Annotation and Retrieval System Using Confabulation Model for ImageCLEF2011 Photo Annotation

Ryo Izawa, Naoki Motohashi, and Tomohiro Takagi

Department of Computer Science
Meiji University
1-1-1 Higashimita, Tama-ku, Kawasaki-shi, Kanagawa, 214-8571, Japan
Email: {rizawa27, motohashi, takagi} @cs.meiji.ac.jp

Abstract. We describe systems we developed and submitted to ImageCLEF2011. These systems are applied to the confabulation model which is based on the human brain structure and are used for image recognition and retrieval. Simply put, these systems involve the co-occurrence of visual words. The visual words approach has recently become a focus of attention in the field of image recognition. We propose a new approach that differs from the ordinary technique and evaluate its effectiveness by participating in this workshop.

Keywords: Bag of Visual Words, Confabulation Model, ImageCLEF, Flickr User Tag, Annotation

1 Introduction

The annotation task [1] at ImageCLEF2011 is a task designed to evaluate the accuracy of "automatic image annotation." In this task, 99 concepts such as "Dog," "Sea," and "flower" are provided to the participants. The concept based retrieval task [1], which is extended from the annotation task, involves retrieving images that are relevant to the provided topics (40 topics). Participation in this task is easy because we can use the resources of the annotation task. The MIR Flickr 1 million image dataset [2] is used as a dataset for these tasks. In addition, we can use the Flickr User Tags attached to each image. These tasks are carried out by using three approaches: "Visual Information only," "Flickr User Tags only," and a "Multi-modal" approach. Then we evaluated which approach was most effective.

There is currently a need for image retrieval technology to search for target images from a large number of images. Thus, a lot of research is being done on image recognition systems. In particular, generic object recognition systems, in which a general object is recognized by searching for its name, have been studied closely for the last several decades. However, such systems have not yet reached a level of practical use. The main reason for this is that there is a semantic gap between the image features and recognition.

Recently, the Bag of Visual Words [3] technique has been very popular in image recognition and retrieval. In this technique, the key points extracted by Scale Invariant

Feature Transform (SIFT) [4] are quantized to patterns that are called visual words, and an image is represented as a frequency histogram of these words. This method performs well when the objective is to simply represent an image. Therefore, many techniques using this approach have been proposed.

We propose a mixed approach that combines the Bag of Visual Words technique with the Confabulation model. The Confabulation model is based on the structure of the human brain. Although recent techniques have used machine learning techniques such as the Support Vector Machine (SVM), several problems may arise. For example, the computational cost may be large if the amount of training data is excessive; additionally, the classifier that is used needs to be constructed to identify many categories. We attempt to resolve these problems by using a recognition approach that is familiar to human beings. Although the Confabulation model was applied to natural language processing in [5] [6] [7] [8], our goal is to evaluate the accuracy when it is applied to image processing.

This paper is organized as follows: Section 2 describes the Confabulation model, and section 3 explains the idea extended from it. Section 4 details our submitted systems, and section 5 explains their results. Finally, a conclusion is given in section 6.

2 Confabulation Model

In this section, we describe the confabulation model that forms the foundation of our study and explain how we apply this model to generic object recognition.

2.1 Confabulation Model

Hecht-Nielsen advocated the theory, based on brain science, that human recognition is caused by the co-occurrence of multiple attributes. He named this model the Confabulation model and conducted an experiment to predict which word would appear next in a consecutive list of words (Fig. 1.).

