

# Image-to-word transformation based on dividing and vector quantizing images with words

Yasuhide MORI, Hironobu TAKAHASHI and Ryuichi OKA

Real World Computing Partnership, Information Basis Function Laboratory,  
Tsukuba Mitsui Building 13F, 1-6-1 Takezono, Tsukuba-shi, Ibaraki 305-0032, JAPAN  
E-mail: {ymori,hironobu,oka}@rwcp.or.jp

**Abstract:** We propose a method to make a relationship between images and words. We adopt two processes in the method, one is a process to uniformly divide each image into sub-images with key words, and the other is a process to carry out vector quantization of the sub-images. These processes lead to results which show that each sub-image can be correlated to a set of words each of which is selected from words assigned to whole images. Original aspects of the method are, (1) all words assigned to a whole image are inherited to each divided sub-image, (2) the voting probability of each word for a set of divided images is estimated by the result of a vector quantization of the feature vector of sub-images. Some experiments show the effectiveness of the proposed method.

## 1 Introduction

To permit complete access to the information available through the WWW, media-independent access methods must be developed. For instance, a method enabling use of an image is needed as a possible query to retrieve images[1] and texts.

So far, various approaches regarding image-to-word transformation are being studied[2]-[4]. But they are very limited in terms of vocabulary or domain of images. In real data, it is not possible to segment objects in advance, assume the number of categories, nor avoid the presence of noise which is hard to erase.

In this paper, a method of image-to-word transformation is proposed based on statistical learning from images to which words are attached. The key concept of the method is as follows: (1) each image is divided into many parts and at the same time all attached words for each image are inherited to each part; (2) parts of all images are vector quantized to make clusters; (3) the likelihood for each word in each cluster is estimated statistically.

## 2 Procedure of the proposed method

### 2.1 Motivation and outline

To find the detailed correlation between text and image (not simply discriminating an image into a few categories), each portion of the image should be correlated to words instead of the whole image to words.

Assigning keywords to images portion by portion would be an ideal way to prepare learning data. However, with the exception of a very small vocabulary, we cannot find such learning data nor can we prepare them. The more the size of the data increases, the more difficult assigning keywords to images portion by portion becomes. So we have to develop another method to avoid this fundamental problem.

To avoid this problem, we propose a simple method to correlate each portion of an image to key words only using key words for the whole image.

The procedure of the proposed method is as follows :

1. Many images with key words are used for learning data,
2. Divide each image into parts and extract features from each part,
3. Each divided part inherits all words from its original image.
4. Make clusters from all divided images using vector quantization,
5. Accumulate the frequencies of words of all partial images in each cluster, and calculate the likelihood for every word,
6. For an unknown image, divide it into parts, extract their features, and find the nearest clusters for all divided parts. Combine the likelihoods of their clusters, and determine which words are most plausible.

The main point of this method is to reduce noise (i.e. unsuitable correlating) by accumulating similar partial patterns from many images with key words.

For example, suppose an image has two words, ‘sky’ and ‘mountain’. After dividing the image, the part which has only the sky pattern also has ‘sky’ and ‘mountain’ due to the inheriting of all words. The word ‘mountain’ is inappropriate for the part. However if an another image has two words, ‘sky’ and ‘river’, accumulating these two images, the sky pattern has two ‘sky’s, one ‘mountain’ and one ‘river’. In such way, we can hope that the rate of inappropriate words are gradually decreased by accumulating similar patterns.<sup>1</sup>

Figure 1 shows the concept of estimating likelihoods of data.

## 2.2 Dividing image, feature extraction, and inheriting key words

Each image is divided equally into rectangular parts because it is the simplest and fastest way to divide images. The number of divisions ranges from  $3 \times 3$  to  $7 \times 7$ . In this paper, the dividing method driven by the contents of images such as region extraction has not been tried.

In parallel with the dividing, all words given for an image are inherited into each of the divided parts. This is a straightforward way to give words to each part because there is no informations to select words at this stage.

Extracted features for the divided images are (1) a  $4 \times 4 \times 4$  cubic RGB color histogram and (2) an 8-directions  $\times$  4-resolutions histogram of intensity after Sobel filtering, which can be calculated by general fast and common operations.

