Bag-of-Visual Words Based Automatic Image Annotation

Biniyam Kebede
HiLCoE, Computer Science Programme, Ethiopia
biniyamkg@gmail.com

Fekade Getahun
HiLCoE, Ethiopia
Department of Computer Science, Addis Ababa University, Ethiopia
fekade.getahun@aau.edu.et

Abstract

Content-based Image retrieval systems extract and retrieve images using their low-level features, such as color, texture, and shape. Nevertheless, these visual contents do not allow a user to formulate semantically meaningful image query. Image annotation systems are a solution to solve the inadequacy of CBIR systems and allow text based image retrieval. There have been several studies on automatic image annotation utilizing machine learning techniques and images' representation with low level features extracted using either global or local methods. However, typically, these approaches suffer from low correlation between the globally assigned annotations and the visual features used to obtain annotations automatically. In this paper, we present an approach to enhance the effectiveness of CBIR using learning based automatic images annotation based on bag of visual word images representation that is created automatically using a set of manually annotated training images. The experimentation is performed with 4,000 annotated images for training, 1000 images for testing from ImageNet. The result has shown 77.5% of performance accuracy. The result of this work is believed to be one step towards enhancing the performance and effectiveness of existing CBIR and minimizing the semantic gap.

Keywords: Content-based Image Retrieval; Automatic Image Annotation; Bag of Visual Words; Classification

1. Introduction

Given an example query image, content-based Image Retrieval approaches return list of images with similar low-level features (or visual content) values. However, these visual contents do not allow a user to formulate semantically meaningful image query [5]. On the other hand, in text-based image retrieval, list of images having words in the query aligned with the textual description of the depicted content are returned. While this approach is best-suited in scenarios where the desired pictorial information can be efficiently described by means of keywords, it demands for translation of the depicted contents into a textual representation (annotation). Manual image annotation is a tedious and time-consuming task [15]. Hence, automatic annotation approach infers automatically the annotations of unseen images from a set of manually annotated training example images.

According to [19], Automatic Image Annotation (AIA) can be regarded as an automatic classification of images by labeling images into one of a number of predefined classes or categories. Technically, the AIA is a two-step approach: 1) image component decomposition and representation: by decomposing an image into a collection of sub-units, which could be segmented regions, equal-size blocks or an entire image; and modeling each content unit based on a feature representation scheme; and 2) content classification: by computing the associations between unit representations and textual concepts; in this stage, higher level semantic can be learned from samples image.

However, automatic image annotation utilizing machine learning techniques and images representation with low level features suffer from low correlation between the globally assigned annotations and the visual features used to obtain annotations automatically. In this paper, we propose a bag-of-visual words, BoVW, (visual dictionary) based image classification approach for automatic image

annotation. SIFT is used for BoVW model construction and SVM multi-class classification technique for training and classification of image instance.

The rest of the paper is organized as follows. Section 2 discusses the related works on AIA. Section 3 describes the proposed approaches. Section 4 presents the experimental results. Conclusions are drawn in Section 5.

2. Related Work

A number of machine learning approaches have been explored for the AIA problem. Using set of annotated training set of images, image annotation learns the annotation of example images using the co-occurrences of words and images low-level features. Learning the correlation between global low-level image features computed per-image level and annotation data has been successfully applied in general scene classification [8, 10]. These approaches provide good results for classifying images when applied to image classes whose discriminative visual properties are spread equally over the whole image surface.

However, global visual features are often insufficient to represent the prominent objects that have been used to annotate images [3]. Hence, recently approaches in [11, 14] automatic segmentation step before the actual learning stage to identify real-world objects within the image has been used. The general assumption is that feature computation based on a potentially strong segmentation better describes the visual objects, depicted in the image, than global features. However, no general and robust automatic segmentation algorithm has been presented [26], and the existing algorithms suffer from low segmentation accuracy.

Partition-based approaches try to overcome this obstacle by decomposing the images into multiple regions of equal shapes [9]. This can be seen as a weak segmentation, which tries not to capture the shape of a visual object, but to produce multiple regions per image each corresponding to a single depicted object. This will result in more redundancy (depending on the patch size), which will help to statistically detect a correlation between global labels and local patches. This approach has been applied in work presented in [18, 19].

