Flickr Distance *

Lei Wu
MOE-MS Key Lab of MCC
University of Science and Technology of China
leiwu@live.com

Xian-Sheng Hua
Microsoft Research Asia
49 Zhichun Road
Beijing 100190, China
xshua@microsoft.com

Nenghai Yu
MOE-MS Key Lab of MCC
University of Science and Technology of China
ynh@ustc.edu.cn

Wei-Ying Ma
Microsoft Research Asia
49 Zhichun Road,Beijing
100190, China
wyma@microsoft.com

Shipeng Li
Microsoft Research Asia
49 Zhichun Road,Beijing
100190, China
shipeng.li@microsoft.com

ABSTRACT

This paper presents Flickr distance, which is a novel measurement of the relationship between semantic concepts (objects, scenes) in visual domain. For each concept, a collection of images are obtained from Flickr, based on which the improved latent topic based visual language model is built to capture the visual characteristic of this concept. Then Flickr distance between different concepts is measured by the square root of Jensen-Shannon (JS) divergence between the corresponding visual language models. Comparing with WordNet, Flickr distance is able to handle far more concepts existing on the Web, and it can scale up with the increase of concept vocabularies. Comparing with Google distance, which is generated in textual domain, Flickr distance is more precise for visual domain concepts, as it captures the visual relationship between the concepts instead of their co-occurrence in text search results. Besides, unlike Google distance, Flickr distance satisfies triangular inequality, which makes it a more reasonable distance metric. Both subjective user study and objective evaluation show that Flickr distance is more coherent to human perception than Google distance. We also design several application scenarios, such as concept clustering and image annotation, to demonstrate the effectiveness of this proposed distance in image related applications.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing-indexing methods; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Perceptual reasoning*

General Terms

Algorithms, Theory, Experimentation.

Keywords

Flickr distance, Visual concept net, visual distance, TagNet, concept relationship, similarity measurement

1. INTRODUCTION

Exploring the Semantic relationship between concepts is a hot research topic recently, since it has wide application on natural language processing, object detection, and multimedia retrieval [6][23][13]. It is important to note that the semantic relationship is more than synonym (football-soccer) and concept similarity (horse-donkey). It also includes relationships such as meronymy (car-wheel) and concurrence (airplane-airport). Here concurrence denotes the two concepts may appear simultaneously in daily life rather than in the textual documents. This semantic relationship connects concepts into a network like human society. Some concepts are more closely related, such as "airplane" and "airport", and some are more alienated, i.e. "acropolis" and "alcohol". However, as the concept relationship is in fact a kind of knowledge based on human cognition, it is not easy to model this relationship.

Some long-term and labor-intensive works, such as the Cyc project [12] and the WordNet project [16], tried to create a relationship network among the common concepts. Great efforts were made to design a structure for the manipulation of knowledge about concept relationship as well as filling up the structures by well trained human experts. Although it costs great efforts to enlarge the scale of the concept network, this concept database is still limited and difficult to update comparing to the overall information on the Web.

Later on, Cilibrasi and Vitányi proposed the Normalized Google similarity Distance (NGD) [5] by exploring the textual information available on the Web. The distance between two concepts is measured by the Google page counts when querying both concept names to the Google search engine. It costs little human effort and can cover far more concepts based on the information from the Web. However, it assumes the concept relationship only depends on the co-occurrence of these concepts in the textual documents on

* *This work was performed at Microsoft Research Asia.*

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.

the Web. This assumption is a little bit simple, and cannot cover the cases of meronymy and concurrence. Here we use "concurrence" to represent the concept co-occurrence or background coherence in visual domain, and use "similarity" to represent the concept co-occurrence in textual documents. In general cases two concepts with meronymy or concurrence relationship may not be described simultaneously in textual documents, and thus their relationship may not be accurately captured by NGD. For example, the "airport" and "airplane" are concurrent, but they may less likely to co-occur in the same textual web page. It is the same with "car" and "wheel". Table 1 shows the Google distance between some concepts. NGD for "airplane–dog" pair is 0.2562. Both NGD for concept pairs "airport–airplane" and for "Car–wheel" are larger than this distance, which is somewhat against the human cognition.

Table 1: Illustration of Google distance.

Concept pair	Google distance
Airplane–dog	0.2562
Football–soccer	0.1905
Horse–donkey	0.2147
Airplane–airport	0.3094
Car–wheel	0.3146

As the relationship between concepts is the knowledge of human perception, and 80% of human cognition comes from visual information, it is more reasonable to generate artificial knowledge about concept relationship by visual correlation rather than by concept co-occurrence in textual documents. While in visual domain, the direct count of the number of returned images to measure visual similarity, does not make sense, since there are many noisy tags that do not describe the semantic meaning of the images. Besides, the correlation in visual domain is not only represented by the occurrence of the concepts, but also their spatial information in the image. That is to say, in order to generate actual concept relationship in visual domain, content analysis is unavoidable.

In this paper, we propose the Flickr distance to measure the semantic correlation between two concepts based on the Flickr photo database, which properly records the concept correlation in daily life. This Flickr distance is calculated by measuring the square root of Jensen-Shannon (JS) divergence between the visual language models corresponding to the concepts. If two concepts are more likely to appear in the same photo, the square root of JS divergence of their visual language models tends to be small; otherwise large. To provide an unbiased estimation of this distance, there are two requirements. First, it requires a sufficiently large and unbiased image dataset, which contains the connection information between images and concepts. This requirement is easily met by taking the Flickr photo collection as the image pool. There are a huge amount of images in Flickr and millions of new photos are uploaded everyday by numerous independent users. To avoid noisy tags, we only use the top 1,000 returned images to represent each query concept. Although the bias is unavoidable for any image database, sampling from this large scale image pool is one of the most reasonable choices. Associated with these photos, great amount of tags are made by users to connect these images with concepts. Thus it is easy to obtain a large collection of images about the targeting concept on Flickr. Comparing with Google/Yahoo! image search results, Flickr data contains less noise and more daily life photos. Second, it requires an efficient visual characteristic modeling method to represent the content of the images. There are lots of effective models in computer vision literature, such as bag of words model [7], regions of interest (ROI) based models [4][14], however, they are unable to handle the large scale database. Visual language modeling (VLM) [21] is an efficient and effective visual concept modeling method for large scale dataset. It captures the statistical semantics of the images by analyzing the spatial dependency between neighboring image patches. Since the statistical semantics in textual domain is defined by words frequency as well as their order of recurrence, in visual domain it is natural to represent it as the image local features and their spatial dependency. Furthermore, we incorporate the latent semantic analysis into VLM to capture the concept appearance variation. Then the distance of the two concepts is easily calculated by the square root of Jensen-Shannon (JS) divergence betweem their VLMs.