Feature (1) is calculated as follows:

1. divide RGB color space into  $4 \times 4 \times 4$ ,
2. count the number of pixels fall in each bin.

As a result, 64 features are calculated.

Feature (2) is calculated as follows:

For 4-resolutions (1, 1/2, 1/4 and 1/8):

1. filtering by vertical ( $S_y$ ) and horizontal ( $S_x$ ) Sobel filters,
2. for each pixel, calculate arguments ( $\tan^{-1}(S_y/S_x)$ ),
3. divide arguments  $[-\pi, \pi)$  into 8 directions,

<sup>1</sup>Moreover, we hope that these inappropriate words may convey the correlation between two different kinds of patterns in a database.

4. sum the intensity ( $\sqrt{S_x^2 + S_y^2}$ ) of each pixel in each direction.

As a result, 32 features are calculated.

As a result of these operations, a total of 96 features are calculated from a divided image.

## 2.3 Vector quantization

The feature vectors extracted from the divided parts of all learning images are clustered by vector quantization in a 96-dimensional space. In this paper, data incremental vector quantization is used. In this method centroids (representative vectors for each cluster) are created incrementally for data input. Each cluster has one centroid and each data belongs to a cluster uniquely.

There is only one control parameter in this method, that is, the threshold of error for quantization (referred to later as *scale*). The less a scale is, the more centroids are created.

The procedure for vector quantization is as follows:

1. Set the scale  $d$ .
2. Select a feature vector as the first centroid,
3. For the  $i$ -th feature ( $2 \leq i \leq \text{total number of feature vectors}$ ):  
if there are centroids such that the distance<sup>2</sup> from the  $i$ -th feature is less than  $d$ , then the  $i$ -th feature vector belongs to the nearest centroid,  
else set the  $i$ -th feature vector as a new centroid.

## 2.4 Probability estimation of key words for each cluster

After centroids  $c_j$  ( $j = 1, 2, \dots, C$ ) are created by the vector quantization, likelihoods (conditional probability)  $P(w_i|c_j)$  for each word  $w_i$  ( $j = 1, 2, \dots, W$ ) and each  $c_j$  are estimated by accumulating their frequency:

$$\begin{aligned} P(w_i|c_j) &= \frac{P(c_j|w_i)P(w_i)}{\sum_{k=1}^W P(c_j|w_k)P(w_k)} \\ &\approx \frac{(m_{ji}/n_i)(n_i/N)}{\sum_{k=1}^W (m_{jk}/n_k)(n_k/N)} \\ &= \frac{m_{ji}}{\sum_{k=1}^W m_{jk}} = \frac{m_{ji}}{M_j}, \end{aligned}$$

where,  $m_{ji}$  is the total of word  $w_i$  in centroid  $c_j$ ,  $M_j (= \sum_{k=1}^W m_{jk})$  means the total of all words in centroid  $c_j$ ,  $n_i$  is the total of word  $w_i$  in all data, and  $N (= \sum_{k=1}^W n_k)$  is the total of words for all data (each word is counted repeatedly each time it appears).