A problem of all the presented approaches is a typically high correlation between different annotations. Various labels that often appear together within the training images cannot be distinguished. For example, if the training images always depict sky- and tree-regions within an image together, those objects are hardly distinguishable using the presented statistical methods. Furthermore, in [17] it is argued that global annotations are more general than a pure region labeling and thus a semantic correspondence between labels and image regions does not necessarily exist e.g., an images globally labeled “wild life” might depict region for foxes as well as region for lions – deducing from both regions to a global label “wild life” is impossible with approaches that are based on color and texture features only.

In summary, the above studies show global and local image features based automatic annotation to tackle the drawback in CBIR systems. However, global features are inefficient for representing and classifying images. In our approach for automatic image annotation, we aim at taking advantage of the bag-of-visual word model extracted using SIFT to reduce drawbacks of global feature and segmentation.

3. The Proposed Approach

AIA can be regarded as an automatic classification [19] of images by labeling images into one of the predefined classes/categories, where each class has keywords or labels which can describe the conceptual content of images in the class. The major components of the proposed approach are Object detection, Machine Learning and Annotation propagation.

3.1 Object Detection

An object detection component extracts identifiable real world objects from the given image, using its descriptive visual features, which are known

a priori. The main difficulty in developing reliable object detection approach arises from the wide range of variations in images of objects belonging to the same object class [14, 25]. Different objects belonging to the same category often have large variations in appearance. A successful object detection approach must therefore be able to represent images in a manner that renders them invariant to such intra-class variations, but at the same time distinguishes images of the object from all other images. In this paper, the BOVW model is used as it is invariance to camera angle, image scale and orientation, as well as, occlusion, and lighting. In this work, learning to detect objects consists of three stages as discussed below.

BoVW Model

BoVW model was first used in text retrieval domain [9] for text document analysis, and it was further adapted for computer vision applications [12, 13]. For image analysis, a visual analogue of a word is used in the BoVW model, which is based on the vector quantization process by clustering low-level visual features of local region or points, such as color, texture, and so forth. Extracting the BoVW features from images involves the following steps as shown in Figure 1 [6].

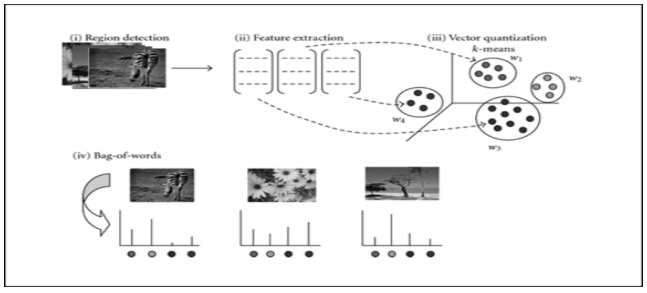


Figure 1: BoVW Construction Steps

a. Interest Point Detection

The first step of the BoVW methodology is to detect local interest keypoints that can be used to represent the objects in the target class. This is done automatically by using an interest point detector operator to extract information-rich patches from each image. The interest point detectors detect the "keypoints", salient patches, in an image. In this paper, the scale-space extrema of Differences-of-Gaussians (DoG) [2] is used for the automatic detection of key points from an image. The DoG algorithm searches over all scales and image locations to identify potential points of interest that are invariant to scale and rotation within a DoG pyramid.

b. Local Descriptors

Feature representation deals on how to represent the patches as numerical vectors or feature descriptors. In this work, a SIFT descriptor is used to extract local features that are reasonably invariant to changes in illumination, image noise, rotation, scaling, and small changes in view-point [3]. Interest points for SIFT features correspond to local extrema of DoG filters at different scale. Once a key point/interest orientation has been selected, the local feature descriptor is computed as a set of orientation histograms on 4x4 pixel neighborhoods. The orientation histograms are relative to the keypoint orientation, the orientation data comes from Gaussian image closest in scale to the keypoint's scale. A histogram contains 8 bins each, and each descriptor contains an array of 4 histograms around the key point. This leads to a SIFT feature vector with 4x4x8=128 elements. This vector is normalized to enhance invariance to changes in illumination [4]. Figure 2 [2] shows the overall SIFT based local descriptors extraction method from an image.

