The Third Eye: Mining the Visual Cognition across Multilanguage Communities

Chunxi Liu[1,4], Qingming Huang[1,2,4], Shuqiang Jiang[2,4], Changsheng Xu[3,4]
[1]Graduate University of Chinese Academy of Sciences, Beijing, 100049, China
[2]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, 100190, China
[3]National Lab of Pattern Recognition, Institute of Automation, CAS, Beijing, 100190, China
[4]China-Singapore Institute of Digital Media, 21 Heng Mui Keng Terrace, 119613, Singapore
{cxliu, qmhuang, sqjiang}@jdl.ac.cn, csxu@nlpr.ia.ac.cn

ABSTRACT
Existing research work in the multimedia domain mainly focuses on image/video indexing, retrieval, annotation, tagging, re-ranking, *etc*. However, little work has been contributed to people's visual cognition. In this paper, we propose a novel framework to mine people's visual cognition across multi-language communities. Two challenges are addressed: the visual cognition representation for a specific language community, and the visual cognition comparison between different language communities. We call it "*the third eye*", which means that through this way people with different backgrounds can better understand the cognition of each other, and can view the concept more objectively to avoid culture conflict. In this study, we utilize the image search engine to mine the visual cognition of the different communities. The assumption is that the image semantic distribution over the search results can reflect the visual cognition of the community. When a user submits a text query, it is first translated into different languages, and fed into the corresponding image search engine ports to retrieve images from these communities. After retrieval, the obtained images are categorized into different semantic clusters automatically. Finally, inter semantic cluster ranking is employed to rank the semantic clusters according to their relationship to the query, while intra cluster ranking is used to rank the images according to their representativeness. The visual cognition difference among these language communities is achieved by comparing the different community image distributions over these semantic clusters. The experimental results are promising and show that the proposed visual cognition mining approach is effective.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: *Information search and retrieval-search process*.

General Terms
Algorithms, Design, Human Factors, Experimentation

Keywords
Visual cognition, image search, language community

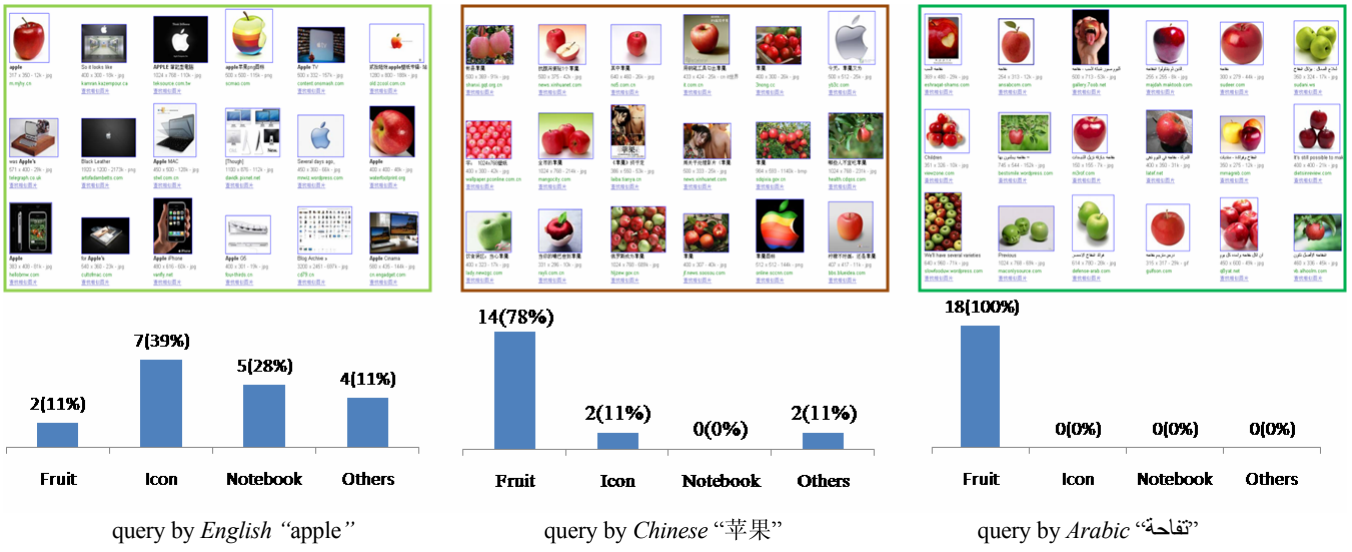
MM'10, October 25–29, 2010, Firenze, Italy.

1. INTRODUCTION
With the rapid development of the modern technologies, the communication between different countries and cultures is getting more and more convenient. However, at the same time due to their different cognitions, the conflicts between different communities/cultures are also getting more and more intense. The 21st century is even called as the century of cultural conflict. According to the *Freudian Projection* effect [23] in psychology, we usually take it for grant that our cognition on a concept or the way of our thinking is similar to others. However, due to the different backgrounds, cultures, ideologies, *etc*. the cognition of the people, who use different languages, for the same concept is different. Figure 1 is an example to show the different understanding on the concept "*dragon*" between the Chinese and the European. The left one is the Chinese dragon, which is a totem, while the right one is the European dragon, which always has two wings. According to the descriptions in the Wikipedia [1], "*Chinese dragons and Oriental dragons generally, can take on human form and are usually seen as benevolent, whereas European dragons are usually malevolent*". Therefore, for these two communities the visual cognition for "*dragon*" is rather different.

Chinese dragon *European dragon*

Figure 1. An example to show the cognition difference between the Chinese and the European on "*dragon*".

According to Wikipedia [1] "*The term cognition refers to a faculty for the processing of information, applying knowledge, and changing preferences. Cognition, or cognitive processes, can be natural or artificial, conscious or unconscious.*" The cognition of people is different from culture to culture. In the multimedia domain, everyday a large number of data, including text, audio, image and video, are generated, edited, and spread. The data are generated by people belonging to a specific community, and are full of subjective cognition of the world. Among these data, image is often used to show people's understanding, cognition, *etc*, with the belief that "*One picture is worth ten thousand words*".



query by *English* "apple" query by *Chinese* "苹果" query by *Arabic* "تفاحة"

Figure 2. The top 18 images of the search result of the query "*apple*" from the most popular image search engine *Google Image*. It is observed that the semantic distribution of "*apple*" is rather different from one language community to another.

