A LEXICA FAMILY WITH SMALL SEMANTIC GAP

Jiemin Liu[1,2,*], *Qi Tian*[1], *Yijuan Lu*[3], *Changhu Wang*[1], *Lei Zhang*[1], *Xiaokang Yang*[2], *Shipeng Li*[1]

[1]Microsoft Research Asia, 49 Zhichun Road, Beijing 100190, China
[2]Department of Electronic Engineering, Shanghai Jiao Tong University
[3]Department of Computer Science, Texas State University, USA

ABSTRACT

Defining a lexicon of high-level concepts is the first step for data collection and model construction in concept-based image retrieval. Differences of semantic gaps among concepts are well worth considering. By measuring consistency in visual space and textual space, concepts with small semantic gap can be obtained. Considering so many diverse concepts in large-scale image dataset, we construct a lexica family of high-level concepts with small semantic gap based on different low-level features and different consistency measurements. In this lexica family, the lexica are independent to each other and mutually complementary. It provides helpful suggestions about data collection, feature selection and search model construction for large-scale image retrieval.

Index Terms— semantic gap, lexica, large-scale, image retrieval

1. INTRODUCTION

With the explosive growth of digital images and videos in recent years, Multimedia Information Retrieval (MIR) techniques have experienced a rapid development. A fundamental challenge in MIR is the *semantic gap* between low-level visual features and high-level semantic concepts. To bridge the semantic gap, generic approaches focus on concept detection which is to learn a concept classifier by machine learning method based on labeled examples. Hence, defining a lexicon of semantic concepts is a necessary pre-task for ground truth data annotation and concept detection modeling.

Most research works were developed on datasets based on a lexicon manually defined like LSCOM-Lite, LSCOM[1] and MediaMill[2]. Large-Scale Concept Ontology for Multimedia (LSCOM) is an ontology of about 1,000 concepts produced based on manually annotating a large corpus of 80 hours of broadcast news video. LSCOM-Lite is a subset of the full LSCOM and it contains 39 concepts with annotations over the entire development set of TRECVID[3] 2005 videos. It was selected based on semi-automatic mapping noun search terms from BBC query logs to WordNet senses. MediaMill is a lexicon of 101 concepts selected by taking LSCOM as leading example and analyzing extended manual annotations. In addition to these lexica produced by manual selection, there are several public image databases such as Caltech-101[4], Caltech-256[5] which contain many images belonging to some manually selected concepts.

However, one important question was not proposed until 2007 by Hauptmann *et al*.[6]. That is, what kind of concepts are most useful? They took the first step to answer this question based on analyzing TRECVID'05 video archive annotated with the 320 LSCOM concepts. They calculate concept utility to measure how each concept contributes to retrieval. In their work, those useful concepts were selected from statistic aspect of concept frequency.

Although the above research proposes some useful concepts, it still ignores inherent semantic gap information of concepts. Difference of semantic gaps among concepts deeply affects performance of corresponding concept detection. Concepts with small semantic gaps are better to be modeled and retrieved. Hence, constructing a lexicon of concepts with small semantic gaps is meaningful for multimedia information retrieval. Based on a large-scale web image dataset, Lu *et al*.[7] proposed a novel way to construct a lexicon of high-level concepts with small semantic gaps (LCSS). These concepts were extracted from textual information of candidate images by measuring their consistency in visual feature space and semantic textual space.

However, a single lexicon is not enough for various types of concepts in large-scale image data. It has two problems. First, in different visual spaces, images of concepts distribute differently. Two concepts may be far away from each other in one visual feature space. But in another space, they might be closer to each other. For example,"wood" and "sand" have similar color features while they are totally different in texture feature space. So a lexicon based on a single visual feature is insufficient and inefficient for presenting concepts. Second, in [7], the authors use a method called as nearest neighbor confidence score (NNCS) to measure image's consistency in the visual and textual spaces. But this NNCS is a visual-central method that only considers visually similar images' consistency in textual space but ignores textually similar images' consistency in visual space. Concepts' consistency should be

* This work was done at Microsoft Research Asia.

measured by considering both visual consistency and textual consistency, and may be more.

