KIDS-NUTN at ImageCLEF 2012 Photo Annotation and Retrieval Task

Been-Chian Chien, Guan-Bin Chen, Li-Ji Gaou, Chia-Wei Ku,
Rong-Sing Huang, and Siao-En Wang

Knowledge, Information, and Database System Laboratory
Department of Computer Science and Information Engineering
National University of Tainan, Tainan, Taiwan
bcchien@mail.nutn.edu.tw

Abstract. The task of visual concept detection, annotation, and retrieval using Flickr photos at ImageCLEF 2012 was organized as two subtasks: concept annotation and concept retrieval. In this paper, we present the effort of KIDS lab for the two subtasks. The proposed approaches combine various visual and textual features, dimension reduction methods, the random forest classification models, and the semi-supervised learning strategy. For the concept annotation subtask, the annotation results show that combination of tags and visual features outperforms visual-only features while using the same classification model. The results also show that semi-supervised learning is not superior to supervised learning in this subtask. Further, it does not seem able to gain more advantage on F-measure when more different visual features were used. For the concept retrieval task, the results illustrate that the textual features contain much richer informatics than visual features in general retrieved concepts.

Keywords: ImageCLEF, Image Annotation, Image Classification, Image retrieval, Random Forest

1 Introduction

The ImageCLEF 2012 visual concept detection, annotation, and retrieval task using Flickr photos arranged two subtasks [13]: the concept annotation task and the concept-based retrieval task. The challenge of the first subtask, concept annotation, is to assign each image to a set of concepts taken from a list of 94 pre-defined concepts automatically. This task takes a subset of the MIRFLICKR-1M collection containing 15,000 images for training and 10,000 images for testing. The second subtask, concept-based retrieval, aims on retrieving target images from a subset of the MIRFLICKR collection comprising of 200,000 photos for 42 concept queries. The queries are provided in the XML format containing title, description, and three annotated images located in the training set of subtask 1. In this paper, we describe the approaches used in the two subtasks including extraction of image features, concept learning models and concept retrieval methods.

The annotated concepts in this task cover a wide range of topics such as natural scene, animal kinds, human gender, human emotions, transportation tools, etc. Some of the concepts are so abstract and ambiguous in semantics that even users can not annotate the images well. In order to annotate images precisely, we worked at the tasks in the following aspects. First, various visual features are extracted from images to investigate the correlation between the visual features and the annotation concepts. Second, the high-dimension and multilingual textual tags need to be analyzed, processed and reduced for improving the efficiency of image annotation and retrieval in large datasets. Third, effective classification models and efficient learning methods are necessary for integrating visual and textual features to generate multi-label classifiers in annotating numerous concepts.

Our concept annotation approaches are the combinations of different techniques including image features extraction, text processing, dimension reduction, the random forest classification models, and the semi-supervised learning strategy. The validation set used 5,000 images selected from the given 15,000 training images and the left 10,000 images were used for training in our evaluation. After tuning the parameters, the final classification models are learned from the total 15,000 training images and the 10,000 testing images were annotated. We also used the modified MBRM [4] method as a baseline method to observe and compare the effectiveness of the annotation approaches. The concept retrieval approaches are based on the concept annotation approaches. This paper will also discuss the ranking method of images retrieval from the given three query images for the subtask 2.

The rest of this paper is organized as follows. In section 2, we describe the extraction and preprocess of visual features and textual features, respectively. The feature reduction method, concept learning methods and classification models are presented in Section 3. The concepts annotation and retrieval methods are also given in this section. Section 4 shows and discusses the experimental results for different submission runs. Finally, we draw a conclusion for our labs in Section 5.

2 Extraction and Preprocess of Image Features

2.1 Feature Extraction

The original image data set in the Flickr photo task consists of JPEG images, EXIFs in image files, and supplementary tags for each image. The main image features are thus considered to be extracted from the JPEG image and the textual part. In this subsection we first introduce the extraction of visual features and textual features, respectively. Then, the process of normalization on visual features is described in the next subsection.