In the multimedia domain, image search is a hot research topic, and many image search engines are available online, such as *Google image* [2], *Bing image* [3], *Yahoo image* [4], *Ask image* [5], *etc*. It is quite convenient for us to retrieve lots of images through image search engine by simply typing some query keywords. Figure 2 shows the top 18 images retrieved by the *Google image* using three languages of the query "*apple*". The algorithm utilized by the *Google image* search is the well known *PageRank(PR)* algorithm [16]. The images with high *PR* values are ranked at the top rank, which means these images are more valuable than others from the viewpoint of the image link quotation.

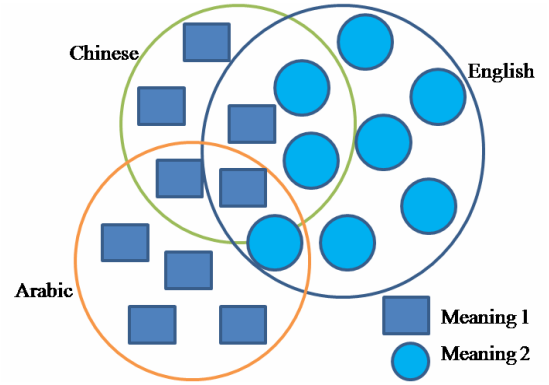


Figure 3. A symbolic example to show the cognition difference among the language communities.

We define the people using the same language as a language community. We also define four concepts for the above example in Figure 2: apple fruit, apple icon, apple notebook, and others. We calculate these concepts distribution over the three retrieval results, which are also shown in Figure 2. It can be observed that the concept distributions over the retrieved images are quite different from one community to another. Usually, people may think that this result is caused by the un-preciseness of the image search algorithm. However, this does not fully reflect the real situation. From Figure 2 we can see that although few of the images are unrelated to the query, most of the images have something to do with the concept "*apple*". From a more objective viewpoint, this phenomenon is not fully caused by the un-preciseness of the search algorithm, but partly by the different data distribution over those language communities. Figure 3 is a symbolic example to explain how the phenomenon in Figure 2 is generated. Figure 3 shows there are two meanings of a concept: meaning 1 and meaning 2. However, due to the different cultures, the coverage over these two meanings is different for the three communities. From Arabic, Chinese to English, the coverage is moving from meaning 1 to meaning 2. The different visual cognition coverage over the meaning space for the same concept results in the different retrieval outcome for these language communities. However, the influence of these language communities is different. We use the number of retrieved images in the search engine as the coarse index of the influence of the community. The influence comparison of the three language communities on the concept "*apple*" is shown in Figure 4. It can be seen that compared with the Chinese and the English community, the influence of the Arabic community is smaller. Even though, the cognition of the small community should be respected, otherwise the conflict between different communities may happen.

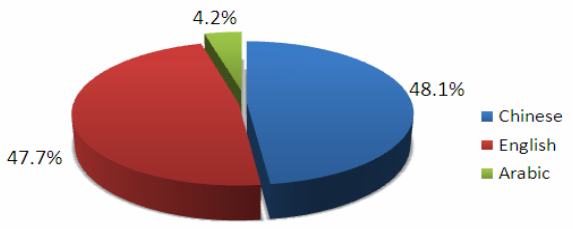


Figure 4. The influence comparison of the three language communities.

In this paper, based on the images retrieved by the image search engine, we propose a novel approach to mine the visual cognition

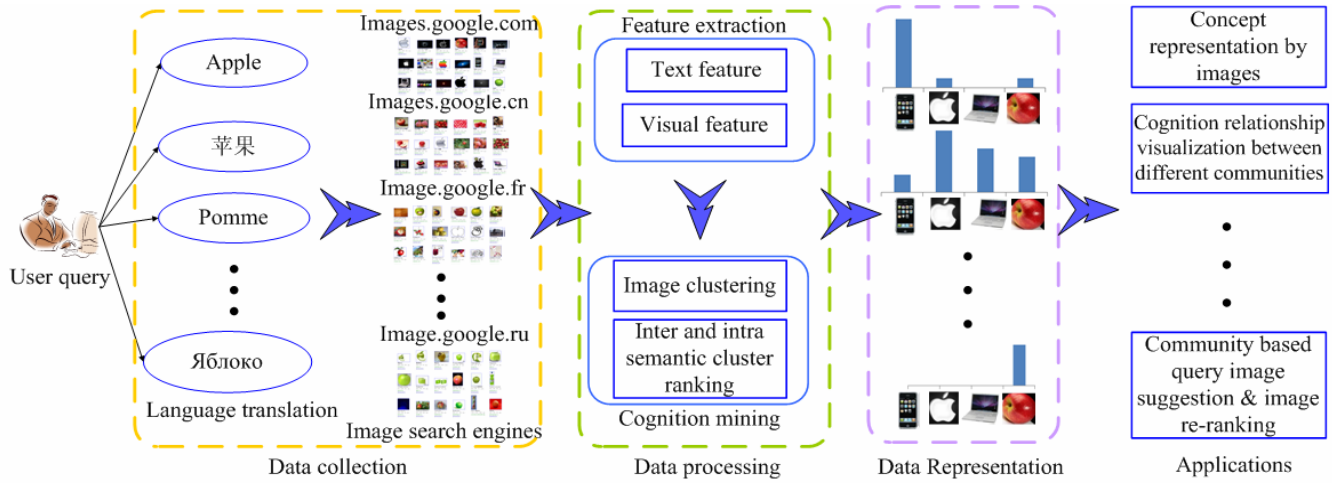


Figure 5. The proposed visual cognition mining framework.

difference among the language communities for the same concept. The assumption is that most of the people in the same language community have similar cognition point of view, while the people belonging to different language communities may hold different cognition viewpoints. The images distributed in the web is full of people's cognition about the world, and the top images retrieved by the image search engine from a language community can reflect the main visual cognition of the people in the community. From Figure 2 it can be seen that the cognition difference of the three language communities can be obtained by comparing the concept distribution over the image data. By comparing the concept distribution over the search results, we can find that the cognition of Chinese and Arabic on the concept "*apple*" is similar. Therefore, if we can build a feature description like the concept distribution probability in Figure 2, the visual cognition comparison between different language communities will become easy.