Therefore, facing to diversity of high-level concepts in large-scale dataset, we should construct a lexica family covering high-level concepts with small semantic gap. In this paper, we analyze semantic gaps of concepts in several different low-level feature spaces and produce mutually independent and complementary lexica. In addition, both visual-central NNCS and textual-central NNCS methods are used in the framework. Semantic gap is measured based on both visual and textual space. By comparing different NNCS-based lexica, we can make suggestions of choosing appropriate search method for specific concepts.

The rest of the paper is organized as followed: In Section 2, we discuss the details about lexica family and framework of its construction. In Section 3, we present experimental procedure and analysis of results. Conclusions and future work are given in Section 4.

2. A LEXICA FAMILY FOR CONCEPT-BASED IMAGE RETRIEVAL

2.1. Different Visual Feature Spaces

In large-scale image dataset, there are hundreds and thousands of high-level concepts. The overall semantic space includes many different types of concepts such as object, scene, state, landscape and so on. People distinguish things of different concepts through species' inherent diversity. It reflects on color, shape, texture and other low-level features. So it's necessary to construct lexica based on different low-level features. A family of feature-based lexica can provide appropriate options for feature selection for specific concepts.

2.2. Semantic Gap

For all images in a large-scale image dataset with rich surrounding textual information, they can be located in two different high-dimensional spaces. One is visual space. The other one is textual space. If nearest neighbors of one image in visual space are the same as the nearest neighbors of it in textual space (as shown in Fig. 1(a)), concept within the image has small semantic gap because of both consistency in two spaces. If nearest neighbors of one image in visual space are disperse around it in textual space (as shown in Fig. 1(b)), concept within the image has loose-textual semantic gap. Similarly, if nearest neighbors of one image in textual space are disperse around it in visual space (as shown in Fig. 1(c)), concept within the image has loose-visual semantic gap. In order to measure images' consistency of visual space and textual space, two algorithms are used to measure two situations as followed:

- Visual-central Nearest Neighbor Confidence Score (VNNCS): for a given image I_q, first find its K neighbors in visual space $\{I_i|i = 1, 2..., K\}$, then calculate average of textual similarity between I_q and every I_i.

$$VNNCS(I_q) = \frac{1}{K}\sum_{i=1}^{K} sim\_textual(I_q, I_i) \quad for\ \ I_i \,\epsilon\, Visual\_neighbors(I_q) \tag{1}$$

- Textual-central Nearest Neighbor Confidence Score (TNNCS): for a given image I_q, first find its K neighbors in textual space $\{I_i|i = 1, 2..., K\}$, then calculate average of visual similarity between I_q and every I_i.

$$TNNCS(I_q) = \frac{1}{K}\sum_{i=1}^{K} sim\_visual(I_q, I_i) \quad for\ \ I_i \,\epsilon\, Textual\_neighbors(I_q) \tag{2}$$

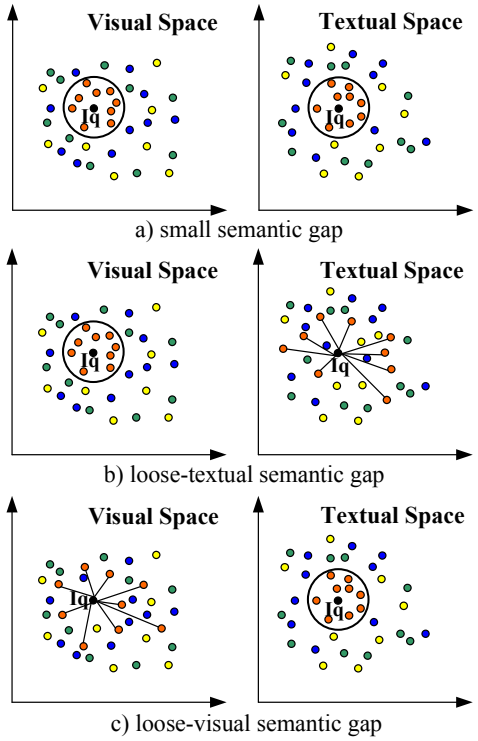

Fig. 1. Consistency in visual space and textual space

$sim\_textual(I_q, I_i)$ can be calculated by measuring their textual descriptions' cosine similarity. $sim\_visual(I_q, I_i)$ can be calculated by measuring their negative visual features' Euclidean distance. The higher the VNNCS value of one image is, the tighter textual consistency the concept concerned in this image has. Similarly, the higher the TNNCS value of one image is, the tighter visual consistency the concept concerned in this image has.

