International Journal of Control and Automation
Vol.7, No.6 (2014), pp.213-222
http://dx.doi.org/10.14257/ijca.2014.7.6.21

Semi-supervised Learning for Automatic Image Annotation Based on Bayesian Framework

Dongping Tian[1,2]

[1] *Institute of Computer Software, Baoji University of Arts and Sciences, Baoji, Shaanxi, 721007, China*
[2] *Institute of Computational Information Science, Baoji University of Arts and Sciences, Baoji, Shaanxi, 721007, China*

{tdp211}@163.com

Abstract

In this paper, we present a new method for automatic image annotation by applying semi-supervised learning based on the Bayesian framework. On the one hand, we employ the semi-supervised learning, i.e., transductive support vector machine (TSVM) to enhance the quality of training image data, which is a promising way to find out the underlying relevant data from the unlabeled ones. On the other hand, a simple yet very efficient Bayesian model is built to implement image annotation by the maximum a posteriori (MAP) criterion. The novelty of our method mainly lies in two aspects: exploiting TSVM to improve the quality of training image dataset and utilizing the Bayesian model to predict the candidate annotations for the unseen images. Experimental results on the standard Corel dataset demonstrate that the proposed method is superior or highly competitive to several state-of-the-art approaches.

Keywords: *Automatic image annotation, TSVM, Bayesian model, Gaussian distribution, Image retrieval*

1. Introduction

Automatic image annotation (AIA) has been an active research topic in recent years due to its potential impact on both image understanding and semantic based image retrieval. AIA refers to the process of learning statistical models from a training set of pre-annotated images in order to generate annotations for unseen images using visual feature extracting technology. In recent years, many methods have been proposed for image annotation by viewing image annotation as a supervised classification problem [1], which treats each semantic concept as an independent class and constructs different classifiers for different concepts. As the representative work, Li *et al.* [2] present a statistical modeling approach based on the two dimensional multi-resolution hidden Markov model to index images linguistically. Cusano *et al.* [3] make use of support vector machine (SVM) to do image segmentation and annotation simultaneously. Goh *et al.* [4] employ one-class and two-class SVMs by constructing a 3-level multiple sets of SVMs for multiclass image annotation. To be specific, level 1 consists of several sets of classifiers trained by different subset of training samples, and its probabilistic outputs are normalized by each set of classifiers before using them in level 2. Level 2 is the fusion process used to find a confidence factor in addition to the highest probabilistic decision. At level 3, the confidence factors of same concept are added together, and the concept with the maximum cumulative confidence is the final decision corresponding to the final annotation. Besides, Rui *et al.* [5] present a novel semi-naive Bayesian approach

ISSN: 2005-4297 IJCA

by incorporating clustering with pairwise constraints for automatic image annotation. In addition, Yang and Dong [6] implement image annotation based on a Bayesian framework, in which the complement components analysis is constructed to estimate the class conditional probabilities in the feature subspace. Followed by they propose asymmetrical support vector machine-based multiple-instance learning (MIL) algorithm to extend the conventional support vector machi- ne through MIL so as to train the SVM in a new space [7]. Li *et al.* [8] formulate the image annotation problem as a joint classification task based on two dimensional conditional random fields (CRF) together with semi-supervised learning, in which the 2D CRF is used to effectively capture the spatial dependency between the neighboring labels while the semi- supervised learning technique is employed to exploit the unlabeled data to improve the joint classification performance. Recently, Qi *et al.* [9] apply a similar framework to [4] for image annotation, but they fuse the decisions classifier by classifier instead of set by set. Meanwhile, they use both global and local features in two different sets of SVMs, which can effectively compensate the limitations of one type of feature by the other. Gao *et al.* [10] present CG- SVM for semantic image annotation by utilizing the clustering result to select the most informative image samples to be labeled and optimizing the penalty coefficient. More recently, Zhao and Zhu [11] construct TSVM-HMM for automatic image annotation, which integrates the discriminative classification with the generative model to mutually complete their advantages for better annotation performance. Subsequently, Feng *et al.* [12] propose an improved transductive multi-instance multi-label (TMIML) learning algorithm with application to automatic image annotation, which aims at taking full advantage of both labeled and unlabeled data to address some annotation issues. In addition, Jiang *et al.* [13] adopt learning vector quantization (LVQ) technique to optimize low-level features extracted from given images, and then select some representative vectors with LVQ to train support vector machine classifier instead of using all feature data.