The proposed cognition mining framework is shown in Figure 5. When the user submitted text query is obtained, it is first translated into different languages. Then, the different language queries are fed into the corresponding image search engine ports to retrieve images from the different language communities. The retrieved images are downloaded together with their file names and surrounding text. After that, features are extracted from the images and text, respectively. For image feature, the Scale Invariant Feature Transform (*SIFT*) based interesting points are detected in each image, and the bag-of-visual-word is extracted to represent each image. Then, the probabilistic Latent Semantic Analysis *(pLSA)* [11] model is employed to further analyze these bag-of-word features, and the topic distribution feature is generated for each image. For text, the bag-of-word text feature is generated. At last, these images are clustered into different semantic clusters. For clustering, the Affinity Propagation (*AP*) [8] algorithm is adopted for its good performance. The final cognition of the communities is represented by the concept distribution over the semantic clusters. The cognition difference is obtained by comparing their concept distributions. For concept distribution comparison, the symmetrical Kullback-Lebler (*KL*) [15] divergence is employed. Three applications are proposed: the visual cognition representation for a specific language community query, the visual cognition relationship visualization among different communities, and the community based query image suggestion and image re-ranking. The experimental results will show the effectiveness of the proposed approach. To the best of our knowledge, this work represents the first attempt toward mining the different visual cognition viewpoints across multi-language communities for the same concept by using the image search engine. The main contributions of this paper can be summarized as follows:

- We propose a novel community visual cognition mining and comparison approach using image search engine. We call it "*the third eye*", the meaning is that by this way we can see a specific concept more objectively, and the people from different communities can better understand each other to avoid culture conflict.
- We propose a novel inter cluster ranking scheme to mine the cognition viewpoint of a community for a specific concept, which considers both the expectation ranking order and the size of the clusters.
- We propose to use the concept distribution over the dataset to evaluate the visual cognition relationships between different communities. The similarity values are visualized through graph, from which we can see the relationships among these communities clearly.

The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 presents the visual cognition mining approach, which includes feature extraction, image clustering, inter and intra cluster ranking for visual cognition representation, and visual cognition comparison. The experimental results and evaluations are provided in section 4. Finally, we conclude the paper with future work in section 5.

2. RELATED WORK

In recent years, many web image search based applications have been proposed. In this section we mainly review the related work on four domains: image search result clustering and representation, image re-ranking using multi-search engines, text-image linking, and video ideology classification. For image search result clustering and representation, several algorithms have been proposed [9][12][24][25], *etc*. A reinforcement clustering algorithm and a bipartite graph co-partitioning algorithm are proposed to integrate visual and textual features in [24] and [9], respectively. Jing [12]

proposes to first mine query related key phrases, and employ the new phrases to retrieve images from the search engine. After that, the image search results are presented as the clusters named by the key phrases. Similar to [12] an image clustering result representation work is proposed by Zha [25] to solve the query ambiguous problem. Firstly the query word related images are retrieved from the *Flicker* [6]. After that, a set of keywords are selected and the representative images for each keywords are generated. Finally, the selected images are used as positive examples to refine the initial search result. For image search re-ranking using multi-search engines, a work is proposed by Liu [17]. In this work, the visual patterns from multi-engines are used to help refining the search result, which is called *CrowdReranking*. For text-image linking some work has been proposed [10][13] etc. Hong [10] proposes an application called *Mediapedia*, which links the web mined images with the text in the Wikipedia. The image data is collected from *Flicker* and is clustered through *AP* [8] algorithm. Another work called Story Picturing Engine is proposed by D. Joshi [13], where the text story is illustrated by images. Firstly, the keywords are selected from the text story. Then, based on these keywords, the images are selected from the image database. After that, the reinforcement ranking is used to rank the images according to their relationships to the story.

Although most of the above work utilizes the image information in the database or from the web, none of them addresses the different cognition across multi-language communities. The work most similar to ours is proposed by Lin [18], where he addresses the different ideological perspective of the news video based on the emphatic patterns. Their assumption is that news broadcasts holding contrasting ideological beliefs appear to emphasize different subsets of visual concepts. However, there are several differences. Our objective is to mine the visual cognition across multi-language communities for any concept, while they only limit their work to some specific news event. In our view, ideology is a subset of cognition. In our approach the cognition is mined using the natural image data distributed in the web, while in their work the data are manually selected. The well trained models cannot be adapted to other events automatically.

3. VISUAL COGNITION MINING ACROSS MULTI-LANGUAGE COMMUNITIES

In this section, we will elaborate the proposed visual cognition mining and comparison approach across different language communities. We will show the visual cognition viewpoint of a specific community for a given concept, and see how to compare the cognition among different communities.

3.1 Semantic Image Clustering

In order to mine the visual cognition for a specific community query, the first step is to cluster the retrieved images into semantic clusters. The proposed image clustering and ranking framework is shown in Figure 6. When the images and the text are obtained, the bag-of-visual-word and bag-of-word features are extracted respectively. Then the bag-of-visual-words are further analyzed to generate the more effective topic distribution feature. After that, the bag-of-word text feature and the topic distribution feature are used to measure the similarity between the images, and the affinity propagation clustering [8] is utilized to cluster the images. Finally, these semantic clusters and images are ranked to represent the visual cognition of the community.

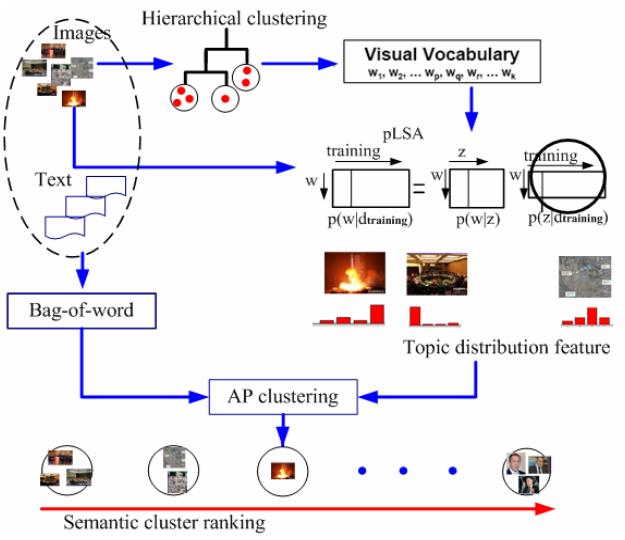


Figure 6. The image clustering framework.

