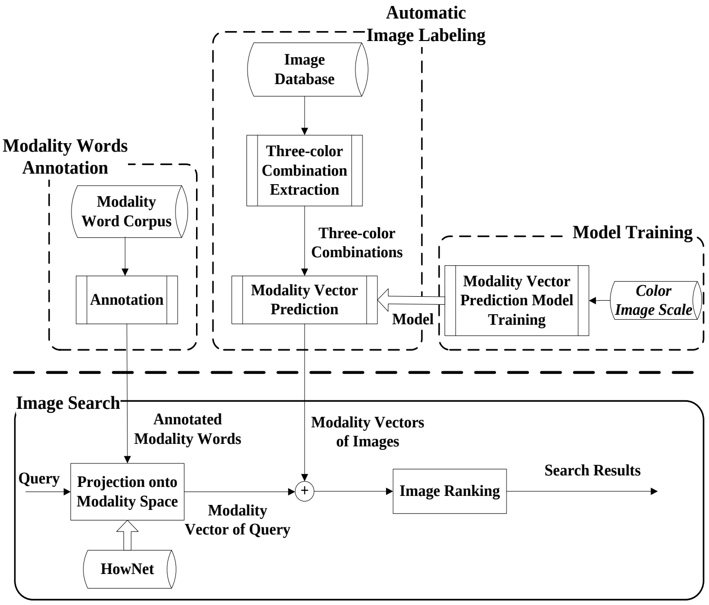

Fig. 2 Framework of the proposed image search system.

IV. Dimensional Description of Modality

We propose seven modality dimensions to construct a modality space. A modality word can be described by a seven-dimensional vector composed by the seven modality dimensions. By dimensional description of modality, we aim to provide a quantitative model to realize semantic computation. We set up a Chinese modality word corpus with 1,200 words and each word is annotated with the seven modality dimensions. Based on the modality dimensions, we can calculate the similarity between two words.

A. Definition of Modality Dimensions

Inspired by cognitive psychology, we propose seven modality dimensions to describe modality quantitatively. Cognitive psychologists find that human cognition, mental processing of understanding and knowing, involves various different kinds of information processing which occur at different stages [10] [28]. The main sequential stages of cognitive processing are attention, perception, memory, decision making and reasoning. Seven modality dimensions are selected to quantify the process of human cognition on images, which are shown in Table I. *Attention* represents the first stage of cognition called attention. *Pleasure*, *Arousal* and *Nervousness* are representations of perception stage. Attention and perception are the initial stages of human cognition and reflect the common understanding and knowing. Since our research is focused on the human cognition after learning, there is no modality dimensions related to the memory stage. *Certainty*, *Dominance* and *Determination* are representations for different aspects of decision making and reasoning stage.

TABLE I
DEFINITIONS OF MODALITY DIMENSIONS

Modality dimensions	Definitions
Attention	Interests in contents of an image
Pleasure	Positive/negative quality of emotions
Arousal	Physical activity and mental alertness
Nervousness	Control of one's own feeling towards the environments or situations described in the image
Certainty	Opinion towards situations described in the image
Dominance	Emotional status in the environments or situations described in the image
Determination	Will and belief in decision making of situations described in the image

B. Modality Word Corpus

Based on previous researches on Chinese words [29, 30], we collect 1,200 most common modality words to build the modality word corpus. The modality words include psychological adjectives, mental verbs and subjective adverbs. For each word, we collect 1) the explanation of its basic sememe which has modality related meaning; 2) the sample sentence using the word (sememe). Table II shows an example of words in the modality word corpus.

TABLE II
AN EXAMPLE OF WORDS IN MODALITY WORD CORPUS

Word	浪漫(romantic)
Explanation	富有诗意，充满幻想 (Poetic and full of fantasy)
Sample sentence	我们营造了一个浪漫的气氛。 (We create a romantic atmosphere.)

Most of the words are disyllable words (65.9%), others are monosyllable words (3.4%), three-syllable words (0.5%) and four-syllable idioms (30.2%). Most of the words have single and stable sememe.

