A Novel JSVM Approach for Automatic Image Annotation and Retrieval

T. Sumadhi
Research Scholar,
Karpagam University,
Coimbatore-21.

M. Hemalatha,
Phd, Head & Professor,
Karpagam University,
Coimbatore-21

ABSTRACT

This paper presents a novel image annotation framework for domains with large numbers of images. Automatic image annotation is such a domain, by which a computer system automatically assigns metadata in the form of captioning or keywords to a digital image. This application of computer vision technique is used in image retrieval system to organize and locate images of interest from a database. Many techniques have been proposed for image annotation in the last decade that has given reasonable performance on standard datasets In this work, we propose a new model for image annotation known as JSVM which treats annotation as a retrieval problem. In this work, we introduce an JSVM model for image annotation that treats annotation as a retrieval problem. The proposed technique utilizes low level image features and a simple combination of basic distances using JEC to find the nearest neighbors of a given image; the keywords are then assigned using SVM approach which aims to explore the combination of three different methods. First, the initial annotation of the data using flat wise and axis wise methods, and that takes the hierarchy into consideration by classifying consecutively its instances through position wise method. Finally, we make use of pair wise majority voting between methods by simply summing strings in order to produce a final annotation. The result of the proposed technique shows that this technique outperforms the current state of art methods on the standard datasets.

Keywords

JEC, SVM, image annotation, image retrieval, Radial Basis Function

1. INTRODUCTION

As high resolution digital cameras become more affordable and widespread the high quality digital image becomes ever more available and useful. With the exponential growth on high quality digital images, there is an urgent need to support more effective image retrieval over large scale archives. However content based image retrieval(CBIR) is still in its infancy and most existing CBIR systems can only support feature based image retrieval. Unfortunately, the naive users may not be familiar with low level visual features and it is very hard for them to specify their query concepts by using low level visual features directly. Thus there is a great need to develop automatic image annotation framework, so that the naive users can specify their query concepts easily by using the relevant keywords. However the performance of image classifiers depends on two inter related issues: (1) suitable framework for image content representation and automatic feature extraction. (2) Effective algorithm for image classifier training and feature subset selection.

To address the first issue there are two widely accepted approaches for image content representation and feature extraction. To address the second issue for automatic image annotation two approaches are widely used to train the image classifiers. (a) Model based approach by using Gaussian mixture model to approximate the underlying distribution of image classes in the high dimensional feature space (b) SVM-based approach by using support vector machine(SVM) to directly learn the maximum margins between the positive images and the negative images. In this work, SVM based approach is used to enable more effective classifier training with small generalization error rate in high dimensional feature space. So, for the annotation process we relied on SVM with a Radial basis function (RBF) kernel due to its outgoing performance. In this paper, we have proposed a hierarchical framework by incorporating the feature hierarchy and boosting to scale up SVM image classifier training. This framework is done in Mat lab using the popular label me web based annotation implementation.

2. RELATED WORK

A large number of techniques have been proposed in the last decade. Most of these treat annotation as translation from image instances to keywords. The translation paradigm is typically based on some model of image and text co-occurrences. Latent Dirichlet Allocation (Corel LDA) considers association through a latent topic space in a generatively learned model [4, 18]. Mori et al. [4, 7] used a Co-occurrence Model in which they looked at the co-occurrence of words with image regions created using a regular grid. Monay and Gatica-Perez [4, 7] introduced latent variables to link image features with words as a way to capture co-occurrence information. The addition of a sounder probabilistic model to LSA resulted in the development of probabilistic latent semantic analysis (PLSA) [4, 6, 7]. Blei and Jordan [4, 18] viewed the problem of modeling annotated data as the problem of modeling data of different types where one type describes the other. Jeon et al. [4] improved on the result of Duygulu et al. by introducing a generative language model referred as Cross Media Relevance Model (CMRM) the same process used by Duygulu et al. was chosen to calculate the blob representation of images [12]. They assumed that this could be viewed as analogous to the cross-lingual retrieval problem to perform both image annotation and ranked retrieval. Lavrenko et al. [4, 14] argued that the process of quantization from continuous image features into discrete blobs, as the approach used by the machine translation model and the CMRM model, will cause the loss of useful information in image regions. While Feng et al. [4, 14] modified the above model such that the probability of observing labels given an image was modeled as a multiple-Bernoulli distribution. In addition, they simply divided images