For a web image two kinds of text information are available: the *URL* information and the content information. For *URL* information we refer to the image name, which is often contained in the site *URL*. The image name is added by the editor of the image to show some specific description of the image. Therefore, the image name may be an important cue to mine the semantic of the image. For content information we refer to the text surrounding the image. In most of the time, the image surrounding text always describes the related information about the image. After these texts are extracted, we adopt the vector model to describe the texts, where the textual feature of a document is defined as:

$$f = (k, w) \tag{1}$$

where $k = (k_1, k_2, ..., k_n)$ is a dictionary of all keywords appearing in the whole document pool, $w = (w_1, w_2, ..., w_n)$ is a set of corresponding weights, n is the number of unique keywords in the document dataset. We employ the term frequency (*tf*) and the inverted document frequency (*idf*) to calculate the importance of a keyword in a document. *tf* is the raw frequency of a given term inside a document. *idf* is the ratio of the total number of the document to the number of documents in which the index term appears. In our approach we adopt the cosine distance to measure the text semantic similarity between the two images. Assuming the text features of the two images are d_x and d_y. The cosine distance between d_x and d_y is shown in equation (2), where $w(d_x)$ denotes the weight of d_x.

$$s(d_x, d_y) = \frac{w(d_x).w(d_y)}{|w(d_x)||w(d_y)|} \tag{2}$$

For image visual feature, we first extract the bag-of-visual-word from each image. In order to use the bag-of-visual-word feature we have to first obtain the visual word vocabulary codebook. In our approach the *SIFT* based interesting point descriptors are used to construct the visual word vocabularies. Similar to the existing work [21][22], we train visual word vocabulary by clustering a large number of *SIFT* descriptors. Most of the existing clustering methods such as one-step K-means, affinity propagation [8] or some other visual vocabulary generation methods [14][19], could also be adopted. However, when the number of the descriptors becomes

bigger and bigger, these algorithms are in general less efficient in terms of either time or space complexity. In our approach we adopt the hierarchical K-means [22] to conduct the clustering, taking its advantages of high efficiency and the ability to organize the generated visual words in the vocabulary tree. With the hierarchical structure, finding the closest visual word for a local feature descriptor can be performed very efficiently. More details about the vocabulary tree and its applications can be found in [21]. After the local feature descriptors are clustered, a vocabulary tree is generated and the leaf nodes (cluster centers) are considered as the visual words. By searching hierarchically in the vocabulary tree, images in each retrieval result are represented as the bag-of-visual-word by replacing their *SIFT* descriptors with the indexes of the corresponding nearest visual words.

There are a lot of works using the bag-of-visual-word for image classification, retrieval, *etc*. However, a recent paper [7] points out that using the *pLSA* to further analyze these bag-of-word features will generate more effective feature for image classification, retrieval, *etc*. Therefore, after the bag-of-visual-word feature is obtained, we further analyze this feature to generate the more effective topic distribution feature to evaluate the similarity between two images. The basis of the *pLSA* is a model referred to as aspect model [11]. It assuming that there exists a set of hidden factors (topics) underlying the co-occurrences among two sets of objects (documents and words). It can reveal the latent topics connotated by the words. The advantages of *pLSA* consist of: 1) *pLSA* provides a probabilistic approach for the discovery of latent topics in the text data, which is flexible and has solid statistical foundation; 2) *pLSA* is an unsupervised approach, which can take advantage of any unlabeled data and build the aspect-based representation from it. *pLSA* uses Expectation-Maximization (*EM*) algorithm to estimate the probability values which measure the relationship between the hidden factors(topic) and the two sets of objects(document and word). In our approach, the two co-occurrence objects are visual words and images. The hidden factors are image categories. Let *d* represent the images, *t* represent the latent image categories, and *w* represent the visual words, then the *E-M* model training procedure is formulated as:

$$E\ step: P(t_k \mid d_i, w_j) = \frac{P(w_j \mid t_k)P(t_k \mid d_i)}{\sum_{l=1}^{K} P(w_j \mid t_l)p(t_l \mid d_i)} \tag{3}$$

$$M\ step:\ P(w_j \mid t_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j)P(t_k \mid d_i, w_j)}{\sum_{m=1}^{M}\sum_{i=1}^{N} n(d_i, w_m)P(t_k \mid d_i, w_m)} \tag{4}$$

$$P(t_k \mid d_i) = \sum_{j=1}^{M} n(d_i, w_j)P(t_k \mid d_i, w_j) / n(d_i)$$

where *N*, *M* and *K* denote the total number of images, visual words and latent image classes respectively. *P(t|d,w)* represents the posterior probability of topic *t* given image *d* and visual word *w*. *P(w|t)* is the generation probability of visual word *w* in the topic *t*. *P(t|d)* is the topic probability given image *d*. *n(d,w)* represents the frequency of visual word *w* in image *d* and *n(d)* is the total visual word number in the image *d*. At the beginning, *P(w|t)* and *P(t|d)* are initialized randomly with numbers between 0 to 1 and are normalized to sum to 1 along the row. In the *E*-step the posterior probabilities are computed for the topics based on the estimates of the two parameters *P(w|t)* and *P(t|d)*. And in the *M*-step these two parameters are updated. After the above probabilities are obtained, we use the topic distribution probability *P(t|d)* to represent the images.

