Meiji University at the ImageCLEF2010 Visual Concept Detection and Annotation Task: Working notes

Naoki Motohashi, Ryo Izawa, and Tomohiro Takagi

Department of Computer Science

Meiji University

1-1-1 Higashimita, Tama-ku, Kawasaki-shi, Kanagawa, 214-8571, Japan

Email: {motohashi, takagi} @cs.meiji.ac.jp

Abstract

This paper describes the participation of the Web Science Laboratory of Meiji University in the ImageCLEF2010 photo annotation task. We devised a system that combines conceptual fuzzy sets with visual words that have become popular in the fields of image retrieval and recognition. The utility of the system was verified by comparing it with the ordinary Bag of Visual words technique. In addition, we constructed a system using Flickr User Tags that was based on the same idea. Because there are many images without tags, the system integrates visual word and tag-based methods. The utility of the integrated system was verified in an experiment.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries;

Keywords

Annotation, Bag of Visual words, Conceptual Fuzzy Sets, Visual Words Combination, ImageCLEF, Flickr User Tag

1 INTRODUCTION

The recent ImageCLEF2010 ran a Visual Concept Detection and Annotation Task [1] [2]. This task follows on from the one in 2009 and involved teams competing to create the most accurate automatic annotations of test images. The test images were annotated with words belonging to the 93 concepts of ImageCLEF. The concepts ranged from vague ones such as "outdoor" to rather concrete ones such as "dog", and they formed an ontology with a tree structure. The MIR Flickr 25.000 image dataset [3] was used as the data collection. 8,000 training images and 10,000 test images were chosen from this dataset. A big change from last year was that the participants could use Flickr User Tags. Hence, this task could be solved by using three approaches: visual only, tag only, and a mixed approach. Seventeen groups participated in this year's task, and altogether, they submitted 63 runs.

Our system combined conceptual fuzzy sets with visual words, which have recently become popular in the fields of image retrieval and recognition. We verified the utility of the system by comparing it with the ordinary Bag of Visual words technique.

This paper is organized as follows. Section 2 describes conceptual fuzzy sets.

Section 3 describes the visual words approach. Section 4 describes the technique of applying conceptual fuzzy sets to visual words. Section 5 describes our overall system. Section 6 describes the five variants that we submitted and the results of our experiments. In Section 7, we discuss the results and their implications.

2 CONCEPTUAL FUZZY SETS

A Conceptual Fuzzy Set (CFS) [4] [5] is a theory to solve the problem of situated cognition of the meaning representation for the concept, and it is based on the theory of meaning propounded by Wittgenstein. Simply put, it is a theory that aims to make a computer understand the meanings of words that change from situation to situation. For example, a dictionary gives the word "Java" three meanings: the "name of an island and region", a "brand or type of coffee", and a "programming language". Thus, a computer cannot understand which meaning to use unless it is supplied with a sufficient context. The vagueness can be resolved by considering the words appearing around Java. For instance, if "Java" co-occurs with the word "computer", Java likely refers to the programming language. A CFS for "Java" is thus the word plus links to other words related to or co-occurring with it that distinguish its meanings. We devised a technique for applying CFS to visual words that are made from clustered keypoints of images.

3 VISUAL WORDS APPROACH

First, the keypoints are extracted from an image by using SIFT [6], Second, they are clustered using k-means clustering, and the center of each cluster is taken to be a visual word of that cluster. Reference [7] indicates that image recognition and retrieval using visual words has been actively researched in recent years. The Bag of Visual words [8] is a popular technique. In this technique, the keypoints are replaced with visual words, and the image is represented by a frequency histogram of visual words. The images are retrieved on the basis of the similarity of this histogram. However, visual words have a drawback in that they are vaguer than actual words. For instance, a word like "Dog" may refer to an object in an image, but it is difficult for visual words to be used to recognize such an object because they are only for one point in the image. Some research [9] has been done in an attempt to resolve this vagueness.

4 VISUAL WORDS COMBINATION APPROACH

The Visual Words Combination Approach (VWCA) is a model that the CFS applies to visual words. Several visual words are combined in order to decrease their vagueness. The Bag of Visual words technique was originally proposed as an analogy of the Bag of Words approach, and visual words have the same meaning as words in sentences. Thus, CFS, which is useful for sentence retrieval, might also be useful for visual words. Moreover, this idea considers human psychology; someone visually recognizes an object by seeing many points at the same time. For instance, we recognize someone's face by seeing the eyes, nose, and mouth at the same time.

