



IMAGE RETRIEVAL FOR VISUALLY IMPAIRED PERSON

S.PADMINI VASAVI

III year ECE

National College of Engineering, Tirunelveli.

ABSTRACT:

Bag-of-visual Words (BoWs) representation has been applied for various problems in the field of multimedia and computer vision. In this paper, descriptive visual words (DVWs) and descriptive visual phrases (DVPs) are proposed as the visual correspondences to the text words and phrases, where visual phrases refer to the frequently co-occurring visual word pairs. The proposed framework will select images and transform it into a text file using descriptive visual words and visual phrases method. The text file will be read by the speech synthesizer which tells to the visually impaired person. Independent navigation is always a challenge to visually impaired person, whenever they learning in the image based books such as science, social science, biology, computer science books. We apply the identified DVWs and DVPs in several applications including large-scale near-duplicated image retrieval,

image re-ranking, object recognition and text to speech synthesis. The proposed image re-ranking algorithm: DWP Rank performs the state-of-the-art algorithm by 12.4% in mean average precision and about 11 times faster in efficiency.

Key Words:

Image retrieval, image search re-ranking, object recognition, visual phrase, visual word, text to speech synthesizer, visually impaired people

1.INTRODUCTION

Visually impaired is a state where the partially blind people have difficulties on seeing things which in some degree can be considered as handicap. People with normal long-sighted and myopic are not considered as handicap or visually impaired. There are an estimated 45 million blind people and 135 million visually impaired people worldwide. On the statistics by gender, female has the significantly higher risk to be visually impaired than male, in every parts of the world and also every age group.



There are many factors contributing to visually impairment. The most common of the blind persons are not higher study. Visually impaired use different ways and methods to make their life easier to learn and many more. In this paper, **descriptive visual words (DVWs) and descriptive visual phrases (DVPs)** are proposed as the visual correspondences to text words and phrases, where visual phrases refer to the frequently co-occurring visual word pairs. The proposed framework will capture images and transform it into a text file using descriptive visual words and visual phrases method. The text file will be read by a speech synthesis that tells to the visually impaired people to selected images. So books used to all image details system is displayed and tells so easily study for higher standard books for all blind students. These observations strongly imply the limited descriptive ability of the classic visual word. A toy example illustrating this finding is presented in Fig. 1. In the figure, SIFT descriptors are extracted on interest points detected by **Difference of Gaussian (DoG)**. The three images are then

represented as BoWs with a visual vocabulary containing 32 357 visual words, by replacing their SIFT descriptors with the indexes of the closest visual words. In the figure 1.1, two interest points are connected with a red line (online version) if they share the same visual word. As we can clearly observe, although the visual appearances of the plane and cat are very different, there are still many matched visual words between them. It can be inferred

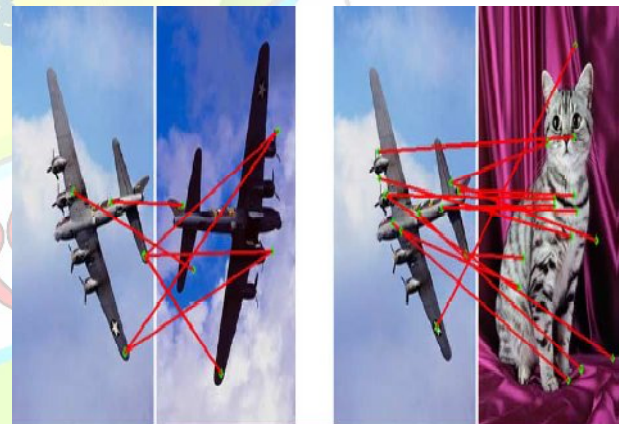


Fig.1.1. **Matched visual words between the same and different objects.**

that the visual word is noisy and indiscriminative, resulting in its ineffectiveness in measuring the similarity between the two images. There are two problems in the classic visual words, which



may be the main causes for their limited descriptive power.