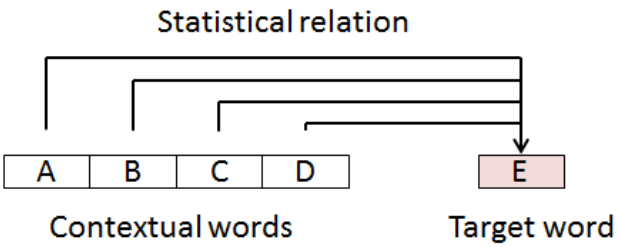

Fig. 1. Overview of confabulation model

Here, we explain the co-occurrence of multiple attributes by offering a concrete example. For instance, humans can identify an apple by the co-occurrence of senses: the color is red, the shape is round, and the surface is smooth. Therefore, if an

unknown object has these attributes, humans can determine that it is an apple. Repeating the same experience makes it easier for them to identify the object. Hecht-Nielsen called this confidence of prediction cogency, and it is defined by (1).

$$cogency(X,Y) = P(X \mid Y) \tag{1}$$

"P" is a backward probability based on a physiological experiment of the brain. A target word can be predicted by multiplying each cogency, and when the cogency of each attribute α, β, γ, and δ is high, the cogency of ε becomes high by calculating (2).

$$conf(\varepsilon, \alpha\beta\gamma\delta) = P(\varepsilon \mid \alpha)P(\varepsilon \mid \beta)P(\varepsilon \mid \gamma)P(\varepsilon \mid \delta) \tag{2}$$

We apply this to visual words in this system, and the system recognizes an object by these co-occurrences.

2.2 Recognition technique using confabulation model

The Bag of Visual Words technique makes it possible to recognize the object by using training in the similarity of visual words. We focus on this feature and assume that images in the same category have common features, and that the category can be identified by the co-occurrence of these features. For example, in the category "city," it is highly possible that the associated images contain both "road" and "building." In other words, if a query has the features of these objects, a "city" may show up in this image. We constructed a training model using visual words such as those in Fig. 2 and use it to recognize categories.

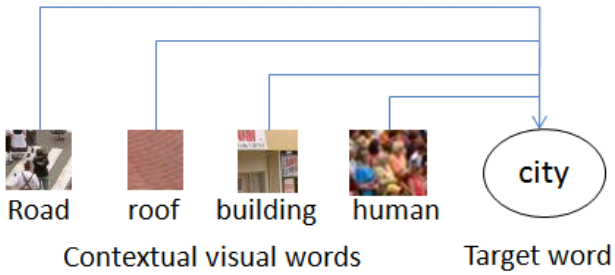

Fig. 2. Learning model using Confabulation model

3 Recurrent Confabulation Model

We propose a recurrent confabulation model as an extension of the Confabulation model described in the previous section. This model is based on the mechanism of human learning and recognition.

When a human tries to understand an unknown situation, he thinks recurrently, starting from the most extreme condition that he can think of. If there is no correspondence between what the person sees and information he already knows, it seems that he gradually relaxes the conditions to attempt to understand the situation. For example, take the case of a person seeing a green apple for the first time. Few thoughts are necessary to recognize this object. The person identifies the object as a green apple because it corresponds closely to the known characteristics of previously experienced apples, except that the color is green. In other words, a human chooses "the thing which seems most likely to be the answer" that corresponds to known information (e.g. apples are red, round, and sweet) that excludes only color information as an answer. That is to say, a human repeats the process of relaxing the conditions if a solution is not found in the most extreme condition and looks for a solution again. Humans naturally judge that something that seems most likely to be the answer is the answer.

The recurrent confabulation model was made by applying this idea to linguistic processing by computer. This model uses both the recognition results under strict and more relaxed conditions. However, because there may only be a few retrieval results under strict conditions, the results obtained in more relaxed conditions, in which the number of combinations of attributes is reduced, compensate for that.

In this case, for example, even though there are only 10 recognition results using color, shape, and taste, the system can make up for this lack by using the results from the relaxed condition, e.g., using just shape and taste, if there are more than 10 recognition results.

4 System Description

We describe our systems in this section.

4.1 Annotation Task

We applied the confabulation model to the system submitted to the annotation task.

We collected training images of every concept in the dataset and tried to represent the concepts by extracting common features from their images.

We used the features of Bag of Visual Words and color. This is because in recent years image recognition using Bag of Visual Words has reportedly had high accuracy, and we thought that color was an important feature depending on the concept.

4.1.1 Flickr User Tag Approach

We describe the approach using Flickr user tags before explaining the visual features approach.

The overall process is shown in Fig. 3.