Visual Features. The annotated concepts in the Flickr photo task are very diverse. Although the total annotated 94 concepts are categorized as natural elements, environment, people, image elements, and human elements, the job of concept annotating is still ambiguous and vague in visual for photos from the viewpoints of different persons. For collecting visual features as many as possible from an image, first, an image was equally segmented into 16 sub-images (in 4 by 4 blocks). The original image and its corresponding 16 sub-images, totally 17 images, are the sources of generating visual features. Basically, the four visual features, AutoColorCorrelogram [5], ColorLayout [1], FCTH [2], Gabor [11], were extracted from each original image and 16 sub-images. Gist [12] feature is only applied to the original images. Each image generates a list of multi-dimensional data.

Except for extracting the five visual features, the region of interest (ROI) in original images are also marked automatically by the visual attention model proposed by Itti, *et al.* [7]. We modified the method by applying 6-level Gaussian pyramid to generate the saliency map representing the degree of concern in an image. Then, the region growing method [13] was used to mark the appropriate ROIs. For the 16 blocks of sub-image, each block of sub-image is marked as foreground if the area of a block is covered over 60% by marked ROIs; otherwise, the block is marked as background. The AutoColorCorrelogram values in the blocks of background sub-image then were averaged as the visual feature of ROI background.

To recognize the number of people in photos, the package of face detection in OpenCV was used to detect and estimate the number of persons in each photo. The numbers of dimensions for the extracted visual features are summarized in Table 1.

The visual features including AutoColorCorrelogram, ColorLayout, FCTH, and Gabor were extracted by applying LIRE (Lucene Image REtrieval) JAVA library.[1] The face detection tool was implemented by using OpenCV.[2] The ROI marking method and the Gist method were designed and implemented by ourselves.

Table 1. The used visual features.

Visual features	Feature dimensions	#Images	Total
AutoColorCorrelogram [5]	1024	17	17408
ColorLayout [11]	120	17	2040
FCTH [2]	192	17	3264
Gabor [11]	60	17	1020
Gist [12]	192	1	192
ROI background [7,13]	16	1	1024

Textual Features. The textual information for the Flickr photos comes from the EXIF (Exchangeable image file format) and announced tags of each image. The EXIF is a standard specification that specifies the formats of media data like images and sounds produced by digital cameras. A given EXIF contains 407 fields totally in each image, but only 24 EXIF fields were selected (e.g. black level, blur warning, brightness, compression, contrast, data and time, zoom, expiration, ISO, noise, etc.).

The other source of textual features is the description file of tags for each image. The tags describe some kinds of related semantic information of the images. Before

[1] http://www.semanticmetadata.net/lire

[2] http://opencv.org

applying the tags information to annotate images, two problems should be solved. First, the practical tags of images are multilingual. Actually, more than 68 different languages are found in the set of tags. Synonyms of terms need to be unified. Second, the problem of high-dimension features must be reduced. To resolve the two problems, the Google translation tools[3] were used to translate the multilingual tags into English, and the stop words then were deleted from the set of tags. The number of the final tags is 60821 terms in English. The term frequency for each tag was also counted and recorded.

Further, in order to support the detection of humans in photos, the package of face detection in OpenCV was used to detect and estimate the number of persons in each photo. The range of the estimated number of people is between 0 and 13. The face number for each photo is marked by binary information as 14 features. The numbers of final textual features are listed in Table 2.

Table 2. The used textual features.

Textual features	original	After extraction
Number of faces	1	14
EXIF	407	24
Tags	69099	60821
Total	69507	60859

2.2 Preprocess of Visual Features

The final extracted visual features and textual features in subsection 2.1 are quite various in dimensions and ranges of values. Furthermore, the high-dimension feature is an important problem for learners to generate annotating models. For reducing the dimensions and combining the extracted visual and textual features in a unified representation, the visual features are processed as follows.

