Probabilistic-Based Semantic Image Feature Using Visual Words

Cheng-Chieh Chiang
Dept. of Info. Tech.
Takming U. of Science & Tech.
No. 56, Sec. 1, Huan-Shan Rd.
Taipei 114, Taiwan, R.O.C.
kevin@csie.ntnu.edu.tw

Jia-Wei Wu
Dept. of Comp. Science and Info. Eng., National Taiwan Normal U.
No. 88, Sec. 4, Ting-Chou Rd.
Taipei 114, Taiwan, R.O.C.
g96470148@csie.ntnu.edu.tw

Greg C. Lee
Dept. of Comp. Science and Info. Eng., National Taiwan Normal U.
No. 88, Sec. 4, Ting-Chou Rd.
Taipei 114, Taiwan, R.O.C.
leeg@csie.ntnu.edu.tw

Abstract

This paper presents a new image feature that is based on a semantic-level perspective in order to bridge the semantic gap between low-level features of images and high-level concepts of human perception. In this work, low-level image features are first quantized into a set of visual words, and then we apply probabilistic Latent Semantic Analysis model to automatically analyze what kinds of hidden concepts between visual words and images are involved. Therefore, we collect discovered concepts of an image and filter a part of unreliable concepts out to build a semantic-based image feature. We also discuss in detail how to define parameters for extracting the proposed feature. Several experiments are presented to show the efficiency of this work.

1 Introduction

Image analysis and understanding has been an active research topic for many years. This is even more so as multimedia information is readily created and available. Image features are mostly low-level, i.e., they are extracted directly from signal information of raw images. However, human cognition of perceiving an image is not directly based on low-level features, but is based on high-level concepts derived from those low-level features. Semantic gap [2] between these two levels is still a challenging problem in image analysis and understanding.

To discover what kinds of semantic information are embedded in an image is potential to bridge the semantic gap in image content. Image representation based on semantic contents of image can be more reasonable than representation based on low-level image features. However, contents embedded in an image are often rich and complex, hence it is difficult to extract semantic contents direct from low-level features.

In this paper, we construct a region- and semantic-based image feature for bridging the semantic gap. Instead of employing a segmentation algorithm to divide an image into a set of non-overlapping areas, we extract some, possibly overlapping, regions of interest in an image to search for salient areas of the image content. In our approach, Scale Invariant Feature Transform (SIFT) descriptor [9] is used for low-level feature extraction from image regions. A set of visual words is generated based on these low-level features to reduce the influence of noises in the feature space. Next, we employ probabilistic Latent Semantic Analysis (pLSA) [6][7] to automatically analyze what kinds of hidden concepts between visual words and images are involved. We collect discovered hidden semantic contents from all regions of interest in an image to be its semantic image features.

The rest of this paper is organized as the follows. Section 2 presents how to extract visual words in our work. Then, we introduce the basic model of pLSA in Section 3. Our proposed semantic image feature is described in Section 4. Finally, we provide experimental results for our work in Section 5 and conclusion and future works in Section 6.

2 Visual Words

An original concept of visual words is derived from the text analysis of documents. The text terms, i.e., words, are appropriate for analyzing documents in information retrieval. However, it is difficult to find a proper unit for better representing images. Thus, some researchers regarded visual features that are extracted from an image as a variant type of "words" in the compound of the image. The most common approach to generate visual words is to extract a set of region features from images and to quantize these feature vectors into a pre-built vocabulary of visual words, e.g., in [1][4][5][13]. Simply speaking, the construction of visual words is to quantize or cluster (most using *K*-means clustering) region features in the feature space.

Figure 1 shows the procedure of generating visual words in our work. This procedure mainly follows the setup in [13] for experimental comparison. For each image, we first use two methods, from [10] and [11], to extract its interest points in order to discover the informative areas in an image. Therefore, ellipses centered by interest points are generated with random radiuses, and SIFT descriptors are extracted from all of ellipse regions.