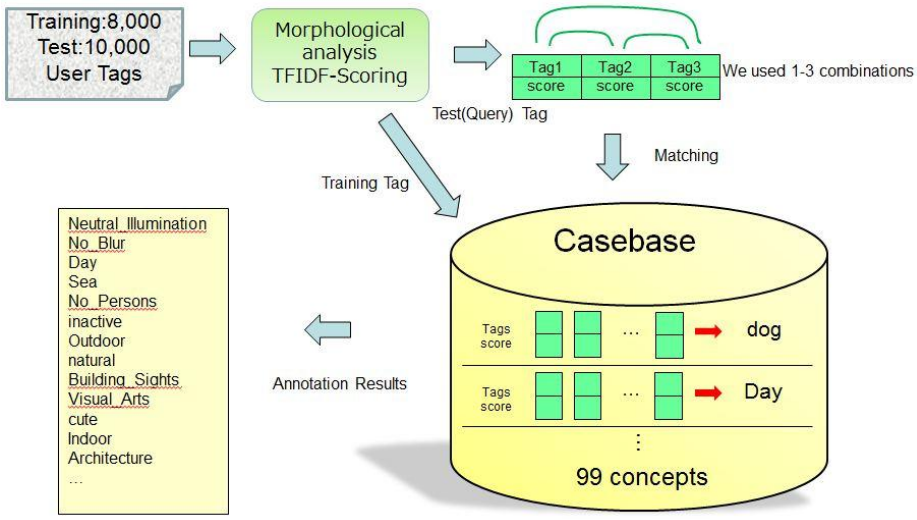

Fig. 3. Schematic of annotation system using Flickr user tags

We perform a morphological analysis of all the user tags of the training images. Next, a TFIDF score is given to each tag. Then, the system stores the correspondence information for the tags and concepts in a casebase.

When using a test image, we perform morphological analysis similarly, and the system matches the tags attached to the test image to the tags of the casebase.

However, this matching is probably vague because the system checks a match by using one word.

In the "apple" example, when we presume that the object is an apple, using combined information such as "red and round" makes it easier to judge whether the object is an apple than by using a single piece of information such as "red" or "round".

Therefore, we establish combinations between words to reduce vagueness. A matching score is calculated as shown in (3).

$$Score = \sum_{C_j \in query} \left\{ \sum_{w_i \in C_j} TFIDF(w_i) \cdot TFIDF(hit) \right\} \tag{3}$$

Here, C is a combination, w is a tag in the combination, and *hit* is a tag in the corresponding concept.

Finally, the system sorts these scores for the concepts and outputs them as results.

In addition, queries are generated by changing the number of combinations (b of aCb) from b=1 to 3. We output the results using these queries and apply the recurrent confabulation model to them (Fig. 4.).