The rest of the paper is organized as follows. Section 2 gives more details about the related work. Section 3 elaborates on the Flickr distance. Section 4 discusses the generation of visual concept network (VCNet) based on Flickr distance. In section 5 we demonstrates that Flickr distance is more coherent to human cognition by both subjective and objective experiments. We also provide some application scenarios of Flickr distance and the associated VCNet. Section 6 concludes the paper.

2. RELATED WORK

2.1 WordNet

WordNet developed by the Cognitive Science Laboratory of Princeton University is a large semantic lexicon for English language. It groups English words into sets and builds various semantic relations between these synonym sets by well-trained experts. The purpose of the work is both to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications. Psychological experiments suggests that human store semantic information in a way that is much like WordNet.

The WordNet database contains about 150,000 words organized in over 115,000 synonym sets for a total of 207,000 word-sense pairs. This dataset is relatively limited comparing to the concept on World-wide-web. For example, in the well-known photo sharing website Flickr, there are more than 130,000,000 tags constructing almost 60,000,000 concepts, and this number is still increasing every day. With the manual effort, the WordNet semantic lexicon can never catch up with the growing of web concepts.

2.2 Google Distance

Google distance is proposed to calculate the relationship between two concepts by their correlation in the search result from Google search engine when querying both concepts. It assumes that the words and phrases acquire meaning from the way they are used in society. Since Google has indexed a vast number of web pages, and the common search term occurs in millions of web pages, this database can somewhat reflect the term distribution in society. Thus the Normal-

ized Google distance (NGD) is defined to approximate the semantic relations governing the search terms by calculating the correlation from Google search results.

$$NGD(x,y) = \frac{max(\log f(x), \log f(y)) - \log f(x,y)}{\log N - min(\log f(x), \log f(y))} \quad (1)$$

where $NGD(x,y)$ represents the Normalized Google distance between concepts x and y. $f(x)$,$f(y)$, and $f(x,y)$ denotes the number of pages containing x, y, both x and y, separately. N is the total number of web pages indexed by Google.

This equation indicates that NGD can only capture certain concept relationship, i.e. synonym and similarity, when the concepts are frequently co-occurred in textual document or web page, while this method is difficult to capture the concept relationship like meronymy and concurrence in daily life. This makes sense, since Google distance is proposed for textual domain applications, i.e. machine translation, author clustering. However, in the multimedia fields, concept relations of meronymy and concurrence play more important role. Thus it is not quite reasonable to directly apply Google distance to multimedia field.

3. FLICKR DISTANCE

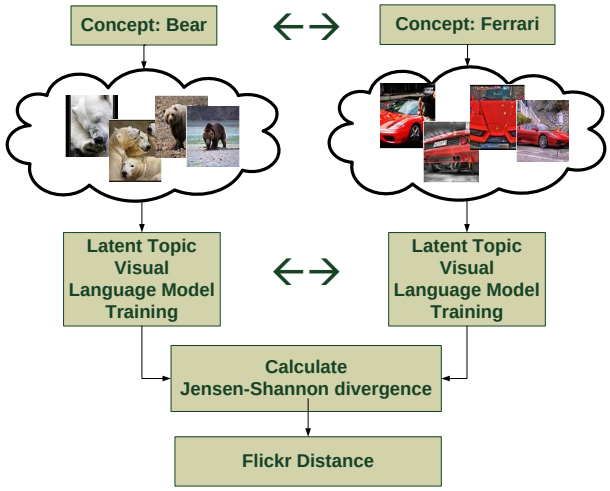


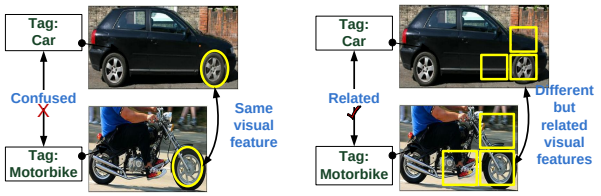
Figure 1: Illustration of the Flickr distance calculation

3.1 Overview of Flickr Distance

Given two different concept names, the algorithm tries to calculate the semantic distance between them. As afore mentioned, semantic distance is generated by human cognition and 80% of the human cognition comes from visual information, it makes sense to measure the semantic distance between concepts by their concurrence in daily life, i.e. objects co-occurrence or the coherence of background.

To simulate concept concurrence in daily life, the concept relation learning process should be performed in daily life environment. Similar to the human observation system, the digital cameras in the world are recording the realistic daily life every day. Statistical semantic relations between concepts can be mined from a large pool of the daily life photos. To achieve a less biased estimation of the statistical concept relations, the image pool should be very large and the source of the images should be independent. Luckily, the on-line photo sharing website Flickr meets both conditions. It has collected more than 10^9 photos, and these photos are uploaded by independent users. Besides, there are also large amount of manually labeled tags, which connect the photos to the semantic concepts. Thus it is an ideal dataset for learning the concept semantic relations.

To analyze the concept correlation in the large Flickr photo pool, visual language model (VLM), an efficient visual statistical analysis method, is adopted. Though other models may be also applied, there are three reasons for choosing VLM to represent the concepts. Firstly, VLM is more discriminative than the well-known bag-of-words (BOW) model. Superior to BOW in computer vision literature, VLM captures not only local appearance features but also their spatial dependence, which is more discriminative in characterizing the concept than the pure visual feature distribution, as illustrated in Figure 2. In Figure 2(a), the same visual feature has frequently co-occurred in "car" and "motorbike" images. Due to the ignorance of spatial information between visual features, these two concepts are more likely to be confused. Figure 2(b) considers the neighboring information of these visual features and the relation between the concepts is more clear. This shows that the arrangement of the visual features is also informative in representing the concept. VLM has also been demonstrated an effective content modeling method in image classification task [21]. Secondly, the training of VLM is fast, which makes the modeling method especially suitable for large scale concept dataset. Finally, the output of VLM is conditional distribution of visual features, based on which a strict distance metric can be easily defined. Thirdly, VLM can depress the noise. The images that actually contain the target concept will share some visual features, which actually contribute to the model. Visual features in noisy samples, which are tagged by mistake, have little influence on the final VLM.



