Flickr Image Classification using SIFT Algorism

Hyun-Woong Jang , Soosun Cho

Department of Computer Science & Information Engineering
Korea National University of Transportation, 157 Cheoldoparkmulgwan-ro, Uiwang-si, Gyeonggi-do, 437-763, Korea
jhwskorg@gmail.com, sscho@ut.ac.kr

Abstract: As a huge image storing and sharing site such as Flickr is getting popular, the amount of image information is also increasing and the users want more accurate image searching system. To increase the accuracy of tag-based image search, a variety of approaches have been tried including usage of semantically related tags. In this paper, we propose a method to classify the Flickr images with bag of visual word approaches using the SIFT algorithm, which has good performance for classifying images based on the image feature extraction. As a result of applying SIFT algorithm to the database of the preliminary retrieved images with semantically related tags, we found that it showed better accuracy than a result from SURF algorithm. Therefore we expect that our method can classify a variety of Flickr images more accurately.
[Jang HW, Cho S. **Flickr Image Classification using SIFT Algorism**. *Life Sci J* 2014;11(7):607-611] (ISSN:1097-8135). http://www.lifesciencesite.com. 84

Keywords: visual-words; SIFT algorithm; semantic tags; image classification

1. Introduction

The basic image contents such as color, shape, texture and the outline of objects have been used to classify images for a long time. However, as a huge image data has being saved in web sites, the users want to get more accurate image information quickly. Thereby lots of technologies of the image classification have been developed. Recently, it is mainly used that extracting the features of images and using them for image classification (Andrea and Brian, 2010) that is robust in changes of colors, positions, scales and rotations.

In this paper, we propose a method to classify the Flickr images with bag of visual word approaches using extracting feature algorithm SIFT(Scale Invariant Feature Transform). The bag of visual word method has good performance for image matching and transforming, to classify images more accurately. The performance of SIFT algorithm used for classifying images has been introduced by lots of researches (Lowe, 2004, Andrea and Brian, 2010)

On the experiment, we do not use the refined image database, but Flickr web images are used as a target image database, which is one of huge web image storing and sharing sites. To train the classifying algorithm, we used the improved retrieval result from Flickr web images with semantically related tags based on Wikipedia.

This paper consists of section 2, related studies, section 3, the proposed methods, section 4, experimental results, and section 5, conclusions

2. Related Studies

2.1 Tag or content-based image classification

Tagging, which is used to attach many keywords to an image, is generally made by users. So it has limits to find an appropriate image based on tags because there are so many tags and they are attached to images by users subjectively. As one of studies for tagging technology, research (Lee et al., 2007) summarized the limits of tagging. Firstly, tagging does not provide the management for synonyms or homonyms. Secondly, for the information retrieval, tagging has become the cause of precision rate decrease.

Therefore, we can agree that the tagging is very useful for locating contents into a wide range of category, but is not efficient for searching exact information satisfying users' desire. And for this reason, the simple matching of tags to a given query for tagged image retrieval is left in serious limitations.

To solve this problem, there were some studies using semantically related tags (Kweon et al., 2008, Lee and Cho 2011). The research (Lee and Cho 2011) proposed the image re-ranking method from the simply retrieved Flickr images using semantically related tags and keywords based on Wikipedia. The retrieved images are re-ranked in the order calculated by using the semantically related tags. These researches are showing that the method of searching image with semantically related tags is better in accuracy compared to the simple matching method.

On the other hand, the image classification based on image feature extraction is getting a lot of attentions in content-based image classification recently. This is an image comparing or searching method based on clustering feature points which are extracted from the separated components of image

contents. This method can overcome the disadvantages that people have to attach image tags themselves by subjective judgments.

The methods based on image feature points use vector quantized values pulled out of image key-point as a salient area contains rich image information. We regard images as one 'bag of visual word' like the documents considered as one 'bag of word' in text-based search, and use preconfigured the visual-word vocabulary to classify images (Yang, 2010).

2.2 SIFT algorithm and image classification

In a wide variety of image application technology such as image recognition, object recognition, face recognition and solving for 3D images based on computer vision, the image matching technology has been developed. To match different images extracts the image features that have lots of properties. The features are invariant to scaling, rotation and changes. Lowe suggested DoG(Difference of Gaussian), which is using difference of Gaussian scale space theory, to apply SIFT algorithm (Lowe, 2004). He proved that SIFT algorithm with DoG showed the best performance in image matching technology. The SIFT algorithm extract image feature vectors. The feature vectors are called keypoints. The keypoints with constructed descriptors can be called visual words as an intermediate image characterization.