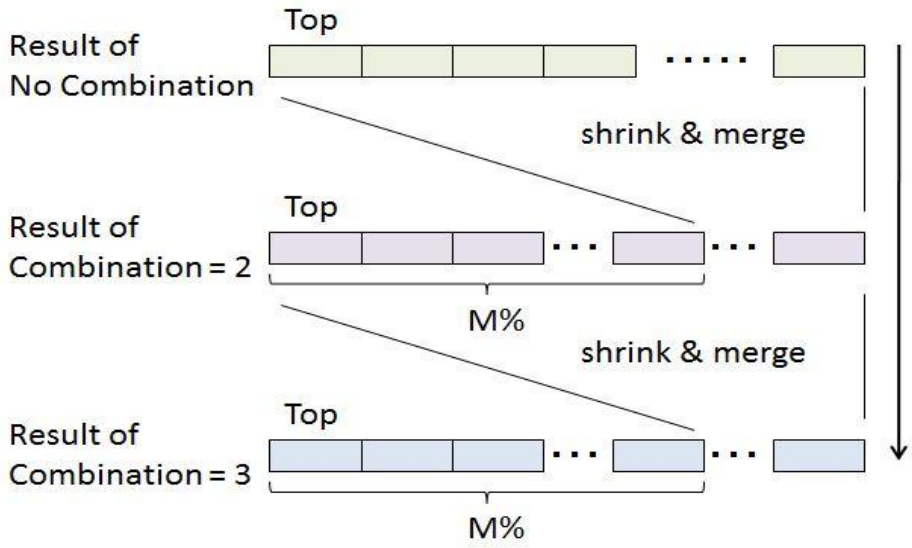

Fig. 4. Overview of recurrent confabulation model

i. First, we output results for different numbers of combinations. Then, concepts that are annotated for both a small and large number of combinations are excluded from the set of results with a small number of combinations. As a result, there are no concepts included that appear in annotation results for many combinations.

ii. Next, we take the top M% of the high-ranked annotation results of N combinations. We use Eq. (4) to convert the score so that the best results for the number of combinations N-1 are ranked highest and the others are integrated with the top results. In other words, we integrate the score of the i rank for the number of combinations = N-1 with the number of combinations = N. We decided to set M = 30% from the experiment; however, it seemed that the optimum value of M will change depending on the corpus.

$$score_N(i) = \frac{score_N(M) \cdot score_{N-1}(i)}{score_{N-1}(top)} \tag{4}$$

iii. We use the annotation results integrated up to the number of combinations = N and the result of the number of combinations = N+1 and follow the procedure in ii. This is recursively repeated.

Therefore, the high-ranked results do not include much noise. The low-ranked results are vague; however, we can expect them to include many candidate concepts.

4.1.2 Bag of Visual Words Approach

i. Construction of visual words

Constructing visual words is an important factor when we represent all images using a histogram. Although many techniques for this have been proposed over the years, the most effective technique to reveal how visual words are constructed has yet to be revealed. In our study, we had to process many images in a short period (about one month), so we only used 1,000 randomly selected images from 8,000 training images. We constructed 500 visual words by clustering SIFT features extracted from these images. We clustered the features using the common k-means clustering algorithm.

ii. The representation of image using visual words

Our goal was to recognize test images by using the Confabulation model. We show the detailed process we used to achieve this (see Fig. 5) in this section.

First, the training images are represented as a histogram just as in step i. Then, the visual words that are used are those obtained in step i.

Second, each visual word is weighted using TFIDF. TFIDF is a popular technique in the field of text retrieval and is used for weighting the words in a text. We used 8,000 training images. In short, the maximum document frequency (DF) is 8,000. We assume that 50 visual words of an image that have high weighted scores have a strong relation to that image, and we keep these words for the third process.

Third, we calculate the occurrence probability of each visual word in each concept. As shown in Fig. 5, for instance, a visual word representing "ear" occurs three times in three images of the concept "horse." Thus, its probability is 1 (3/3).

$$P(t|vw_i) = \frac{df_i}{num_t} \tag{5}$$

The occurrence probability of each visual word is calculated by (5). Here, "num" represents the number of images of each concept *t,* and "df" shows the number of images having i-th visual words. However, this probability cannot deal with rare cases. For instance, when comparing one image with 100 images, the importance of a visual word differs between 1 of 1 time probability and 100 of 100 times. These probabilities concurrently become 1. However, the former is obviously a rare case, and its degree of confidence is probably low. On the other hand, it is highly possible that the visual word in the latter case is important because it comes out 100 times in 100 images. To deal with this problem, we decided to multiply the logarithm of the "num" by the occurrence probability as a weight and calculate the score using (6). The rare case can be removed by using this equation, and the occurrence probability of visual words appearing in many images can be increased.

$$P(t|vw_i) = \frac{df_i}{num_t} \cdot \log(num_t) \tag{6}$$

By using the above process, a training pattern such as that in Fig. 5 is constructed. Finally, 99 training patterns are made.