C. Annotation of Modality Words

The modality word corpus is manually annotated with the seven modality dimensions. The annotators were asked to rate the modality words on each of the seven modality dimensions using a five-point Likert scale, which is described as follows:

- 2 - Obviously showing the positive state on the modality dimension;
- 1 - Generally showing the positive state on the modality dimension;
- 0 - Nothing is related to the modality dimension;
- -1 - Generally showing the negative state on the modality dimension;
- -2 - Obviously showing the negative state on the modality dimension.

We choose this 5-point scale to evaluate each of the modality dimensions because of the fuzziness of human perception of natural language.

Seven university students with Chinese as mother-tongue were invited to annotate the 1,200 modality words. The annotators discussed the standard of annotation to come to a consensus. An online annotation system was built to collect annotations from the seven annotators. During annotation, the system provided the explanation and sample sentence of the target word, and annotators needed to evaluate the modality dimensions with a five-point Likert scale according to his/her understanding of the word meaning and the standard of annotation. After the annotation, each modality dimension of a word was annotated seven times and the value was the one with the maximum number of annotation times. Each modality word in the corpus is described by a seven-dimensional vector containing values of the seven modality dimensions, which is the modality vector of the word.

We define annotation frequency to characterize the descriptive abilities of each modality dimension. Annotation frequency for a modality dimension is the number of non-zero annotations divided by the number of annotated words in the final annotation result. The results of annotation frequencies are shown in table III.

D. Modality Similarity

Given the modality vectors of two modality words, denoted as $\boldsymbol{X} = (x_1, x_2,\ldots, x_7)^T$ and $\boldsymbol{Y} = (y_1 , y_2,\ldots, y_7)^T$, we define modality similarity between two modality words as follows:

$$S_{\text{modality}} = \frac{1}{\sqrt{\sum_{i=1}^{7} \omega_i (x_i - y_i)^2}} \qquad (1)$$

where x_i and y_i are the values of the ith modality dimension of the two modality words. ω_i is the weight of the ith modality dimension, calculated as the annotation frequency of the ith modality dimension divided by the sum of annotation frequencies of all the dimensions. The weight of each modality dimension is shown in table III.

TABLE III
ANNOTATION FREQUENCY AND WEIGHT OF EACH MODALITY DIMENSION

Modality dimension	Attention	Pleasure	Arousal	Nervousness	Certainty	Dominance	Determination
Annotation frequency	59.57%	76.25%	88.32%	70.26%	41.51%	66.44%	32.48%
Weight	0.14	0.18	0.20	0.16	0.10	0.15	0.07

V. Color-Based Modality Vector Prediction of Images

Three-color combinations are extracted from images as visual signatures using K-means clustering algorithm. The decision tree algorithm is adopted to predict modality vectors from three-color combinations based on *Kobayashi's Color Image Scale* to bridge the gap between high-level semantic concepts and visual signatures. CIE $L^*a^*b^*$ color space is utilized in both three-color combination extraction and the prediction model. To predict the modality vector of an image, three-color combination is extracted first and the modality vector can be predicted from the three-color combination with the prediction model.

A. Extraction Algorithm of Three-color Combination

For an image I, K-means algorithm is adopted to calculate its three-color combination. Before starting K-means algorithm, I is first resized to a 100×100 pixels image I_r. In this way, the efficiency of the computation can be greatly improved and the cluster centers will not be affected. Colors of the 10,000 pixels would form a set $\boldsymbol{S}=\{I_r(i, j) \mid 1\leq i, j\leq100\}$, where $I_r(i, j) = (I_r^L(i, j), I_r^A(i, j), I_r^B(i, j))$ is the color of the $(i, j)^{\text{th}}$ pixel in the resized image in CIE $L^*a^*b^*$ color space.

K-means clustering algorithm is then run on $\boldsymbol{S}$ with three random initial cluster centers. The distance function between colors is the Euclidean distance in CIE $L^*a^*b^*$ color space [32]. After clustering, three cluster centers $\boldsymbol{C_1}$, $\boldsymbol{C_2}$, $\boldsymbol{C_3}$, where $\boldsymbol{C_i}$ is (L_i, A_i, B_i) (i=1, 2, 3), would form the three-color combination $\{\boldsymbol{C_1}, \boldsymbol{C_2}, \boldsymbol{C_3}\}$ of the image.