Let $\mathcal{I}$ be an image set with n images and I_i be an image in $\mathcal{I}$. The $\mathbf{x}$ denotes the vector of a specified visual feature and $\mathbf{x}_i$ represents the multi-dimension vector of the specified visual feature for the image I_i. We have $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$, an m dimensional vector, and x_{ij} is the value of jth-dimension of the $\mathbf{x}_i$ for the image I_i, $1 \le j \le m$. We assume that $C_1, C_2, \ldots, C_K$ represent the K possible annotated concepts in the system. $|C_k|$ denotes the number of images belonging to the concept C_k in the image set $\mathcal{I}$. We first calculate the *mean* vector $\boldsymbol{\mu}_k$ and *deviation* vector $\boldsymbol{\sigma}_k$ of the visual feature for each concept C_k, as follows:

$$\boldsymbol{\mu}_k = \frac{\sum_{I_i \in C_k} \mathbf{x}_i}{|C_k|}, \quad \boldsymbol{\sigma}_k = \sqrt{\frac{\sum_{I_i \in C_k} (\mathbf{x}_i - \boldsymbol{\mu}_k)^2}{|C_k|}}, \quad 1 \le k \le K. \tag{1}$$

[3] http://translate.google.com

Then, the *concept similarity* of the visual feature $\mathbf{x}_i$ corresponding to the concept C_k can be defined as

$$y_{ik} = \prod_{j=1}^{m} \exp\left[-\left(\frac{x_{ij} - \mu_{kj}}{\sigma_{kj}} \right)^2 \right], \quad 1 \le k \le K, \tag{2}$$

where μ_{kj} and σ_{kj} are the jth-dimension values of the vectors $\boldsymbol{\mu}_k$ and $\boldsymbol{\sigma}_k$, respectively, for the image $I_i \in \mathcal{I}$, $1 \le i \le n$. Hence, a m-dimension visual feature $\mathbf{x}_i$ of an image I_i will be normalized as a K-dimension features $\mathbf{y}_i = [y_{ik}]$, $0 \le y_{ik} \le 1$, $1 \le k \le K$. Since multidimensional visual features of an image, as shown in Table 1, were transformed into 94 dimensions (the number of concepts), the total number of features is 6580 after the processing.

3 Feature Reduction and Concept Learning Models

Although we had reduced part of the number of features, the extracted visual features and textual features in the previous section still have very high dimensions (67439 features in total). Generally, it is not easy for any classification model to learn effective classifiers from high dimensional datasets efficiently. For dealing with the high dimensional datasets, we applied a feature reduction method, discriminant coefficient [9, 10], to reduce dimensions before learning the classifiers. The submitted runs are mainly based on two learning models: the random decision trees [3] and the Multiple Bernoulli Relevance Models (MBRM) [4]. Except for the supervised learning strategy, the semi-supervised learning strategy is also considered for investigating the feasibility in image annotation. In this section, we briefly present the main methods used in this task including the feature reduction method, the concept classification models and the leaning strategies in the following subsection.

3.1 Features Reduction

The reduction method is based on the discriminant coefficient proposed by Lin & Chien [9, 10]. In the method, the discriminant coefficients are calculated by the difference between the statistics of two classes. Before calculating the discriminant coefficients, the image features need to be normalized according to the class of concept. Let $\mathbf{y}_i$ be the concept similarity of visual features $\mathbf{x}_i$ as defined in Section 2.2 and y_{ij} is the jth dimension of transformed concept similarity for a visual feature in the image I_i. For textual features, $\mathbf{y}_i$ is the term frequency of textual features and y_{ij} is the term frequency of the term j for the image I_i. The normalization of visual and textual features are defined as

$$f_{kj} = \frac{\sum_{I_i \in C_k} y_{ij}}{|C_k|}, \quad \text{for } 1 \le k \le K. \tag{3}$$

The normalized features can be denoted as a matrix $\boldsymbol{F} = [f_{ij}]_{K \times P}$, K is the number of conceptual classes, and P is the number of all transformed visual features and all final extracted textual features. The feature reduction method in [10] first calculates the relative discriminant variables of each feature for all K conceptual classes. Then, the discriminant variables are normalized to be the *log-scaled discriminant coefficient matrix* $\boldsymbol{J} = [J_{ij}]_{K \times P}$. The range of J_{ij} is between 0 and 1. A large J_{ij} represents that the jth feature has high discrimination on the concept C_i. On the contrary, a small J_{ij} value means that the jth feature provides less discernable information for the concept C_i.