The Visual Words Combination (VWC) is a word set generated by combining several visual words. For example, ten pairs (combinations) can be generated from five visual words. Because we cannot judge which visual words are related to each other, we generated various combinations by using this method.

5 SYSTEM DESCRIPTION

This section explains the system that we made for this task. First, we will explain the system using only visual information. After that, we will explain the system that uses Flickr User Tags.

5.1 *Visual Dictionary*

The Visual Dictionary (VD) was our dictionary that contains visual words. Visual words were generated by quantizing keypoints extracted from images. We randomly chose 1,200 images from the 8,000 training images, and made 4,000 visual words from about 830,000 keypoints [10][11].

5.2 *CFS System (VisualOnly_CFS_of_meiji)*

The CFS system was constructed by using VWCs. The procedure, from preprocessing to annotation, is as follows.

1) The test images are converted into a frequency histogram of visual words.

2) The VWCs are generated from the result of step 1).

3) Compare the test image's VWCs with the visual words in the casebase made from the training images, and identify similar images.

4) The test images are annotated.

5.2.1 *Frequency histogram*

The test images were converted into a frequency histogram of visual words. The procedure is as follows and is illustrated in Figure 1.

1) The keypoints are extracted from the test image by using SIFT.

2) Each keypoint is assigned into a visual word in the Visual Dictionary by using the nearest neighbor

algorithm [12].

3) A histogram is made from the visual words.

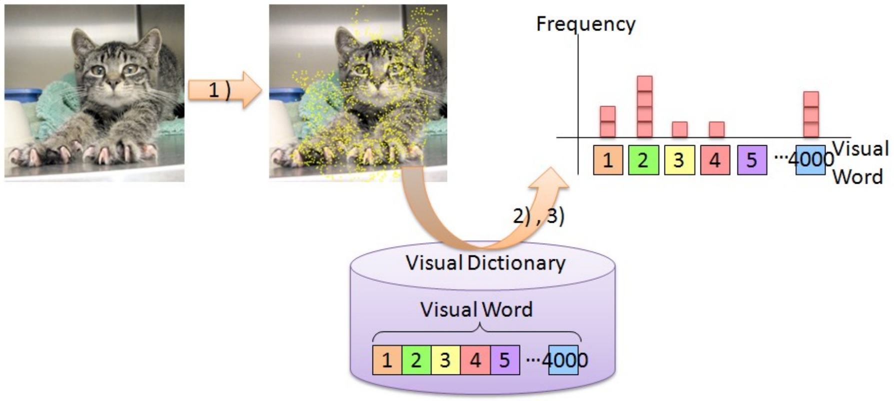

Fig. 1 Frequency Histogram

5.2.2 *Generation of VWC*

Next, the VWCs of the test image were generated. We needed to use a small number of visual words. For instance, the VWCs generated from 300 visual words in combinations of 3 would amount to $300C3$. As a result, 4,455,100 combinations would be generated. The calculation cost increases when using a lot of visual words. Hence, we used only 20 visual words that had high frequencies in the histogram and made 1,140 VWCs from them.

5.2.3 *Construction of the Casebase*

The casebase for the CFS system to annotate the 10,000 test images was made from 8,000 training images. Figure 2 shows an example of the casebase. The casebase stored relations such as "If this image has these visual words, then this image has these concepts". The training images were also converted into histograms in the same way and stored as combinations of 20 visual words.

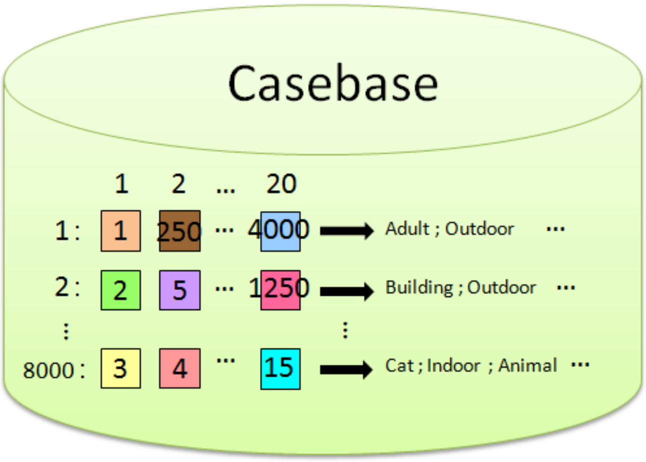
Casebase
1 2 ... 20
1 : 1 250 ··· 4000 → Adult ; Outdoor ···
2 : 2 5 ··· 1250 → Building ; Outdoor ···
⋮ ⋮
8000 : 3 4 ··· 15 → Cat ; Indoor ; Animal ···

Fig. 2 Example of Casebase (CFS System)

5.2.4 *Retrieval*

We retrieved similar images in the casebase by matching VWCs generated from the test image and the training images. If a training image had a VWC of the test image, the matching produced a hit. If it didn't have a VWC, the matching was considered to be a miss. If the number of hits for a certain training image exceeded ten, the images were considered to be similar. An example of matching is shown in Fig. 3.