1) Single visual word contains limited spatial contextual information, which has been proven important for visual matching and recognition. Thus, it is not effective in presenting the characteristics of objects and scenes. This can be explained by an analogy between basic English alphabets and single visual words.

2) Previous K -means-based visual vocabulary generation cannot lead to very effective and compact visual vocabulary. This is because simply clustering the local descriptors in unsupervised way generates lots of unnecessary and nondescriptive visual words in the cluttered background, e.g., the noisy mismatched visual words.

In order to overcome the above two shortcomings and generate visual vocabulary that is as comparable to the text words as possible, descriptive visual words (DVWs) and descriptive visual phrases (DVPs) are proposed in this paper. DVWs are defined as the individual visual words specifically effective in describing certain objects or scenes. Similar to the semantic

meaningful phrases in documents, DVPs are defined as the distinctive and commonly co-occurring visual word pairs in images. Intuitively, because DVWs and DVPs only keep the descriptive visual words and visual word pairs, they would be descriptive, compact, and clean. Once established, they will lead to compact and effective BoWs representation. Generating DVW and DVP set seems to be a very difficult problem, but statistics in large-scale image datasets might provide us some help. Because images are carriers of different visual objects or visual scenes, classic visual words and their combinations that are descriptive to certain objects or scenes could be selected as DVWs and DVPs, respectively. The corresponding DVWs and DVPs will function more similar to the text words than the classic visual words because of the reasons given here.

1) Only unique and effective visual words and combinations are selected. Thus, the selected set would be compact to describe specific objects or scenes. In addition, this significantly reduces the negative effects of visual words generated



from the cluttered background. Therefore, the DVWs and DVPs would be more descriptive.

2) Based on the large-scale image training set containing different scenes and objects, DVWs and DVPs might present better descriptive ability to the real word and could be scalable and capable for various applications. Consequently, our algorithms identify and collect DVWs and DVPs from a large number of object and scene categories.

3) Text to voice convertor: It is use to convert text to speech which will tell the visually impaired or the blind person about the image in a descriptive manner .

A novel visual-word-level ranking algorithm:

Visual WordRank which is similar to the **PageRank** and **VisualRank** is proposed for identifying and selecting DVWs efficiently. Based on the proposed ranking algorithms, DVWs and DVPs for different objects or scenes are discriminatively selected. The final DVW and DVP set is generated by combining all of the selected DVWs and DVPs across

different categories. Extensive experiments on image retrieval tasks show that the DVW and DVP present stronger descriptive power than the classic visual words. Furthermore, DVW and DVP show promising performance in image search re-ranking and object recognition tasks. In summary, the contributions of our work are given here. The drawbacks of classic visual words are discussed. A novel large-scale web image-based solution is proposed for generating DVWs and DVPs. The idea of PageRank and VisualRank is leveraged in VisualWordRank for DVWselection. Experiments validate the effectiveness and efficiency of VisualWordRank.