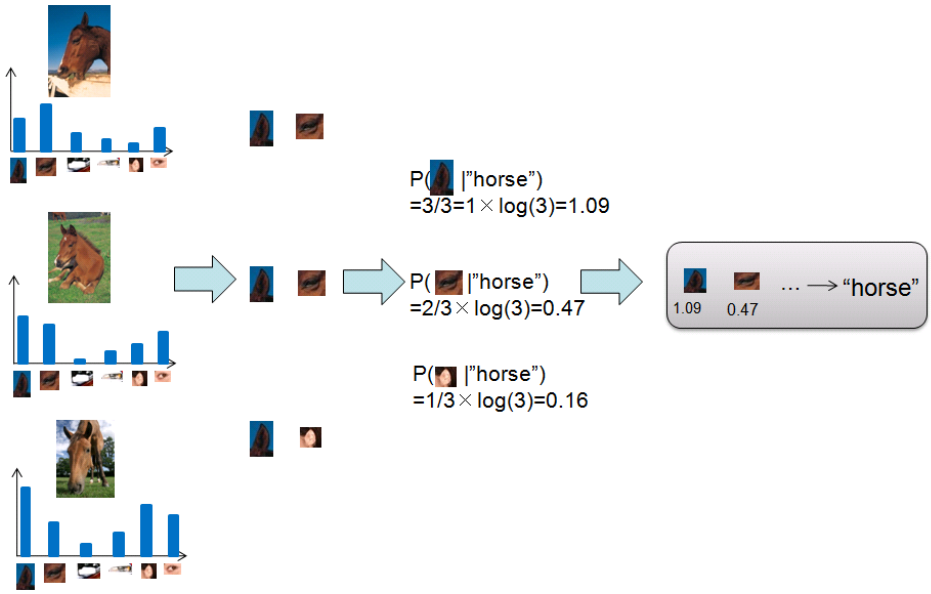

Fig. 5. Process of producing the characteristic visual words of each concept.

iii. Recognition

Here, we show how to calculate the similarity between the test image and each concept. Fig. 6 represents the process flow.

First, a test image is transformed into a visual word histogram, and each word is weighted by TFIDF the same way as in the training step. Visual words contained by many training images are obviously noise and are not suitable for use with recognition. Therefore, by using TFIDF, the weights of these visual words can be decreased.

Second, 18 visual words having a high score are used for matching. This number was obtained experimentally. Then, the matching takes place by using the recurrent idea described in chapter. 3. In other words, the number of visual words used varies from 18, 15, and 7. The parameter of M in 4.1.1 is set as 20.The matching equation is shown as (7).

$$sim(concept) = \prod_{i=1}^{18} P(i) \tag{7}$$

Here, “i” represents the index of 18 visual words, and P(i) shows the occurrence probability of the i-th visual word in the concept. As a result, the top 20 concepts having high similarity are annotated to a test image.