The method building keypoint to vector quantization, DoG detector and SIFT algorithm are usually adopted in image classification based on feature points. Using SIFT algorithm as an image keypoint descriptor, we can extract robust features of images in scalable and rotational changes. The keypoints derived from images are used to find the nearest-neighbor indexes. These indexes can generate the histogram of visual words. The histogram can be used to make a kernel map to classify the image categories.

There were studies to classify the image categories with extracting feature algorithm SURF(Speeded Up Robust Feature) constructing image visual words (Cho et al., 2012, Kim et al., 2012). The SURF is a feature detector and first presented by Herbert Bay (Bay et al., 2008). It uses an integer approximation.

In image classification performance between SIFT and SURF, even the SIFT method is slower compared to SURF (Lim and Lee, 2009), but we use SIFT algorithm because of the high accuracy of SIFT method. SIFT method's main focus is not speed, but the high accuracy. The image matching process with SIFT is advantageous because it produces robust features of images in scalable and rotational changes (Choi et al., 2012).

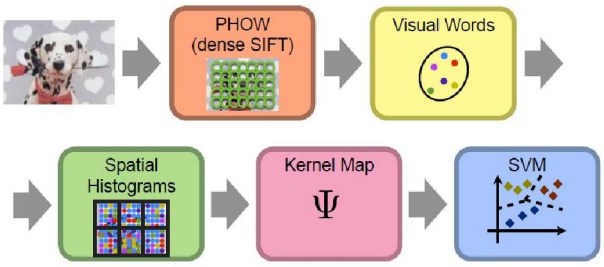

Figure 1. Image detection process

We used open library VLFeat (Andrea and Brian, 2010) in the implementation of image classification system. Figure 1 shows the image classifying process based on SIFT in VLFeat.

PHOW: This process is extracting the features of image. Isotropic Gaussian kernel is used to blur the images. The Gaussian kernel uses DoG filter, which followed by the Gaussian smoothing, removes fine details of images and increases execution time. Edges and corners of image are used for extracting features in images. Frist, Gaussian smoothing filter is applied to a single grayscale image in order to reduce its sensitivity to noise and then detects the image regions of rapid intensity change and the regions are used for edge and corners detection. And then SIFT algorithm extracts a dense set of features from image VLFeat (Andrea and Brian, 2010). It returns columns storing the center (x, y) of a keypoint frame and a number of keypoints matrix with one descriptor per column. The extracted keypoints detect the keypoint gradients through the detector. The gradients construct 4x4 array of histogram which has the vectors of eight directions.

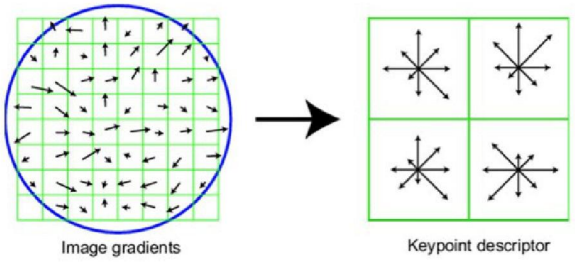

Figure 2. The image gradients construct 4x4 array of Keypoint descriptor which has the vectors of eight directions.

Visual Words: To make visual words, k-means algorithm is applied to the result of PHOW process. k-means algorithm clusters the columns of the matrix in the number of centers using k-means algorithm (Richard, 2010). It returns a set of vectors

around common mean vector. This process is the preparation to make the vocabulary of visual words(a set of common features) matrix with the central vectors accumulatively (Jeong, 2013).

Spatial Histograms: The vocabulary used for building KD-tree which is a data structure used to quickly find the nearest data. The KD-tree data structure computes the nearest column in Euclidean distance. Using returned data indexes from KD-tree, the visual word histograms are generated for each scene of image.

Kernel Map: To train the SVM, the SVM needs the homogeneous kernel map and it is generated by the histograms. The histograms generate homogeneous kernel map.

SVM: A linear Support Vector Machine (SVM) is trained by the data vector from the histograms (Kwon et al., 2011). After training the SVM, test SVM to classify input image into one of categories.