We adopt the affinity propagation (*AP*) algorithm [8] to cluster the images. The *AP* algorithm makes use of information propagation between data points. It has high quality clustering capability and is very efficient. Moreover, it can automatically choose the right cluster number. These features make *AP* a suitable choice for our clustering. For AP the number of cluster is automatically decided. However, in our approach we prefer the final clustering number to be less than a threshold. Therefore, we adopt an iterative manner to cluster the images. The image clustering procedure is described as follows:

The iterative semantic image clustering

1. Set a cluster number threshold: T
2. Calculate the similarity $s(i,k)$ for all the image pairs (i, k), and set $r(i,k) = 0,\ a(k,i) = 0$
3. Responsibility and availability updates:
$$r(i,k) = s(i,k) - \max_{j:j\neq k}(a(i,j) + s(i,j))$$
$$a(k,k) = \sum_{j:j\neq k} \max\{0, r(j,k)\}$$
$$a(k,i) = \min(0, r(k,k) + \sum_{j:j\notin\{k,i\}} \max\{0, r(j,k)\})$$
4. Making assignments: $c_i^* = \arg\max_k \{r(i,k) + a(k,i)\}$
5. If the clustering number is bigger than T, then constitute the new image set by all the exemplars and go to step 2, else clustering over

In the algorithm, $r(i,k)$ represents the responsibility of cluster *k* for image i, $a(i,j)$ denotes the availability of image x_i as a candidate exemplar for image x_j. More details can be found in [8]. The similarity between two images *i* and *j* is measured by the linear fusion of the text and visual similarities, which is:

$$s(i,j) = \alpha s_T(i,j) + (1-\alpha)s_V(i,j) \tag{5}$$

where $s_T(i,j)$ is the text similarity between the image *i* and *j*, $s_V(i,j)$ is the visual similarity, and α is a tradeoff between the text and the visual similarity, with the value in [0,1].

3.2 Inter and Intra Semantic Image Cluster Ranking for Visual Cognition Representation

After the images are clustered, we obtain a few image clusters. In this subsection we propose to rank these clusters and the images in each cluster to mine the visual cognition viewpoint for each language community query. Assuming these clusters are $\bar{c} = \{c_1, c_2, ..., c_n\}$, and the initial query is Q. The inter cluster ranking function is defined as:

$$F = f(Q, \bar{c}) \tag{6}$$

In our approach, for clustering the following two properties are considered:

- **Relatedness**: the image cluster with closer relationship with the initial user query should have a higher rank.
- **Importance**: the image cluster with more images should have a higher rank.

We represent the relationship between the semantic image cluster and the initial user query as $r(c_i,Q)$, and the importance of the semantic cluster as $p(c_i)$. Then, the final inter cluster ranking score for semantic c_i is reformulated as:

$$F_{\mathrm{i}} = \beta p(c_i) + (1-\beta) r(c_i,Q) \tag{7}$$

where β is a tradeoff between $p(c_i)$ and $r(c_i,Q)$, with the value in [0,1].

In order to obtain the final inter semantic cluster ranking list, we have to estimate the value of $p(c_i)$ and $r(c_i,Q)$ for each semantic cluster. The values of $p(c_i)$ and $r(c_i,Q)$ are estimated as follows:

$$p(c_i) = \frac{N_i}{\sum_{j=1}^{n} N_j} \tag{8}$$

$$r(c_i,Q) = \frac{1}{N_{\mathrm{i}}} \sum_{j=1}^{m_i} P_j^i R_j^i \tag{9}$$

where N_i is the number of images in cluster i, R_j^i is the ranking value of the jth image in cluster i, P_j^i is the probability density of image j in cluster i, m_i represents the number of images in image semantic cluster *i, and n* is the number of clusters. $p(c_i)$ represents that the cluster with the larger number of image will get higher ranking score, which means that the bigger the cluster is, the more importance the cluster is. $r(c_i,Q)$ is the rank expectation for cluster c_i, and denotes the higher the rank expectation is, the higher the semantic cluster ranking score is. The key point in equation (9) is how to calculate the value of R_j^i and P_j^i. Assuming the rank of the jth image in cluster i in the image search result is k, and the rank of the image search result is $\{0,1,...,N\}$ .Then, the value of R_j^i is calculated as bellow:

$$R_j^i = 1 - \frac{k}{N} \tag{10}$$

The values of the P_j^i is estimated by the k nearest neighbor estimation, which is:

$$P_j^i = \frac{N}{\max_{\mathrm{g},h}(|\, k_g^N - k_h^N \,|)} \tag{11}$$

where $\max_{g,h}(|\, k_g^N - k_h^N \,|)$ represents the maximum rank distance of the N neighbors of image j in cluster i.

After the semantic image clusters are ranked, instead of using the exemplars in each cluster to represent the cognition viewpoint for each language community, we propose to use the reinforcement based algorithm [13] to rank the images in each cluster, and select the image with the biggest rank value to represent the semantic of that cluster. The reinforcement based ranking algorithm is described as bellow. Let $I_1^c, I_2^c, ..., I_{\mathrm{n}}^c$ represent the set of images in the cth cluster after clustering. We define the rank of image I_i^c as G_i^c, which is the solution to the equation:

$$G_i^c = \sum_{j=1}^{n} s_{ij} G_j^c \tag{12}$$

where s_{ij} represents the similarity between image i and j. The above equation can be solved iteratively by using the method described bellow.

The reinforce based representative cluster image selection

1. Initialize $\vec{r^0} = (r_1^0, r_2^0, ..., r_N^0)$ for the images randomly such that $\sum_{i=1}^{N} r_i^0 = 1$ and $r_i^0 > 0 \ \forall i$.
2. Set $t = 1$.
3. $r_i^t = \sum_{j=1}^{N} s_{ij} r_j^{t-1} \ \ \forall i \in 1,...,n.$
4. $r_i^t = r_i^t / \| \vec{r_i^t} \|$, $\| \vec{r_i^t} \| = \sum_{i=1}^{N} r_i^t$
5. $t = t + 1$.
6. Repeat steps 3 to 5 till convergence.
7. Select the image with the biggest ranking score to represent the cluster.

More discussion of the reinforce algorithm can be found in [13].