into rectangular tiles instead of applying automatic segmentation algorithms. Their Multiple Bernoulli Relevance Model (MBRM) achieved further improvement on performance. Liu. et. al. [4, 18], they estimated the joint probability by the expectation over words in a pre-defined Lexicon. It involves two kinds of critical relations in image annotation. First is the word-to-image relation and the second is the word-to-word relation. Torralba and Oliva [4, 15] focused on modeling a global scene rather than image regions. This scene-oriented approach can be viewed as a generalization of the previous one where there is only one region or partition which coincides with the whole image. Yavlinsky et. al. [4, 16] followed an approach using global features together with robust non-parametric density estimation and the technique of kernel smoothing. Jin et.al [4, 10] proposes a new frame work for automated image annotation that estimated the probability for language model to be use for annotation an image.

3. DATA SET DESCRIPTION

In this method we have utilized flicker dataset which contains 550 images of which 90% has been considered as training dataset and 10% as testing dataset.

4. METHODOLOGY

Annotation of images in this work undergoes several stages: first we extract information from the images and form a feature vector; hence we train several SVM's to create a model from the data for annotation accordingly to the mentioned approaches, flat and axis-wise, and position wise approaches herein tested. Finally we use majority voting, by summing strings, for a pair wise fusion between all three methods. We treat image annotation as a process of transferring keywords from nearest neighbors. The neighborhood structure is constructed using simple low-level image features resulting in a rudimentary model. A general flowchart of our procedure can be found in Fig. 1.

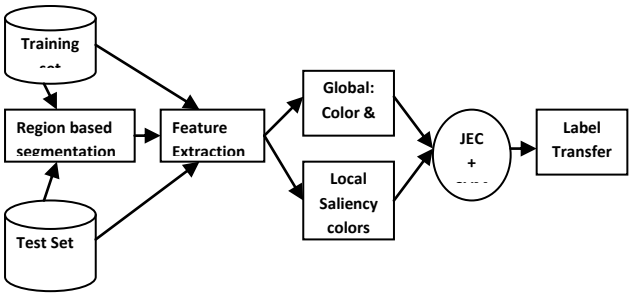

Fig:1 A Frame work of our proposed system

4.1 Feature Extraction

To extract information from the images we used both global and a local image descriptor in a JEC approach. Feature selection was made accordingly to the desired image properties that we aimed to discriminate: color, texture and shape. All global descriptors were extracted using the Local and Web Image Retrieval Engine.

A. Color

RGB is the default color space for image capturing and display, both HSV and LAB isolate important appearance characteristics not captured by RGB. The RGB, HSV, and LAB features are 16-bin-per-channel histograms in their respective color spaces. To determine the corresponding L1 distance measures, as it performed the best for RGB and HSV, while KL-divergence was found suitable for LAB distances.

B. Combining distances

Joint Equal Contribution (JEC). If labeled training data is unavailable, or the labels are extremely noisy, the simplest way to combine distances from different descriptors would be to allow each individual distance to contribute equally (after scaling the individual distances appropriately). First the keywords are selected from the nearest neighbor. If more keywords are needed, they are selected from neighbors 2 through N based on co-occurrence and frequency. Each feature contributes equally towards the image distance. Let I_i be the i^{th} image, and say we have extracted N features $f_i^1, f_i^2, \ldots\ldots, f_i^N$. Let us define $d_{(i,j)}^k$ as the distance between f_i^k and f_j^k. We would like to combine the individual distances $d_{(i,j)}^k$, k = 1. . . N to provide a comprehensive distance between image I_i and I_j. Since, in JEC, each feature contributes equally towards the image distance, we first need to find the appropriate scaling terms for each feature. These scaling terms can be determined easily if the features are normalized in some way (e.g., features that have unit norm), but in practice this is not always the case. We can obtain estimates of the scaling terms by examining the lower and upper bounds on the feature distances computed on some training set. We scale the distances for each feature such that they are bounded by 0 and 1. If we denote the scaled distance as $d_{(i,j)}^k$ we can define the comprehensive image distance between images I_i and I_j as $\sum_{k=1}^{N} \frac{d_{(i,j)}^k}{N}$. We refer to this distance as Joint Equal Contribution (JEC).