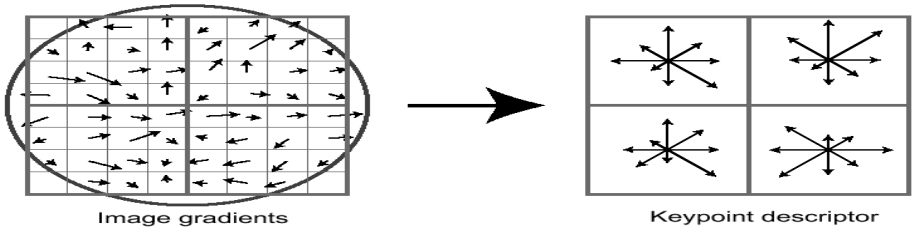


Figure 2: SIFT Descriptor

c. Visual Words Construction

The final step, visual words construction, converts vectors representing patches to "visual words" which produces a BoVW. In this paper, the K-Means algorithm [1] is exploited to cluster the vectors. The K-Means clustering aims to group *n* descriptors from all the training images into *k* clusters in which each descriptor belongs to the cluster with the nearest mean and the center of each cluster corresponding to a unique visual word. Suppose there are *n* descriptors $X = (x_1, \ldots, x_n)$, descriptor x_i is a d-dimensional vector. K-Means clustering aims to group the *n* descriptors into *k* sets $(k \leq n)M = \{m_1, \ldots, m_k\}$ and find *k* centers (descriptors) in M. It operates as follows: it starts with randomly chosen descriptor $\boldsymbol{x_i}$ assigning to the most similar cluster m_j (i.e., has shortest Euclidean distance to the center). Then, the cluster head/center is recomputed to represent information of all members. Continue the process for each descriptor.

3.2 Learning a Classifier

After the BoVW feature is extracted from images, it is entered into a classifier for training or testing. The construction of discriminative models for AIA is based on the supervised machine learning principle for pattern recognition. In this paper, a multi-class SVM* algorithm for supervised machine learning is used. SVM was originally designed for binary classification [16]. It uses the vector space model for the documents' representation and assumes that documents in the same class shall fall into separable spaces of representation. Upon this, it looks for a hyperplane that separates classes. This hyperplane should maximize the distance between it and the nearest documents. The following function is used to define the hyperplane [24] (see Figure 3).

$f(x) = w \bullet x + b$, where: $\bullet$ denotes the dot product, *w is* the normal vector to the hyperplane, *x* is a real vector space

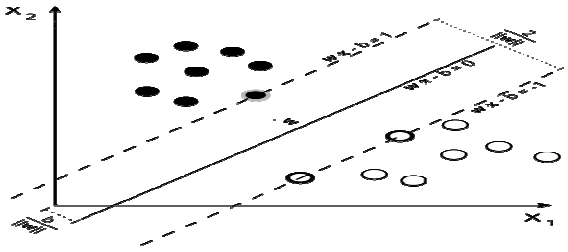


Figure 3: An Example of Binary SVM Classification, Separating Two Classes (Black Dots from White Dots)

In order to resolve this function, all the possible values should be considered and, after that, the values of w and b that maximize the margin should be selected. Moreover, SVM implements the "one-against-one" (OAO) approach for multiclass classification [17]. In Multiclass classification each training point belongs to one of N deferent classes. The goal is to construct a function which, given a new data point, will correctly predict the class to which the new point belongs.

The OAO approach works as follows: if k is the number of classes, $\frac{k(k-1)}{2}$ classifiers are constructed; and each classifier trains data from two classes, i and j. To employ multiclass (see the process in Figure 4) image classifier in SVM, each image instance has to be represented as a vector of real numbers. Each image instance of training and testing classes is represented by the BoVW model as follows.

A training dataset D containing *n* images is represented as $D = \{d_1, \ldots, d_n\}$, where each image d_i is represented by the extracted visual descriptors, supervised learning algorithm K-means, is used to group D into a fixed number of visual words W (or categories) represented as $W = \{w_1, \ldots, w_v\}$, where *v* is the number of clusters. Then, the data is summarized into a VxN co-occurrence table of counts $N_{ij} = n(w_i, d_j)$, where $n(w_i, d_j)$ denotes how often the word w_i occurred in an image d_j.

In Addition, the BoVW representation ignores spatial information of detected objects in an image and may result in inferior classification performance [6]. To integrate the spatial information, we include the images bounding box information to their corresponding BoVW.