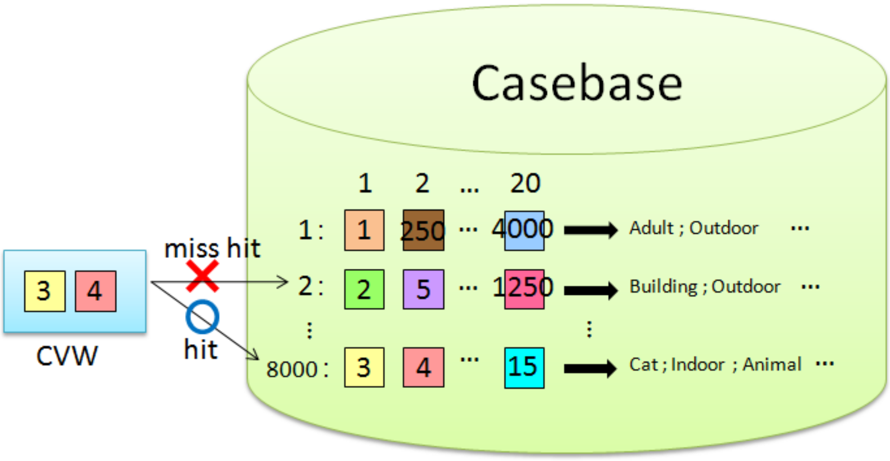

Fig. 3 Example of Matching

5.2.5 *Annotation*

The confidence was the word frequency in the retrieval results divided by the number of retrieval results. In addition, when confidence was converted into a binary value of either 0 or 1, if the concept had a confidence of 0.2 or more, the concept's binary value was taken to be 1. If the value was less than 0.2, the binary value was 0.

5.3 *Bag of Visual words system (VisualOnly_BagOfVisual wordss_of_meiji)*

The Bag of Visual words system was used as a baseline for evaluating the CFS system. The procedure to annotate the test images is almost the same as in the CFS system. The difference is that left side of the casebase of the Bag of Visual Words system in Fig. 2 is not 20 visual words but the frequency histogram in 5.2.1. Twenty training images were retrieved from the casebase, and each test image was annotated in the same way as described in section 5.2.5. The cosine measure was used to determine the degree of similarity between the test image and the training image.

5.4 *Flickr User Tag System*

We describe about a system using Flickr User Tag. The procedure from preprocessing to annotation is as follows.

1) The casebase is constructed in advance.
2) The confidence between the test image and concept is calculated.
3) The test images are annotated.

5.4.1 *Construction of the Casebase*

The casebase here is different from the casebase explained in 5.2.3. It was constructed as follows.

1) The training images for each concept are collected.
2) The tags of the collected images are collected.
3) The term frequency (TF) and the document frequency (DF) of each tag are calculated, and the tags are weighted by using the TF-IDF method.
4) The casebase (Fig. 4) paraphrasing concepts with tags is constructed by using tags having the calculated TF-IDF values. In the figure, the numerical values in parentheses are the TF-IDFs.

We shall illustrate this procedure by using the concept "Dog" as an example. The training images come annotated with concepts from ImageCLEF. Of the 8000 images, there were 211 images with the concept "dog" attached to them. Moreover, the training images had tags plus the concept. The tags of the 211 images were collected. The TF-IDF value of each tag was then calculated and the tag and its value were stored.

The casebase described in sec. 5.2.3 stored the 8,000 cases, but the casebase here stored 93 cases. The idea behind this casebase comes from the theory of meaning propounded by Wittgenstein. According to the theory of meaning, the meaning of a word can be represented by another word. In the example of Figure 4, “partylife” is paraphrased by “concert” and “fireworks”.