3.3 Visual Cognition Comparison between Different Language Communities

In this subsection, we mainly address the problem of how to compare the visual cognition among the communities. After image clustering, if we define each image cluster as a concept, the retrieved images for a specific language community could be naturally represented as the concept distribution feature like in Figure 2. Assuming there are m language communities $\{L_1,L_2,\ldots,L_m\}$, K semantic clusters $\{1,2,\ldots,K\}$, and the total number of the image for language community L_i is N_i and the number of image for language community L_i in cluster j is N_i^j. Then, the concept distribution feature for language community L_i can be represented as $\{\frac{N_i^1}{N_i}, \frac{N_i^2}{N_i}, ..., \frac{N_i^K}{N_i}\}$. In this paper, we adopt the *Kullback-Leibler* (*KL*) divergence [15] to calculate the similarity between the language community L_i and L_j, which is:

$$KL(L_i,L_j) = \sum_{k=1}^{K} \frac{N_i^k}{N_i} \log\left(\frac{N_i^k}{N_i} \times \frac{N_j}{N_j^k}\right) \tag{13}$$

From function (13) we can see that the *KL* divergence is asymmetric measure, i.e. $KL(L_i,L_j)$ is not equal to $KL(L_j,L_i)$. However, in our application we prefer the similarity measure to be symmetric. Therefore, instead of using the original version of *KL* divergence, we adopt the symmetric version which is calculated as:

$$A\_KL(L_i,L_j) = KL(L_i,L_j) + KL(L_{\mathrm{j}},L_i) \tag{14}$$

where the value of $KL(L_i,L_j)$ and $KL(L_j,L_i)$ are calculated as in equation (13). For visual cognition comparison, another condition should be considered. If the distribution over the clusters is similar for the two datasets, while the expectation ranking order of these clusters in the two sets is rather different, the cognition for these two communities should be different. Therefore, the visual cognition distance between two communities is calculated as:

$$Cd(L_i, L_j) = \sum_{k=1}^{K} e^{\|r_{L_i}(c_k,Q) - r_{L_j}(c_k,Q)\|} \times \left(\frac{N_i^k}{N_i} - \frac{N_j^k}{N_j}\right) \log\left(\frac{N_i^k}{N_i} \times \frac{N_j}{N_j^k}\right) \quad (15)$$

where $r_{L_i}(c_k, Q)$ is calculated as in equation (9). After the similarities between these language communities are obtained, we use them to visualize the relationship between the different language communities about a specific concept.

4. EXPERIMENTAL RESULTS

In this section, extensive experiments and evaluations are conducted, including visual cognition representation evaluation, and visual cognition relationship evaluation. The results will demonstrate the effectiveness of the proposed method.

4.1 Data Collection and Parameter Setting

To collect the images which can reflect the real visual cognition of the multi-language communities, we adopt the *Google image* as our tool. We select *Google image* [2] for the reasons that *Google image* is the most popular image search engine cross many language communities available at hand, and the number of images indexed by *Google* is relatively bigger than other image search engines. Many of the applications for image processing in the multimedia domain employ the *Flicker* [6] as their image database, where the images are uploaded by different users with high quality. However, the Flicker data is not suitable for our application. As most of the images in *Flicker* are tagged with only English word and they mainly reflect the visual cognition of the English language community. For *Google image*, after retrieving only the top1000 images could be assessed, which contain all kinds of image. With the assumption that the top retrieved images can reflect the main visual cognition of the community, for each query the top 500 images with good quality are retrieved.

Table 1: The queries used in the experiments

Initial query	
Apple	Tiger
Paris	Great wall china
Car	Beijing 2008

Table 2: The query languages used in the experiments

Query language		
Chinese	English	Russian
Japanese	Spanish	French
German	Arabic	Persian

To facilitate the visual cognition representation and comparison evaluation, we select 6 popular queries, which are shown in Table 1. These queries belong to different types, such as scene and object. For the language communities, 9 representative language communities are selected based on their influence on the world, which is shown in Table 2. When the user submitted query is obtained, it is translated into these languages and fed into the corresponding engines, such as the *image.google.com*, *etc*, to retrieve the images from the corresponding communities.

In the experiment, the parameter α in equation (5) and the parameter β in equation (7) are both set as 0.5. For image clustering the cluster number threshold is set as 30. For the k nearest density estimation in equation (11) the number of neighborhood is set as 10. Similar to [7] about 1500 visual words are generated, and for *pLSA* the latent topic number is set as 25.

4.2 Experimental Result for Visual Cognition Representation Evaluation

In this subsection, we use the proposed method above to mine the visual cognition representation of the six queries for each community. The final visual cognition representation results are shown in Figure 7(which lies in the last page), where the top five semantic representative cluster images are shown for each group. In Figure 7, some interesting patterns should be noted. The first one is the representation for "*apple*". There are mainly four meanings as been defined in Figure 2. It is quite interesting to note that the visual cognition for Chinese is from *fruit apple* to *iphone* and *notebook*, while the visual cognition for English community is just the opposite from *notebook* to *iphone* and f*ruit apple*. For the Japanese community, there are no visual appearances about the animal Tiger. This may partly due to that there are no wild tiger living in Japan ever before, and the concept of animal tiger is not deeply in mind. Also there are some unique visual patterns in some of the communities, which are marked by the red little circles, like the *Great wall wines* in the Chinese community, apple (Germany: *Iris Apfel*) in the German community, and the first tiger image in Arabic community, *etc*.

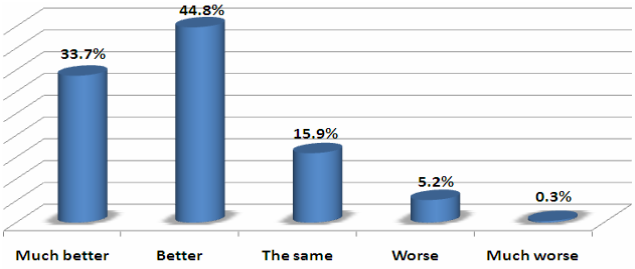


Figure 8. The cognition representation results evaluation.

In order to evaluate the cognition representation quantitatively, a subjective user study is conducted. Totally, 12 users, 6 males and 6 females aging from 23 to 32, are invited to take part in the study. These users are familiar with the image search engine and use *Google image* every day. To avoid any bias on the evaluation, they have no knowledge about the visual mining approach, but just to give scores according to the task definition. We ask each user to give score by comparing each representation result with the top 5 images in the initial search result. The comparison is based on the data coverage of these two group images, which means the more meanings of one concept the 5 images covers, the better the result is. Five scores are defined: much better, better, the same, worse, much worse than the original top 5 images. The final evaluation result is shown in Figure 8, which is the average evaluation ratio of the total community queries. From Figure 8 we can see that on average 33.7% of the evaluations are much better than the original result and are better to the remaining 44.8% evaluations. This result shows the proposed method can effectively mine out the visual cognition for each community.