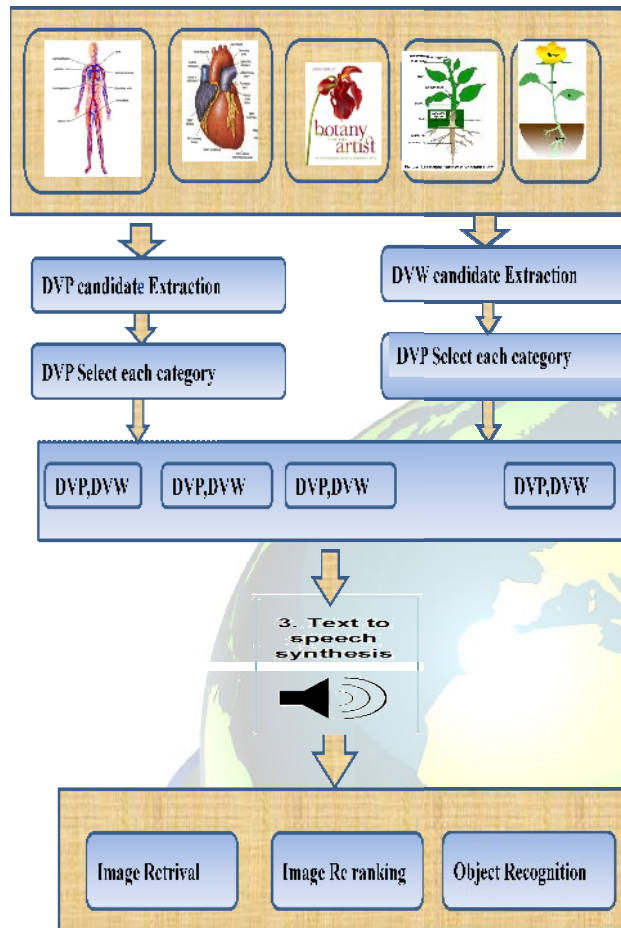


Fig 1.2 Proposed framework for DVW and DVP generation.

The proposed DVWs and DVPs are general and perform impressively in four applications: large-scale near-duplicated image retrieval, web image search reranking, object recognition and text to speech synthesis with simple nonparametric algorithms

2. RELATED WORK

To improve the descriptive power of visual vocabulary, many approaches have been proposed. In the following two paragraphs, we will review these algorithms in detail.

For visual vocabulary generated from unsupervised clustering, lots of noisy visual words can be generated from the local features in the cluttered background and large quantization error could be introduced. To overcome these shortcomings, many works have proposed novel feature quantization algorithms targeting for more effective and discriminative visual vocabularies, the authors generate discriminative vocabularies according to the Information Loss Minimization theory. In Extremely Randomized Clustering Tree is proposed for visual vocabulary generation, which shows promising performance in image classification. The visual word ambiguity and the influences of visual vocabulary size on quantization error and retrieval performance are studied. To reduce the quantization error introduced in feature space partition, soft-quantization quantizes a



SIFT descriptor to multiple visual words. In addition, to generate the visual vocabulary from single image local descriptors, the - means clustering commonly employs a general distance metric, such as Euclidean distance, to cluster or quantize the local features. This is unsatisfactory since it largely neglects the semantic contexts of the local features. With a general distance metric, local visual features with similar semantics may be far away from each other, while the features with different semantics may be close to each other. As a result, the local features with similar semantics can be clustered into different visual words, while the ones with different semantics can be assigned into the same visual words. This defection results in some incompact and noisy visual words, which are also closely related with the mismatches occurred between images. The classic visual vocabulary is used as the basis, and a semantic distance metric is learned to generate more effective high-level visual vocabulary. In a recent work , the authors capture the semantic contexts in each object category by learning a set of effective

distance metrics between local features. Then, semantic-preserving visual vocabularies are generated for different object categories. Experiments on large-scale image database demonstrate the effectiveness of the proposed algorithm in image annotation. However, the codebooks in are created for individual object categories, thus they are not universal and general enough, which limits their applications. It has been illustrated that a single local feature cannot preserve enough spatial information in images, which has been to be proven important for visual matching and recognition To combine BoWs with more spatial information, spatial pyramid matching is proposed to capture the hierarchical spatial clues of visual words in images . Video Google utilizes structure-free spatial clues in neighboring visual words to remove the mismatched visual words between images .

Recently, many works have been conducted to seek visual word combinations to capture the spatial information among visual words]. This may be achieved, for example, by using feature pursuit algorithms



such as AdaBoosting, as demonstrated by Liu *et al.* Visual word correlogram and correlation, which are leveraged from the color correlogram, are utilized to model the spatial relationships among visual words for object recognition. In a recent work, visual words are bundled and the corresponding image indexing and visual word matching algorithms are proposed for large-scale near-duplicated image retrieval. Defined as descriptive visual word combination in, collocation pattern captures the spatial information among visual words and presents better discriminative ability than the traditional visual vocabulary in object categorization tasks (image annotation), thus these proposed visual vocabularies are still not comparable with the text words, which could be used as effective features and perform impressively in various information retrieval tasks. Our proposed algorithm is different from the previous ones in the following aspects.