We assume that the matrix $\boldsymbol{Y} = [y_{ij}]_{n \times P}$, y_{ij} is the visual and textual features for the image I_i and n is the number of images in the training set. Finally, the goal of feature reduction is to find a transformation matrix $\boldsymbol{T}$ such that the number of visual and textual features is much smaller than the original features. The transformation of feature reduction can be completed by the following equation:

$$\boldsymbol{T} = \boldsymbol{Y} \cdot \boldsymbol{J}^{t}, \qquad (4)$$

where $\boldsymbol{J}^{t}$ is the transpose of matrix $\boldsymbol{J}$. After transforming of the equation (4), the $\boldsymbol{T}$ is a $n \times K$ matrix which is used to replace the matrix $\boldsymbol{Y}$ as the reduced features of training set for learning models.

3.2 Random Forest

The random decision tree method [3] is an ensemble classifier that first builds a number of decision trees randomly. Each decision tree is constructed by selecting a non-tested feature randomly as the decision node at each level. The training data are not used in the tree construction and is independent from the tree structures completely. After the decision trees are built, the training data are used to update the statistics of the classes at each node for all random decision trees. While classifying an unknown example, the predicted class is estimated by trees voting or averaging the possibilities of all decision trees to determine the classification result.

The Matlab code[4] of the Random Forests was used in the task. For dealing with the multi-label problem, a two-class classifier was learned for each annotation concept. Although totally 94 classifiers should be learned, the random forest method is still efficient because the reduced matrix $\boldsymbol{T}$ is used to be the training set.

3.3 Semi-supervised Learning

Generally, the training set containing only labeled data are applied to build classifiers by supervised learning method. In this paper, we also investigated the feasibility of applying semi-supervised learning. Semi-supervised learning uses both labeled data and unlabeled data to perform the learning process. The goal is to integrate the unlabeled data to improve the effectiveness of classification.

[4] http://code.google.com/p/randomforest-matlab/

The first step of using semi-supervised learning is to use all ground truth labeled data to learn classifiers as Section 3.2. Next, the unlabeled data are classified and ranked by their voting ratios of random decision trees. The top 10 positive examples and top 10 negative examples from the unlabeled data are then added to the training set, and new classifiers are re-trained. Such a learning process proceeds *k* times iteratively. The final classification models are used to annotate the concept of images.

3.4 Multiple Bernoulli Relevance Models method

The Multiple Bernoulli Relevance Models method (MBRM) was proposed by Feng, *et al.* 4] to solving the problem of automatic image annotation. This method and its modified weighting version were implemented as the baseline methods in the annotation task. We briefly introduce the method in the following.

Let $\mathcal{J}$ denote the training set of annotated images, and I_i be an image of $\mathcal{J}$. Every image I_i were cut into 16 blocks in 4 by 4 rectangular sub-images. We obtain one original image r_0 and the 16 sub-images r_1, r_2… r_{16}. We then extract features from the 17 regions separately and assume that the features are denoted as $f_0 \ldots f_{16}$.

Now let I_j be a test image, and $f'_0 \ldots f'_{16}$ denote the features of image I_j. The joint probability $P(I_j, w)$ is computed for each word w in the annotation vocabulary. The annotations of image I_j would be the top several words which have maximum probabilities. The joint probability $P(I_j, w)$ is defined as following equation:

$$P(I_j, w) = \sum_{I_i \in \mathcal{J}} \left\{ P_T(I_i) \times \prod_{p=0}^{16} \sum_{q=0}^{16} Sim(f_p, f'_q) \times P(w \mid I_i) \right\}. \tag{5}$$

We assume that the distribution of the training set is an uniform distribution, the probability $P_T(I_i) = 1/n$, where n is the number of images in the training set $\mathcal{J}$. The $Sim(f_p, f'_q)$ stands for the similarity degree between the features f_p and f'_q, and the $P(w|I_i)$ is defined as following equation:

$$P(w \mid I_i) = \frac{\mu \cdot N_{(w,I_i)} + N_{(w,\mathcal{J})}}{\mu + n}, \tag{6}$$

where $N_{(w,\, Ii)}$ denotes the number of times the annotation w occurs in the image I_i, $N_{(w,\mathcal{J})}$ denotes the number of times the annotation w occurs in the training set $\mathcal{J}$, and μ is the smooth parameter.

