NII, Japan at ImageCLEF 2011 Photo Annotation Task

Duy-Dinh Le and Shin'ichi Satoh

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo, Japan 101-8430
ledduy@nii.ac.jp, satoh@nii.ac.jp
Demo available at
http://satoh-lab.ex.nii.ac.jp/users/ledduy/Demo-ImageCLEF

Abstract. In this paper, we describe the participation of NII, Japan in the ImageCLEF 2011 Photo Annotation Task. Using NII-KAORI-SECODE framework, we evaluate the performance of global features and local features for this task. As for global features, color moments, color histogram, edge orientation histogram, and local binary patterns are used. As for local features, keypoint detectors such as Harris Laplace, Hessian Laplace, Harris Affine, Dense Sampling are used to extract SIFT-descriptor. The results obtained by our runs are presented.

1 Introduction

We have developed NII-KAORI-SECODE, a general framework for semantic concept detection, and use it to participate several benchmarks such as IMAGE-CLEF, PASCAL-VOC and TRECVID. The purpose is to evaluate performance of various visual representations for concept detection-like task. In this framework, first features are extracted from keyframes, then concept detectors using these features are learned by using SVM with χ^2RBF kernel. The probability output scores of the learned concept detectors are used for ranking.

2 Feature Extraction

We evaluate both global features and local features. The global features include color moments, color histogram, edge orientation histogram, and local binary patterns. The local feature is based on the BOW model in which the SIFT descriptor is extracted at interest points detected by Harris Hessian Laplace and multi-scale dense sampling detector. The details are described below.

2.1 Color Moments (CM)

Color moments have been successfully used in retrieval systems [1] and proved to be efficient and effective in representing color distributions of images [2]. The first order (mean), the second order (variance) and the third order (skewness) color moments are defined as:

$$\mu_i = \frac{1}{N} \sum_{j=1}^{N} f_{ij}$$

$$\sigma_i = (\frac{1}{N} \sum_{j=1}^{N} (f_{ij} - \mu_i)^2)^{\frac{1}{2}}$$

$$s_i = (\frac{1}{N} \sum_{j=1}^{N} (f_{ij} - \mu_i)^3)^{\frac{1}{3}}$$

where f_{ij} is the value of the i-th color component of the image pixel j, and N is the number of pixels in the image.

2.2 Color Histogram (CH)

Given a color space C , the color histogram H of image I is defined

$$H_C(I) = \{N(I, C_i) | i \in [1, .., n]\}$$

where $N(I, C_i)$ is the number of pixels of image I that fall into cell C_i and i indicates the color level of color space C. $H_C(I)$ shows the proportion of pixels of each color within the image.

2.3 Edge Orientation Histogram (EOH)

Edge orientation histogram has also been used widely in object detection and recognition [3]. The basic steps to compute edge orientation histogram feature are as follows [4]:

- Extract edges from the input image by using Canny edge detector.
- Compute a $k+1$-bin histogram of edge and non-edge pixels. The first k bins are used to represent edge directions quantized at 5^o interval and the remaining bin is used for counting non-edge pixels. The histogram is normalized by the number of all pixels to compensate for different image sizes.

2.4 Local Binary Patterns (LBP)

The LBP operator proposed by Ojala et al. [5] is a powerful method for texture description. It is invariant with respect to monotonic grey-scale changes, hence no grey-scale normalization needs to be done prior to applying the LBP operator. This operator labels the pixels of an image by thresholding the neighborhoods of each pixel with the center value and considering the result as a binary number.

2.5 Local Feature

The local feature based on BoW model is the state of the art feature for object categorization and concept detection. We adapt the implementation described in [6] for extracting local feature. Harris-Hessian-Laplace and multi-scale dense sampling keypoint detector is used to extract interest points from which SIFT descriptor is extracted. These descriptors are quantized in visual words. To form the codebook, approximate 1 million descriptors randomly selected from all descriptors extracted from the training set are used for clustering. We use k-mean clustering method to group these descriptors into 500 clusters. The codebook is formed by picking 500 cluster centers computed from the 500 clusters. Soft assignment is used to form the feature vector.

3 Feature Configuration

3.1 Granularity

Since global features do not capture spatial information, to overcome this problem, a grid $n \times m$ is usually used to divide the input image into non overlapping sub-regions. The features extracted from these regions are concatenated to form the feature vector for the image.