1) We identify the DVWs and filter the noisy visual words, thus the shortcomings of unsupervised -means clustering

are depressed. Additionally,

2) The DVWs and DVPs are capable to handle large-scale image datasets and show promising performance in three applications, i.e., large-scale image retrieval, objection recognition, image search and reranking.

3) The text file that has been created will be fed into text to speech audio descriptor to tell the user what the signage is as voice output.

3. CANDIDATE GENERATION:

3.1 Classic Visual Vocabulary Generation

Similar to existing works, we train classic visual vocabulary by clustering a large number of SIFT descriptors. We adopt hierarchical -means to conduct the clustering for its high efficiency. Though some other clustering methods such as Affinity Propagation or some recent visual vocabulary generation methods, could also be adopted, they are expensive to compute, in terms of either time or space complexity. Another advantage of hierarchical-means is that the generated visual words can be organized in the vocabulary tree and the leaf



nodes are considered as the classic visual words. Thus, with the hierarchical structure, searching the nearest visual word for a local feature descriptor can be performed efficiently.

3.2 Descriptive Visual Phrase Candidate Generation

The DVW candidates in each category are portions of the total visual vocabulary (i.e., the blue line, 32 357 classic visual words). It can be inferred that only parts of the entire visual vocabulary are descriptive to the corresponding categories. Thus, selecting DVWs from their candidates would be more efficient and reasonable than from the entire visual vocabulary.

3.3 Descriptive Visual Phrase Candidate Generation

In literature, different algorithms are proposed for capturing the spatial clues among visual words, e.g., the spatial histogram proposed in. However, these algorithms are expensive to compute, additionally, capturing complicated spatial relationships commonly causes the

sparseness of the generated visual word combinations and accumulates the quantization error introduced in the visual vocabulary. Therefore, we capture the simple co-occurring clues between two visual words, and the corresponding DVP candidates for a certain category are defined as the co-occurring classic visual word pairs in this category. Suppose visual word and co-occur in an image category. Then, the DVP candidate containing the two visual words for this category can be denoted as where is the overall average frequency of co-occurrence computed between the visual word and in image category.

4. DVW AND DVP SELECTION

4.1 DVW Selection

DVWs are defined as the representative visual words that are descriptive to certain objects or scenes. It is designed to describe certain categories, thus several unique features are desired in them.

1) If one object or scene appears in some images, the DVWs descriptive to it should appear more frequently in these



images. Also, they should be less frequent in images that do not contain such object or scene.

2) They should be frequently located on the object or scene, even though the scene or object is surrounded by cluttered background.

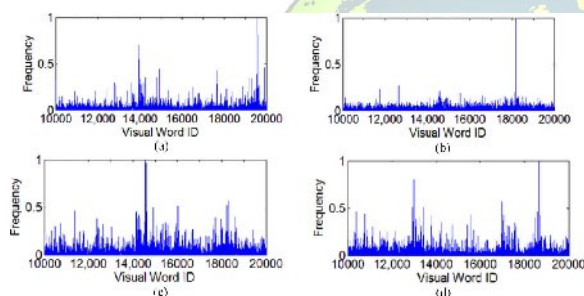


Fig. 4.1.1 shows the DVW candidates in image categories: *butterfly*, *ceiling fan*, *ant*, and *crab*. The selected DVWs in the corresponding categories are presented in Fig. 4.1.2(b). Obviously, although there are many candidates (i.e., classic visual words) on the cluttered background, most of the selected DVWs appear on the object. In order to show the descriptiveness of the selected