2.3. Lexicon Construction Framework

The framework of lexicon construction procedure contains four steps[7]:

1) Data collection and preparation: About 2.4 million web images with rich surrounding text information are collected from 5 online photo forums. For all images, we build indexes based on low-level visual feature and surrounding textual feature respectively.

2) Confidence map construction: By definitions of NNCS, we calculate Visual-NNCS and Textual-NNCS of each image. Then, top 30000 images with higher scores are selected as candidate ones. They are considered to contain high-level concepts with small semantic gaps.

3) Affinity propagation clustering: We cluster candidate images by using affinity propagation method [8].

4) Text-based keyword extraction: From well clustered images sets, we extract the most representative keywords of clusters by sorting relational degree between keywords and clusters. The final sorted list of keywords is a lexicon of high-level concepts with small semantic gaps.

3. EXPERIMENTS AND ANALYSIS

3.1. Data and Features

We collected about 2.4 million web images from 5 online public photo forums including Photosig[1], Usefilm[2] and so on. Rich surrounding textual information of images such as title, category, tag, description, comments were extracted and built as a textual index. These textual information are almost actually good semantic description of the images. We extracted 4 low-level features for 2.4 million web images, as shown in Table 1.

3.2. Feature-based Lexica

Concepts have different semantic gaps in different low-level feature spaces. By comparing different lexica with small semantic gaps, we can select more appropriate low-level feature as representation, that is, feature selection for concepts. Given a specific low-level feature and Visual-NNCS algorithm, we extract top 101 concepts as a corresponding lexicon. Then we compare and contrast these lexica. After removing some noisy concepts, there are totally 104 meaningful concepts. Within them, 65 concepts belong to the lexicon based on color feature and 81 concepts belong to the lexicon based on texture feature. Meanwhile, 42 concepts both belong to two lexica. All of these concepts can be found in Fig. 2.

As shown in Fig. 2, the lexicon based on color feature consists of concepts shown in the box with solid line and the lexicon based on texture feature consists of concepts shown in the box with dotted line. For the concepts within the overlapping part, they have inherently small semantic gaps based on either color feature or texture feature.

[1]http://www.photosig.com/.
[2]http://www.usefilm.com/.

Fig. 2. Concepts concerned in Feature-based Lexica

For those concepts which have small semantic gap only based on color feature, like sunflower, peacock, orchid (as shown in Fig. 3(a)), color feature is the best choice for feature selection. For those concepts which have small semantic gap only based on texture feature, like window, butterfly, rock (as shown in Fig. 3(b)), texture feature is the best choice. For some concepts which have small semantic gap based on either color feature or texture feature, such as sunset, rose, firework (as shown in Fig. 3(c)), we can choose color and texture combined feature as appropriate representation.

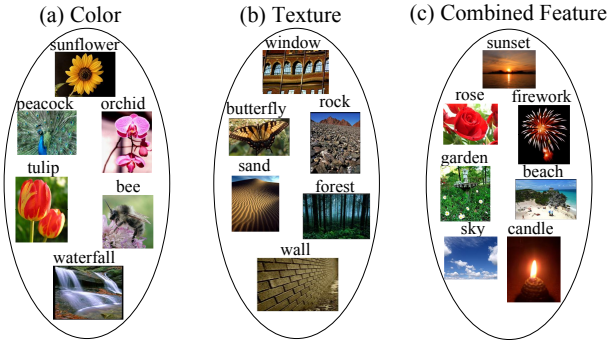

Fig. 3. Examples of feature selection for concepts

3.3. NNCS-based Lexica

Definitions of two NNCS algorithms as mentioned in Section 2.2 reflect two different search methods, content-based search (visual-based) and context-based (textual-based) search. Lexica with small semantic gap based on different NNCS algorithms can provide suggestion of search method for concepts. Therefore, we use 50-dim Color feature as visual representations of image and construct two lexica by using VNNCS and TNNCS algorithms respectively. After calculating NNCS of each image, selecting candidate images, clustering and extracting concepts, two different lexica are obtained. Similarities and differences between these two lexica are shown in Table 2.