As briefly reviewed above, most of these approaches can achieve state-of-the-art performance and motivate us to explore better image annotation methods with the help of their excellent experiences and knowledge. So in this paper, we present a new method for automatic image annotation by using semi-supervised learning based on the Bayesian framework. On the one hand, the semi-supervised learning is applied to enhance the quality of training image data, which can meet the requirement of a large number of labeled images to guarantee the annotation model's feasibility. On the other hand, a simple yet very efficient Bayesian model is constructed to implement image annotation by the maximum a posteriori (MAP) criterion. Finally, we evaluate our method on the standard Corel dataset and the experimental results are comparative to several state-of-the-art approaches.

The rest of the paper is organized as follows. Section 2 describes semi-supervised learning, especially the transductive support vector machine. Section 3 elaborates the procedure for image annotation based on the Bayesian framework. Experimental results on the standard Corel dataset are reported and analyzed in Section 4. Finally, we end this paper with some important conclusions and future work in Section 5.

2. Semi-Supervised Learning

Semi-supervised learning, which is a family of algorithms that take advantage of both labeled and unlabeled data, has been extensively studied for a couple of years [14, 15]. Among which the transductive support vector machine (TSVM), also called semi-supervised support vector machine (S^3VM) located between supervised learning with fully-labeled training data and unsupervised learning without any labeled training data, is a promising way to find out the underlying relevant data from the unlabeled ones. TSVM works as follows: given a keyword *w*, several labeled regions are taken as the relevant examples and the initial non-relevant examples are randomly sampled from the remaining regions. A two-class SVM classifier is trained firstly. Then based on the learnt SVM classifier, the most confident relevant regions and the most non-relevant ones are added into the relevant and non-relevant training set respectively. With the expanded training set, SVM classifier will be re-trained until the maximum time of iteration is reached. Finally, an expanded set of labeled regions can be obtained to benefit for modeling the visual feature distribution of the keyword *w*. So in this paper, we adopt TSVM to explore more relevant image regions so as to enhance the quality of training data. The procedure of it for mining more relevant labeled data is shown as follows.

Algorithm 1 Procedure of the TSVM for Mining Relevant Regions

```
Input: $R_L^0$ and $R_U^0$ denote the sets of labeled and unlabeled regions
for the keyword $w$; $S$ is a SVM classifier; $m$, $n$ and $K$ denote
control parameters
Output: $R_L^k$ an expanded set of labeled regions
Process:
for $k$=1 to $K$ do
    Learning a SVM classifier $S$ from $R_L^k$
    Using $S$ to classify regions in $R_U^k$
    Selecting $m$ most confidently predicted regions from $R_U^k$ which
    are labeled as relevant examples
    Selecting $n$ most confidently predicted regions from $R_U^k$ which
    are labeled as non-relevant examples
    Adding $m$+$n$ regions with their corresponding labels into $R_L^k$
    Removing these $m$+$n$ regions from $R_U^k$
end for
```

3. Bayesian Framework for Automatic Image Annotation

For automatic image annotation, the Bayesian model works by finding the posterior probability that a concept belongs to an image. Given prior probabilities of annotation words and the corresponding class conditional probabilities for the image, then it is possible to assign annotation words to an image according to the posterior probability. Let *J* denote the testing set of images, and let *T* denote the training collection of annotated images. Each testing image $I \in J$ is represented by its regional visual feature $r = \{r_1, \ldots, r_m\}$, and each training image $I \in T$ is represented by both a regional visual feature $r = \{r_1, \ldots, r_m\}$ and an annotation word list $w = \{w_1, \ldots, w_n\}$, where r_j (j =1,2,…,m) is the set of visual features for region *j* and w_i(i =1,2,…,n) is the *i*th annotation word in the vocabulary *V*. If we treat each annotation word w_i as a distinct class label, then the annotation problem can be formulated as a supervised classification problem under Bayesian framework [6]. The posterior probability $P(w_i|I)$ can be calculated based on the class conditional probability $P(I|w_i)$ and prior probability $P(w_i)$ as follows:

$$P(w_i \mid I) \propto P(I \mid w_i) P(w_i) \tag{1}$$

Since a testing image is divided into many regions, the class conditional probability can be computed by the following formula,

$$P(I \mid w_i) = \sum_{j=1}^{m} P(r_j \mid w_i) \tag{2}$$

By substituting Eq. (1) into Eq. (2), we can get,

$$P(w_i \mid I) \propto \sum_{j=1}^{m} P(r_j \mid w_i) P(w_i) \tag{3}$$

Note that the posterior probability is served as a degree of confidence and can be used as the criterion to select the top *N* words as the annotation for image *I*.

$$\begin{aligned} \hat{w} &= \arg\max\{P(w_i \mid I)\} \\ &= \arg\max\{\sum_{j=1}^{m} P(r_j \mid w_i) P(w_i)\} \end{aligned} \tag{4}$$

Here the prior probability $P(w_i)$ needs to be calculated from the training image set *T*, which is usually estimated as follows:

$$P(w_i) = \frac{|T_i|}{|T|} \tag{5}$$

where $|T_i|$ denotes the number of training images containing the annotation word w_i. $|T|$ is the size of training image data.