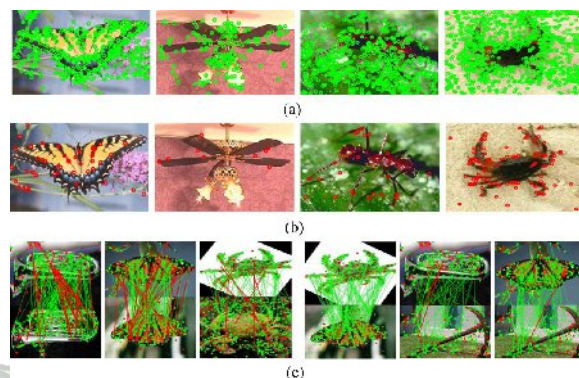


Fig 4.1.2 The DVW Selections

Fig. 4.1.2 DVW candidates, the selected DVWs, and the matched DVWs (red lines) and matched visual words (green lines) between the same and different objects. (a) DVW candidates before VisualWordRank. (b) Selected DVWs in corresponding categories. (c) Matched DVWs and visual words between same and different objects.

DVWset, the matched classic visual words and matched DVWs between same and different objects are compared in Fig. 4.1.2(c). In the figure, visual words and DVWs are denoted by green dots and red dots, respectively. The identical visual words and DVWs across images are connected by green lines and red lines, respectively. In the left three images, matches are conducted



between same objects. It can be observed that, though some DVWs exist on the background, most of the matched ones locate on the object. In the right three figures, which show the matched DVWs and classic visual words between different objects, lots of classic visual words are wrongly matched. Nonetheless, there are very few mismatches occurred between DVWs. Thus, it can be observed that DVWs are more descriptive and more robust than classic visual words.

4.2 Descriptive Visual Phrase Selection

In Fig. 4.2.1(a), the visual words are denoted as green dots and the dots connected by red lines denote the selected DVPs. Because there are dense visual words on the background in each image, it can be inferred that there would be a lot of DVP candidates generated on the object and background. As we can clearly observe, most of the selected DVPs appear on the object and maintain obvious spatial characteristics of the corresponding object.

Fig. 4.2.1(b) shows the matched DVPs across same and different objects. All of the

DVPs in the example images are denoted as red lines and the matched ones are connected by blue lines. It can be seen that, many DVPs are correctly matched between the same objects, while between images containing different objects, none of the DVPs is matched. Therefore, it can be concluded that the selected DVPs are valid and descriptive. After selecting DVWs and DVPs in each category, the final DVW and DVP set can be created by combining all of the selected candidates across different categories. Since the DVWs and DVPs are descriptive for certain objects or scenes, the final DVW and DVP sets are desired to be descriptive and general

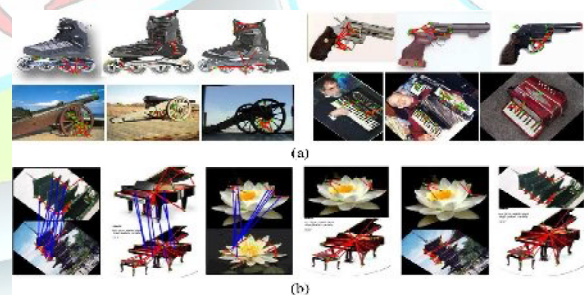


Fig.4.2.1 Selected DVPs and matched DVPs between the same and different object. (a) Selected DVPs “inline skate”, “revolve” and “cannon” (b) Matched DVPs same and different object.



4.3 Text to Speech Synthesis:

With these shortcomings of conventional audio description in mind, text-to-speech audio description (TTS AD) is proposed here in order to increase the Audio Descriptor output and to make AD more available. The idea behind it is that instead images convert into text document using DVP and DVW .the text words reading out the AD script , TTS AD can be read by speech synthesis software. Modern text-to-speech applications convert text input into a speech waveform with the use of special algorithms (Cryer and Home 2008: 5), producing an effect far more natural than speech synthesizers did a few years ago. Figure 1 below the text words convert into voice using speech synthesis software. The resulting text file with synchronised time-codes is then read by speech synthesis software while the audiovisual material is simultaneously provide the voice. Ideally, the script should be developed in consultation with a visually impaired person.