<sup>2</sup>Euclid distance in the feature space

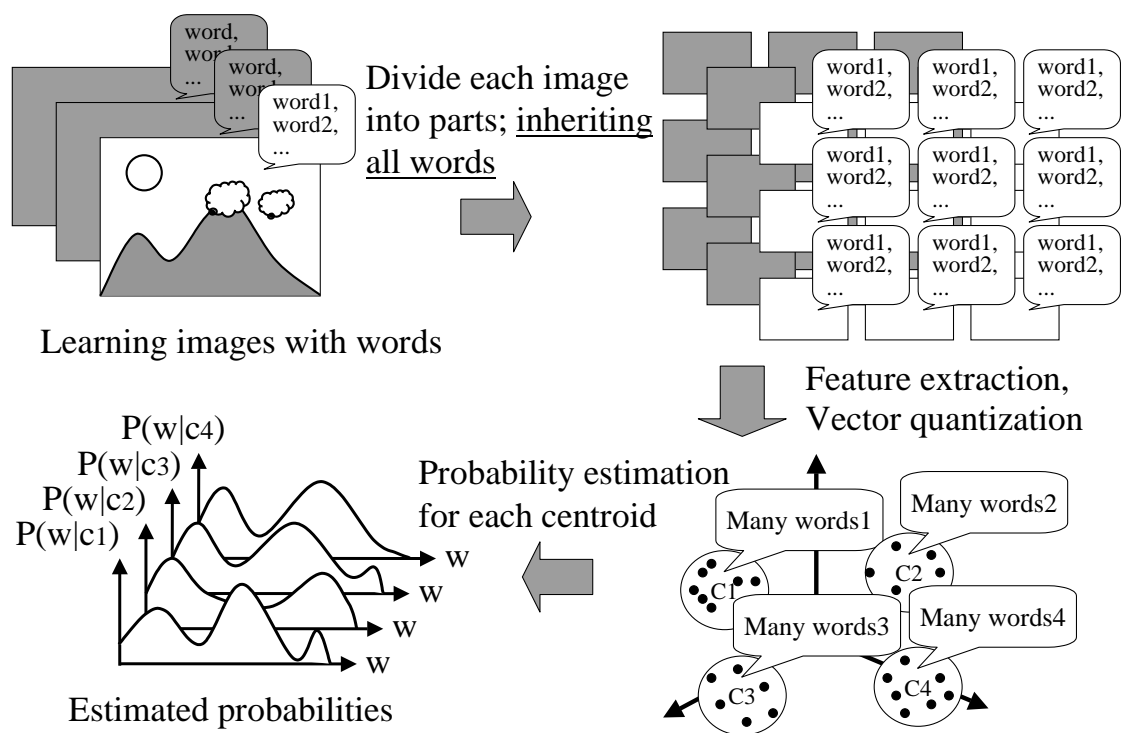


Figure 1: Concept of the proposed method.

## 2.5 Determining correlated words from an unknown image

Using estimated likelihood  $P(w_i|c_j)$ , correlated words are determined for an unknown image as follows: First, an unknown image is divided into parts and its features are extracted in the same way as for the learning data. Second, the nearest centroids are found for all divided parts in the feature space. Third, an average of the likelihoods of the nearest centroids is made. Then, words which have the largest average value of the likelihoods are output.

Figure 2 shows the concept of determining correlated words from an unknown image.

## 3 Experiment and results

### 3.1 Data used in the experiment

In the experiment, a multimedia encyclopedia<sup>3</sup> is used as an original database. The encyclopedia contains about 60,000 items and about 10,000 images in total. About 10,000 items which have citations to images are selected from all items. Therefore, the data for the experiment consists of about 10,000 pairs of images and corresponding documents (in Japanese).

There are 9,681 images in the data. There are various kinds of images for the experiments; landscapes, architecture, historical materials, plants (photographs and sketches), portraits, paintings, etc. About 80% of the images are in color. They have 256 grades of brightness and their sizes average 400×280 pixels.

Next, a set of words is extracted from the documents using the following procedure:

1. Divide documents in all items into words<sup>4</sup> and determine each word's part-of-speech (noun, verb, adjective, etc.) using a morpheme analysis program called "*Chasen*" (resulting in about 100,000 vocabularies),
2. Select only (common and proper) nouns and adjectives (resulting in about 51,708 vocabularies),
3. Eliminate rare words (those which less than frequency of 64), finally 1,585 words are remain (their frequencies range from 5,150 to 64)

After the extraction on average 32 words are attached to each image in average.

As a result of the operation, 9,681 pairs of images and several words are prepared for the experiment.

### 3.2 Procedure used in the experiment

Two-fold cross validation is used in the experiment. The data is randomly divided into two groups (4,841 and 4,840). One group is used for learning (i.e. estimating the likelihood for words), and the other group is used for recognition (i.e. for the output of words), and the same process is performed once again after swapping the two groups.

A unit of the scale for vector quantization is defined as the standard deviation of 'one-dimensional pull-downed data'. The 'one-dimensional pull-downed data' is a set of scalar data which is composed of all component of feature vectors in the original set of data.