As for the class conditional probability $P(r|w_i)$, similar to [6], we assume that the low-level features in set R_i follow a Gaussian distribution (R_i denotes the set of regions extracted from the training images annotated by the word w_i), thus it can be defined as below:

$$P(r \mid w_i) = \frac{1}{\sqrt{2^k \pi^k |\Sigma|}} e^{-(r-\bar{r})^T \Sigma^{-1} (r-\bar{r})} \tag{6}$$

where $\bar{r}$ is the mean of the low-level features in R_i, Σ denotes the covariance matrix and *k* is the dimension of the feature vector in each region. To simplify our calculation, we assume each feature component is independent for the given keyword and Σ is reduced to a diagonal matrix. The framework of the proposed model is illustrated in Figure 1.

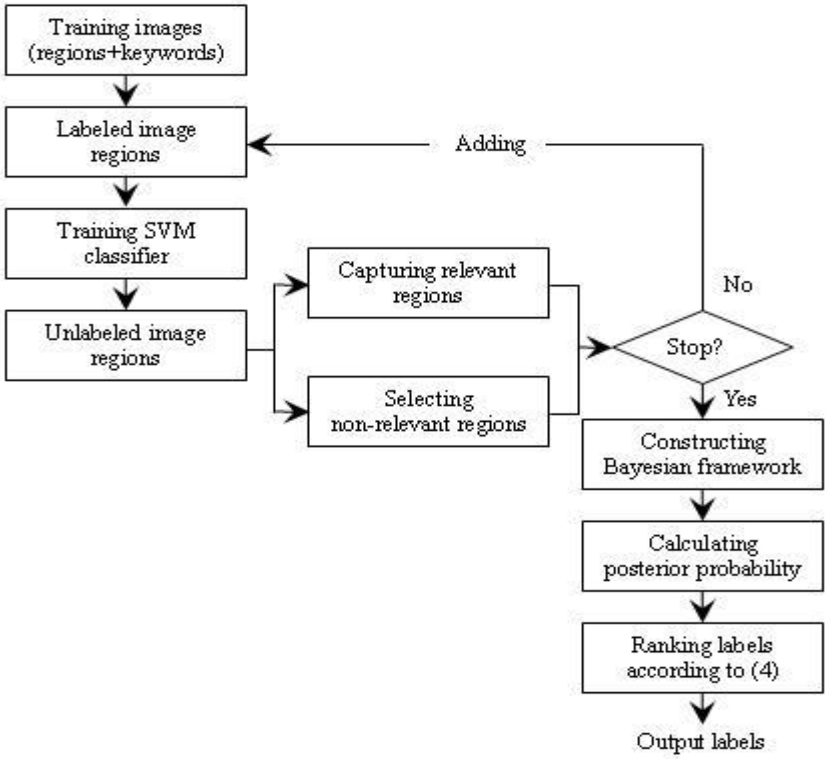
Training images (regions+keywords)
Labeled image regions
Adding
Training SVM classifier
Capturing relevant regions
No
Unlabeled image regions
Stop?
Selecting non-relevant regions
Yes
Constructing Bayesian framework
Calculating posterior probability
Ranking labels according to (4)
Output labels

Figure 1. The Proposed Image Annotation Scheme in this Paper

4. Experimental Results and Analysis

To make a fair comparison with the state-of-the-art approaches, we employ Corel5k as our experimental dataset, which consists of 5000 images from 50 Corel Stock Photo CD's provided by [16]. Each CD contains 100 images with a certain theme, of which 90 are designated to be in the training set and 10 in the test set, resulting in 4500 training images and a balanced 500-image test collection. In addition, the normalized cuts (Ncuts) algorithm [17] rather than JSEG [18] is utilized to segment images into 1 to 10 regions. The reason lies in that JSEG is only focusing on local features and their consistencies while Ncuts aims at extracting the global impression of an image data. So Ncuts, to some extent, can get a better segmentation result than JSEG (As can be seen from Figure 2). For each image, only the regions larger than a threshold value are selected with the following visual features [19], which include 81 dimen- sional grid color moment features, 59 dimensional local binary pattern (LBP) texture features, 120 dimensional Gabor wavelets texture features, 37 dimensional edge orientation histogram features and 512 dimensional GIST features. As a result, each region is represented by a 809 dimensional feature vector. Subsequently, a Bayesian model is learned to implement image annotation by the maximum a posteriori criterion.