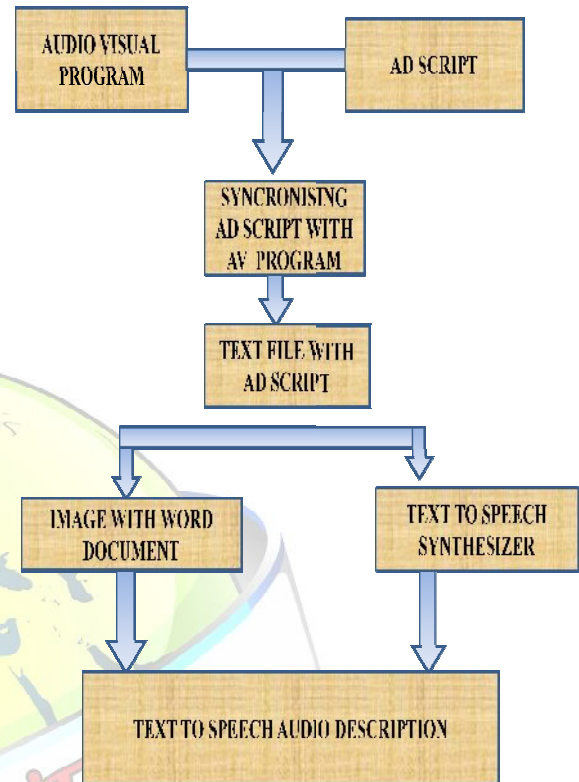


Fig.1 Text-to-speech audio description preparation process

4.3.1 Text to speech synthesis framework

5.APPLICATION AND EVALUATION

5.1Image Dataset Collection

Download the all images and then stored the data base .The downloading task is finished within one month by 10 downloading processes. In the collected raw database, categories with images less than 100 are removed. Then,from the remaining



images, we carefully select 100 categories with visually consistent single objects or scenes, by viewing the thumbnails in each category. Finally, we form a dataset composed of about 1000 images. The final dataset sufficiently covers the common visual objects and scenes. Thus, extracting DVWs and DVPs based on it would be statistically reasonable.

5.2 Large-Scale Image Retrieval Based on DVW and DVP

BoWs image representation has been proven promising in large-scale image retrieval by leveraging the classic information retrieval algorithms such as inverted file indexing and TF-IDF weighting. In this part, experiments are carried out to compare the state-of-the-art algorithms with the proposed DVWs and DVPs on large-scale near-duplicated image retrieval tasks. Near-duplicated image retrieval differs with common image retrieval in that the target images are usually obtained by editing the original image with changes in color, scale, or partial occlusion. In near-duplicated images, different parts are often cropped from the original image and

pasted in the target image with modifications. The result is a partial-duplicated version of the original image with different appearances. Our large-scale image dataset in Each image in the database is first represented as BoWs, with the classic visual word ,DVW, and DVP, respectively. Then, the images are indexed using inverted file structure. In the retrieval process, TF-IDF weighting is applied for similarity computation. All of the images with ground truth i.e., the 315 images, are used as queries. For each query, we compute the MAP, which takes the average precision across all different recall levels in the first 30 returned images.



Some results of near-duplicated image retrieval based on DVW and DVP

Fig 5.2.1 near duplicate image retrieval on DVP and DVW

5.3. Image Re-Ranking



Image search re-ranking is a research topic catching more and more attentions in recent years. The goal is to resort the images returned by text-based search engines according to their visual appearances to make the top-ranked images more relevant to the query. Generally, image re-ranking can be considered as identifying the common visual concept (i.e., scene, object, etc.), in the returned images and re-ranking the images based on how well each one fits the identified concept. DVWs and DVPs are effective in describing the objects and scenes where they are selected. Therefore, they can be utilized to measure the relevance between images and the concept. Based on this idea we proposed the DWPRank, which is detailed in Algorithm 2. We first carry out DWPRank on our database where each category contains the top 250 images returned from Google Image.