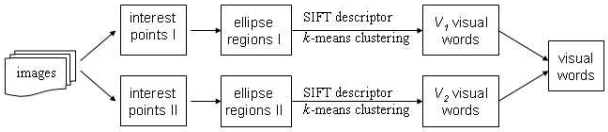

Figure 1. The flow chart to extract visual words in this work.

Let there exist N images denoted I_i, i=1 to N. The two methods of extracting interest points, from [10] and [11], generate numbers R_1 and R_2 of ellipse regions, respectively. These ellipse regions may be overlapped. Next, we extract 128-D SIFT descriptors for each ellipse region. We apply K-means clustering algorithm to quantize the feature space of SIFT descriptor. We set K as V_1 and V_2 associated with the two methods of extracting interest points, respectively. That is to say, we have V_1 and

V_2 visual words, respectively, for representing ellipse regions in images. Finally, we combine them to be $V=V_1+V_2$ visual words in the following extraction of the proposed semantic image feature.

3 Probabilistic Latent Semantic Analysis

Probabilistic latent semantic analysis (pLSA), first proposed by T. Hofmann [6][7], is an unsupervised method to automatically index document based on a statistical latent class model for factor analysis of count data. Considering a set of documents D, a set of words W, and a set of latent topics Z, we have:

- p(d): the prior probability of selecting a document d in D.
- p($z|d$): the probability of an unobserved topic $z \in Z$ given a document d.
- p($w|z$): the likelihood of a word $w \in W$ appeared in a given topic z.

Hence, the joint probability of document d and word w can be modeled by

$$p(d,w) = p(d)p(w \mid d) = \sum_z p(d)p(w \mid z)p(z \mid d) \quad (1)$$

In the above equation, one could apply EM algorithm to determine the density of p($w|z$) and p($z|d$) by maximizing the log-likelihood function:

$$\mathcal{L} = \sum_{d \in D} \sum_{w \in W} n(d,w) \log p(d,w) \quad (2)$$

where n(d, w) indicates the term frequency, i.e., the number of times w occurred in d. Therefore, image d is classified in to the concept z^* with highest probability:

$$z^* = \arg\max_z p(z \mid d) \quad (3)$$

pLSA automatically links together visual words and images through tuning hidden categories z. Several related works employ pLSA to image annotation [8][12] and object recognition [5][13]. We have implemented pLSA for object recognition according to the description in [13], in which D indicates images and W implies visual words described in the previous section. Figure 2 illustrates several images that are classified into the same categories by use of pLSA. Analyzing the data of pLSA in detail, these images are miss-classified because of a large region of uniform background. Indeed, they are classified to the same category for their plain backgrounds, not for their foregrounds. Because the contents of images may be various, the classification of images, i.e., p($z|d$), may be dominated by similar visual words but, in fact, unrelated topics.

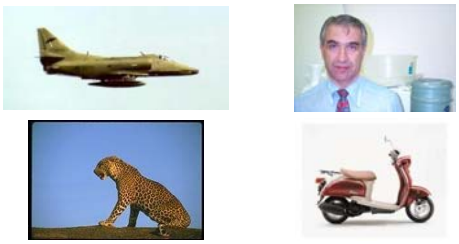

Figure 2. Examples of wrong classification using pLSA. All of the four images are classified into the same category.

4 Feature Extraction

Instead of directly analyzing to which category z an image d belongs, we propose estimating in which semantic concept z a visual word w is involved. A visual word which contains many regions associated with similar low-level features may have a bigger possibility than an entire image to be accurately classified using pLSA. Hence, our idea is to compute p($z|w$) instead of p($z|d$) to estimate the semantic concepts of visual words in images.