4.3 Visual Cognition Comparison across Multi-language Communities

In this subsection, we evaluate the visual cognition on two aspects: the visual cognition compactness of a specific community, and the cognition relationships between different language communities. Assuming the communities are $\{L_1,L_2,\ldots,L_m\}$, the images retrieved for query x in community i are $x_1^i, x_2^i, \ldots, x_n^i$, and the features are $X_1^i, X_2^i, \ldots, X_n^i$. The visual cognition compactness for query x of the community i is calculated as:

$$Compact(i,x) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\cos ine(X_j^i, X_k^i) \qquad (16)$$

The bigger the compactness value is, the more compact the cognition is. The compactness results are shown in Figure 9. It can be observed that the concept "*apple"* almost has the lowest compactness score over the communities. In our view, this is partly due to the diverse meanings of the query "*apple*", and partly may due to the fact that the features used in the paper may not well represent the object *apples*. The second lowest is the query "*Beijing 2008*". This query represents a large concept, and there are no corresponding objects. It can also be seen that the compactness of the query "*tiger*" for Japanese community is quite lower than other communities. This result corresponds to the explanation in section 4.2 that there are no unified visual cognition for animal tiger in the Japanese community.

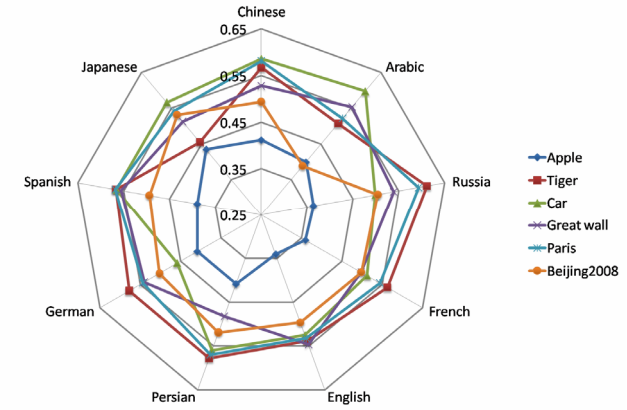


Figure 9. The visual cognition compactness comparison between the nine communities for the six queries.

After the compactness of these community queries are evaluated, we further evaluate the visual cognition relationship between different language communities for the same query concept. After query related images are collected from these communities, they are clustered into different clusters. Then, we use the equation (15) to evaluate the visual cognition difference between two communities. The final comparison results for the six queries among the nine communities are visualized in Figure 10, where the relationships between the communities are visualized through the grid intensities. The darker the grid is, the more similarity the two communities are. In order to show the relationships more clearly, in Figure 10 the grid intensities are the rescaled values of the cognition similarities.

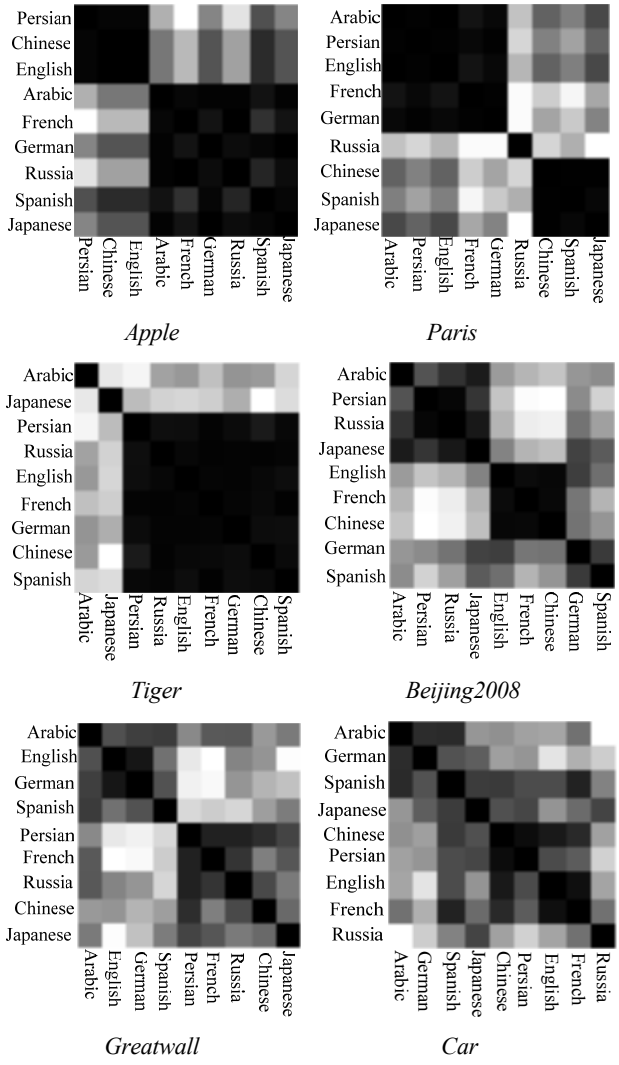


Figure 10. The visual cognition comparison visualization results for the six queries.

From Figure 10 we can see that the relationships between these communities are quite clear. For query "*apple*" the communities can be classified into two classes. Persian, Chinese and English consist of one class, while the other communities consist of another class. After checking the community images we find the difference between the two classes is that there are nearly no iphone and notebook images in the bigger class, which can also be seen from the visual representation results in Figure 7. From the results, we can see it has some conflict with the observation in Figure 2 that Arabic and Chinese are similar. It is because for the total 500 images, there are nearly no iphone and notebook images in the Arabic community, while for Chinese there are lot of these images. For the query "*Paris*", there are three classes. Arabic, Persian, English, French, and German consist of the first class, Russia itself consists of the second class, while Chinese, Spanish, and Japanese consist of the third class. The first class contains the figure of *Paris Hilton*, while the other classes not. The reason why Russia itself consists of a class may be that, except the most popular locations there are several map images in the result, which

discriminate it from others. Besides, other queries also show very interesting structure patterns.