Low-level features	Dimension	Descriptions
Color	50	6-dim color moment(LUV) and 44-dim banded auto-color correlogram(HSV)
Co-occurrence Texture(COT)	16	16-dim normalized vector as measurement of global grey-level co-occurrence matrix
Wavelet Texture(WT)	128	128-dim vector of wavelet parameters
Color+Texture	64	50-dim color vector concatenated with 14-dim color texture moments

Table 1. Four low-level features of image dataset

Part I: Visual-central NNCS	
Category	**Concepts**
Scene	firework, sunrise, rain, wild
Landscape	bay, field, home, house, coast, ocean, pier, hill
Color	yellow, green, pink, purple, orange, golden
Object	candle, moon, drop, boat, saw
Plant	rose, sunflower, orchid, tulip, daisy, lily, irid, leaf, bloom, glass
Animal	bee, peacock, fish, bird
Season	spring, summer, autumn
Part II: Textual-central NNCS	
Category	**Concepts**
People	girl, man, woman, model, angel, nude, sister, children, male, face
Animal	cat, tiger, dog, wolf
Water	creek, valley, canyon, stream
Place	street, road, church, castle, cemetery, market, square, metro, studio, village, town
object	stone, chain, crater
Part III: Either of Visual-central NNCS and Textual-central NNCS	
Category	**Concepts**
Scene	sunset, sky, shadow, city, water snow, storm, ice, cloud
Landscape	fall, lake, river, garden, beach, mountain, bridge, waterfall, island
Color	red, blue, dark
Object	eye, rock, key, window, flower, tree

Table 2. Similarities and differences between two lexica based on two NNCS algorithms

In the Part I of Table 2, there are 7 categories of concepts with small semantic gaps only based on VNNCS but not TNNCS. Hence, for these concepts like firework and rose, content-based search is preferred than context-based search. By contrast, 5 categories of concepts within the Part II of Table 2 have small semantic gaps only based on TNNCS but not VNNCS. For these concepts like girl and street, it's better to search them based on textual information than visual information. In addition, the Part III of Table 2 consists of concepts which have small semantic gaps based on either of VNNCS and TNNCS. For these concepts such as sunset, content-based search and context-based search both have good performance. Above analyses of experimental results are very useful for choosing search method for concepts.

4. CONCLUSION AND FUTURE WORK

In this paper, from a large-scale web image dataset, we constructed a lexica family of high-level concepts with small semantic gaps which contains feature-based lexica and NNCS-based lexica. Feature-based lexica provide feature selections for image retrieval with concepts. And NNCS-based lexica obtained by measuring semantic gap from both visual and textual space make suggestion of choosing search model for concepts. However, this is still a preliminary attempt on quantitatively analyzing semantic gaps of high-level concepts. More systematic approaches on modeling semantic gaps are still worth probing.

5. REFERENCES

[1] M. Naphade, J.R. Smith, J. Tesic, S.F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia", *IEEE MultiMedia*, vol. 13, no. 3, pp. 86–91, 2006.

[2] C.G.M. Snoek, M. Worring, J.C. van Gemert, J.M. Geusebroek, and A.W.M. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia", in *Proceedings of the 14th annual ACM international conference on Multimedia (MULTIMEDIA '06)*, New York, NY, USA, 2006, pp. 421–430, ACM.

[3] A.F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid", in *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (MIR '06)*, New York, NY, USA, 2006, pp. 321–330, ACM Press.

[4] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories", *Comput. Vis. Image Underst.*, vol. 106, no. 1, pp. 59–70, 2007.

[5] G. Gregory, H. Alex, and P. Pietro, "Caltech-256 object category dataset", *Caltech Technical Report*, 2007.

[6] A. Hauptmann, R. Yan, W.H. Lin, M. Christel, and H. Wactlar, "Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news", *IEEE Transactions on Multimedia*, vol. 9, no. 5, pp. 958–966, Aug. 2007.

[7] Y. Lu, L. Zhang, Q. Tian, and W.Y. Ma, "What are the high-level concepts with small semantic gaps?", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, 2008.

[8] B.J. Frey and D. Dueck, "Clustering by passing messages between data points", *Science*, vol. 315, pp. 972–976, 2007.