In general, we can view an image as a compound of regions, which may be mapped into different semantic concepts. Therefore, these concepts of regions construct a global view for the image in human perspectives. Figure 3 draws the probabilistic structure mentioned above. The perspective of observing the semantic concepts of visual words is region-based to compute p($z|w$), not image-based to compute p($z|d$). If we can precisely estimate to which semantic concept a visual word belongs, we can collect p($z|w$) associated with all visual words w appeared in an image d to be a semantic-based feature. Figure 4 shows the relationship among visual words w, image d, and hidden concept z using pLSA following the probabilistic structure shown in Figure 3. Note that it is a transposed version of the original pLSA.

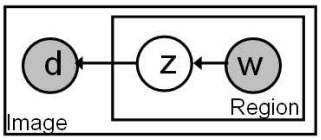

Figure 3. The probabilistic structure in an image.

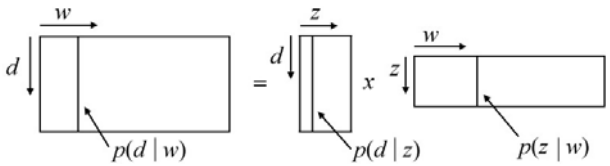

Figure 4. The concept of our proposed feature.

Following the structure shown in Figure 3 and Figure 4, we have

$$p(d \mid w) = \sum_z p(d \mid z)p(z \mid w) \quad (4)$$

It is also clearly that we can simply transpose matrix p($w|d$) in Eq. (1) to get Eq. (4). Then, we can summarize E-Step and M-Step of EM algorithm, according to [7], to estimate Eq. (4) as the follows.

E-Step:

$$p(z_k \mid d_i, w_j) = \frac{p(d_i \mid z_k)p(z_k \mid w_j)}{\sum_l p(d_i \mid z_l)p(z_l \mid w_j)} \quad (5)$$

M-Step:

$$p(d_i \mid z_k) = \frac{\sum_j n(d_i, w_j)p(z_k \mid d_i, w_j)}{\sum_m \sum_j n(d_j, w_m)p(z_k \mid d_j, w_m)}$$

$$p(z_k \mid w_j) = \frac{\sum_i n(d_i, w_j)p(z_k \mid d_i, w_j)}{n(w_j)} \quad (6)$$

,

where $n(w_j) = \sum_i n(d_i, w_j)$ refers to the visual word length. Our goal is to compute p(z| w) and to find the most possible concept z^* which a visual word w_j involves:

$$z^* = \arg\max_z p(z \mid w_j) \tag{7}$$

Let Z be the number of concepts to represent the semantic contents of visual words. Considering all of V visual words w_j, j=1 to V, we compute p($z_i|w_j$) for each concept z_i, i=1 to Z, and take the first and second high values of p($z_i|w_j$) with denoted p($z^*|w_j$) and p($z^+|w_j$), i.e.,

$$z^+ = \arg\max_{z \neq z^*} p(z \mid w_j) \tag{8}$$

Therefore, we define the score of semantic assignment, denoted as $s_Z(w_j)$, of visual word w_j associated with Z as:

$$s_Z(w_j) = p(z^* \mid w_j) - p(z^+ \mid w_j) \tag{9}$$

Here $s_Z(w_j)$ means the difference between the first two high probabilities of recognition for visual word w_j. The larger the score s(w_j), the more confident the concept z^* is correctly involved in visual word w_j. Thus, we define a function $\kappa(w_j, h)$ to filter most ambiguous visual words out using a threshold value h:

$$\kappa(w_j, h) = \begin{cases} 1, \text{if } s_z(w_j) \geq h \\ 0, \text{if } s_z(w_j) < h \end{cases} \tag{10}$$

Given an image I_i, we can compute p($z^*|w_j$) and p($z^+|w_j$) for all visual words w_j that are associated with regions involved in image I_i. Thus, the semantic-based feature of image I_i, a Z-dimensional feature denoted $f_{I_i} = \{f_1, ..., f_Z\}$, is defined as the propagation of the significant concepts of visual words passing the filtering function $\kappa(w_j)$, i.e.,

$$f_i = \sum_{w_j \in I_i} \delta(i, z^*) \cdot \kappa(w_j) \cdot p(z^* \mid w_j) \tag{11}$$

where δ is the Kronecker's delta function:

$$\delta(x, y) = \begin{cases} 1, \text{if } x = y \\ 0, \text{if } x \neq y \end{cases} \tag{12}$$