* http://www.csie.ntu.edu.tw/~cjlin/libsvm/

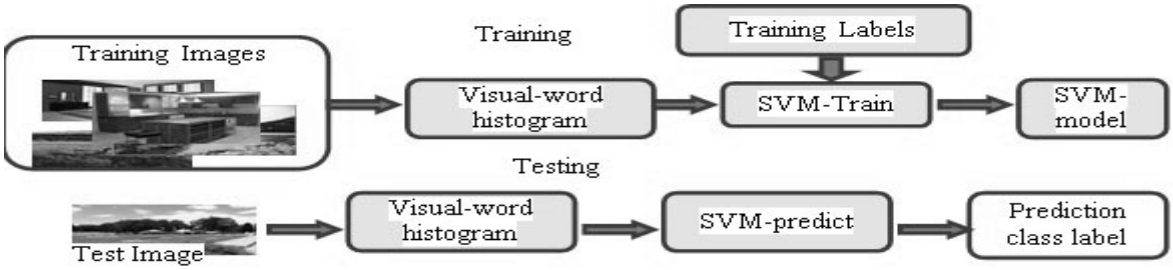


Figure 4: Learning Based Object Detection Processes

3.3 Automatic Image Annotation

The detail method for AIA is briefed as follows. As we briefly discussed earlier the training set T contains n images X = (x_1, …, x_n} of C classes, {1, …, m}, where m is the number of classes, each instance image in training class T_C represented by: visual-word histogram x_i, its spatial information S_i, and their individual class label C_i. The training dataset T formalized as:

$$T = \{(C_1, X_1, S_1), \ldots, (C_n, X_n, S_n)\}$$

This training data T, will be given to the classifier to recognize all classes and their attributes values, which are visual word histogram, class identifier which are class that image belongs to and spatial information. SVM model is generated to represent the training data that will be used to classify new images depending on their attribute value. The OAO classification method used in our work constructs C(C-1)/2 classifiers, one for each possible class pair.

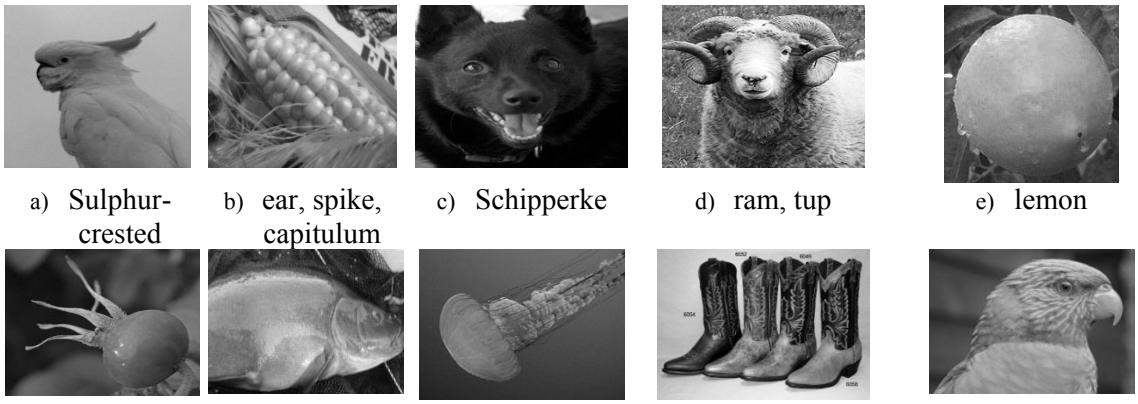
Once the model is generated, AIA is performed as follows, for an input image J. First, J will be represented by visual-word histogram. The classifier predicts the possible object category or class J belongs to. Thus, any annotation of the training images will be used as the annotation of J. More specifically, it classifies an input image J by using all the classifiers C(C-1)/2, where a vote is added for the winning class over each classifier. The method will propose the class with more votes as the result and its class label will be assigned to J.

4. Experimentation and Dataset

In order to validate the proposed approach, a prototype is developed using NetBeans IDE (version 6.9), Java SE Development Kit (JDK) 6.0 platform. Lucene (Version 3.4.0) is used to organize image's textual features. LIRe (Version 9.0, http://www.semanticmetadata.net/lire/) [21] is used to extract and wrap visual features [21]. SVM (version 3.17) [22] is used for machine learning purpose.

ImageNet [7, 23] is our primary images source having on average of 400 images along with bounding box that represent spatial location of salient objects per synset [http://www.image-net.org/].