4 The Concept Retrieval Method

The concept retrieval task gave 42 queries containing a concept title, text description, and three images for each query. The query images are all in the training set and the test database comprises 20,000 photos selected from the MIRFLICKR collection. The approaches of retrieving the images with the same concept as the query are based on the concept annotation approaches in the subtask 1.

We first analyzed the concept ratios of the three images for each query. Next, we applied the concept annotation approaches used in the subtask 1 to annotate images in the test database. Finally, the images were ranked by the concept ratios and the voting ratios of random decision trees. Formally, we assume that the three query images are annotated by a few concepts individually. Let τ_{ij} be the voting ratio of the jth concept on the image I_i , and w_j be the ratio of the concept C_i annotated by the three images, $1 \le j \le K$, where K is the number of concepts. The similarity degree of the image I_i for the query Q are defined as

$$Sim(Q,I_i) = \sum_{j=1}^{K} \tau_{ij} \cdot w_j \, . \tag{7}$$

5 Experimental Results and Discussions

5.1 The Concept Annotation Task

First, we would like to introduce the methods for the runs we submitted. In concept annotation subtask, we totally submitted five runs which based on two methods. One is the approach based on the feature reduction and random decision trees described in Section 3.1 to 3.3. The other is based the Multiple Bernoulli Relevance Models (MBRM) described in Section 3.4. The run_a1, run_a2, and run_a3 use the former method. Especially, the run_a2 applies the semi-supervised learning instead of super-vised-learning. The run_a4 and run_a5 take the latter one. The run_a4 used annotation scores to weight the probabilities of words in images. The run_a5 only considered binary annotations. The used features and methods are summarized in Table 3.

Table 3. The features and methods used in the submission runs.

	Features	run_a1	run_a2	run_a3	run_a4	run_a5
Visual features	AutoColorCorrelogram	○	○	○		
	ColorLayout	○	○	○	○	○
	FCTH	○	○	○		
	Gabor	○	○	○		
	Gist	○	○	○		
	ROI	○	○	○		
Textual features	Face detection	○	○	○		
	EXIF	○	○	○		
	Tags			○		
Classification models	Random Forest	○	○	○		
	MBRM				○	○
Learning methods	Semi-supervised learning		○			
	Feature reduction	○	○	○		
	Weighting features				○	

As Table 3 shows, the used features in run_a1 and run_a2 are the same including visual features and part of the EXIF metadata. Except for the visual and EXIF features, the run_a3 employed tags as a part of textual features. Since the MBRM method is time consuming and the time complexity is dependent on the number of features, the run_a4 and run_a5 only considered the visual feature ColorLayout. The two runs are used to be the baselines while comparing with the other runs having multi-features.

The results of evaluation are shown in Table 4. The measures are low in MiAP and GMiAP because our methods are only concerned with binary annotation for each image. The confidence scores in the submission runs were produced by the voting ratios of ensemble random decision trees. We use the voting ratio 0.5 to be threshold of annotating images for a concept. However, the voting ratio generally cannot stand for the annotation confidence score of an image. Actually, we think that neither MiAP nor GMiAP measure is appropriate to be the metric in this subtask.

Table 4. The results of the concept annotation task.

Measures	run_a1	run_a2	run_a3	run_a4	run_a5
MiAP	0.1022	0.1018	0.1717	0.0947	0.0985
GMiAP	0.0470	0.0472	0.0984	0.0495	0.0537
Precision	0.6257	0.5860	0.6313	0.6339	0.6414
Recall	0.2588	0.2153	0.3384	0.2422	0.2385
F-ex	0.3662	0.3149	0.4406	0.3505	0.3478

Table 5. The results of F1-measure for different concept categories.