Our proposed semantic feature f_{I_i} collects confident concepts of visual words to characterize image contents. Because pLSA is an unsupervised approach to cluster visual words w to hidden concepts z, we may not exactly depict what kinds of semantic names can describe concept z. However, regions containing the same hidden concepts can be regarded as the same components in an image. Hence, the proposed feature extracts and propagates semantic-based components in the compound of image contents.

5 Experimental Results

5.1 Dataset

We adopted one dataset of objects, Caltech 101-Object [3], for our experiments. In order to perform the experiments, we took four larger categories of datasets: airplanes, faces, leopards, and motorbikes, from Caltech 101-Object. 200 images are randomly chosen for each category, and therefore we have a total of N=800 images. Figure 5 illustrates the four categories of our dataset. These images contain a semantic subject as foreground and variant contents as background. For the dataset, visual words are constructed according to the procedure in Figure 1. Then, we have R_1 and R_2 are 122,364 and 60,244, respectively, and V_1 and V_2 are 300 and 150, respectively.

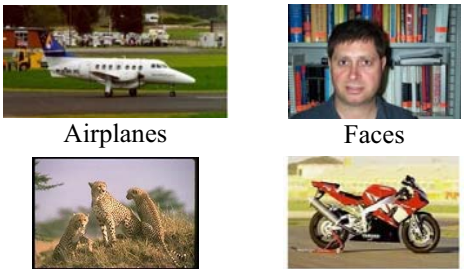

Figure 5. Illustration of the dataset.

5.2 Parameters

In the extraction of the proposed feature, two parameters need to be pre-defined: (i) Z for the number of concepts to represent semantic contents of visual words, and (ii) h for the percentage of most ambiguous visual words that are filtered out. In order to understand how to define the parameters, we perform a simple test that follows the setups in [13] except using our proposed semantic feature instead of visual words. Table 1 shows the recognition rates with different parameters Z and h. Thus, we took h and Z to be 0.05 and 30, respectively, in the following experiments for the best recognition results. Note that about 25% unreliable visual words are filtered out with h=0.05 in Eq. (10).

Table 1. Recognition rates with different h and Z.

h	Z=10	Z=20	Z=30	Z=40	Z=50
0.03	0.6388	0.6488	0.6837	0.6863	0.6838
0.04	0.6462	0.6863	0.7133	0.6950	0.7013
0.05	0.5487	0.6975	**0.8100**	0.6975	0.7725
0.06	0.6638	0.6925	0.8025	0.6887	0.7737
0.07	0.6262	0.7450	0.7263	0.6950	0.6838
0.08	0.6462	0.7737	0.7138	0.6575	0.5900

5.3 Results

To achieve the quantitative comparison to evaluate the performance for our proposed feature, we implemented the method in [13] that used visual words to directly represent an image and classify images using pLSA. We show their recognition rates in Table 2. Similarly, the results of using the methods in [13] expect our proposed feature instead of visual words are shown in Table 3. In the two tables, each row counts the recognition rates for

an image category. Since pLSA is an unsupervised method, we use "Cat. 1" to "Cat. 4" instead of the names of categories in the row titles of the two tables. Comparing the results in the two tables, our proposed feature gets a little improvement than [13].

In order to analyze the performance of the proposed feature in deep, we apply *k*-NN with leave-one-out strategy to image classification by use of our proposed feature. Table 4 shows the good results of recognition rates with different *k* in *k*-NN. Most of recognition rates are higher than 0.9.

Table 2. Recognition results of Sivic et al. [13] using our dataset. The average of recognition rates is 0.79.