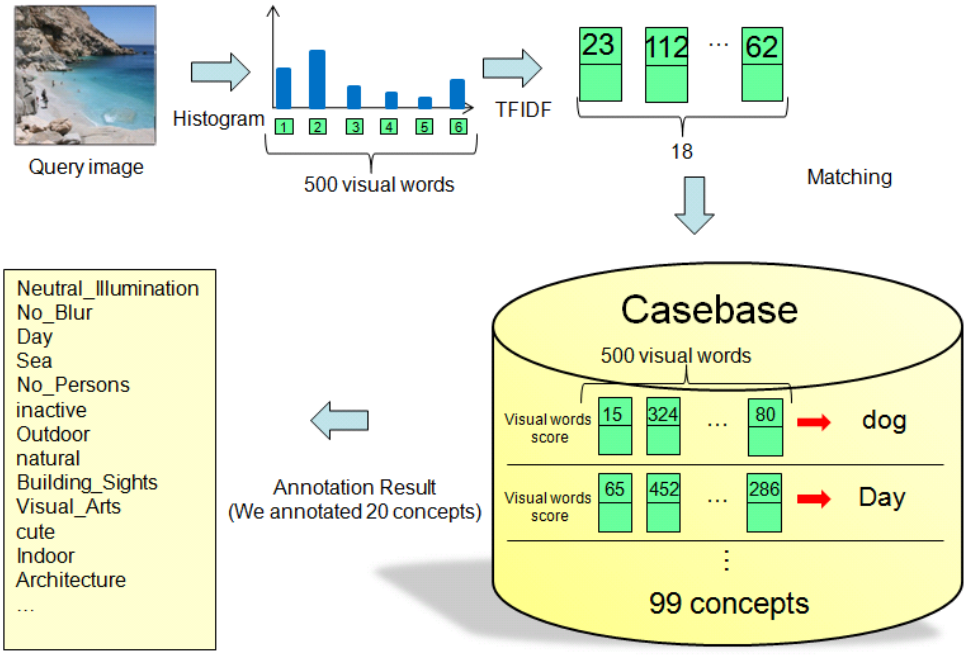

Fig.6. Recognition process. We construct 99 training patterns for each concept. Each pattern is kept as one case in the casebase. Matching takes place by using 18, 15, and 7 visual words having a high TFIDF weight for the test image. As a result, the top 20 concepts having high similarity are annotated to this test image.

4.1.3 Color Approach

i. Representation of images using color feature

The idea for this is basically the same as visual words. The color features representing a concept are trained as one pattern. We use RGB as the color feature in this task. However, each color (for example, R) has 256 tones; thus, the histogram has about 16 million dimensions (256×256×256). To avoid an expensive calculation cost, we reduce each color from 256 tones to 4. Thus, the histogram can be represented in 64 dimensions.

With the color feature, we cannot weight each color (here, each dimension) by TFIDF. Hence, we use the top 10 colors out of the 64 that have high occurrence frequency. The training patterns are created using the same procedure as for visual words.

ii. Recognition for use with color feature

The recognition process is the same as for visual words. However, the weighting cannot be carried out for the color feature (as described above). Thus, 15 color features having high frequency are used, and color features are not applied in the recurrent approach. This number is also determined experimentally. The matching equation is shown as (8).

$$sim(concept) = \prod_{i=1}^{15} Log_e(P(i)) \quad (8)$$

In this case, P(i) is not probability but occurrence frequency. This score is usually a few thousand. Thus, we transform it into a logarithm because an overflow occurs if P is multiplied many times.

4.1.4 Integration

We also construct the integration system using many feature values (In short, tag, visual words, color feature). The training and recognition are conducted by using co-occurrences of each annotation result, which is based on Confabulation. This can be simply implemented because each similarity of the annotation result is just multiplied respectively. When this method is applied, the multiplied result may be zero if there is a feature that has zero similarity. Thus, we need to make improvements in order to resolve this problem in the future.

4.2 Concept Based Retrieval Task

The overall flow of the concept based retrieval task is shown in Fig. 7.

First, all 200 thousand images are annotated to concepts by using the annotation task system.

Second, the case base is constructed by arranging the similarity of these images in descending order (see Fig. 7(B)). Only 5,000 of the 200 thousand images are memorized because only 1,000 retrieval results are needed.

The topic images are also annotated to concepts respectively. The matching is conducted between the concept of the test image and the one in the casebase. We use 20 concepts from the annotation result of the test image (see Fig. 7(A)).

Here, we describe a retrieval example. It is assumed that the result of the test image has two concepts (“cat” and “Day”), and Img0111 is retrieved. The similarity of the concept “cat” of Img0111 and the test image is 0.8 and 0.5 respectively, and that of “Day” is 0.4 and 0.7, respectively. Then, the similarity of the retrieved Img0111 is 0.68 ($0.5\times0.8 + 0.7\times0.4$). In fact, the topic has several images, and these also retrieve the relevant images in the case base. As a topic result, the final similarity for Img0111 is the maximum in the result of each topic image. In short, if the value of topic image 1 is 0.87, and it is 0.98 for topic image 2, the final similarity of Img0111

becomes 0.98. This process is used because the annotation result may be unstable. Even if several images are contained in the same topic, the appearances of these images may be different. Thus, a difference regarding annotated concepts or similarity occurs. Then, concepts such as ”dog” that have only a few training images may not have a high ranking in the annotation result of all images in one topic, but concepts such as ”Day” may be at the top. Thus, in all topics, the images relevant to “Day” are retrieved. This is a negative effect from the logarithm used to calculate the occurrence probability. By using this approach, we can avoid this problem. Finally, 1,000 images having high similarity are retrieved.