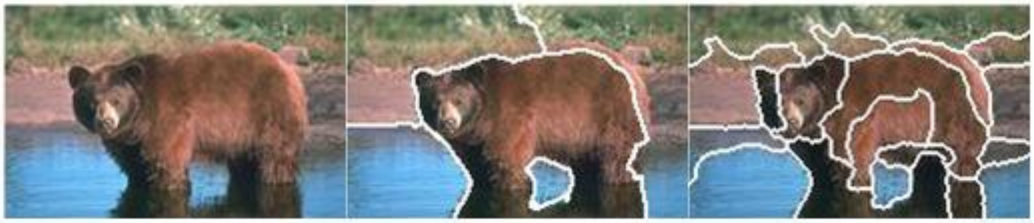

Figure 2. The Segmentation Results Using NCuts (mid) and JSEG (right)

To demonstrate the effectiveness of our model proposed in this paper, we compare it with several previous approaches [16, 20, 21, 22, 23]. Without loss of generality, we compute the re- call and precision of every word in the test set and use the mean of these values to summarize the model performance. recall=*B*/*C* and precision=*B*/*A*, where *A* is the number of images automatically annotated with a given keyword in the top 5 returned word list, *B* is the number of images correctly annotated with that keyword in the top 5 returned word list, and *C* is the number of images having that keyword in the ground truth annotation. Generally, the number of keywords selected to annotate the test image is determined by the number of regions. In our case, the keywords of five-top ranked regions with the largest area are determined to annotate the test image. The experimental results listed in Table 1 are based on two sets of words: the subset of 49 best words and the complete set of all 260 words that occur in the training set. From Table 1, it is easy to see that our model outperforms all the others, especially the first three approaches. Meanwhile, it is also superior to CRM and PLSA-WORDS by the gains of 6 and 8 words with non-zero recall, 4% and 3% mean per-word recall together with 14% and 20% mean per-word precision on the sets of 49 best words respectively. Correspondingly, the improvements of 21% and 15% mean per-word recall as well as 13% and 29% mean per-word precision on the sets of 260 words can be achieved.

Table 1. Performance Comparison of AIA on Corel5k

Models	Co-occurrence	Translation	CMRM	CRM	PLSA-WORDS	Our Model
#words with recall>0	19	49	66	107	105	113
Results on 49 best words						
Mean per-word recall	-	0.34	0.48	0.70	0.71	0.73
Mean per-word Precision	-	0.20	0.40	0.59	0.56	0.67
Results on all 260 words						
Mean per-word recall	0.02	0.04	0.09	0.19	0.20	0.23
Mean per-word Precision	0.03	0.06	0.10	0.16	0.14	0.18

Table 2 shows some annotating results (only four cases are listed here due to the limited space) generated by PLSA-WORDS and our model respectively. As can be seen from Table

2, the performance of our model is superior or highly competitive to that of PLSA-WORDS (Note that the enriched and re-ranked annotations compared to those of the ground truth and PLSA-WORDS are underlined and italic respectively), which further demonstrates the effectiveness of our model proposed in this paper.

Table 2. Annotation Comparison Between PLSA-WORDS and Our Model

Images				
Ground Truth Annotation	tiger, forest, cat, trees	garden, flowers, landscape, trees	mountain, water, sky, clouds	polar, bear, snow, tundra
PLSA-WORDS Annotation	tiger, trees, leaves, forest, cat	flowers, trees, garden, plants, farm	sky, mountain, water, clouds, trees	snow, bear, polar, tundra, ice
Our Model	tiger, trees, leaves, forest, cat	flowers, *garden*, *trees*, plants, farm	*mountain*, *sky*, water, clouds, trees	snow, bear, *tundra*, *polar*, ice

In addition, to validate the retrieval performance of our model proposed in this paper, mean average precision (*m*AP) is also employed as a metric to evaluate the performance of single word retrieval, which has been a standard measure for the retrieval of text document for years and it has the ability to summarize the retrieval performance in a meaningful way. Here, we only compare our model with CMRM, CRM and PLSA-WORDS due to *m*AP of other methods cannot be accessed directly. As shown in Table 3, our model is obviously superior to CMRM. Compared with PLSA-WORDS, it can get 9% and 27% improvements on 260 words and words with positive recall respectively. In addition, even though our model can get the same *m*AP as that of CRM model, we can obtain 22% improvement of the mean average precision value on the words with positive recall.