Table 1: Selected Images Class Description



a) Sulphur-crested b) ear, spike, capitulum c) Schipperke d) ram, tup e) lemon

f) hip, rose hip, rosehip g) tench, Tinca tinca h) jellyfish i) cowboy boot j) lorikeet

Table 2: Number of Images for Training and Testing

Class Description	*Sulphur-crested*	*Schipperke*	*Ram, tup*	*Lemon*	*Hip rose*	*Jellyfish*	*Cowboy boot*	*Lorikeet*	*Ear, spike*	*Tench, Tinca tinca*	*Total*
Training	400	400	400	350	420	400	430	400	400	400	*4,000*
Test	400	400	400	350	420	400	430	400	400	400	*1,000*
										Total	*5,000*

In order to test our approach, all experiments are carried out on Windows 2007 Operating system having 2.17 GHz dual Intel Pentium CPU and 2.0 GB RAM. We have selected 10 image classes (i.e., wordnet synset – collection of synonymous words) as shown in Table 1 and from each class an average of 400 and 100 images are used for training and testing respectively as shown in Table 2.

4.1 Evaluation of Experimental Results

A total of 253,195 keypoints are extracted from the training dataset. For each concept to be predicted, we have subjectively selected 60 sample images having the concept presented clearly from the training set. Parameter values are fixed experimentally after conducting a series of grid search using 5-fold cross-validation in SVM [20]. The RBF kernel with gamma parameter set to 0.125 and the trade-off parameter C is set to 512. The performance of the prototype system is evaluated using statistical precision and recall methods extracted after building confusion matrix shown in Table 3.

Table 3: Confusion Matrix on Test Data

	Sulphur-crested	*Schipperke*	*Ram, tup*	*Lemon*	*Hip rose hip*	*Jellyfish*	*Cowboy boot*	*Lorikeet*	*Ear, spike, capitulum*	*Tench, Tinca tinca*
Sulphur-crested	77	3	4	5	0	3	0	5	0	3
Schipperke	0	78	5	2	1	0	2	4	3	5
Ram, tup	3	5	74	1	0	4	4	0	2	7
Lemon	4	2	3	76	3	2	2	1	1	6
Hip rose hip	2	0	1	4	75	5	0	5	3	5
Jellyfish	0	0	0	5	4	71	3	4	2	11
Cowboy boot	2	0	0	0	2	2	82	4	5	3
Lorikeet	3	3	1	2	4	5	1	77	4	0
Ear, spike	0	2	2	2	2	0	3	0	78	11
Tench,	3	0	4	3	3	2	4	2	0	79

Table 4: Classifier Precision and Recall Results Computed on Confusion Matrix

	Sulphur crested	*Schipperke*	*Ram, tup*	*Lemon*	*Hip rose*	*Jellyfish*	*Cowboy boot*	*Lorikeet*	*Ear, spike*	*Tench, Tinca tinca*	*Mean*
Precision	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.77
Recall	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.76

*Figure 5: A*nnotation for example image 1

Figure 6: Annotation for example image 2

Figure 7: Annotation for example image3

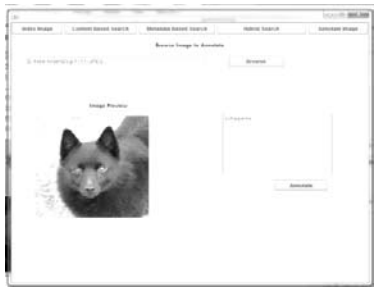


Figure 8: Annotation for example image 4

We have constructed a confusion matrix, shown in Table 4 for each class used in the test data. This is performed by providing separately each class of test data to the classifier and recording the prediction result of each instance from the trained model. Some of the running examples are shown in Figure 5 to Figure 8.

5. Conclusion and Future work

Bridging the semantic gap for image retrieval is not easy to overcome. In order to overcome the well-known problem, associating text to the image is a solution which is done in this work using automatic annotation approach. In this paper, AIA is proposed with the intention of extending the effectiveness of the current CBIR systems. In this system, we employed an approach for learning to detect objects in images using a bag of visual words model. A vocabulary of distinctive object patches is automatically constructed from a set of sample images. The images are then represented in visual histogram of these patches, together with spatial information. In the future, we would like to extend the work in the following directions. First, we will do more experiments on bag of visual words based image representation and clustering. Second, we would like to experiment the approach on more image classes and annotation tagging to individual objects in the image than to the whole image. Finally we would like to extend this work to annotate video.