B. Prediction Model for Modality Vectors

The proposed prediction model is built for automatic image labeling, which is based on Kobayashi's *Color Image Scale*. The C4.5 decision tree algorithm is adopted and a feature vector is generated for each three-color combination as the input of the decision tree algorithm. Thus the prediction model consists of two parts:

1. *Feature vector generation*. For the three-color combination $\{\boldsymbol{C_1}, \boldsymbol{C_2}, \boldsymbol{C_3}\}$ of an image I, the feature vector is a 21-dimension vector $(L_1, A_1, B_1, L_2, A_2, B_2, L_3, A_3, B_3, L_{mean}, L_{var}, L_{max}, L_{min}, A_{mean}, A_{var}, A_{max}, A_{min}, B_{mean}, B_{var}, B_{max}, B_{min})^T$, where (L_i, A_i, B_i) is the color $\boldsymbol{C_i}$ in the three color combination; $L_{mean} = (L_1+ L_2+ L_3)/3$, $L_{var} = ((L_1- L_{mean})^2+ (L_2- L_{mean})^2+ (L_3- L_{mean})^2)/3$, $L_{max} = \max(L_1, L_2, L_3)$ and $L_{min} = \min(L_1, L_2, L_3)$ are the mean, variance, maximum and minimum of L component respectively; A_{mean}, A_{var}, A_{max}, A_{min}, B_{mean}, B_{var}, B_{max} and B_{min} are the mean, variance, maximum and minimum of A and B components, respectively. The feature vector represents the individual color component and their relationship in three-color combination.
2. *Modality vector prediction*. For each modality dimension in the modality vector, a decision tree is built and would predict discrete values (namely -2, -1, 0, 1 or 2) by using the 21-dimensional feature vector of the three-color combinations as input.

The model contains seven decision trees for the seven modality dimensions. In this paper, the collection of Kobayashi's *Color Image Scale* is adopted as the training and testing set, which contains 1170 three-color combinations and their corresponding 180 keywords. All the 180 keywords are in the modality word corpus and annotated by the modality dimensions. Validation of the model would be discussed in sub-section VII.A, where effects of different color spaces are also discussed.

VI. Image Search System

To demonstrate the performance of the proposed algorithm and model, we establish the image search system on an image database with 8,131 images. The system allows modality words as its query, and images with their corresponding visual signatures (three-color combinations) as its output. For each image in the database, the modality vector is labeled automatically using the prediction model proposed in sub-section V.B. In image search process, for each query, we first project the query onto the modality space to get the modality vector. Then all the images are ranked by the similarities between the modality vector of query and those of all the images in the database. The search results are the top ranking images.

A. Image Database Construction and Labeling

We construct an image database containing 8,131 images. These images are downloaded from flickr.com and google.com and divided into 11 categories: textiles, printed, fine art, poster, web design, product design, demonstration, interior decor, clothes, photography and architecture. The numbers of images in these categories are 686, 593, 474, 448, 800, 592, 200, 2312, 1132, 327 and 567 respectively. The classification of the image database is based on art and design. According to the image tag on the Internet, we assign each image to one category manually. The contents of the database contain almost all of the common picture themes of daily life. It is rich in the color composition, from using a few kinds of colors in art and design (e.g. poster, web design, product design) to containing a larger variety of colors in natural landscape.

This image database has been labeled automatically with modality vectors by the prediction model introduced in the previous section.

B. Project query onto the modality space

Considering the variety of queries, we need to annotate the words which don't belong to the modality word corpus. We accomplish this task by finding the word in the modality word corpus which is most similar to the query with the help of the HowNet, and using the modality annotation of the word found as the modality vector of the query.

HowNet [22] is a bilingual general knowledge base describing relations between concepts and the attributes of concepts. It is also an electronic lexical database for words, which contains 100,168 Chinese words by April 2010. The similarities between any two words can be calculated by HowNet knowledge system [22].

All the words in the modality word corpus are in HowNet. If the query is not in the modality word corpus, we calculate the HowNet similarities [22] between the query and all the words in modality word corpus, and then find the most similar word. The modality vector of the query is that of the most similar word.