4.2. Annotation

For the annotation process we relied on SVM's with a Radial Basis Function (RBF) kernel due to their performance in the Image CLEF medical image annotation tasks. We have set up a framework in MATLAB using the popular label me web based implementation. We performed an extensive grid-search on the common approaches to this problem, flat and axis-wise strategies, to optimize the kernel parameters using 10-fold cross validation. Each image is classified one axis at the time but, unlike the axis-wise method, conceptualization of the image content does not take the full meaning of the axis into consideration. Instead, we first consider the highest hierarchical position of the axis, its root, and use the whole training set to perform an initial classification. Afterwards, we reduce the training set to those images which match the initial classification, a semantic reduction of the training set, and classify the hierarchically subsequent inferior position. We undergo this top-down process thorough the axis tree until it is completely classified. We undertake the same methodology for all axes and assemble the final annotation. After the annotation from the three methods separately we make pair wise fusions of these by summing strings. The chart given below shows the percentage of keywords being annotated in our flicker dataset.

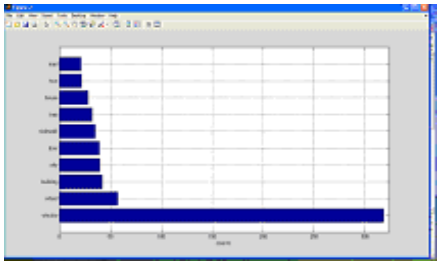

Fig:2 Chart showing the annotation statistics

5. Evaluation and Discussion

5.1 Evaluation of annotation.

To evaluate annotation, we query images from the test dataset using 20 frequent keywords from the vocabulary. The image will be retrieved if the automatically established annotation contains the query keyword. We evaluate the result using P% and R% denotes the mean precision and the mean recall, respectively, over all keywords in percentage points. N+ denotes the number of recalled keywords. Note that the proposed simple baseline technique (JEC) outperforms state-of-the-art techniques in all datasets [2]. The precision, recall and common E measure which are defined as

P= NUM $_{correct}$ / NUM $_{retrieved}$ **(1)**

R= NUM $_{correct}$ / NUM $_{exists}$ **(2)**

$$E(p,r) = 1 - 2/([1/p] + [1/r]) \quad (3)$$

5. 2 Query Results

Label me tool is used to query the flicker dataset and label me dataset from different perspectives. This method has been implemented in mat lab by incorporating Label me tool. The figures below shows the query results of our proposed method.

5. 2. 1 Query 1

Query 1is used to retrieve all the information for concept cars from the dataset.

>>**LMdbshowobjects (LMquery (D, 'object. name', 'car'), HOMEIMAGES)**;

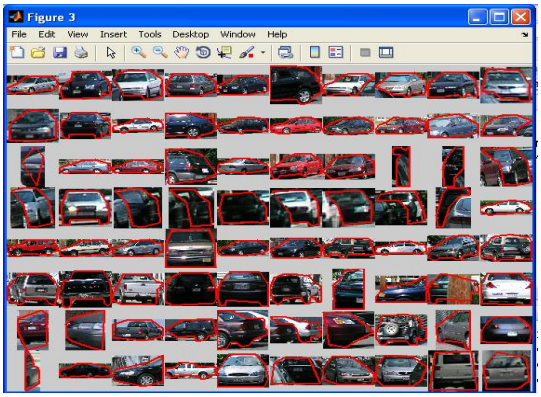

Fig:3 Query and Output for all type of cars

5. 2. 2 Query 2

>> drawXML(filename, HOMEIMAGES)

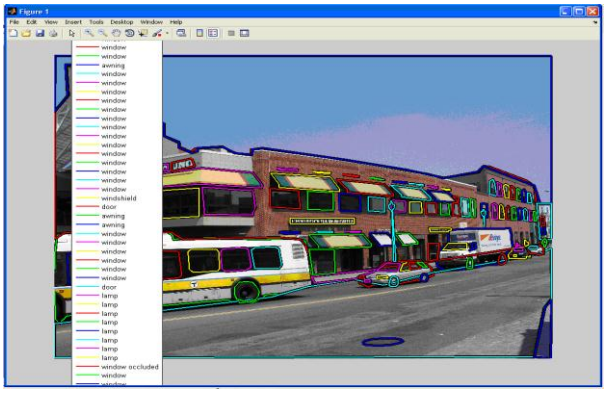

Fig:4 Query and annotation results for the sample input image

5.2.3 Query 3

Figure 54 shows output for query on "car side view, building, road and tree".