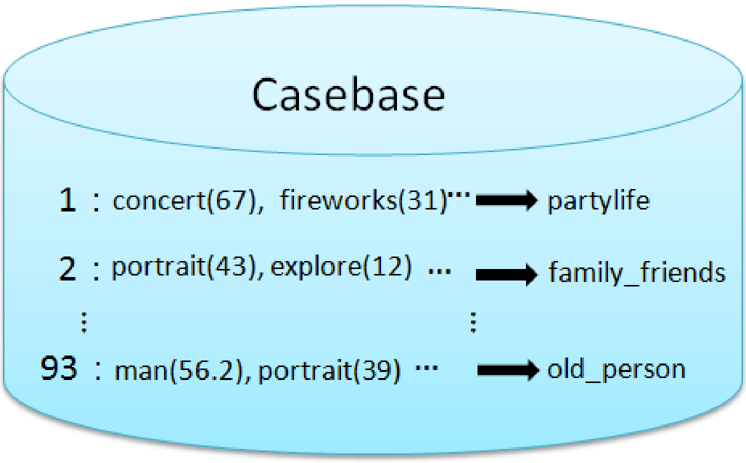
Casebase
1 : concert(67), fireworks(31)··· → partylife
2 : portrait(43), explore(12) ··· → family_friends
⋮
93 : man(56.2), portrait(39) ··· → old_person

Fig. 4 Example of Casebase (Flickr User Tag System)

5.4.2 *Confidence Calculation*

Next, the confidence between the test image and each concept is calculated. The procedure is as follows (an example is shown in Fig. 5).

1) The tags of the concept and the test image are matched, and the TF-IDFs of the tags that become hits are added together. The total TF-TDF is stored in the concept.
2) Step 1) is repeated for all concepts.
3) All concepts are regularized by the maximum value of each concept, and this value is assumed to be the confidence.
4) The steps from 1) to 3) are repeated for all test images.

When the tag of the test image and concept were compared, combinations like those in section 5.2.2 were not always generated because there were many unsuitable tags. For instance, there were spelling mistakes like "Mweeting" and sentences like "Girlecstaticallydancingondnbtuneswhileholdingaglass ofbeerinherhand".
Moreover, about 1,000 images of the 10,000 test images didn’t have any tags; i.e., only about 9,000 images were annotated.

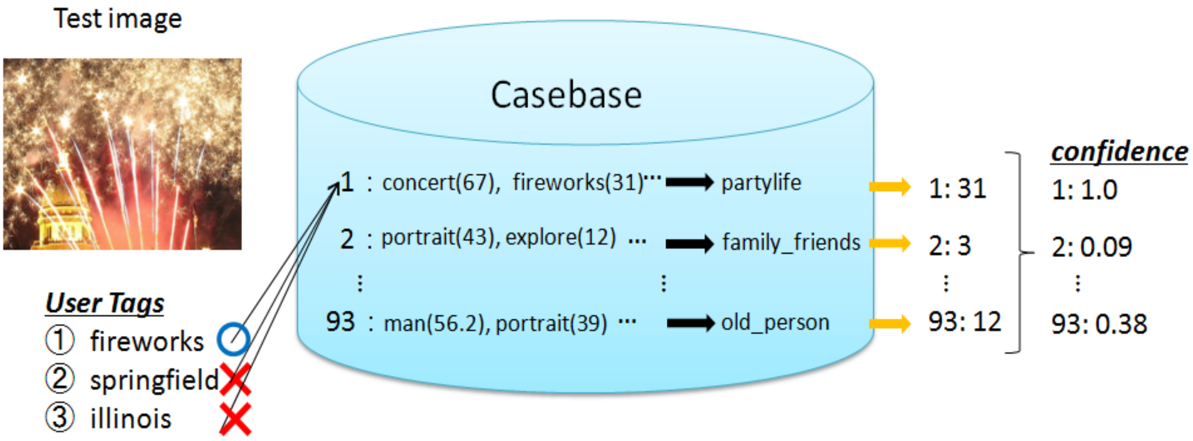
Test image
User Tags
① fireworks ○
② springfield ✗
③ illinois ✗
Casebase
1 : concert(67), fireworks(31)··· → partylife
2 : portrait(43), explore(12) ··· → family_friends
⋮
93 : man(56.2), portrait(39) ··· → old_person
1: 31
2: 3
93: 12
confidence
1: 1.0
2: 0.09
93: 0.38

Fig. 5 Example of Calculation Confidence

5.4.3 *Conversion to Binary*

Every concept, not every image, was considered when the confidence was converted into a binary value.
For every concept, the confidences of all test images were sorted in descending order. The binaries of the test images that had the top n confidences became 1 because these images may have had a strong relationship with each concept. The binaries of the other test images were 0.
The value of n was determined stochastically. For instance, in the example described in 5.4.1, the concept

"Dog" was annotated on 211 of the 8000 training images. Note that when calculating the number of test images that are annotated to 9,000 images stochastically, the concept "Dog" is sure to be annotated to 237 test images in 9000 training images.