C. Image Ranking based on Modality Similarity

For each query, we first calculate the modality similarities, defined in sub-section IV.D, between modality vector of the query and those of all the images in the image database. Then, all the images are ranked in descending order of the similarities. We choose the top ranking images as the search results.

VII. Evaluation and Results

In order to examine the validity of the proposed image search system and its components, three experiments were conducted on the database mentioned in VI.A. First, evaluation on modality vector prediction was to examine the accuracy of the proposed prediction model. Then, evaluation on three-color combination extraction was to test whether the extracted three-color combinations of images were in agreement with the human perception. Finally, evaluation on image search system was to validate the consistency of the search results and queries. Some search results were demonstrated at last.

A. Evaluation on Modality Vector Prediction

The prediction model aims to predict the modality vector of a given image from its three-color combination. The experiment setup was as follows: we separated the collection of Kobayashi's *Color Image Scale*, which contains 1170 three color combinations and their corresponding 180 keywords, into five subsets. Then we use K-fold cross validation (K = 5, 4 for training and 1 for testing) to get the prediction accuracy of the proposed model.

The prediction accuracy $A_{\text{prediction}}$ is defined as:

$$A_{\text{prediction}}(i) = \frac{\#Correct\_Prediction(i)}{\#ColorComb\_Test} \cdot 100\% \qquad (2)$$

where *#Correct_Prediction*(*i*) is the number of correct prediction for *i*th modality dimension. *#ColorComb_Test* is the number of three-color combinations in testing set, which is 234 in this experiment.

To further test whether color spaces have impact on the prediction accuracy, we trained and tested the prediction model in four color spaces: CIE $L^*a^*b^*$, HSV, RGB and LCH [33]. The feature vectors used in the model were calculated in each color space.

The prediction of each modality dimension is actually five-category (namely -2, -1, 0, +1 and +2) classification. The prediction accuracies of five-category classification in the four color spaces are shown in Table IV.

Table IV
The Prediction Accuracy $A_{\text{prediction}}$ of Five-Category Classification

Modality Dimensions	CIE $L^*a^*b^*$ (%)	HSV (%)	RGB (%)	LCH (%)
Attention	86.32	86.15	86.32	85.47
Pleasure	89.40	87.61	87.70	88.97
Arousal	87.01	85.64	86.58	86.67
Nervousness	86.15	85.47	83.50	86.15
Certainty	90.51	89.91	90.26	89.83
Dominance	90.00	89.91	90.51	90.77
Determination	92.05	90.85	91.54	91.54
Average	88.78	87.93	88.06	88.49

Since positive/negative judgment of modality dimensions gives great impact on the image semantic while the degree of a positive/negative judgment, like judging +1 or +2, has less impact, we combined the +1, +2 categories and the -1, -2 categories in the above result into one positive category (P) and one negative category (N) respectively. The prediction accuracies of three-category (namely P, N and 0) classification in the four color spaces are shown in Table V.

Table V
The Prediction Accuracy $A_{\text{prediction}}$ of Three-Category Classification

Modality Dimensions	CIE $L^*a^*b^*$ (%)	HSV (%)	RGB (%)	LCH (%)
Attention	91.71	92.48	92.05	91.97
Pleasure	91.97	90.85	91.03	91.88
Arousal	89.32	88.55	89.40	88.63
Nervousness	90.34	90.26	88.72	90.51
Certainty	92.05	91.79	91.79	91.88
Dominance	90.94	90.34	90.77	91.45
Determination	92.91	92.48	92.74	92.99
Average	91.32	90.96	90.93	91.33

The average prediction accuracy of five-category classification and three-category classification for each modality dimension reach 80% and 90% respectively. It indicates that the prediction model achieves high prediction accuracy and is effective.

Results also show the accuracies with different color spaces have subtle difference. Due to the high accuracy of the three-color combination extraction in CIE $L^*a^*b^*$ color space, which will be shown in the next sub-section, we use CIE $L^*a^*b^*$ space in modality vector prediction to reduce the calculation of color space conversion.

We also chose other popular classifiers, such as Support Vector Machines (SVM), the Naïve Bayes classifier, to train the prediction model for modality vectors. But in our case the decision tree classifier proved to be the best in the prediction accuracy.