>>**LMdbshowobjects (LMquery (D, 'object. name', 'car + side, building, road, tree'), HOMEIMAGES);**

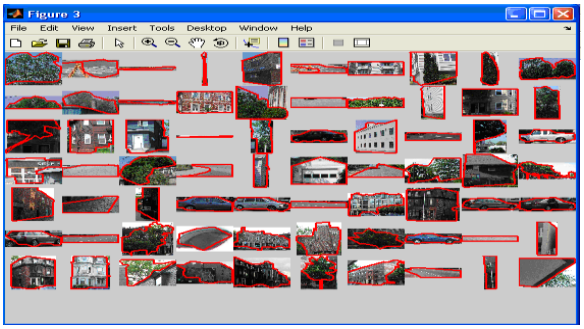

Fig:5 Query and output for side view of car, building, road and tree

5.2.4 Query 4

Figure 6 shows output for query on "flower".

>>**LMdbshowobjects(LMquery(DFLOWER,'object.name','flower'),HOMEIMAGES)**

2 matches out of 2

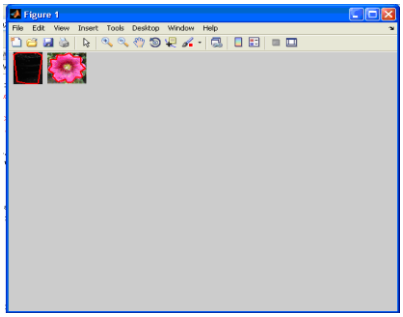

Fig:6 Query and output for flower.

5.3 Discussion

We have evaluated our method based on various features like RGB, HSV and LAB. The table 1 below shows their performance it can be stated that JEC when combined with RGB feature performs well. The comparison of JEC with various features has been illustrated in figure 7. The obtained results show that JEC when combined with RGB features works well than other features. Table 2 shows the results of comparison of our method with other feature extraction

methods like lasso, group lasso, least square, L2 regularization. From the results it is clear that JEC when combined with SVM gives better results than other feature extraction methods. Table 3 shows the results of comparison of our method with other two methods like new base line method which makes use of greedy approach for annotation and hierarchical model which makes use of bag of words for feature extraction. This comparative analysis of our method with other methods has been clearly illustrated using the line chart in figure 8. As per obtained results, our JSVM method has higher precision and recall rate compared with the other two methods.

TABLE.1PERFORMANCE OF JEC WITH VARIOUS FEATURES

Methods	**P%**	**R%**	**N+**
RGB	**18**	**22**	**110**
RGB16	12	14	94
HSV	17	19	80
HSV16	14	16	108
LAB	12	13	102

TABLE.2.COMPARISON OF OTHER FEATURE EXTRACTION METHODS WITH SVM

Methods	**P%**	**R %**	**N+**
JEC+SVM	**19**	**22**	**110**
Lasso+SVM	12	19	94
group lasso+SVM	10	18	87
Least Square+SVM	10	13	88
L2 -regularization+SVM	11	14	93

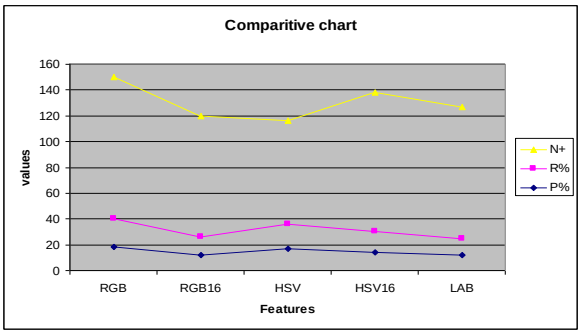

Fig.7. Chart showing the results of comparison with other features

TABLE.3. COMPARISON WITH OTHER METHODS

Methods	**Precision**	**Recall**	**E-measure**
Proposed JSVM Model	**0.77**	**0.35**	**0.513**
JEC + KNN Model	0.54	0.32	0.60
New Baseline Method	0.20	0.23	0.786
Hierarchical Model	0.34	0.29	0.636