Figure 6 shows an example of converting the concept "1: partylife". In this case, n is 2.

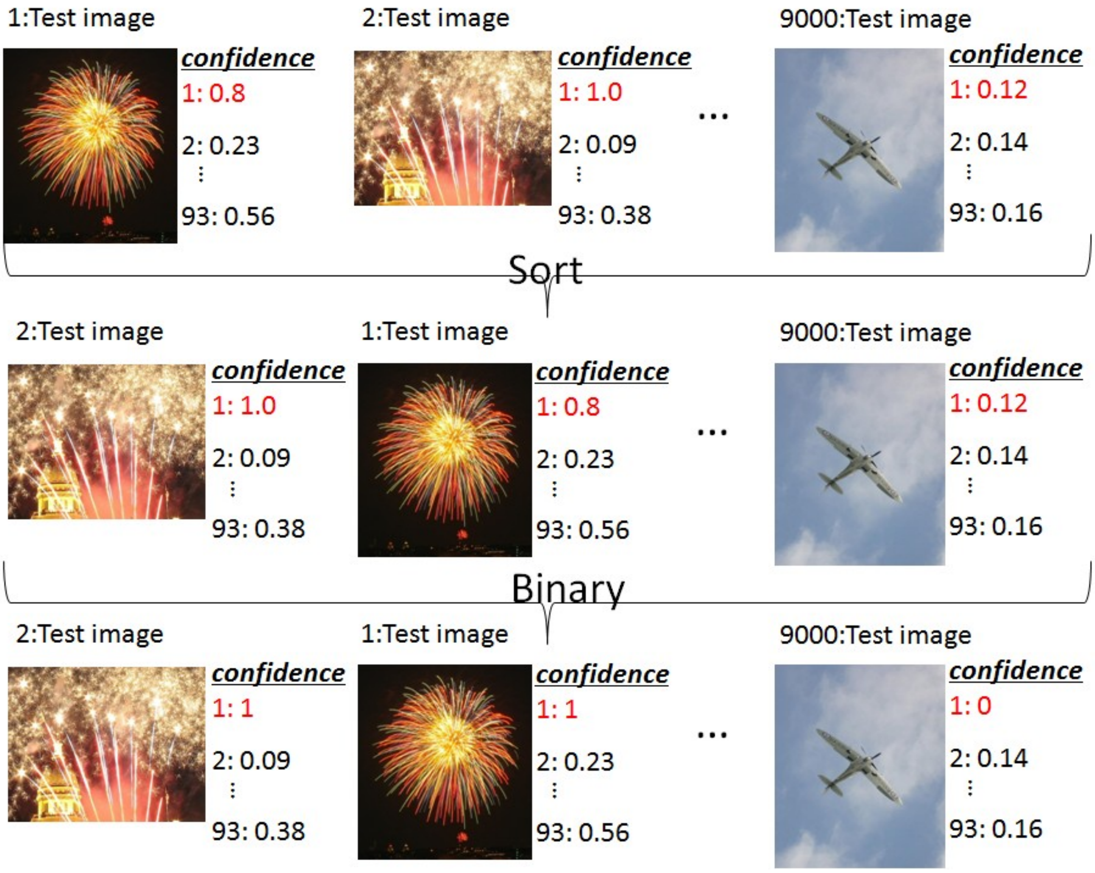

Fig. 6 Example of converting a confidence score into a binary value

6 SUBMISSIONS AND RESULTS

Below, we describe the five systems that we submitted and their results.

6.1 *Submitted Systems*

1. VisualOnly_CFS_of_meiji
   This system is the CFS system.
2. VisualOnly_BagOfVisual words_of_meiji
   This system is the Bag of Visual words system.
3. Mixed_CFS_and_Tags_of_meiji
   9,000 images that have tags attached to them are annotated by using the Flickr User Tag System. The remaining 1,000 images are annotated with the CFS system. This system integrated the results of systems 1 and 2.
4. Mixed_Tag_Based_CFS_of_meiji.txt
   This system attempts to improve the result of system 1 by using the Flickr User Tag System.
5. Mixed_Text_Based_BagOfVisual_words_of_meiji
   This system attempts to improve on the result of system 2 by using the Flickr User Tag System.