### 3.3 Results

Tables 1 and 2 show the examples of output words (the top 3 words) for images in the recognition group (i.e. 'unknown' images). In Tables 1 and 2, bold words indicate 'hit' words (i.e. originally attached words for the image). Tables 1 and 2 show that words output change depending on images input. However it is difficult with our method to output suitable words in the same manner humans do because of too large a variety of images in this data. In Table 1 and 2, hit words appear more than in the case of random selection (if it is a random selection from a set of words with *uniform frequencies*, the probability is about 3/1585). However the frequencies of words in our data is *not* uniform, the words which have high frequencies tend to appear many times (ex. 'year', 'Japan').

Table 3 shows the numerical results of the experiment for various scales in vector quantization. In Table 3, the hit rate means the rate of originally attached words in output words. This table shows that scale 4 has the best hit rate. The difference between the hit rate in scale 4 and the hit rate in scale 0 shows that vector quantization is effective.

As the scale increases above 4, the hit rate decreases gradually.

The scale 22 in Table 3 has only one centroid. In this case, features from images are not considered at all. In other words, this case corresponds to the random selection from the set of words which has a biased frequencies.

Therefore, the difference between scale 22 and other smaller scales shows the effect when features from images are considered. Compared to the case for scale 4, there is a 7% advantage.

Table 4 shows the result of the experiment for various number of divisions. Table 4 shows that the more the image is divided, the better the hit rate is. This

<sup>3</sup> *Mypaedia*, Hitachi Digital Heibonsha, 1998

<sup>4</sup> Japanese is not divided into words originally.

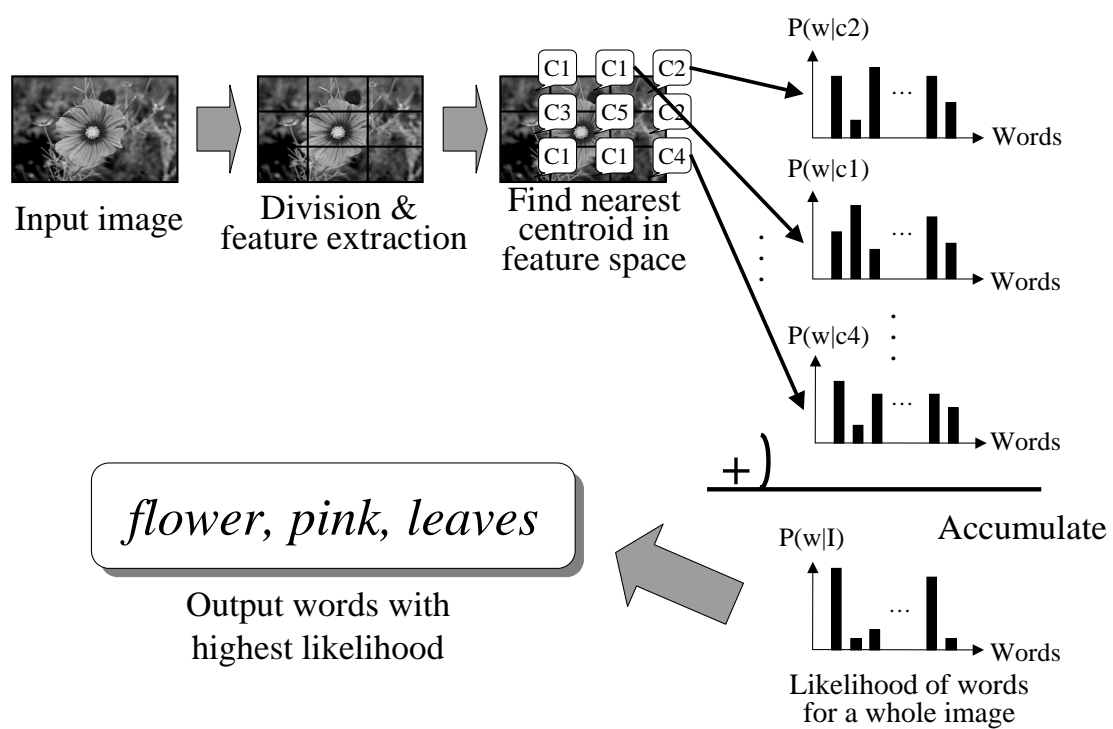


Figure 2: Concept of determining correlated words from an unknown image.

Table 1: Examples of output words for unknown images – part 1. Bold words shows ‘hit’ words. The image is divided into 3×3, scale = 4.0.