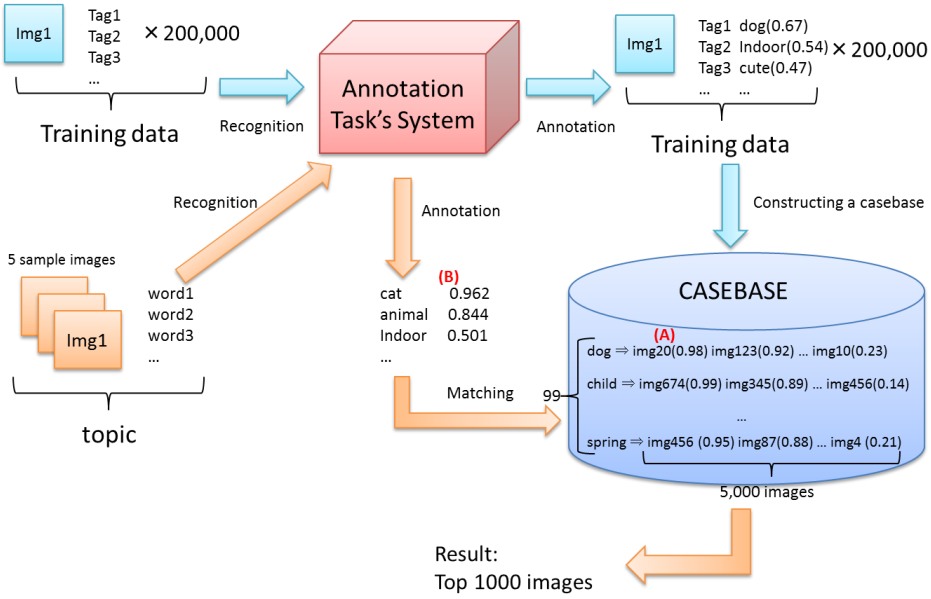

Fig. 7. Detailed system of concept based retrieval task

5 Submissions and Results

This section explains our submitted systems and the results.

5.1 Annotation Task

We submitted the following five systems to the annotation task.

i. meiji_tag_r
   Method using only Flickr user tags
ii. meiji_vw_r
   Method using only visual words

iii. meiji_vw_color_r
Method using co-occurrence of visual words and color
iv. meiji_vw_color_tag_r
Method using co-occurrence of visual words, color, and tags
v. meiji_vw_tag_r -vw_r
Method using co-occurrence of visual words and tags
However, there were about 1,000 images that did not have any tags. Therefore, if their co-occurrence is used as described in 4.1.4, we cannot annotate concepts to these images. To make up for this, this system reflects the result using only visual words in the images that have few annotations.

The results of the annotation task are listed in Table 1.

Table 1. Results of Annotation Task

Run	MAP	F-ex	SR-Precision
meiji_vw_r	0.188589	0.471923	0.43197626
meiji_tag_r	0.303823	0.458523	0.49061212
meiji_vw_color_r	0.20408	0.451822	0.4520852
meiji_vw_color_tag_r	0.287871	0.423201	0.4799387
meiji_vw_tag_r -vw_r	0.287695	0.495162	0.46919692

The best execution result in each evaluation measure was “meiji_tag_r” in MAP and SR-Precision and “meiji_vw_tag_r-vw_r” in F-ex. The result using only tags was best, and visual information reduced the precision. Because a tag is a word, its meaning is clear. Thus, the high-ranked concepts using tags had good reliability. (For example, let us consider an image annotated to the concept “sea.” In visual words, a visual word representing “sea” is vague; however, in tags, “ocean” is close enough to “sea” to be correct.) Nevertheless, although the precision was low for visual information when the system integrated each feature by using co-occurrence, we integrated them by equivalent weight.