3.2 Color space

Local binary patterns and edge orientation histogram are extracted from gray scale image. For color moments and color histogram, color spaces including HSV, RGB, Luv, and YCrCb are used.

3.3 Quantization

For color histogram, we only use 8-bin histogram for each channel. For edge orientation histogram, we quantize orientations into histograms of 12+1 bins, 18+1 bins, 36+1 bins, and 72+1 bins. For local binary patterns, we quantize binary patterns into histograms of 10, 30, and 59 bins.

Each combination of feature type, granularity, quantization, and color space forms one feature configuration. The feature configurations evaluated in this study are described in Table 2.

4 Classifier Learning

LibSVM [7] is used to train SVM classifiers. The extracted features are scaled to $[0, 1]$ using the svm-scale tool of LibSVM. In order to handle the problem of imbalanced training sets (more than 99% of training samples are negative), we select three different training sets that are subsets of the original training sets and train three different classifiers. For each training set, we randomly select

Table 1. Feature configurations.

Feature Type	Granularity	Color Space	Quantization (#Bins)	Total Configs
Color moments (CM)	2x2, 3x3, 4x4, 5x5, 6x6	HSV, Luv, RGB, YCrCb	3x3	20
Color histogram (CH)	2x2, 3x3, 4x4, 5x5, 6x6	HSV, Luv, RGB, YCrCb	8x3	20
Local binary patterns (LBP)	2x2, 3x3, 4x4, 5x5, 6x6	GRAY	10, 30, 59	15
Edge orientation histogram (EOH)	2x2, 3x3, 4x4, 5x5, 6x6	GRAY	12, 18, 36, 72	20
Local features (harhes, harlap, heslap, haraff, hesaff, dense, phow)	1x1, 2x2, 1x3, 3x1	GRAY	500 visual words	44

a maximum of 10,000 samples for the positive set and 20,000 samples for the negative set. Since the number of positive samples in the original training set is small, usually less than 1,000, the three positive sets are usually the same.

The χ^2RBF kernel is used as similarity measure since its good performance for this task is proved in [6]. The optimal (C, g) parameters for learning SVM classifiers are found by conducting a grid search with 5-fold cross validation on a subset of 3,000 samples stratified selected from the original dataset.

5 Photo Annotation

We use probabilistic output returned by learned classifiers when applied to keyframes as scores. We use late fusion in which the keyframe scores of classifiers are averaged to form the fused score for each keyframe. These scores are used to form the ranked list for evaluating detection performance.

6 Results

Our implementation of local features has bug in concatenating feature vectors from sub-regions of the grid. We reported in Table 2 performance of both runs with bugs and runs without bugs.

Although performance of global features is worse than that of local features, the computational cost for feature extraction is significantly low. The result can help to decide the balance in using local features and global features in each context.

More details of performance of feature configurations are available at http://satoh-lab.ex.nii.ac.jp/users/ledduy/Demo-ImageCLEF

Table 2. Performance of feature configs.

Feature Config	MAP (%)	Note
Local Feature Full + Global Feature	35.8854	Bug free
Local Feature Full + Global Feature - Bug	33.7069	Submitted run NII.R1 but having bugs
Local Feature Full	34.4895	Bug free
Local Feature Full - Bug	33.0984	Submitted run NII.R2 but having bugs
Local Feature Light	33.4979	Submitted run NII.R4 but having bugs
Local Feature Light +Global Feature	32.3061	Submitted run NII.R3 but having bugs
Global Feature	27.8498	Submitted run NII.R5

References

1. Flickner, M., Sawhney, H.S., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., Yanker, P.: Query by image and video content: The qbic system. IEEE Computer **28** (1995) 23–32
2. Stricker, M.A., Orengo, M.: Similarity of color images. In: Proc. of SPIE, Storage and Retrieval for Image and Video Databases III. Volume 2420. (1995) 381–392
3. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004) 91 – 110
4. Vailaya, A., Jain, A., Zhang, H.: On image classification: city images vs. landscapes. Pattern Recognition **31** (1998) 1921–1935
5. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. on Pattern Analysis and Machine Intelligence **24** (2002) 971987
6. Jiang, Y.G., Yang, J., Ngo, C.W., Hauptmann, A.: Representations of keypoint-based semantic concept detection: A comprehensive study. IEEE Transactions on Multimedia **12** (2010) 42–53
7. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. (2001) Software available at `http://www.csie.ntu.edu.tw/" "cjlin/libsvm`.