(a) model without spatial information (b) model with spatial information

Figure 2: Illustration of different concept models. VLM modeling is more discriminative than the BOW modeling.

In order to measure the distance, the square root of Jensen-Shannon (JS) divergence between the VLMs is calculated. Both JS and Kullback-Leibler (KL) divergence are the commonly used similarity measurement between two distributions. KL divergence is unsymmetrical and is not a strict metric. JS divergence is demonstrated symmetric and satisfies the triangle inequality. It is also shown that the square root of the Jensen-Shannon divergence is a metric. Since we aim to define a distance metric between concepts, using the square root of JS divergence is more appropriate.

3.2 Concept Modeling by VLM

For each concept, we can easily extract a collection of images from Flickr with the help of user created tags. We assume that images about the same concept share similar appearance features as well as their arrangement patterns, which form the statistical semantic of the concept. For reasons discussed in Section 3.1, we mainly model the concept based on the framework of visual language model (VLM) [21]. It divides the image into equal-sized patches, and calculates the conditional dependence between these patches to capture the visual statistical semantics.

However, the original VLM cannot handle the concept appearance variation, i.e. close/perspective shots, side/front views. In this paper, we further incorporate the latent topic analysis into the VLM, and assume each appearance variation of a concept corresponds to a latent topic z_i. Then pLSA is adopted to analyze the visual characteristics of the concept under each variation. Thus we can model a concept more precisely by latent topic VLMs. In the following, we will discuss the details about the concept modeling process in two phases: feature extraction and latent topic VLM generation.

3.2.1 Feature Extraction

An image is divided into uniformly sampled equal-sized patches, since the uniform sampling requires little computational cost and its performance is comparable to the methods with the salient detection or segmentation based local regions [7]. For each patch, we use an 8-dimensional texture histogram to describe it. Each dimension corresponds to the texture gradient along one of eight quantized directions. The calculation of the texture histogram is the same as previous work on similarity search [22]. We use Bin's hash coding scheme [19] to quantize the texture histogram of a patch to visual word w_{xy}.

3.2.2 Generate Latent Topic VLM

This section incorporates the latent topic (appearance variation) analysis into the VLM [21] to characterize each concept C_i from the low level visual features arrangement. It provides an efficient way to model the concept. Each VLM is presented in the form of the conditional distributions, which describe the spatial dependence between the low level visual features given its neighbors and the latent topic.

The visual language model is sub-categorized into unigram, bigram and trigram models, according to the number of neighboring visual words considered. The unigram model assumes that the visual words are independent of each other. The model actually captures the visual word distribution. The bigram model assumes that the visual word is dependent on one of its neighboring features, i.e. the nearest left neighbor. This model calculates the conditional probability of each visual word given one of its neighboring words. The trigram model further assumes that the visual word is dependent on two of its neighboring words, i.e. the nearest left and nearest up neighbors. Theoretically, this dependency assumption can be extended to higher order models (ngram model, n>3). However, as the order of the model increases, the number of parameters will increase exponentially. Since the parameters are estimated from the occurrence of n-gram in the training set, if the order n is too large, the comparatively limited training set will suffer the sparseness problem. There is a tradeoff between the discrimination and sparseness. In this paper, we choose the trigram model to capture the concepts.

Considering the variation of each concept, we further propose the latent topic VLM to estimate the visual feature conditional distribution given each latent topic. Take the trigram model as an example. The original trigram model is to estimate the conditional distribution $P(w_{xy}|w^2_{x-1,y}, C)$, where C is the concept, and $w^2_{x-1,y}$ represents the bigram $w_{x-1,y}w_{x,y-1}$. Since the visual concept may have various appearances, i.e. close/perspective shots, side/front views, using a single model to represent a concept is not accurate. We further introduce a latent variable z to represent the concept variation. Since this variable is hidden, the probabilistic latent semantic analysis (pLSA) [8] is incorporated into VLM to model the concept under each variation.

The latent topic VLM further estimates $P(w_{xy}|w^2_{x-1,y}, z^C_k)$, where z^C_k represents the k^{th} appearance variation of concept C. This latent topic trigram modeling process can be formulated as follows.

$$P(w_{xy}|w^2_{x-1,y}, d_j) = \sum_{k=1}^{K} P(w_{xy}|w^2_{x-1,y}, z^C_k)P(z^C_k|d_j) \quad x = 1, \cdots, m;\ y = 1, \cdots, n;\ j = 1, \cdots, N. \tag{2}$$

where d^C_j represents the j^{th} image in concept C. z^C_k is the k^{th} latent topic in concept C. K is the total number of latent topics, which is determined by experiment. EM algorithm is adopted to estimate both parameters $P(w_{xy}|w^2_{x-1,y}, z^C_k)$ and $P(z^C_k|d_j)$. The object function of the EM algorithm is to maximize the joint distribution of concept and its visual word arrangement A_w.

$$maximize \quad p(A_w, C) \tag{3}$$

$$p(A_w, C) = \prod_{d_j \in C} \prod_{x,y} P(w_{xy}|w_{x-1,y}w_{x,y-1}, d_j) \tag{4}$$

In order to obtain analytically tractable density estimation, we propose a cross-updating scheme, in which we simultaneously estimate $P(w^3_{xy}|z^C_k)$ and $P(w^2_{x-1,y}|z^C_k)$. Then we calculate $P(w_{xy}|w^2_{x-1,y}, z^C_k)$ by these two estimations (Eq. (11)). The E step and M step are performed as follows.

E step:

$$Q_2(z^C_k|d^C_j, w^2_{x-1,y}) \leftarrow P(z^C_k|d^C_j)P(w^2_{x-1,y}|z^C_k) \tag{5}$$

$$Q_3(z^C_k|d^C_j, w^3_{xy}) \leftarrow P(z^C_k|d^C_j)P(w^3_{xy}|z^C_k) \tag{6}$$

$$Q(z^C_k|d^C_j, w^3_{xy}) \leftarrow P(z^C_k|d^C_j)P(w_{xy}|w^2_{x-1,y}, z^C_k) \tag{7}$$