Input image	Output words (top 3)	Input image	Output words (top 3)
	year, <b>Japan</b> , family		year, <b>age</b> , white
	year, <b>many</b> , family		area, east, shore
	year, <b>park</b> , family		<b>park</b> , <b>national</b> , center
	year, <b>ten thousand</b> , <b>city</b>		<b>city</b> , god, layer

Table 2: Examples of output words for unknown images – part 2. Bold words shows ‘hit’ words. The image is divided into 3×3, scale = 4.0.









Input image	Output words (top 3)	Input image	Output words (top 3)
	<b>year</b> , Japan, China		architecture, <b>shrine</b> , represent
	<b>family</b> , year, leaf		<b>family</b> , leaf, flower
	year, <b>many</b> , <b>Japan</b>		<b>year</b> , <b>century</b> , age
	<b>year</b> , age, <b>work</b>		<b>year</b> , <b>Japan</b> , age

Table 3: Results for various scales. Two columns in the ‘number of centroids’ correspond to the two learnings in the two-fold cross validation. The hit rate is a mean value for all tests in the two-fold cross validation.

Scale	# of centroids		Hit rate (top 3 words)
0	43181	43190	0.31
2	20732	21082	0.35
4	2373	2361	0.40
6	407	404	0.39
12	40	39	0.37
22	1	1	0.33

Table 4: Hit rates for various numbers of divisions. Each scale is the best one for each number of dividing. Two columns in the ‘number of centroids’ correspond to the two learnings in the two-fold cross validation. The hit rate is a mean value for all tests in the two-fold cross validation

Divisions	Scale	# of centroids		Hit rate (top 3 words)
$1 \times 1$	6	164	144	0.37
$3 \times 3$	4	2373	2361	0.40
$5 \times 5$	4	3941	4010	0.40
$7 \times 7$	4	6588	6564	0.40



result cannot verify the optimal number of divisions for the data. However this result at least shows that the hit rate for not dividing ( $1 \times 1$ ) is inferior to the other.

## 4 Discussion

effect when features from images are considered.

The results of the experiment show that the contribution of vector quantization is about 9%. And they also show that the effect of considering images comparing to random selection is about 7%. These percentages cannot be said to be 'significant or not' unconditionally because it depends on the data used in experiments.

As one can see from Table 1, a considerably wide domain of images and vocabulary are used in the experiment. It is easy to improve these percentages by restricting the domain of images and/or vocabulary. However it is desirable to make this method useful even though the domain of data is wide to make real-world data tractable. We hope that the results achieved by this method become useful when it is combined with a good human-interface system to help users in mining data.

## 5 Conclusion

In this paper, we have proposed a new method for correlating images with key words based on two kinds of processes, that is, dividing images and vector quantization. The result of experiments using encyclopedia data confirmed the contribution of these two processes.

Future work is needed to select a set of words depending on a given task and to determine the optimum size for dividing the images depending on the characteristics of image-word databases.

## Acknowledgment

The authors would like to thank Hitachi Digital Heibonsha for permission to use the data in *Mypaedia* for this work.

## References

- [1] M.Flickner, H.S.Sawhney, J.Ashley, Q.Huang, B.Dom, M.Gorkani, J.Hafner, D.Lee, D.Petkovic, D.Steele and P.Yanker: "Query by Image and Video Content: The QBIC System," IEEE Computer, 28-9, pp. 23-32, 1995.
- [2] T.Kurita, T.Kato, I.Fukuda and A.Sakakura: "Scene retrieval on an image database of full color paintings," Trans. of Information Processing Society of Japan, Vol. 33, No. 11, pp. 1373-1383, 1992 (in Japanese).
- [3] A.Ono, M.Amano, M.Hakaridani, T.Satou and M.Sakauchi: "A flexible content-based image retrieval system with combined scene description keyword," Proc. IEEE Computer Society, International Conference on Multimedia Computing and Systems '96, pp. 201-208, 1996.
- [4] Y.Watanabe and M.Nagao: "Image analysis using natural language information extracted from explanation text," J. of Japanese Society for Artificial Intelligence, Vol. 13, No. 1, pp. 66-74, 1998 (in Japanese).