3. Implementation and Experiment

In this paper, we designed the classification engine using VLFeat library. VLFeat is a popular open source library in the field of computer vision. It supports HOG, SIFT, MSER, k-means, hierarchical k-means, agglomerative information bottleneck, SLIC superpixels, SVM, and quick shift algorithms to be easily implemented in MATLAB (Andrea and Brian, 2010).

At first, we evaluated the performance of SIFT algorithm using Caltech101 dataset (Caltech 101 Dataset Web Site), which is a universal classification of image data set used for testing. Caltech101 dataset is a set of refined images created in September 2003, for intending to facilitate computer vision techniques at the California Institute of Technology. In Caltech101, a total of 9,146 images and 101 categories are included. As you see in the table 1, the performance of SIFT using Caltech101 dataset recoded very high accuracy (94%) because Caltech101 images already refined.

Table 1. Classification of 10 test images with 30 training images using Caltech101 dataset (Accuracy: 94.00%)

Category	classified images	Accuracy
accordion	10	100%
airplanes	10	100%
anchor	7	70%
ant	10	100%
barrel	10	100%
Total	47	94.00%

In our experiments, retrieved Flickr images are used as training and testing data for the real target of the images on the internet. At this time, we used the result from the previous study, which collects images with the semantically related tags using semantic information based on Wikipedia, to use more accurately retrieved images for training data (Lee and Cho, 2011).

We collected a total of 3,003 images using the image data from the previous study (Lee and Cho, 2011). In each category, bird, car, or sea, for training the classifier, 200 more images than the previous study are used. So 700 per each category, the total of 2,100 Flickr images are randomly selected and used as training images. We pulled out 30 images per each category as test data and tried the experiments for classifying images with the total of 90 test images.

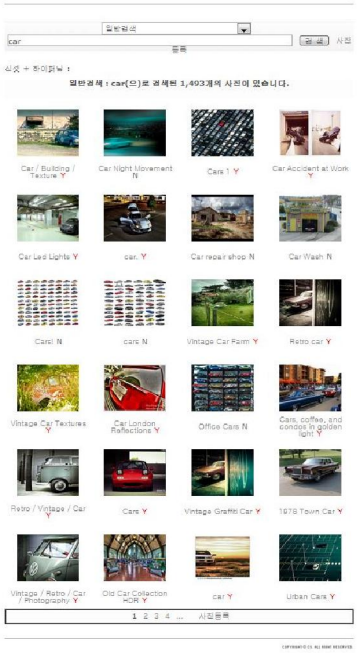

Figure 3. 1,001 car Flickr images selected from retrieval results using Wikipedia-based semantic relatedness

SIFT detector is used for extracting features, and K-means clustering algorithm is used for image vocabulary learning. Elkan algorithm (Charles, 2003) is implement for K-means algorithm. Elkan algorithm is for these tasks several times faster than the standard Lloyd K-means implementation, it has good performance ordering faster than the MATLAB basic k-means function. K-means clusters a few ten millions all of the image visual descriptors into a vocabulary of 700 visual words. It calculates the K central axis from the extracted features, repeats clustering the dots close to their central axis, the

center values make a visual word. KD-tree is used to quickly accumulate the visual words into a spatial histogram and uses it as image descriptors and fed to a linear SVM classifier.

SVM classifier uses pegasos algorithm (Shai, 2007) which is a modified stochastic gradient method for SVM training. Test the trained SVM to classify image data.

4. Analysis and Evaluation

Table 2 shows the experimental result of classification. The 30 randomly selected images per category, a total of 90 images were used as test data and 700 images for each category from the 1,001 images were used as training data. In the 'bird' category, 21 images were classified correctly from the total of 30 images, and it showed 70% accuracy. In the 'car' category, 22 images were classified with 73.33% accuracy, and 19 images in the 'sea' category were classified with 68.33% accuracy. As a result, the average accuracy indicated 68.89%.

Table 2. Classification of 30 test images with SIFT algorithm, which is trained with 700 training images from the Flickr retrieval results using Wikipedia-based semantic relatedness (Accuracy: 68.89%)

Category	classified images	Accuracy
bird	21	70%
car	22	73.33%
sea	19	63.33%
Total	62	68.89%

As the experimental result, the accuracy of 68.89% is lower than the accuracy of 94.00% from the Caltech101 data set. The reason is that the Caltech101 is the set of image data already refined. But we used the 2,100 retrieved real images from the Flickr site for training classifier. However, higher accuracy than the previous result of 65.6% (Cho et al., 2012) was recorded. So this result represents that in the bag of visual word approaches, using SIFT algorithm for a feature detector is better than SURF algorithm. The 3.4% gab may be thought as small, but as the amount of images to be retrieved is increasing quickly, the gab is expected to grow.