Table 3. Ranked Image Retrieval Results based on one Word Queries

Mean Average Precision for Corel Dataset		
Models	All 260 words	Words with recall >0
CMRM	0.17	0.20
CRM	0.24	0.27
PLSA-WORDS	0.22	0.26
Our Model	0.24	0.33

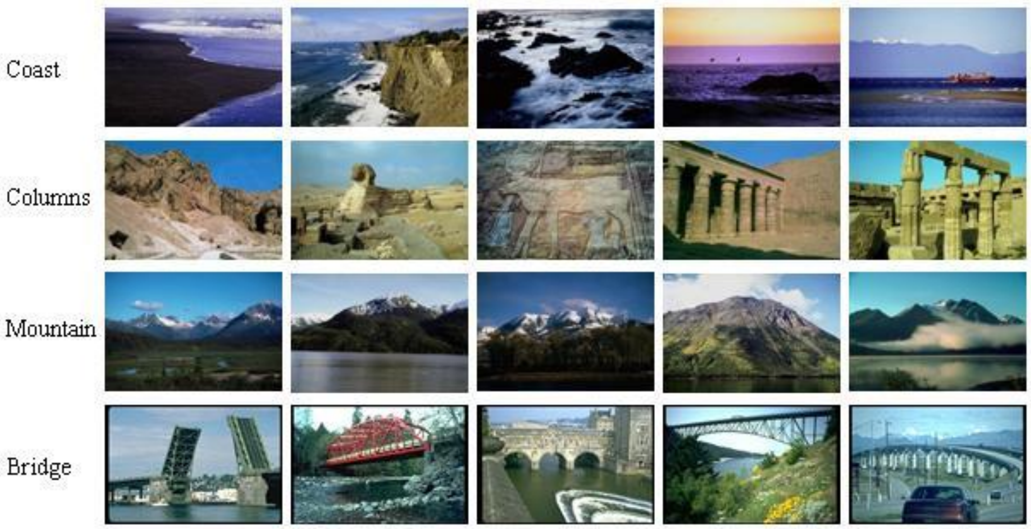

Figure 3. Semantic Retrieval Results on Corel5k Data Set

To further illustrate the effect of our model proposed in this paper, Figure 3 presents the retrieval results obtained with single word queries on several challenging visual concepts being queries. Each row displays the top five matches to the semantic query “coast”, “columns”, “mountain” and “bridge” from top to bottom, respectively. The diversity of visual appearance of the returned images bears out that our model also has good generalization ability.

5. Conclusions

In this paper, we present a novel method for automatic image annotation by using semi-supervised learning based on the Bayesian framework. We first employ the semi-supervised learning, *i.e.*, transductive support vector machine to enhance the quality of training image data, which utilizes the labeled and unlabeled data simultaneously and alleviates the harsh requirements of a large number of labeled training images during the image annotation model construction. And then the simple yet efficient Bayesian model is constructed to implement image annotation by the maximum a posteriori (MAP) criterion. Experimental results on the Corel5k dataset demonstrate the effectiveness of our model. In the future, we plan to employ other more complicated real-world image datasets, such as NUS-wide and Mirflickr to further evaluate the scalability and robustness of our model comprehensively.

Acknowledgements

This work is supported by the National Natural Science Foundation of China(No.61035003, No.61072085, No.60933004, No.60903141) and the National Program on Key Basic Resear -ch Project (973 Program) (No.2013CB329502).

References

[1] G. Carneiro, A. Chan, P. Moreno and N. Vasconcelos, “Supervised learning of semantic classes for image annotation and retrieval”, IEEE Transactions on Pattern Analysis and Machine Intellig- ence, vol. 29, no. 3, **(2007)**, pp. 394-410.

[2] J. Li and J. Wang, “Automatic linguistic indexing of pictures by a statistical modeling approach”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 9, **(2003)**, pp. 1075- 1088.

[3] C. Cusano, G. Ciocca and R. Schettini, “Image annotation using SVM”, In: Proceedings of the International Society for Optical Engineering, California, USA, **(2003)**, pp. 330-338.