B. Evaluation on Three-color Combination Extraction

A subjective experiment was conducted to test that how many colors in the extracted three-color combination of an image by the extraction algorithm were consistent with the human perception. Since HSV color space and CIE $L^*a^*b^*$ color space have shown to achieve good performance in dominant color extraction [34], we also compared the accuracy of extracted three-color combinations between using CIE $L^*a^*b^*$ color space and using HSV color space.

Participant. Seven university students (3 females and 4 males) with normal visual acuity and normal color vision participated in this experiment. They were all novel to the test.

Stimuli. 153 images were randomly chosen from the image database and divided into two groups according to the complexity of color composition: 52 images with simple color composition, containing a few kinds of colors, (e.g. poster, web design, product design) and 101 images with complex color composition, containing a larger variety of colors, (e.g. photography, interior decor). For each of the 153 images, three-color combinations in CIE $L^*a^*b^*$ space and HSV space were extracted. Totally 306 images and their corresponding three-color combinations were obtained to be evaluated.

Procedure. An image and its corresponding three-color combination were displayed on the screen simultaneously. The middle of the screen was the image and the bottom of the screen was the three-color combination. For each image and its three-color combination, participants were asked to judge how many colors in the three-color combination were prominent in the image. The images and their three-color combinations in the two color spaces were displayed randomly. All participants completed the 306 trials in this experiment.

Results and Discussion. We use accuracy to measure the consistency of the extracted three-color combinations and the human perception of images. The accuracy $A_{\text{extracted}}$ is defined as:

$$A_{\text{extracted}}(m,G)=\frac{\#Good\_Extraction(m,G)}{\#Img(G)}\cdot 100\% \qquad (3)$$

where G stands for the two groups in the testing set: the simple color composition group and the complex color composition group. $\#Good\_Extraction(m, G)$ is the number of Good Extraction in G, and a Good Extraction is defined as the extracted three color combination have at least m ($m \in \{1,2,3\}$) colors judged as prominent by a participant. And $\#Img(G)$ is the number of images in G.

The average accuracies over all participants are shown in Table VI. The results demonstrate that the extracted three-color combinations of images are high consistent with human perception in CIE $L^*a^*b^*$ color space. Fig. 3 is an extraction example by using CIE $L^*a^*b^*$ color space and HSV color space. From the visual comparison of extraction results, we can also conclude that using the CIE $L^*a^*b^*$ color space is more close to human perception.

TABLE VI
THE ACCURACY ($A_{\text{EXTRACTED}}$) OF THREE-COLOR COMBINATION EXTRACTION ALGORITHM

Number of consistent colors (m)	1	2	3
Simple color composition (CIE $L^*a^*b^*$)	98.08%	96.15%	82.69%
Simple color composition (HSV)	78.85%	67.31%	63.46%
Complex color composition (CIE $L^*a^*b^*$)	95.05%	90.10%	88.12%
Complex color composition (HSV)	78.22%	72.28%	60.40%

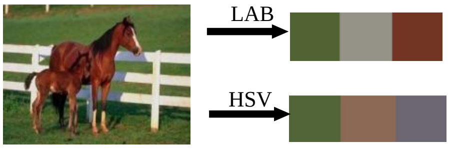

Fig. 3 A three-color combination extraction example by using CIE $L^*a^*b^*$ color space and HSV color space.

C. Evaluation on Image Search System

Objective evaluation measures for image search performance (for example, precision and recall, etc.) are hard to design since color-based high-level semantic concepts are hard to define [4]. So we designed a subjective experiment to evaluate the precision of the proposed image search system instead.

Participant. Ten university students (5 females and 5 males) with normal visual acuity and normal color vision participated in this experiment. They were all novel to the test.

Stimuli and Procedure. Twenty-two different words were used as query. Among the 22 words, 17 words were in the modality word corpus and the other 5 words were only in HowNet. For each of the 22 words, three best results with top ranking were obtained. And totally 66 query-image pairs were collected.

In this experiment, the participants provided the judgment (Yes or No) on whether the image matched to the query. The query and the image of a query-image pair were displayed on the screen simultaneously. The top of the screen was the query and the middle of the screen was the image. The 66 query-image pairs were displayed randomly. All participants completed the 66 trials in this experiment.