The improvements to system's 4 and 5 were to refine the annotation results of systems 1 and 2 by using the results of the Flickr User Tag System. The confidence of each concept was calculated by using the Flickr User Tag System on all test images. Seven concepts that had high confidence were employed to improve the results of systems 1 and 2. The confidences of these seven concepts of systems 1 and 2 added up to 0.5. They were converted into a binary value of 1 because they exceeded 0.2.

6.2 Results

First, we can see that system 2, which uses the original Bag of Visual words, is more accurate than system 1 that uses CFS. The CFS result was not as good as we expected because the number of visual words was limited in order to reduce the processing cost. In short, the image was represented by only 20 visual words, and this is not enough information. Moreover, there is a research result showing that visual words whose frequencies are high are not so important [13]. Thus, the visual words were improperly selected. On the other hand, the accuracy went up when the tags were used. However, the improvement was small. The reason is that the seven concepts were compulsorily annotated when the results of system 1 and 2 improved and this led to the possibility that noise concepts were annotated. System 5 was the most accurate of the systems that we submitted. We ranked 23rd out of 63 runs and 6th among 17 teams.

Table 1: Results of Meiji Systems

	MAP	*Average F-ex*	*OS with FCS*
1: VisualOnly_CFS_of_meiji	0.1776	0.5489	0.3563
2: VisualOnly_BagOfVisual wordss_of_meiji	0.2221	0.5587	0.3634
3: Mixed_CFS_and_Tags_of_meiji	0.3131	0.5103	0.4276
4: Mixed_Tag_Based_CFS_of_meiji	0.3047	0.5657	0.3603
5: Mixed_Tag_Based_BagOfVisual words_of_meiji	**0.3258**	**0.5724**	**0.3663**

7 Conclusion

The system that improved the results of the Bag of Visual words method by using Flickr User Tags was the most accurate one that we submitted. This means that annotations using text information like tags are effective. Unfortunately, the system did not have as good a result as we expected. We think the problem was that the selection and the number of visual words were not suitable (see sec. 6.2). It will be necessary to verify which visual words to select in the future.

References

[1] ImageCLEFphotoAnnotation2010, http://www.imageclef.org/2010/PhotoAnnotation

[2] Stefanie Nowak and Mark Huiskes. “New Strategies for Image Annotation: Overview of the Photo Annotation Task at ImageCLEF 2010.In the Working Notes of CLEF 2010”, Padova, Italy, 2010.

[3] MIR Flickr 25.000 image dataset, http://press.liacs.nl/mirflickr/

[4] T. Takagi, A. Imura, H. Ushida, and T. Yamaguchi, “Conceptual fuzzy sets as a meaning representation and their inductive construction”. International Journal of Intelligent Systems, Vol. 10, pp. 929–945, 1995.

[5] T. Takagi, A. Imura, H. Ushida, and T. Yamaguchi, “Multilayered reasoning by means of conceptual fuzzy sets”. International Journal of Intelligent Systems, Vol. 11, pp. 97–111, 1996.

[6] D.G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints”, International Journal of Computer Vision, Vol. 60, No. 2, pp. 91-110, 2004.

[7] J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," *Proc. Int'l Conf. Computer Vision,* 2003.

[8] G. Csurka, C. Bray, C. Dance and L. Fan, “Visual categorization with bags of keypoints”, In Proc. of ECVWCorkshop on Statistical Learning in Computer Vision, pp. 59-74, 2004..

[9] Jan C. van Gemert, Cor J. Veenman, Arnold W.M. Smeulders, Jan-Mark Geusebroek, "Visual words Ambiguity", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 7, pp. 1271-1283, June 2010, doi:10.1109/TPAMI.2009.132

[10] Koen E. A. van de Sande, Theo Gevers and Arnold W. M. Smeulders, "The University of Amsterdam's Concept Detection System at ImageCLEF 2009", In CLEF working notes 2009, Corfu, Greece, 2009.

[11] Alexander Binder and Motoaki Kawanabe, "Fraunhofer FIRST's Submission to ImageCLEF2009 Photo Annotation Task: Non-sparse Multiple Kernel Learning", In CLEF working notes 2009, Corfu, Greece, 2009.

[12] Nearest neighbour algorithm, http://en.wikipedia.org/wiki/Nearest_neighbor_algorithm

[13] F. Jurie and B. Triggs, “Creating Efficient Codebooks for Visual Recognition,” Proc. Int’l Conf. Computer Vision, pp. 604-610, 2005.