References

[1] Silvan Andreas Saxer, “Region Based Image Similarity Search”, Swiss Federal Institute of Technology, Zurich, 2002.

[2] F. Estrada, A. Jepson, and D. Fleet, “Local Features Tutorial”, 2004, pp.12.

[3] Kraisak Kesorn, “Multi-Modal Multi-Semantic Image Retrieval”, School of Electronic Engineering and Computer Science, Queen Mary, University of London, 2010.

[4] Chin-Fong Tsai, “Bag-of-Words Representation in Image Annotation”, Hindawi Publishing Corporation, 2012.

[5] Thomas Deselaers and Henning Muller, "Combining Textual- and Content-based Image Retrieval”, 2008.

[6] Jun Yang, Yu-Gang Jiang, Alexander Hauptmann, and Chong-Wah Ngo, “Evaluating Bag-of-Visual-Words Representations in Scene Classification”, Carnegie Mellon University, 2007.

[7] http://www.metadataworkinggroup.org, last accessed on April 2013.

[8] Vogel J., "Semantic Scene Modeling and Retrieval", In Selected Readings in Vision and Graphics, Hartung-Gorre Verlag, Konstanz 2004.

[9] Christian H., Sebastian S., Andreas N. and Marcin D., "Automatic Image Annotation Using a Visual Dictionary Based on Reliable Image Segmentation", Otto-von-Guericke-University, Magdeburg, 2008.

[10] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "TagProp: Discriminative Metric Learning in Nearest Neighbor Models for Image Auto Annotation", In ICCV 2009.

[11] Frigui, H., and Caudill, J., "Unsupervised Image Segmentation and Annotation for Content-Based Image Retrieval", In Fuzzy Systems, 2006 IEEE.

[12] L. Zhang, and J. Ma, "Image Annotation by Incorporating Word Correlations into Multi-Class SVM," Soft Computing - A Fusion of Foundations, Methodologies and Applications, 2010.

[13] W. Yi and H. Tang, "Experimental Analysis on Classification of Unmanned Aerial Vehicles Images Using the Probabilistic Latent Semantic Analysis", International Symposium on Spatial Analysis, Modeling, and Data Mining, Wuhan, China, 2009.

[14] C. Yang, M. Dong, and F. Fotouhi, "Region Based Image Annotation through Multiple Instance Learning", ACM International Conference on Multimedia, New York, USA, 2005.

[15] Lucie Molková, "Indexing Very Large Text Data", Faculty of Informatics, Masaryk University, Czech Republic, 2011.

[16] Xin Li and Yuhong Guo, "Active Learning with Multi-label SVM Classification", Department of Computer and Information Sciences, Temple University Philadelphia, 2011.

[17] Julia EunJu Nam, Mauricio Maurer, and Klaus Mueller, "A high-Dimensional Feature Clustering Approach to Support Knowledge-Assisted Visualization", Stony Brook, NY 11794-4400, 2009, USA.

[18] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis", Journal of American Society for Information Science, Vol. 41, No. 6, pp. 391–407, 1990.

[19] Chih-Chung Chang and Chih-Jen Lin, "A Library for Support Vector Machines", Department of Computer Science, National Taiwan University, Taipei, Taiwan, 2013.

[20] National Library of New Zealand, http://natlib.govt.nz/librarians/digital-library-tools, last accessed on June 2013.

[21] Mathias Lux, http://www.semanticmetadata.net/lire/, last accessed on June 2013.

[22] Department of Computer Science, National Taiwan University, http://www.csie.ntu.edu.tw/~cjlin/libsvm/, last accessed on June 2013.

[23] Stanford Vision Lab, http://image-net.org/ last accessed on June 2013.

[24] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," Machine Learning, Vol. 42, No. 1-2, pp. 177–196, 2001.

[25] Itheri Yahiaoui, Bernard Merialdo, and Benoit Huet, "Image Similarity for Automatic Video Summarization", Institute Eurecom, Department of Communication Multimedia, Sophia – Antipolis- France, 2002.

[26] I. K. Sethi and I. L. Coman, "Mining Association Rules Between Low-Level Image Features and High-Level Concepts", Science Academy Publisher, United Kingdom, 2006.