M step:

$$P(w^2_{x-1,y}|z^C_k) \leftarrow \frac{\sum_j n(d^C_j, w^2_{x-1,y})Q_2(z^C_k|d^C_j, w^2_{x-1,y})}{\sum_{x,y,j} n(d^C_j, w^2_{x-1,y})Q_2(z^C_k|d^C_j, w^2_{x-1,y})} \tag{8}$$

$$P(w^3_{xy}|z^C_k) \leftarrow \frac{\sum_j n(d^C_j, w^3_{xy})Q_3(z^C_k|d^C_j, w^3_{xy})}{\sum_{x,y,j} n(d^C_j, w^3_{xy})Q_3(z^C_k|d^C_j, w^3_{xy})} \tag{9}$$

$$P(z^C_k|d^C_j) \leftarrow \frac{\sum_{x,y} n(d^C_j, w^3_{xy})Q(z^C_k|d^C_j, w^3_{xy})}{\sum_{x,y,k} n(d^C_j, w^3_{xy})Q(z^C_k|d^C_j, w^3_{xy})} \tag{10}$$

$$P(w_{xy}|w^2_{x-1,y}, z^C_k) \leftarrow \frac{P(w^3_{xy}|z^C_k)}{P(w^2_{x-1,y}|z^C_k)} \quad (11)$$

$$P(z^C_k|C) \leftarrow \sum_{d^C \in C} P(z^C_k|d^C, C)P(d^C|C) \quad (12)$$

The outputs are the conditional distributions of trigrams for each latent topics, $P(w_{xy}|w^2_{x-1,y}, z^C_k)$, $k = 1, \cdots, K$.

3.3 Concept Distance Measurement

In this section, we aim to define a reasonable distance to measure the relationship between the concepts. Each concept corresponds to a visual language model, which consists of the trigram conditional distributions under different latent topics. Kullback-Leibler divergence (KL divergence) is a common measurement of the difference between two probability distributions. However, as it does not meet the constraint of triangular inequality, it is in fact not a strict metric. Based on KL divergence a more strict metric JS divergence is defined. This divergence is symmetric and its square root is demonstrated a strict metric. The Flickr distance is defined as the average square root of the JS divergence between the latent topic VLMs.

Let $P_{z_i^{C_1}}$ and $P_{z_j^{C_2}}$ be the trigram distributions under latent topic $z_i^{C_1}$ and $z_j^{C_2}$ respectively. $z_i^{C_1}$ represents the i^{th} latent topic of concept C_1. The K-L divergence between them is defined to be

$$D_{KL}(P_{z_i^{C_1}}|P_{z_j^{C_2}}) = \sum_l P_{z_i^{C_1}}(l) \log \frac{P_{z_i^{C_1}}(l)}{P_{z_j^{C_2}}(l)} \quad (13)$$

where $P_{z_i^{C_1}}(l)$,$P_{z_j^{C_2}}(l)$ correspond to the probability density of the l^{th} trigram in these two distributions respectively. In the view of information theory, the KL divergency is in fact a measurement of the mutual entropy between the two visual language models.

$$\begin{aligned} &D_{KL}(P_{z_i^{C_1}}|P_{z_j^{C_2}}) \\ &= -\sum_l P_{z_i^{C_1}}(l) \log P_{z_j^{C_2}}(l) + \sum_l P_{z_i^{C_1}}(l) \log P_{z_i^{C_1}}(l) \\ &= H(P_{z_i^{C_1}}, P_{z_j^{C_2}}) - H(P_{z_i^{C_1}}) \end{aligned} \quad (14)$$

where $H(P_{z_i^{C_1}}, P_{z_j^{C_2}})$ is the cross entropy of the two distributions, and $H(P_{z_i^{C_1}})$ is the entropy of $P_{z_i^{C_1}}$. According to the Gibbs' inequality, $D_{KL}(P_{z_i^{C_1}}|P_{z_j^{C_2}}) \geq 0$. It is zero if and only if $P_{z_i^{C_1}}$ equals $P_{z_j^{C_2}}$.

JS divergence is defined based on KL divergence to define the distance metric between these visual language models (Eq. (15)).

$$D_{JS}(P_{z_i^{C_1}}|P_{z_j^{C_2}}) = \frac{1}{2}D_{KL}(P_{z_i^{C_1}}|M) + \frac{1}{2}D_{KL}(P_{z_j^{C_2}}|M) \quad (15)$$

$$M = \frac{1}{2}(P_{z_i^{C_1}} + P_{z_j^{C_2}}) \quad (16)$$

where M is the average of $P_{z_i^{C_1}}$ and $P_{z_j^{C_2}}$. It is demonstrated that the square root of the Jensen-Shannon divergence is a metric. Thus the Flickr distance between two concepts C_1 and C_2 is calculated as the average square root of the JS divergence between the latent topic VLM of concept C_1 and that of concept C_2.

$$D_{Flickr}(C_1, C_2) = \sqrt{\frac{\sum_{i=1}^K \sum_{j=1}^K D_{JS}(P_{z_i^{C_1}}|P_{z_j^{C_2}})}{K^2}} \quad (17)$$

4. VISUAL CONCEPT NET

In this section, we discuss the construction of the visual concept net by Flickr distance. A visual concept net (VCNet) is a graph $G(V, E, W)$, where concepts are nodes $v_i \in V, i = 1, \cdots, N$ and the semantic relationship between two concepts is the edge $e(v_i, v_j) \in E, i, j = 1, \cdots, N$. The Flickr distance between the nodes is represented by the length (weight) of the edge $w \in W$. If two concepts have large Flickr distance, the edge between them is long; otherwise short.

To visualize the VCNet, NetDraw [2] is adopted. To avoid the overlapping of the concept nodes, force-directed graph layout algorithm is adopted. Generally, edges between nodes are represented as an attractive force, while nodes that do not share a tie are pushed apart by some constraint to help prevent overlap. In order to give a clear perspective of the VCNet, we only visualize 1,000 concepts on Flickr. We calculate the Flickr distance between every pair of these concepts, and use this distance as the edge length between each pair of nodes. In Figure 3, only the top 10% of the shortest edges are shown. For a brief view of the VCNet, we find the concepts "mp3",'MSN',"download", and "napster" form a clique in the network. This is coherent to our perception, since the same as MSN, napster is a kind of P2P software to help download mp3 from Internet, and it also provides messaging and online community services. More detail about the evaluation of the VCNet is discussed in the experiment section.