Top and last 20 re-ranked images by DWPRank

Fig 5.3.1 Top Reranking images

5.4. Object Recognition

Since DVWs and DVPs are designed to effectively describe certain objects or scenes. It is straightforward that the DVWs and DVPs of each image category should be discriminative for the corresponding object. Consequently, we utilize the object recognition task to test their discriminative ability. Moreover, this experiment is also carried out to test the validity of our algorithm in improving the discriminative power of original visual words, from which DVWs and DVPs are generated.

Classic visual word histogram is computed in each image, and histogram intersection is used as the distance metric.

DVWs perform better than the classic visual words, from which they are selected. This shows the validity of our Visual Word Rank. From the figure, it can be concluded that the combination of DVW and DVP shows the best performance and achieves improvement over the baseline by 80% in average.



6. CONCLUSION

The Image based learning into two parts, image to text process using the Descriptive Visual Words and Phrases . method to differentiate the background and foreground object, and text to speech process using the SAPI. This work shows the proof of concept of our proposed image based for assisting visually impaired people to independently learn the all image based books . The framework has successfully used voice synthesizer to tell the as select images. A novel framework is proposed to generate DVWs and DVPs for various applications utilizing a representative training set collected from web images. Comprehensive tests on large-scale near-duplicated image retrieval, image search re-ranking, and object recognition show that our selected DVWs and DVPs are more informative and descriptive than the classic visual words.

REFERENCE

[1] World Health Organization WHO: Visual Impairment and Blindness.

<http://www.who.int/mediacentre/factsheets/fs282/en/>

[2] W. I. Grosky and R. Mehrotra, "Index-based object recognition in pictorial data management", *Computer Vision, Graphics, and Image Processing*, Vol. 52, No. 3, 1990, pp. 416-436.

[3] R. Jain, R. Kasturi and B. G. Schunck, *Machine Vision*, McGraw-Hill, 1995.

[4] A. Lakshmi Ratan, O. Maron, W. E. L. Grimson and T. Lozano-Pérez, "A framework for learning query concepts in image classification", *Computer Vision and Pattern Recognition*, 1999, pp. 423-429.

[5] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. CVPR*, 2008, pp. 1-8.

[6] B. Russell, A. Torralba, K. Murphy, and W. Freeman, "LabelMe: A database and web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, no. 1-3, pp. 157-173, May 2008.

[7] Z. Si, H. Gong, Y. Wu, and S. Zhu, "Learning mixed templates for object



recognition,” in *Proc. CVPR*, 2009, pp. 272–279.

[8] J. Sivic and A. Zisserman, “Video google: A text retrieval approach to object matching in videos,” in *Proc. ICCV*, 2003, pp. 1470–1477.

[9] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT, 1998.

[10] A. Gionis, P. Indyk, and R. Motwani, “Similarity search in high dimensions via hashing,” in *Proc. Conf. Very Large Data Bases*, 1999, pp. 518–529.

[11] W. Hsu, L. Kennedy, and S. Chang, “Video search reranking through random walk over document-level context graph,” *ACM Multimedia*, pp. 971–980, 2007.

[12] H. Jegou, H. Harzallah, and C. Schmid, “A contextual dissimilarity measure for accurate and efficient image search,” in *Proc. CVPR*, 2007, pp. 18.

[13] H. Jegou, M. Douze, C. Schmid, and P. Petrez, “Aggregating local descriptors into a compact image representation,” in *Proc. CVPR*, 2010, pp. 3304–3311.

[14] B.L. Meijer. The vOICE Next-Generation Human Vision Technology.

<http://www.seeingwithsound.com/about.htm>

via multi-graph propagation,” *ACM Multimedia*, pp. 208–217, 2007.

[15] J. Liu, Y. Yang, and M. Shah, “Learning semantic visual vocabularies using diffusion distance,” in *Proc. CVPR*, 2009, pp. 461–468.

[16] D. Nister and H. Stewenius, “Scalable recognition with a vocabulary tree,” in *Proc. CVPR*, 2006, pp. 2161–2168.