Concept categories	run_a1	run_a2	run_a3	run_a4	run_a5
time of day	0.2751	0.2828	0.3447	0.1419	0.1065
celestial bodies	0.0000	0.0000	0.0767	0.0000	0.0000
weather	0.1043	0.1043	0.1154	0.0000	0.0000
combustion	0.0000	0.0000	0.0000	0.0000	0.0000
lighting effects	0.0052	0.0039	0.0423	0.0000	0.0000
scnery	0.0156	0.0175	0.1376	0.0000	0.0000
water	0.0029	0.0029	0.0587	0.0000	0.0000
flora	0.1064	0.1111	0.2656	0.0000	0.0000
fauna	0.0000	0.0024	0.2557	0.0000	0.0000
quantity	0.6847	0.5905	0.7362	0.6902	0.6902
age	0.1239	0.0994	0.4011	0.0425	0.0385
gender	0.0622	0.0598	0.3423	0.0035	0.0081
relationship	0.0000	0.0022	0.0574	0.0000	0.0000
quality	0.6707	0.5959	0.6902	0.6622	0.6619
style	0.0000	0.0074	0.0618	0.0000	0.0000
view	0.1829	0.1931	0.3428	0.0964	0.0933
type	0.0093	0.0043	0.1920	0.0000	0.0000
impression	0.0144	0.0137	0.1048	0.0003	0.0003
transportation	0.0027	0.0000	0.1509	0.0000	0.0000

Table 4 shows that all runs have high precision and low recall. The two baseline methods, MBRM and weighted MBRB, used only one visual feature to classify the concept. The run_4 applying weight scores is not improved much in comparison with the run_5. The results of supervised learning on random trees (run_a1) are better than semi-supervise learning (run_a2) in this task. Further, the results of run_a3 using supervised learning, visual features, and tags are the best. The evaluation results of F1-measure for different concept categories are also shown in Table 5.

From the results some remarkable characteristics are discussed as follows:

- The information of tags does improve the performance of automatic image annotation no matter what measures are in general.
- All our approaches have higher precision rates and lower recall rates. For run_a1, run_a2 and run_a3, the features extracted by the feature reduction method based on discriminant coefficient are high discernible. The lower discernible features are eliminated. These effects should be the main reason of high precision rates and low recall rates for such a kind of method. As run_a4 and run_a5, the selected high threshold of probability $P(I_j, w)$ might be the cause of low recall for the MBRM method.
- The semi-supervised learning did not outperform supervised learning in this subtask. The reason might be caused by the ranking of voting ratio in random trees. As above mentioned, the voting ratio cannot reflect the confidence score of image annotation. The classified test images did not ranked and added to the training set correctly.
- Generally, the concept categories with high annotation rates like *quality* and *quantity* have much more positive examples and obvious visual features in the training set and testing set.
- The concept categories with very low annotation rates such as *combustion* and *relationship* are usually few examples, abstract concept or highly dependent on semantics. It is difficult for image analyzers to find a general model for different kinds of special visual concepts.

5.2 The Concept Retrieval Task

The results of concept retrieval are listed in Table 6. The three submission runs, run_r1, run_r2 and run_r3, are based on the annotation methods used in run_a1, run_a2 and run_a3, respectively. Since the methods in the concept retrieval subtask were accomplished by the annotation results of subtask 1, the performances are highly dependent upon the effectiveness of annotation results. It is obvious that run_r3 has the best results because of the higher annotation rate in run_a3. The others get low precisions. The results also show that the tags are the important factors of retrieving relevant images. Using visual features only may not retrieve correct concepts from a large amount of general images. The semantics inside images still need appropriate textual notation.

Table 4. The results of the concept retrieval task.