5. Summary

There is a limitation to find an appropriate image only based on tags because so many tags and they are attached to images by users subjectively. To solve this problem, we suggest an integrated method using semantically related tags and bag of visual word approach to retrieve and classify the Flickr images.

In this study, to train the classifier with real internet images of higher accuracy, we used the retrieved images with semantically related tags based on Wikipedia. For training data we used the retrieved result from Flickr web images. And we showed that using SIFT algorithm to extract the image features can improve the accuracy of image classification.

The accuracy could be improved compared to the previous study (Cho et al., 2012) by applying SIFT algorithm, using more categories, and high number of training images. By using semantic relatedness of tags and image content itself, we plan to continue research for improving the accuracy of image classifier more effectively in the future.

Acknowledgements:

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (2010-0013307).

Corresponding Author:
Dr. Soosun Cho
Department of Computer Science & Information Engineering
Korea National University of Transportation
157 Cheoldoparkmulgwan-ro, Uiwang-si, Gyeonggi-do, 437-763, Korea
E-mail: sscho@ut.ac.kr

References

1. Lowe DG. Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 2004;60(2): 91-110.
2. Cho SW, Lee S, Cho S. Visual word-based Classification of Images Including Background Objects. Proceedings of the Korea Information Processing Society Fall Conference, Jeju University, 2012.
3. Kweon DH, Hong JH, Cho S. Re-ranking using WordNet in Tag-based Web Image Retrieval. Proceedings of the Korea Multimedia Society Fall Conference, 2008.
4. Lee SJ, Cho S. Tagged Web Image Retrieval Re-ranking with Wikipedia-based Semantic Relatedness. Journal of Korea Multimedia Society, 2011;1(11): 1491-99
5. Kim SR, Yoo HJ, Son J, Oh CB, Sohn KH. A Scale-Space based on Bilateral Filtering for Robust Feature Detection in SIFT. Proceedings of the Korea Society of Broadcast Engineers Fall Conference, Jeju University, 2012.

6. Caltech 101 Dataset Web Site http://www.vision.caltech.edu/Image_Datasets/Caltech101/.
7. Andrea V, Brian FK. VLFeat - An open and portable library of computer vision algorithms. Proceedings of the international conference on Multimedia, NY, 2010.
8. Charles E. Using the triangle inequality to accelerate k-means. Proceedings of the Twentieth International Conference on Machine Learning, Washington DC, 2003.
9. Yang Y, Newsam S. Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classification, ACM GIS 2010, San Jose, CA, USA.
10. Lee K, Kim D, Kim HJ. A Survey on Tagging in the Web 2.0 Environment, Communications of the Korea Information Science Society 2007;25(10):36-42.
11. Shai SS, Yoram S, Nathan S. Pegasos: Primal Estimated sub-GrAdient SOlver for SVM. Proceedings of the 24th International Conference on Machine Learning, 2007.
12. Bay H, Ess A, Tuytelaars T, Gool LV. Speeded-Up Robust Features (SURF), Computer Vision and Image Understanding 2008;110(3): 346-59
13. Lim HS, Lee KS. Comparison of Sift-based feature matching using outline extraction method in 3D image. Proceedings of the Korea Institute of Electronic Communication Sciences Fall Conference, 2009.
14. Jeong HJ, Lee JM, Nang JH. Image Categorization Using SIFT Bag of word. Proceedings of the Korea Computer Congress, 2013.
15. Choi GR, Jung HW, Lee JH. Contents-based Image Retrieval System Design of Shopping Mall using SIFT Matching. Proceedings of the KIIS Spring Conference, 2012.
16. Richard S. Computer Vision: Algorithms and Applications, New York, 2010;289-95.
17. Kwon BJ, Kim SN, Lee KJ, Yun ID, Lee SU. Rotation-invariant Object Categorization using Bag-of-features with Angular Pyramid. Proceedings of the Korean Society of Broadcast Engineers Fall Conference, 2011.

5/26/2014