It is possible that the vagueness of the visual words reduced the correct high-ranking concepts obtained using tags. However, when we used color information and only visual words, precision was not good either. It is therefore necessary to continue studying techniques that apply the Confabulation model to image processing.

5.2 Concept Based Retrieval Task

We submitted the following 10 systems to the concept based retrieval task.

i. meijiTr
Method using only Flickr user tags.
ii. meijiVr
Method using only visual words

iii. meijiVTr
Method using co-occurrence of visual words and tags

iv. meijiVCTr
Method using co-occurrence of visual words, color, and tags

v. meijiVTVr
Method using co-occurrence of visual words and tags
The additional process of this system is the same as "meiji_vw_tag_r - vw_r" which was submitted to the annotation task.

vi – x. The files in which the end of i – v was changed from "r" to "n"
The execution results in which the recurrent confabulation model was not implemented. These were submitted to confirm the effectiveness of this model.

The results of the concept based retrieval task are listed in Table 2.

Table 2. Results of concept based retrieval task

Run	MAP	P@10	P@20	P@100	R-Prec
meijiVn	0.0017	0.015	0.015	0.0197	0.0151
meijiVr	0.0013	0.0125	0.0125	0.0185	0.0122
meijiTn	0.0213	0.0675	0.0862	0.0865	0.0648
meijiTr	0.0227	0.09	0.0962	0.0865	0.0628
meijiVTn	0.0408	0.175	0.1513	0.1432	0.1053
meijiVTr	0.0325	0.1425	0.125	0.114	0.0867
meijiVCTn	0.0444	0.1625	0.165	0.1465	0.1053
meijiVCTr	0.0333	0.13	0.12	0.113	0.0824
meijiVTVn	0.042	0.1725	0.1437	0.1417	0.1061
meijiVTVr	0.0327	0.1275	0.1138	0.1135	0.0847

In our experiment, we knew that precision improves when text is processed by using the recurrent confabulation model. Therefore, one of our purposes was to investigate how effective the model was when applied to visual information.

First, we compared a result using the recurrent confabulation model to a result in which the model was not used and found that the precision of the results using the model decreased except for "meijiTr". Because the meanings of visual words are vague, there is a lot of noise when the number of combinations is small. Thus, the noise increased when different numbers of combinations were integrated. It will be necessary to derive the appropriate number of combinations and the proper generation method in the future.

Second, precision improved using the co-occurrence of plural attributes. Because we constructed a casebase from a large collection of 200,000 images, the system probably learned sufficiently and precision improved.

6 Conclusion

We developed systems for image recognition and retrieval based on the Confabulation model. It is difficult to demonstrate at this time exactly how effective our system was due to the low precision of the system using the Confabulation model. However, we confirmed that integrating plural attributes was effective, although it is necessary to improve the method of co-occurrence of features. We were also able to evaluate the recurrent confabulation model.

The points that need to be improved were clarified from these results and we can make use of them in a future study. We think if we can imitate the function of the brain when a human recognizes an image, it may be possible for the system to recognize it faster and with better precision.

References

1. ImageCLEFphotoAnnotation2011, "http://www.imageclef.org/2011/Photo"
2. MIR Flickr 1million image dataset, http://press.liacs.nl/mirflickr/
3. G.Csurka, C.Bray, C.Dance and L.Fan: Visual categorization with bags of keypoints, In Proc. of ECCV Workshop on Statistical Learning in Computer Vision, pp. 59-74, 2004.
4. D.G.Lowe: Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision, Vol. 60, No. 2, pp. 91-110, 2004.
5. R. Hecht-Nielsen.: A Theory of Cerebral Cortex, UCSD Institute for Neural Computation Technical report #0401, 2004.
6. R. Hecht-Nielsen.: A Theory of Cerebral Cortex, UCSD Institute for Neural Computation Technical report #0404, 2004.
7. R. Hecht-Nielsen.: Cogent confabulation. Neural Networks, Vol. 18, pp. 111–115, 2005.
8. R. Hecht-Nielsen. : Confabulation theory. UCSD Institute for Neural Computation Technical Report #0501, 2005.