The VCNet is useful in many multimedia related tasks, i.e. knowledge representation [1], multimedia retrieval [9],etc. This concept net models the concept relationship in a graph manner like the WordNet. Besides, it can maintain a much larger and ever increasing corpus. Here we briefly give some discussion about the potential functions of VCNet, and more detailed experiments and application scenarios are elaborated in Section 5. One of the most directly application of VCNet is the concept clustering. This task aims to cluster concepts in the image tags or descriptions to help discover the main topic and summarization about the image. Using VCNet, concepts with semantic connections are more likely to be clustered together, which will be demonstrated in the experiment section. Another common application of VCNet is the content based web image/video annotation [18][3], where the common paradigm is to annotate the images or video frame by classification. This is done by either considering the concepts are independent to each other, i.e. CMRM [10], CRM [11], or incorporating the concepts relation into the model by using Google distance, i.e. DCMRM [15]. DCMRM [15] is demonstrated outperforming other previous methods. However, if the concept relation is measured by Flickr distance, the annotation results will be further boosted (shown in Sect. 5). VCNet also has many other potential applications, such as query enpension [17],annotation refinement [20], etc, which are out of the scope of this paper.

Since each concept corresponds to a tag on Flickr, we also

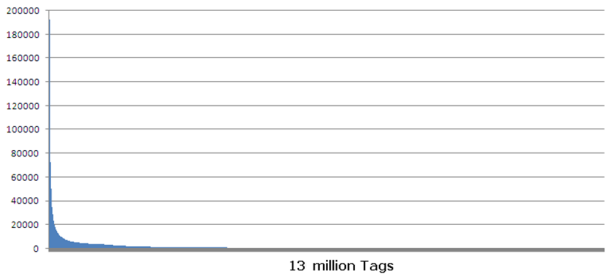


Figure 4: The tag frequency distribution.

call the concept network TagNet. However, there is slight difference between TagNet and VCNet. TagNet can be generated by both Google distance and Flickr distance. If the TagNet is based on Google distance, it is denoted as Google TagNet in the following discussion, and the TagNet based on Flickr distance is denoted as Flickr TagNet or VCNet. Figure 3 shows the output VCNet over the 1,000 concepts on Flickr. The generation of the 1,000 concepts are discussed in Section 5.1.

5. EXPERIMENTS AND APPLICATIONS SCENARIOS

In this section, we first demonstrate that the Flickr distance is more coherent to human cognition than the Google distance which is derived from textual correlation. This is illustrated by comparing the Flickr distance with the Google distance in both subjective and objective experiments. Then we apply the Flickr distance in several scenarios, and show this distance metric in visual domain is applicable and useful in multimedia related tasks.

5.1 Data set description

We use Flickr as the source of the concept and image database, and collected 6,400,000 images with 130,000,000 associated tags. However, among these tags there is much noise, such as the misspelling words, combination of words, and affix variation. The noise as well as some specific words form a long tail in the tag frequency distribution, as shown in Figure 4, where the horizontal axis is the tags indexed by their frequency, and vertical axis is the frequency of the tags used in the data set. For simplicity, we consider the tags whose frequency lies in the range of 1,000 to 50,000 as informative concepts. With this assumption, there are around 10^7 concepts generated. Absolutely, this simple assumption filters out many rarely used concepts. However, they are not critical in the following experiments. To facilitate evaluation and comparison with other concept relationship measurement, we randomly selected 1,000 concepts from the concept pool as the test samples. Taking each concept as query, the first 1,000 related images from the Flickr tag based search results are used to represent the concept. There is a parameter K to represent the number of variations of a concept. This parameter is determined by experiment. In the following experiments, we set K to 6.

5.2 Coherence to Human Perception

In this experiment, we demonstrate that the Flickr TagNet is more coherent to human perception than the Google TagNet. To achieve this, we first generate both Flickr TagNet and Google TagNet based on the 1,000 concept dataset. Then we evaluate these two TagNets by comparing them with the human knowledge on the concept structures.

The evaluation can be conducted both subjectively and objectively. For subjective evaluation, we conduct a user study, in which users are required to score each pair of related concepts according to their knowledge. The higher score means the automatically generated concept relation is more coherent to human perception. Since large scale concept network is too complex and expensive for human evaluation, we also conduct the objective evaluation. We take the WordNet as the ground truth,and compare the detection precision and recall of concept relationship between Flickr TagNet and Google TagNet. Since the ground truth is built by human experts and WordNet is a suitable ground truth to evaluate the coherence to human perception.

5.2.1 *Subjective user study*

As an illustration, a portion of the 1,000 testing concepts are listed in Table 2. We would like to generate the relations between these concepts by both Google distance and Flickr distance measurement. Taking each concept as a node and connecting each related concept pair if their distance is below a threshold, we can obtain a concept network (TagNet).

Table 2: Illustration part of the test concepts.

Concept List				
asian	eva	paris	simpsons	tinkerbell
baseball	ferrari	rainbows	skateboard	titanic
basketball	flower	roses	smallville	trees
bears	football	saturn	snake	troy
birds	friends	scarface	snow	tsunami
bmw	fruit	sharks	soccer	tulips
bowling	golf	shower	softball	turtle
butterflies	hawaii	shrek	space	usher
charmed	horses	mercedes	spiderman	venus
choppers	hurricane	minnesota	spiders	volleyball
cobra	kiss	moon	spongebob	washington
cross	lamborghini	motorcycles	sunset	whale
dolphin	landscape	mountains	superman	wolf
donkey	lingerie	mustang	surfing	wrestling
egypt	love	ocean	tatu	yoga
eminem	marijuana	paintball	tennis	...

To generate the Google TagNet, we calculate the NGD between each pair of concepts, and those pairs whose NGD is below the average level is considered as connected; otherwise disconnected. To generate the Flickr TagNet, we randomly collected 1,000 related images from Flickr for each concept, and build the VLM for the concept. Then we calculate the Flickr distance between every pair of concepts. The concept pair whose Flickr distance is below the average level is connected in the graph. The result Google TagNet and Flickr TagNet are shown in the following figures separately.