A subjective user study is also conducted to show the effectiveness of the relationship mining result. All the users described in section 4.2 take part in the evaluation. They are asked to give score on whether the relation results clearly reflect the true relation between the community data. Three scores are defined: good, neutral and bad. The evaluation result is shown in Figure 10, from which we can see that on average 72% of the evaluators think the result is good and valuable.

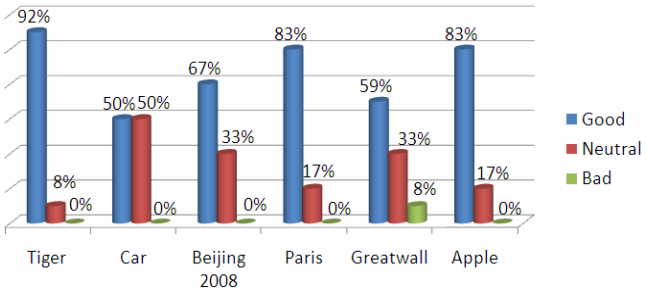


Figure 10. The cognition relation mining results evalua-

5. CONCLUSION AND DISSCUSSION

In this paper, we propose a novel approach to mine and compare the visual cognition across multi-language communities for the same concept. We call it "*the third eye*", which means that through this way people from different language communities can better understand each other to avoid culture conflict. The two problems addressed are the visual cognition representation for a specific language community query, and the cognition comparison between different language communities. The image data is collected through *Google image* with the assumption that the image distribution in the search result can reflect the visual cognition of the community. The different community data are obtained from the different portion of the image search engine by using the corresponding language query. After that, the images are clustered. Through inter and intra semantic clustering ranking, the visual cognition of a specific community is represented by the typical images of the clusters. The cognition difference among the language communities is achieved by comparing the image, which is retrieved by the search engine, distribution over these semantic clusters. The experimental results show that the proposed approach is effective.

It is natural for us to apply the cognition representation for visual query suggestion and image re-ranking. Compared with the work in [25], our suggestion will be more personalized in that the suggested images for different language communities are decided according to their different visual cognition. Another interesting issue may be that how to recommend proper advertising according to the cognition context [20]. It is also very interesting to recommend images to the user according to their difference to the user's cognition, and let the user have change to see the cognition of the people with different backgrounds. It is valuable to conduct a subjective user study by asking people in each community to vote about the concepts to obtain the people's real cognition about the queries and to compare it with the cognition representation in our approach. However, it is hard to find lots of people from each community in a short time, thus this comparison is still pending. In the future, we will try to finish this task.

6. ACKNOWLEDGEMENT

This work was supported in part by National Natural Science Foundation of China: 60773136 and 60833006, in part by National Basic Research Program of China (973 Program): 2009CB320906, and in part by Beijing Natural Science Foundation: 4092042.

7. REFERENCES

[1] http://www.wikipedia.org/
[2] Google image search: http://images.google.com/
[3] Microsoft bing image search http://www.bing.com/?scope=images.
[4] Yahoo! image search: http://images.search.yahoo.com/.
[5] Ask image search: http://www.ask.com/?tool=img.
[6] Flickr: http://www.flickr.com/ .
[7] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pLSA. ECCV, pages 517-530, 2006.
[8] B. J. Frey and D. Dueck. Clustering by passing messages between data points. Science, 16(315): 972-976, 2007.
[9] B. Gao, T. Y. Liu, T. Qin, X. Zhang, Q.-S. Cheng, and W.-Y. Ma. Web image clustering by consistent utilization of visual features and surrounding texts. ACM Multimedia, 2005.
[10] R. Hong, J. Tang, Z.-J. Zha, Z. Luo, T.-S. Chua. Mediapedia: mining web knowledge to construct multimedia encyclopedia. MMM, pages 556-566, 2010.
[11] T. Hofmann and J. Puzicha. Unsupervised learning from dyadic data. Technical report, UC, Berkeley, 1998.
[12] F. Jing, C. H. Wang, Y. H. Yao, K. F. Deng, L. Zhang, and W. Y. Ma. IGroup: web image search results clustering. ACM Multimedia, pages 377-384, 2006.
[13] D. Joshi, J. Z. Wang, and J. Li. The story picturing engine: finding elite images to illustrate a story using mutual reinforcement. International Multimedia Conference, Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval, 2004.
[14] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. ICCV, pages 17-21, 2005.
[15] S. Kullback and R. A. Leibler. On information and sufficiency. Annals of Mathematical Statistics, 22(1): 79-86, 1951.
[16] P. Lawrence, B. Sergey, M. Rajeev and W. Terry. The PageRank citation ranking: bring order to the web. Technical report, Standford infolab, 1999.
[17] Y. Liu, T. Mei, X.-S. Hua. CrowdReranking: exploring multiple search engines for visual search reranking. SIGIR, pages 500-507, 2009.
[18] W.-H. Lin, A. G. Hauptmann. Identifying news videos' ideological perspectives using emphatic patterns of visual concepts. ACM Multimedia, pages 261-270, 2009.
[19] S. Lazebnik and M. Raginshy. Supervised learning of quantizer codebook by information loss minimization. PAMI, 31(7): 1294-1309, 2009.
[20] T. Mei, X.-S. Hua, S. Li. Contextual in-image advertising. ACM Multimedia, pages 439-448, 2008.
[21] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. CVPR, pages 2161-2168, 2006.
[22] S. L. Zhang, Q. Tian, G. Hua, Q. M. Huang and S. P. Li. Descriptive visual words and visual phrases for image applications. ACM Multimedia, pages 75-84, 2009.
[23] S. Simon. Basic psychological mechanisms: neurosis and projection. The Heretical Press, 2008.

[24] X. J. Wang, W. Y. Ma, Q. C. He and X. Li. Grouping web image search result. ACM Multimedia, pages 436-439, 2004.

[25] Z.-J. Zha, L. Yang, T. Mei, M. Wang, Z. Wang. Visual query suggestion. ACM Multimedia, pages 15-24, 2009.

Figure 7. The visual cognition mining result representation for the six queries "*Paris*", "*Beijing 2008*", "*Greatwall China*", "*Tiger*", "*Apple*", and "*Car*" from the nine language communities, which are *Arabic, Chinese, Persian, French, German, Russian, English, Spanish,* and *Japanese.*