Results and Discussion. We use precision to measure the consistency of the search results and the queries. The precision P is defined as:

$$P=\frac{\#relevant\_Img}{\#Img}\cdot 100\% \qquad (4)$$

where *#relevant_Img* is the number of search results, which are relevant to query by participant perception. And *#Img* is the number of images, which is 66 in our experiment.

The average search precision (the percentage of “Yes” judgment) over all participants is 89.1%. This indicates that the search results are consistent with human perception and our system can meet the demands of users.

D. Demonstrations

Fig. 4 gives examples of the results of image search system. For each query, we give its modality vector, three best matches and their corresponding three-color combinations. Fig. 5 illustrates more search results based on modality words.

VIII. CONCLUSIONS

Modality is the extension of emotions in describing human perception and understanding of images. In this paper, we propose a novel image search framework that allows modality words as query. Our main contributions include: 1) The dimensional approach is adopted to quantitatively analyze modality, which realizes the high-level semantic concepts involving in parametric computation; 2) Most existing related works, such as Kobayashi's *Color Image Scale*, are originally designed for a finite number of colors and a finite number of semantic concepts. By the proposed prediction model for modality vectors, we realize that colors in three-color combinations are in a continuous color space, which leads to more precise description of color features in images. Furthermore, with the help of HowNet, the high-level semantic concepts can be any modality word, which makes the users' search intention to be better represented; 3) we implement an image search system, which enables users to intuitively indicate the search goal by modality words. This image search system provides a new search experience to users. And its performance has been proven by both objective and subjective experiments.

Future research focuses on incorporating our framework into content-based image retrieval. It allows more flexible inputs to meet the various demands of users.

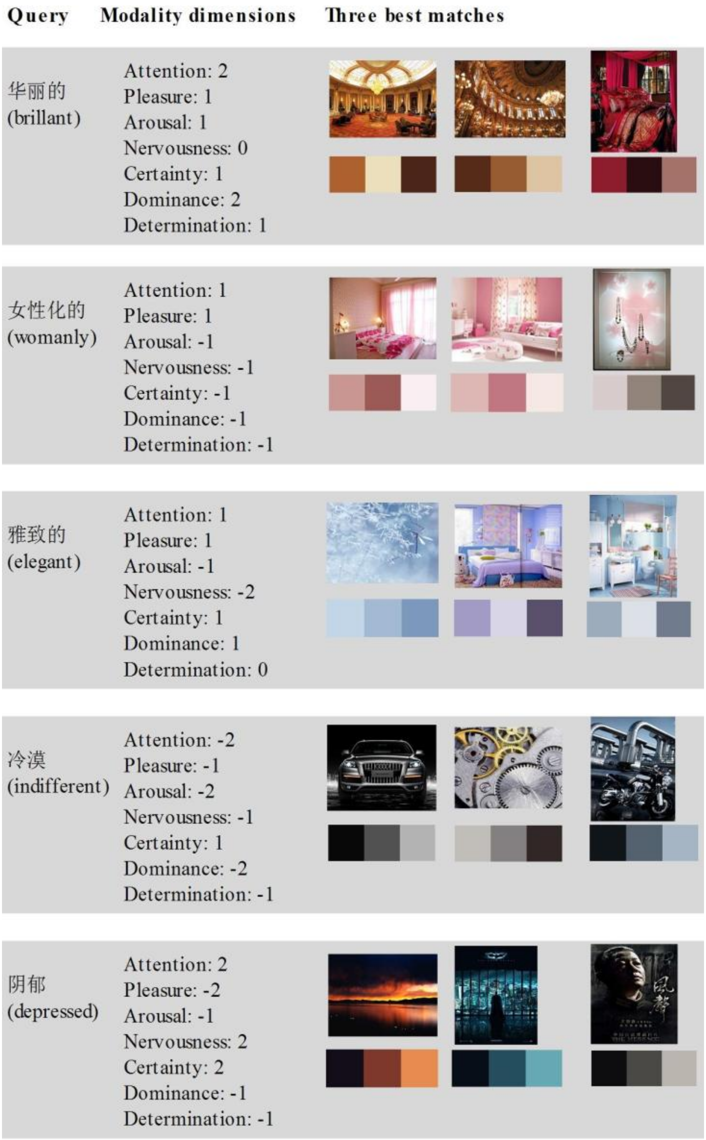

Fig. 4 Results of image search system, including modality words, modality vectors, three best matches and corresponding three-color combinations.