To facilitate observation, we only display the strong relations in the TagNet. Thus although the TagNets are generated based on the same concept corpus, the concepts dis-

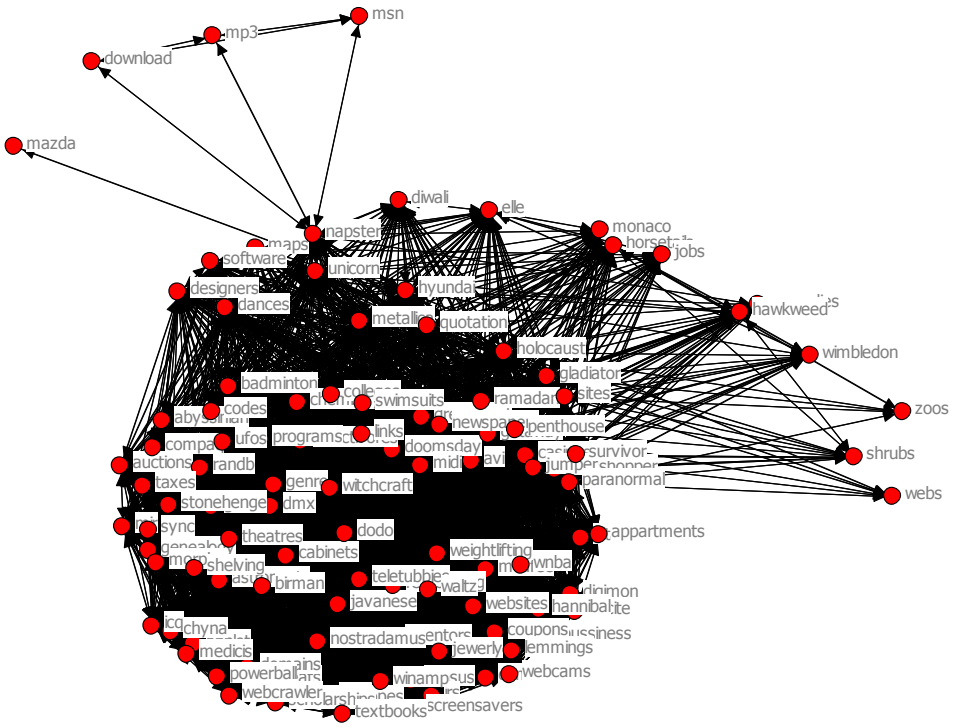


Figure 3: The VCNet over 1,000 Flickr concepts

played in two TagNets may not be the same. Since Google distance between two concepts is not symmetric, we adopt the directed graph to represent the concept network. The arrows "$a \rightarrow b$" in the network represent that concept a is related to concept b.

In the manual evaluation, we generate a list of concept pairs representing all the relations from the TagNets and present them to 12 independent users. For each concept pair, the user is required to give a score ranging from 1 to 5. The higher the score is, the more related the two concepts are according to the user's knowledge. By averaging the score from all the users, we get the final score for each of the concept pairs, which are shown in Table 3. There are four columns in the table. The left two columns are the concept pairs generated by Google TagNet and their corresponding average user score. The right two columns are the concept pairs from Flickr TagNet and the corresponding average user score. The last row of the table gives the global average score for the two kinds of TagNets. The standard deviation of the rating over all testing pairs is 0.15 and 0.11 respectively for the two approaches.

This subjective user study shows that the coherence of Flickr TagNet achieves 29.97% gain over the Google TagNet. This, on the other hand, also demonstrates that human visual perception contributes more in generating knowledge about concepts relationship.

5.2.2 *Objective Experiments*

Superior to the subjective study, the objective evaluation can be conducted in relatively larger scale concept data set. We take the concept relationship in WordNet as ground truth. Although the corpus in WordNet is limited and cannot update easily, it is still reasonable to be taken as the ground truth in the experiment, since WordNet is generated by well-trained experts and demonstrated largely coherent to human cognition.

We filter these concepts by WordNet corpus. After the filtering, there are only 497 concepts left. Google TagNet and Flickr TagNet are generated on these 497 concepts in the same way discussed in the above subsection. The choice of dataset scale is based on the current scale of ground truth as well as the evaluation cost. This dataset is much larger than commonly used dataset. Given enough ground truth and resource, the evaluation can be performed on even larger one. Precision and recall are calculated to evaluate the performance. If the concept distance is below the average level, we connect them. We denote the concept connection set generated by Google distance (NGD), Flickr distance (FD), and wordNet (WN) as E_{NGD}, E_{FD}, and E_{WN};

$$Precision = \frac{T_+}{T_+ + T_-} \tag{18}$$

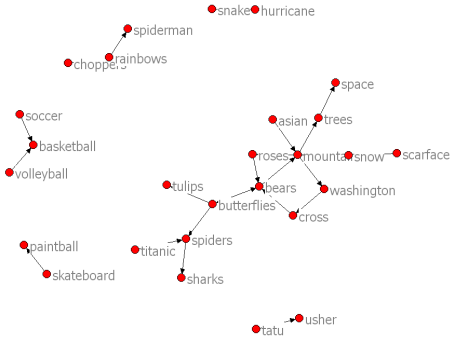
$$Recall = \frac{T_+}{T_+ + F_+} \tag{19}$$

T_+ is the number of connections in both E_{NGD}/E_{FD} and E_{WN}; T_- is the number of connections in E_{NGD}/E_{FD} but not in E_{WN}; F_+ is the number of connections which are not in E_{NGD}/E_{FD} but in E_{WN}.

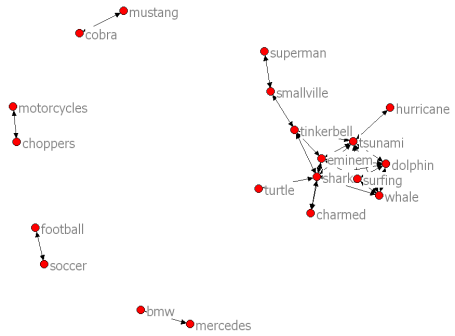
This comparison (shown in Figure 6) illustrates that Flickr TagNet outperforms Google TagNet by 17.2% in precision and 1.6% in recall. The experiment demonstrates that Flickr TagNet is more coherent with WordNet, which embodies by large the human cognition.

Table 3: Subjective study result.

Google Distance		Flickr Distance	
Concept Pair	Score	Concept Pair	Score
asian–mountains	1.33	BMW–Mercedes	4.33
bears–butterflies	2.25	charmed–Eminem	3.17
bears–mountains	2.83	charmed–sharks	1.58
butterflies–bears	2.25	choppers–motorcycles	4.92
butterflies–spiders	3.08	Cobra–Mustang	3.42
butterflies–tulips	2.92	dolphin–Eminem	1.42
choppers–rainbows	1.25	dolphin–sharks	3.67
cross–bears	1.17	dolphin–surfing	3.08
mountains–roses	2.00	dolphin–tsunami	2.5
mountains–snow	2.92	dolphin–whale	4.08
mountains–trees	4.00	Eminem–sharks	1.25
mountains–Washington	1.42	Eminem–surfing	1.25
rainbows–spiderman	1.58	Eminem–Tinkerbell	1.75
roses–bears	1.5	Eminem–tsunami	1.33
scarface-snow	1.33	football–soccer	5.00
skateboard–paintball	2.42	hurricane–tsunami	4.75
snake–hurricane	1.42	sharks–surfing	3.25
snow–mountains	3.08	sharks–Tinkerbell	1.25
soccer–basketball	4.08	sharks–tsunami	2.5
spiders–sharks	1.83	sharks–turtle	3.5
TATU–Usher	3.25	sharks–whale	4.17
titanic–spiders	1.75	Smallville–Superman	4.75
trees–space	1.67	Smallville–Tinkerbell	2.25
volleyball–basketball	4.08	surfing–tsunami	3.25
Washington–cross	1.33	surfing–whale	3.08
-	-	tsunami–whale	2.83
-	-	Tinkerbell–tsunami	1.25
Average	2.27	Average	2.95