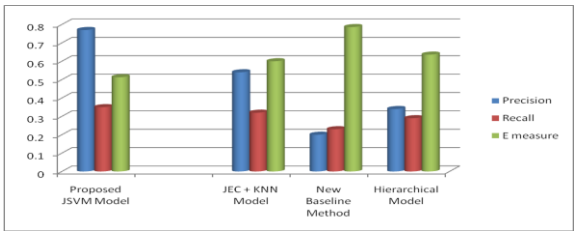

Fig.8. Chart showing the results of comparison with other methods

6. CONCLUSION

To be able to solve the image annotation problem at the human level, perhaps one needs to first solve the problem of scene understanding. The goal of our work was to develop a new annotation method JSVM by combining the JEC distance measure with that of the hierarchical method for image annotation. It could be concluded from the results that our system with JEC feature is efficient for this image annotation purpose. Experiments on these dataset reaffirm the enormous importance of considering multiple sources of evidence to bridge the gap between the pixel representations of images and the semantic meanings. The proposed JSVM algorithm is found to have 74% higher precision than baseline algorithm, 56% than Hierarchical algorithm, and the overall accuracy of JSVM model is 77% which is higher than all other methods.

The obtained result shows that JSVM model outperforms than all other existing algorithms.

7. REFERENCES

[1] Bowman Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar "A New Baseline for Image Annotation, 2009 international conference on electrical engineering and informatics 5-7 August 2009, IEEE 978-1-4244-4913-2/09.

[2] Igor F. Amaral, Filipe Coelho, Joaquim F. Pinto da Costa and Jaime S. Cardoso "Hierarchical Medical Image Annotation Using SVM-based Approaches" 2010 IEEE -978-1-4244-6561-3/10

[3] Sumathi T., Hemalatha M. "An Empirical Study on Performance Evaluation in Automatic Image Annotation and Retrieval" published in International journal of advanced research in computer science, Vol.1., No.4.,Nov-Dec-2010

[4] Sumathi T. et.al, " An Overview of Automated Image Annotation Approaches" International Journal of Research and Reviews in Information Sciences Vol. 1, No. 1, March 2011

[5] Nasullah Khalid Alham, Maozhen Li1, Suhel Hammoud and Hao Qi,' Evaluating Machine Learning Techniques for Automatic Image annotations'ieeeexplore.ieee.org/iel5/pp 53-58

[6] Syaifulnizam Abd Manal, MDJan Nordin 'Review on statistical approaches for automatic image annotation, International conference on electrical engineering and informatics, 978-1-4244-4913-2/2009 IEEE

[7] Nasullah Khalid Alham, Maozhen Li, Suhel,'evaluating Machine learning techniques for automatic image annotations'978-0-7695-3735-1/09 2009 IEEE

[8] YuliGao et.al, 'Automatic image annotation by incorporating feature hierarchy and Boosting to scale up SVM classifiers, ACM Multimedia, October 22-28,2006

[9] Herve Glotin, H., Zhao, Z.Q., Ayache, S.,'Efficient image concept indexing by harmonic and arithmetic profiles entropy', 2009 IEEE international conference on image processing, Nov 7-11, 2009.

[10] Yang, C., Dong, M., Hua, J.: Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. (2006)

[11] Carneiro, G., Chan, A.B., Moreno, P.J., Vasconcelos, and N.: Supervised learning of semantic classes for image annotation and retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence (2007)

[12] Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, and D.A.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: European Conference on Computer Vision. (2002) 97–112

[13] V. Lavrenko, R. Manmatha and J. Jeon, "A Model for Learning the Semantics of Pictures," in Proceedings of Advance in Neutral Information Processing, 2003.

[14] S. Feng, R. Manmatha and V. Laverenko, "Multiple Bernoulli Relevance Models for Image and Video Annotation," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004, p. 1002-1009.

[15] Torralba, A., Oliva, A.: "Statistics of natural image categories. Network: computation in neural systems. 14(3) (2003) 391-412.

[16] Yavlinsky, A., Schofield, E., Riiger, and S: "Automated image annoatation: ACM Press (2003)127-134.

[17] Pan J., Yang H., and Faloutsos C., Duygulu P.: GCap: Graph based Automatic Image Captioning, In Proceedings of the 4th International Workshop on Multimedia Data and Document Engineering (MDDE 04), in conjunction with Computer Vision Pattern Recognition Conference (CVPR 04), 2004.Washington DC, July 2nd 2004.

[18] Blei, D.M. Jordan, M.I.: Latent Dirichlet Allocation. Journal of machine learning research (2003)993-1022