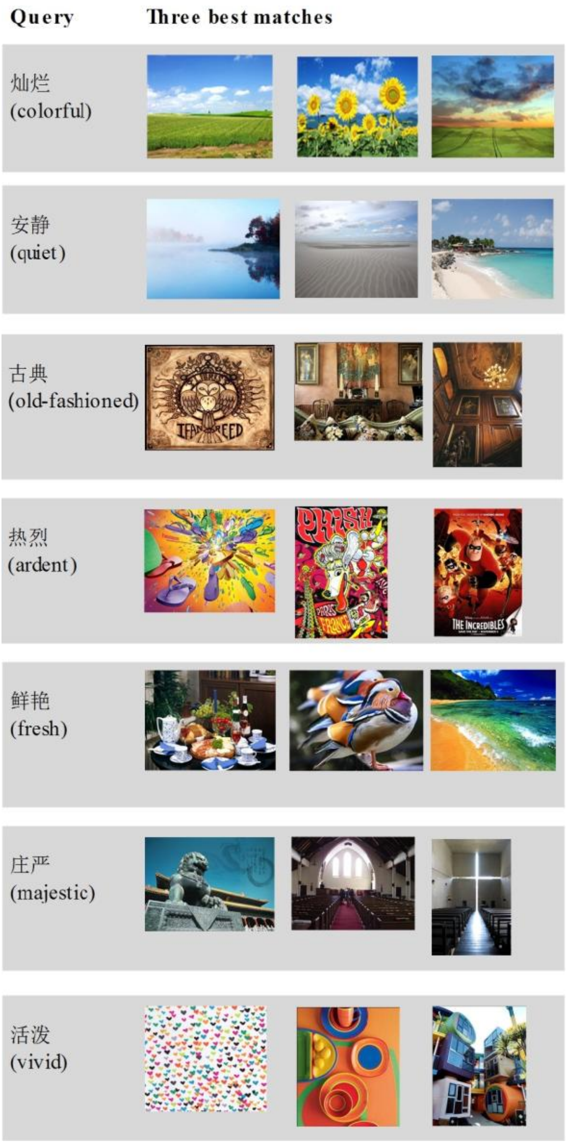

Fig. 5 Some other results of image search based on modality words (three best matches).

ACKNOWLEDGMENT

This work is supported by the National Basic Research Program (973 Program) of China（2011CB302201）, and National Natural, and Science Foundation of China (61003094， 90820304).

REFERENCES

[1] Y. Zaheer, “Content-based image retrieval,” *Proceedings of SPIE*, Vol. 7546, 2010.

[2] R. Datta, D. Joshi, J. Li, and J. Z. Wang, “Image retrieval: Ideas, influences, and trends of the new age,” *ACM Computing Surveys (CSUR)*, Vol. 40, 2008.

[3] T. Chen, M. M. Cheng, P. Tan, A. Shamir, and S. M Hu, “Sketch2Photo: internet image montage,” *ACM Trans. Graph*, 28, 5, 124: 1–10. 2009.

[4] M. Solli and R. Lenz, “Color semantics for image indexing,” in *Proceedings CGIV 2010, 5th European Conference on Colour in Graphics, Imaging, and Vision*, pp. 353-358, 2010.

[5] H. Xu, J. D. Wang, X. S. Hua and S. P. Li, “Image Search by Concept Map, ” *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010.

[6] J. Machajdik and A. Hanbury, “Affective image classification using features inspired by psychology and art theory,” *Proceedings of the international conference on Multimedia*, 2010. DOI:10.1145/1873951.1873965.

[7] M. Solli, *Color Emotions in Large Scale Content Based Image Indexing*. LinkÄoping University Electronic Press, 2011.

[8] W. N. Wang, Y. L. Yu, S. M. Jiang, “Image retrieval by emotional semantics: A study of emotional space and feature extraction,” *IEEE International Conference on Systems, Man and Cybernetics*, 4(8-11): 3534-3539, 2006.