(a) TagNet generated by Google distance



(b) TagNet generated by Flickr distance

Figure 5: Illustration of different TagNets. The two figures are generated based on the same concepts database. For clarity, we set a strict threshold, and only show the strong connections among the TagNet. This automatic sampling comparison is relatively more justified than manually selecting the same subset of samples to evaluate both methods, since the performance on top related samples are generally more important and the manually sampling method is tricky. The concept does not have strong relations with others are not shown in the figure.

5.3 Concept Clustering

In this application scenario, we apply the Flickr distance on concept clustering. Concept clustering is widely used for topic detection and summarization in textual domain. Since there are lots of tags and descriptions associated with web images, we are able to use concept clustering to detect the main topic or summarization of these images. However, the focus of the topic summarization in image may not be the same with that for the text. For example, the image is more likely to focus on the main object or scene, while in the text document it focuses more on the story or point of view of the author. Thus an applicable concept distance measurement for textual domian, i.e. Google distance, may not perform as well as the specific distance measurement for visual domain. Here we compare the concept clustering results of Flickr distance and Google distance to illustrate the difference.

Three groups of concepts are selected: Space related terms (4 concepts), Ball games (10 concepts), and Animals (9 concepts). We choose these concepts because all users agree that these concepts are grouped without ambiguous in the user study. In total there are 23 concepts in the experiment. The task is to group these concepts into clusters automatically. One of the key issues with concept clustering is the concept distance measurement. In most cases, WordNet is used to measure the concept distances. However, due to the limitation of WordNet lexicon, a large portion of concepts, i.e. famous movies, brand, game, sport star, singer, etc. are inextricable. In this experiment, we build two different networks between these concepts with Google distance and Flickr distance separately. Based on these two concept networks, spectral clustering is adopted to generate the concept clusters. We adopt spectral clustering rather than the commonly used K-means algorithm, because it is hard to calculate the cluster centers of these concepts in K-means algorithm, while the spectral clustering only use the relationship between the samples. The results of the concept clusters are shown in Table 4.

Table 4 shows that the Flickr distance based spectral clustering can effectively generate the concept clusters. After

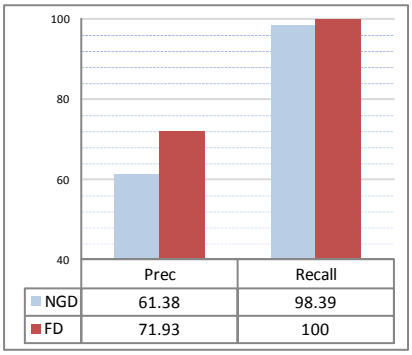


Figure 6: Comparison of precision and recall.

Table 4: Result of concept clustering.

Concept clusters by NGD			Concept clusters by FD		
Group 1:	Group 2:	Group 3:	Group 1:	Group 2:	Group 3:
bears	**bowling**	baseball	moon	bears	baseball
horses	dolphin	basketball	saturn	dolphin	basketball
moon	donkey	football	space	donkey	football
space	**saturn**	golf	venus	**golf**	**snake**
-	sharks	soccer	-	horses	soccer
-	snake	tennis	-	sharks	bowling
-	**softball**	volleyball	-	spiders	softball
-	spiders	-	-	**tennis**	volleyball
-	turtle	-	-	turtle	-
-	**venus**	-	-	whale	-
-	whale	-	-	wolf	-
-	wolf	-	-	-	-

the concept clustering, we know the three categories of images are about space, animals, and ball games. 6 out of the 23 total concepts are mistakenly clustered by Google distance, which are marked in bold in Table 4, and only 3 mistakes occur in the result of Flickr distance. Comparing with the clustering results based on Google distance, the results by Flickr distance is more promising.

5.4 Image Annotation

Automatically annotating concepts for images is critical in web image retrieval and browsing. Most of the state-of-the-art image annotation approaches detect multiple semantic concepts in an isolated manner, which neglect the fact that the concepts may correlated to each other. The generative model of the annotation process can be represented as Eq. (20).

$$w^* = \arg\max_{w \subset V} P(w, I_u) \tag{20}$$

where w is the annotation keywords, and w^* is the best suitable keyword. I_u represents the unlabeled image. This annotation process equals to the maximization of the joint probability $P(w, I_u)$. The annotation performance may be further boosted with consideration of the concept relations.

Based on this motivation, the Dual Cross-Media Relevance Model (DCMRM) [15] is proposed, and achieved well performance. This model assumes that the probability of observing the annotation keyword w and the images I_u are mutually independent given a keyword v, and the relevance model is represented as follows.

$$w^* = \arg\max_{w \subset V} \sum_{v \in V} P(I_u|v)P(w|v)P(v) \tag{21}$$

where w and v are two annotation keywords. $P(I_u|v)$ denotes the probability of an untagged image I_u given a word v. $P(w|v)$ denotes the probability of a word w given a word v. In Jing's work [15], Google distance is adopted to calculate the $P(w|v)$, which is denoted by NGD-DCMRM for short in the following discussion. By contrast, we apply the Flickr distance to calculate the conditional probability $P(w|v)$, and the corresponding annotation method is denoted by FD-DCMRM.

For simplicity, we adopt the Flickr images associated to the 79 sample concepts (Table 2) used in previous experiments as the dataset. Since there are 1,000 images for each concepts, the dataset contains 79,000 images. 80% of the images are used for training and the rest images for testing. The performance is shown in Table 5.

Table 5: The comparison by Precision @ N.