	Cat. 1	Cat. 2	Cat. 3	Cat. 4
Airplanes	0.62	0.025	0.085	0.27
Faces	0.075	0.8	0.11	0.015
Leopards	0.055	0.05	0.87	0.025
Motorbikes	0.02	0	0.025	0.89

Table 3. Recognition results of our proposed feature. The average of recognition rates is 0.81.

	Cat. 1	Cat. 2	Cat. 3	Cat. 4
Airplanes	0.905	0.005	0.05	0.04
Faces	0.105	0.725	0.12	0.05
Leopards	0.05	0.005	0.93	0.015
Motorbikes	0.125	0.005	0.19	0.68

Table 4. Recognition rates: our proposed feature using *k*-NN with different *k*.

	k=1	*k*=5	*k*=9	*k*=11
Airplanes	0.83	0.88	0.875	0.88
Faces	0.92	0.91	0.915	0.91
Leopards	1	1	1	1
Motorbikes	0.91	0.915	0.93	0.92
Avg. Recog.	0.915	0.9262	0.930	0.9275

6 Conclusion and Future Works

This paper presents our design to construct a probabilistic-based semantic image feature using visual words in the space of SIFT descriptor. We first apply pLSA that is an unsupervised approach to extract hidden concepts between visual words and images. Then, the discovered concepts associated with visual words are propagated to be a semantic image feature. This paper describes the details of the design for the feature and provides the convincing evaluation to show the performance. We are planning several tasks to extend this work. The first task is to design a theoretical method to determine the value *Z* for the number of hidden concepts to describe visual words in images. Also, this feature can be applied to a real application, e.g., image retrieval, in the future.

Acknowledgement

This work was supported by National Science Council, Taiwan, under Grant No. NSC 97-2511-S-003- 007-MY3 and NSC 97-2631-S-003-003 and by Ministry of Economic Affairs, Taiwan, under Grant No. 97-EC-17-A -02-S1-032.

References

[1] G. Csurka, C. Bray, C. Dance, and L. Fan, "Visual Categorization with Bags of Keypoints," in Proceedings of Workshop on Statistical Learning in Computer Vision, ECCV, pp. 1-22, 2004.

[2] R. Datta, J. Li, and J. Z. Wang, "Content-Based Image Retrieval - Approaches And Trends of the New Age," in Proceedings of ACM SIGMM International Workshop on MIR, 2005.

[3] L. Fei-Fei, R. Fergus, and P. Perona, "Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories," in Proceedings of Workshop on Generative-Model Based Vision, CVPR, 2004.

[4] L. Fei-Fei and P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories," in Proceedings of CVPR, pp. 524-531, 2005.

[5] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning Object Categories from Google's Image Search," in Proceedings of ICCV, 2005.

[6] T. Hofmann, "Probabilistic Latent Semantic Indexing," in Proceedings of ACM SIGIR, 1999.

[7] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis", Machine Learning, Vol. 42, 177–196, 2001.

[8] D. Liu, and T. Chen, "Semantic-Shift for Unsupervised Object Detection," in Proceedings of Workshop on Beyond Patches, in conjunction with CVPR, 2006.

[9] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," IJCV, 60(2):91-110, 2004.

[10] J.Matas, O. Chum, M. Urban, and T. Pajdla, "Robust Wide Baseline Stereo from Maximally Stable Extremal Regions," in Proceedings of BMVC, pp. 384-393, 2002.

[11] K. Mikolajczyk and C. Schmid, "Scale and Affine Invariant Interest Point Detectors," IJCV, 60(1):63-86, 2004.

[12] F. Monay and D. Gatica-Perez, "PLSA-Based Image Auto-Annotation: Constraining the Latent Space," in Proceedings of ACM MM, New York, Oct. 2004.

[13] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering Objects And Their Location in Images," in Proceedings of ICCV, 2005.