[9] W. Wang and Q. He, “A survey on emotional semantic image retrieval,” *15th IEEE Int. Conf. on Image Processing*, pp: 117-120, 2008.

[10] R. J. Sternberg, *Cognitive Psychology*, *6th*. Wadsworth Publishing, 2011.

[11] P. Ekman, W. V. Friesen, *Manual for the Facial Action Coding System. PaloAlto [M]*.Consulting Psychologists Press, 1977.

[12] A. Mehrabian, “Pleasure-arousal-dominance: A General Framework for Describing and Measuring Individual Differences in Temperament,” *Current Psychology: Developmental, Learning, Personality, Social*, Volume 14, 261-292, 1996.

[13] J. Tanaka, D. Weiskoph, P. Williams, “The role of color in high-level vision,” *Trends in Cognitive Science* 5, 5, 211–215, 2001.

[14] R. Arnheim, *Art and visual perception: A Psychology of the Creative Eye [M]*. University of California Press, 2004.

[15] N. Bianchi-Berthouze, “K-dime: an affective image filtering system,” *IEEE Trans. Multimedia*, Volume 10, no.3, pp.103 - 106, 2003.

[16] T. Hayashi, M. Hagiwara, “Image query by impression words-The IQI System,” *IEEE Multimedia*, vol 6, No.3, pp.38-53, 1999.

[17] K. Yoshida, T. Kato, “Image Retrieval System Using Impression Words,” In *Proc. IEEE Systems, Man and Cybernetics*, Tokyo, 1998, vol. 3, no. 11-14, pp. 2780-2784.

[18] J. Itten, *The art of color: the subjective experience and objective rationale of color*. John Wiley, New York, 1973.

[19] P. Portner, *Modality*, Oxford Surveys in Semantics and Pragmatics, 2009.

[20] S. Kobayashi, *Color Image Scale*. Kodansha Intern., 1991.

[21] S. Kiranyaz, S. Uhlmann, T. Ince, et al., “Perceptual Dominant Color Extraction by Multi-Dimensional Particle Swarm Optimization,” *EURASIP Journal on Advances in Signal Processing*, Vol. 2009.

[22] Z. D. Dong and Q. Dong, *HowNet And The Computation Of Meaning*. World Scientific, 2006.

[23] C. J. Fillmore, The Case for Case. In Bach & Harms, R. T. (Eds.), *Universals in Linguistic Theory* (pp. 1-88). New York: Holt, Rinehart and Winston Inc, 1968.

[24] J. Lyons, *Semantics* (Vol. I), Cambridge University Press, 1977.

[25] F. R. Palmer, *Mood and Modality*. Cambridge: Cambridge University Press, 2001.

[26] J. Nuyts, *Epistemic Modality, Language, and Conceptualization: A Cognitive-Pragmatic Perspective* (Vol. 5). Amsterdam: John Benjamins Publishing Company, 2001.

[27] S. Kobayashi, “Aim and method of the color image scale,” *Color Res Appl*, 6(2):93–107, 1981.

[28] D. Groome, *An Introduction to Cognitive Psychology: Processes and Disorders*. Psychology Press, 1999.

[29] J. Lan, *The Quantification Features of the Mental Verb in Modern Chinese*. Fudan University, Shanghai, China, 2008.

[30] J. Mei, *Synonyms Set*. Shanghai CiShu Press, Shanghai, China, 1983.

[31] J. L. Hu, J. B. Deng and S. S. Zou, “A Novel Algorithm for Color Space Conversion Model from CMYK to LAB,” *Journal of Multimedia*, Vol 5, No 2, pp. 159-166, 2010.

[32] G. Sharma (2003), *Digital Color Imaging Handbook* (1.7.2 ed.). CRC Press.

[33] A. Hanbury, “Constructing cylindrical coordinate colour spaces,” *Pattern Recognition Letters*, 29 (4), 494 ‑ 500, 2008.

[34] Y. TABII, R. O. H. Thami, “A framework for soccer video processing and analysis based on enhanced algorithm for dominant color extraction,” *International Journal of Image Processing (IJIP)*, vol. 3, pp. 187-245, 2009.