	Prec@1	Prec@2	Prec@3	Prec@4
CRM	5%	4%	3.63%	3.38%
NGD-DCMRM	5%	10.5%	7%	6%
FD-DCMRM	5%	17.5%	14%	19.2%

Table 5 gives the precision@N (N=1, 2, 3, 4) of three different annotation methods. precision@N measures the precision of annotation in the first N words. CRM does not use the word relations and only use the correlation between images to annotate. NGD-DCMRM uses concept relations in the annotation process and measures the concept distance by normalized Google distance. FD-DCMRM adopts the same annotation algorithm with NGD-DCMRM, but FD-DCMRM adopts Flickr distance to measure the concept distance. The result shows that FD-DCMRM outperforms both NGD-DCMRM and CRM on Prec@1,2,3,4. To further illustrate the advantage of using Flickr distance rather than Google distance in image annotation, we compare the total correct keywords annotated on the test dataset, consists of 1,000 test images. The result is shown in Figure 7. The number in horizontal axis denotes the evaluation is conducted on the first N keywords for each image (N=1, 2, 3, 4). The vertical axis represents the total number of correct keywords annotated to all the test images. This comparison shows Flickr distance based annotation method can generate more correct keywords, especially when annotating multiple keywords to each image. This also, in some aspects, demonstrates that Flickr distance is more helpful than Google distance in multimedia related tasks.

The computational cost of the proposed method lies on the training of VLM. The average time cost for each concept is about 1 minute. So 1,000 concepts cost around 17 hours (Intel(R) Xeon(R) CPU 2.66GHz; 16G RAM). As the similarity matrix can be calculated offline, the computational time is not critical in this process. Given the similarity matrix, the rest cost is the same with alternative approaches.

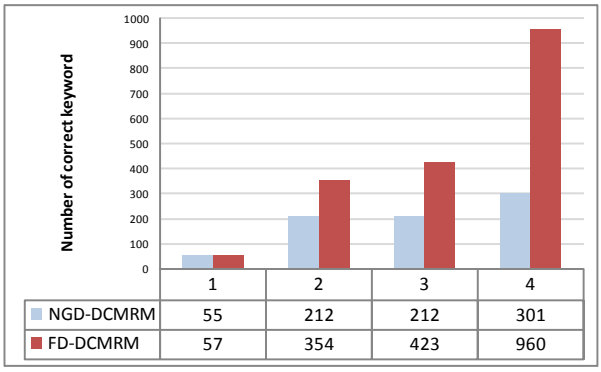


Figure 7: The number of correct annotation words at first N words (N=1,2,3,4).

6. CONCLUSIONS

In this paper, we propose the Flickr distance, which measures the concept relationship by the visual correlation between concepts. We also proposed a latent topic based visual language model to capture the visual characteristic of the concepts. Based on this novel distance measurement, we construct Flickr TagNet (VCNet) to model the relationship between the popular tags on Flickr. Comparing with traditional WordNet, which is limited in concept corpus, Flickr TagNet can deal with far more concepts on the web and can be easily updated. Both subjective user study and objective experiment show that Flickr TagNet is more coherent to human cognition than the Google TagNet. Furthermore, we apply the Flickr distance to concept clustering and image annotation. The results demonstrate that Flickr distance and the corresponding TagNet are more helpful to multimedia applications than Google distance.

7. REFERENCES

[1] W. Al-Khatib, Y. F. Day, A. Ghafoor, and P. B. Berra. Semantic modeling and knowledge representation in multimedia databases. *IEEE Trans. on Knowl. and Data Eng.*, 11(1):64–80, 1999.

[2] S. Borgatti. Netdraw. *http://www.analytictech.com/Netdraw/netdraw.htm*, 2008.

[3] S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. C. Loui, and J. Luo. Large-scale multimodal semantic concept detection for consumer video. In *Proc. of the international workshop on Workshop on multimedia information retrieval*, 2007.

[4] Y. Chen and J. Z. Wang. Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research*, 5:913–939, 2004.

[5] R. Cilibrasi and P. M. B. Vitanyi. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19:370, 2007.

[6] R. Datta, J. Dhiraj, L. Jia, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 2008.

[7] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[8] T. Hofmann. Probabilistic Latent Semantic Indexing. In *Proc. of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, pages 50–57, Berkeley, California, August 1999.

[9] T. S. Huang, C. K. Dagli, S. Rajaram, E. Y. Chang, M. I. Mandel, G. E. Poliner, and D. P. W. Ellis. Active learning for interactive multimedia retrieval. In *Proc. of the IEEE*, 2008.

[10] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proc. of the 26th annual international ACM SIGIR conference, SIGIR '03*, 2003.

[11] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Proc. of NIPS'03.*, 2003.

[12] D. B. Lenat. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.

[13] L. Leslie, T.-S. Chua, and J. Ramesh. Annotation of paintings with high-level semantic concepts using transductive inference and ontology-based concept disambiguation. In *Proc. of the 15th international conference on Multimedia*, 2007.

[14] H. Liu, S. Jiang, Q. Huang, C. Xu, and W. Gao. Region-based visual attention analysis with its application in image browsing on small displays. In *Proc. of the 15th international conference on Multimedia*, 2007.

[15] J. Liu, B. Wang, M. Li, Z. Li, W.-Y. Ma, H. Lu, and S. Ma. Dual cross-media relevance model for image annotation. In *Proc. of the 15th international conference on Multimedia*, pages 605–614, 2007.

[16] G. A. Miller and et.al. Wordnet, a lexical database for the english language. *Cognition Science Lab, Princeton University*, 1995.

[17] A. P. Natsev, A. Haubold, J. Tešić, L. Xie, and R. Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *Proc. of the 15th international conference on Multimedia*, 2007.

[18] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang. Correlative multi-label video annotation. In *Proc. of the 15th international conference on Multimedia*, 2007.

[19] B. Wang, Z. Li, M. Li, and W.-Y. Ma. Large-scale duplicate detection for web image search. In *Proc. of IEEE International Conference on Multimedia & Expo (ICME'06)*, 2006.

[20] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang. Content-based image annotation refinement. 2007.

[21] L. Wu, M. Li, Z. Li, W.-Y. Ma, and N. Yu. Visual language modeling for image classification. In *Proc. of 9th ACM SIGMM International Workshop on Multimedia Information Retrieval, (MIR'07)*, 2007.

[22] L. Wu, J. Liu, M. Li, and N. Yu. Query oriented subspace shifting for near-duplicate image detection. In *Proc. of IEEE International Conference on Multimedia & Expo (ICME'08)*, 2008.

[23] J. Yu and Q. Tian. Semantic subspace projection and its application in image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology (CSVT)*, pages 544–548, 2008.