Measures	run_r1	run_r2	run_r3
MnAP	0.0009	0.0007	0.0313
AP@10	0.0003	0.0006	0.0051
AP@20	0.0010	0.0014	0.0077
AP@100	0.0096	0.0081	0.0729

6 Conclusion

This is the first time to participate the photo annotation task for our lab. Owing to many abstract concepts cannot be described by general visual features, the innovating effective visual features for representing various concepts is important to annotate images precisely. In this paper we present the annotation methods based on precise feature reduction and the random decision trees model. All the used visual features and textual features can be extracted from images generally and easily. The best result is the model combining general visual features and tags. The combination of various visual features does not seem to improve the performance much more than only one visual feature. We also found that the different visual features usually worked well in specific concepts. The performance should be able to be improved if the appropriate visual features could be selected and used in the specific concept.

After submission of the task, some analyses on general visual features and the learning strategies were made. Special features extracting models are necessary for learning classifiers to annotate concepts correctly. For example, the visual concepts, like shadow and refection in the lighting effects category, can be marked or modeled as specific regions or representations. We believe that the representative features, effective feature selection methods and machine learning models will be the solution of annotating specific concepts. However, it should be no direct answer for general visual features to detect concepts effectively.

Acknowledgements. This research was supported in part by the National Science Council of Taiwan, R. O. C. under contract NSC101-2221-E-024-026.

References

1. Chang, S. F., Sikora, T., Puri, A.: Overview of the mpeg-7 standard. IEEE Transactions on Circuits and Systems for Video Technology, pp. 688-695 (2001)
2. Chatzichristofis, S. A., Boutalis, Y. S.: FECH: Fuzzy color and texture histogram a low level feature for accurate image retrieval. In: the 9th International Workshop on Image Analysis for Multimedia Interactive Services, Klagenfurt, Austria, pp. 191-196 (2008)
3. Chen, C., Liaw, A., Breiman, L.: Using random forest to learn imbalanced data. Technical Report. no. 666, Department of Statistics, University of Berkeley (2004)
4. Feng, S. L., Manmatha, R., Lavrenko, V.: Multiple Bernoulli reference models for image and video annotation. In: IEEE International Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, pp. 1002-1009 (2004)

5. Huang, J., Kumar, S.R., Mitra, M., Zhu, W.J., Zabih R.: Image Indexing Using Color Correlograms. In: the International Conference on Computer Vision and Pattern Recognition. San Juan, Puerto Rico, pp. 762-768 (1997)
6. Huiskes, M. J., Lew, M. S.: The MIR Flickr retrieval evaluation. In: ACM International Conference on Multimedia Information Retrieval (MIR'08). Vancouver, Canada (2008)
7. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transaction on Pattern Analysis and Machine Intelligence. vol. 20, pp. 1254-1259 (1998)
8. Lienhart, R., Maydt, J.: An extended set of Haar-like features for rapid object detection. In: IEEE ICIP 2002. vol. 1, pp. 900-903. (2002)
9. Lin, Y. X., Chien, B. C.: A discriminant based document analysis for text classification. In: the International Computer Symposium, Workshop of Artificial Intelligence, Knowledge Discovery, and Fuzzy Systems, Dec. 16-18, 2010, Tainan, Taiwan, pp. 594-599.
10. Lin, Y. X., Chien, B. C.: Efficient feature reduction for high-precision Text classification. In: the National Computer Symposium on Databases, Data Mining, and Information Retrieval. Chia-Yi, Taiwan (2011)
11. Manjunath, B. S., Ma, W. Y.: Texture features for browsing and retrieval of large image data. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18 (8), August, pp. 837-842. (1996)
12. Siagian, C., Itti, L.: Rapid biologically-inspired scene classification using features shared with visual attention. IEEE Transactions on Pattern Analysis and Machine Intelligence. vol. 29, no. 2, pp. 300-312. (2007)
13. Thomee, B., Popescu A. : Overview of the ImageCLEF 2012 Flickr Photo Annotation and Retrieval Task. In: CLEF 2012 working notes, Rome, Italy, 2012.
14. Zhai, Y., Shah, M.: Visual attention detection in video sequences using spatiotemporal cues. In: 14th Annual ACM International Conference on Multimedia. pp. 815-824. (2006)