[4] K. Goh, E. Chang and B. Li, “Using one-class and two-class SVMs for multiclass image anno- tation”, IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 10, **(2005)**, pp. 1333- 1346.

[5] S. Rui, W. Jin and T. Chua, “A novel approach to auto image annotation based on pairwise cons- trained clustering and semi-naive Bayesian model”, In: Proceedings of the 11th International Con- ference on Multimedia Modeling (MMM’05), Melbourne, Australia, **(2005)**, pp. 322-327.

[6] C. Yang, M. Dong and F. Fotouhi, “Image content annotation using Bayesian framework and complement components analysis”, In: Proceedings of the 12th International Conference on Image Processing (ICIP’05), Genoa, Italy, **(2005)**, pp. 1193-1196.

[7] C. Yang, M. Dong and J. Hua, “Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning”, In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR’06), New York, USA, **(2006)**, pp. 2057-2063.

[8] W. Li and M. Sun, “Semi-supervised learning for image annotation based on conditional random fields”, In: Proceedings of the 5th International Conference on Image and Video Retrieval (CIVR’ 06), Arizona, USA, **(2006)**, pp. 463-472.

[9] X. Qi and Y. Han, “Incorporating multiple SVMs for automatic image annotation”, Pattern Recog- nition, vol. 40, no. 2, **(2007)**, pp. 728-741.

[10] K. Gao, S. Lin, Y. Zhang and S. Tang, “Clustering guided SVM for semantic image retrieval”, In: Proceedings of the 2nd International Conference on Pervasive Computing and Applications (ICP- CA’07), Birmingham, United Kingdom, **(2007)**, pp. 199-203.

[11] Y. Zhao, Y. Zhao and Z. Zhu, “TSVM-HMM: transductive SVM based hidden Markov model for automatic image annotation”, Expert Systems with Applications, vol. 36, no. 6, **(2009)**, pp. 9813- 9818.

[12] S. Feng and D. Xu, “Transductive multi-instance multi-label learning algorithm with application to automatic image annotation”, Expert Systems with Applications, vol. 37, no. 1, **(2010)**, pp. 661- 670.

[13] Z. Jiang, J. He and P. Guo, “Feature data optimization with LVQ technique in semantic image annotation”, In: Proceedings of the 10th International Conference on Intelligent Systems Design and Applications (ISDA’10), Cairo, Egypt, **(2010)**, pp. 906-911.

[14] O. Chapelle, B. Schölkopf and A. Zien, "Semi-supervised learning", MIT Press, **(2006)**.

[15] X. Zhu, “Semi-supervised learning literature survey”, Computer Sciences TR 1530, University of Wisconsin-Madison, **(2008)**.

[16] P. Duygulu, K. Barnard, N. de Freitas and D. Forsyth, “Object recognition as machine translation: learning a lexicon for a fixed image vocabulary”, Proceedings of the 7th European Conference on Computer Vision (ECCV’02), Copenhagen, Denmark, **(2002)**, pp. 97-112.

[17] J. Shi and J. Malik, “Normalized cuts and image segmentation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, **(2000)**, pp. 888-905.

[18] Y. Deng and B. Manjunath, “Unsupervised segmentation of color-texture regions in images and video”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 8, **(2001)**, pp. 800-810.

[19] J. Zhu, S. Hoi, M. Lyu and S. Yan, “Near-duplicate keyframe retrieval by nonrigid image mat- ching”, In: Proceedings of the 16th ACM International Conference on Multimedia (MM’08), Van- couver, Canada, **(2008)**, pp. 41-50.

[20] Y. Mori, H. Takahashi and R. Oka, “Image-to-word transformation based on dividing and vector quantizing images with words”, In: Proceedings of the 1st International Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM’99), Florida, USA, **(1999)**, pp. 405-409.

[21] L. Jeon, V. Lavrenko and R. Manmantha, “Automatic image annotation and retrieval using cross- media relevance models”, In: Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’03), Toronto, Canada, **(2003)**, pp. 119 -126.

[22] R. Manmatha, V. Lavrenko and J. Jeon, “A model for learning the semantics of pictures”, In: Proceedings of the 17th International Conference on the Advances in Neural Information Process- ing Systems (NIPS’03), Vancouver, Canada, **(2003)**, pp. 553-560.

[23] F. Monay and D. Gatica-Perez, “Modeling semantic aspects for cross-media image indexing”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 10, **(2007)**, pp.1802- 1817.