International Journal of advancement in electronics and computer engineering (IJAECE)
Volume 4, Issue 1, April 2015, pp.405-408, ISSN 2278 -1412

AUTOMATIC IMAGE ANNOTATION USING L1-PENALIZED LOGISTIC REGRESSION

Arun Kumar[1], Nishchol Mishra [2], Sanjeev Sharma[3]
[1]***M.Tech Scholar,*** *School of IT, RGPV,Bhopal, E- Mail-arunkkt29@gmail.com, India;*
[2]***Faculty,*** *School of IT, RGPV,Bhopal, E- Mail- nishchol@rgtu.net.com, India;*
[3]***Faculty,*** *School of IT, RGPV,Bhopal, E- Mail- sanjeev@rgtu.net, India;*

Abstract *– The task of automatic image annotation is of great interest because it can play a crucial role in building an effective engine for image retrieval. Assigning descriptive keywords to images allows users to search for images using only text-based queries. Evaluating the performance of an image retrieval engine is different than that of an annotation engine because in retrieval we are only interested in the quality of the first few images associated with a given keyword. Following (Carneiro et al., 2007), we have reported the average retrieval precision over all keywords, as well as just the recalled keywords, for the first 10 retrieved images. It is widely acknowledged that image annotation is an open and very difficult problem in computer vision. Solving this problem at the human level may, perhaps, require that the problem of scene understanding be solved first. However, identifying objects, events, and activities in a scene is still a topic of intense research with limited success.*
Keywords*: Image Annotation, JEC, PENALIZED LOGISTIC REGRESSION, HAAR*

I. Introduction

AUTOMATIC image annotation (AIA) has been studied extensively for a several years. AIA is defined as the process by which a computer system automatically assigns metadata in the form of text description or keywords to a digital image. This process is used in image retrieval systems to organize and locate images of interest from a database. This task can be regarded as a type of multi-class image classification with a number of classes equal with vocabulary's size. AIA can be seen also as a multi-class object recognition problem which is a challenging task and an open problem in computer vision. The importance of this task has increased with the growth of the digital images collections.

This image search is based on text retrieval because the content of the image is ignored. For this reason sometimes the search performed does not lead to satisfactory results. In order to avoid this drawback the researchers are looking for another way to search for images. A possible approach is to obtain a textual description from the image and

then use text retrieval for searching. A different approach is to combine two modalities for example text and visual features when indexing images. Image retrieval based on text is sometimes called Annotation Based Image Retrieval (ABIR) [Inoue (2004)]. The systems based on ABIR can have some draw-backs. Researchers working in CBIR have identified two limitations. The first limitation is that ABIR requires manual image annotation which is time consuming and costly. The second limitation is that human annotation is subjective and sometimes it is difficult to describe image contents by concepts. An AIA system can solve the first limitation. The second limitation remains a general question and a unsolved problem for computer vision. AIA is situated on the frontier of different fields: image analysis, machine learning, media understanding and information retrieval. Usually image analysis is based on feature vectors and the training of annotation concepts is based on machine learning techniques. Automatic annotation of new images is possible only after the learning phase is completed. General object recognition and scene understanding techniques are used to extract the semantics from data. This is an extremely hard task because AIA systems have to detect at least a few hundred objects at the same time from a large image database.

Object recognition and image annotation are very challenging tasks. For this reason a number of models using a discrete image vocabulary have been proposed for the image annotation task. One approach to automatically annotating images is to look at the probability of associating concepts with image regions.

[Mori Y., et al. (1999)] used a Co-occurrence Model in which they looked at the co-occurrence of concepts with image regions created using a regular grid. To estimate the correct probability this model required large numbers of training samples. Each image is converted into a set of rectangular image regions by a regular grid.

The keywords of each training image are propagated to each image region. The major drawback of the above Co-occurrence Model is that it assumes that if some keywords are annotated to an image, they are propagated to each region in this image with equal probabilities. [Duygulu P., et al. (2002)] described images using a vocabulary of blobs. Image regions were obtained using the Normalized-cuts segmentation algorithm. For each image region 33 features such as color, texture, position and shape information were computed. The regions were clustered using the K-means clustering algorithm into 500 clusters called "blobs". The vector quantized image regions are treated as "visual words" and the relationship between these and the textual keywords can be thought as that between two languages, such as French and German. The training set is analogous to a set of aligned bitexts – texts in two languages. Given a test image, the annotation process is similar to translating the visual words to textual keywords using a lexicon learned from the aligned bitexts. This annotation model called Translation Model was a substantial improvement of the Co-occurrence model. [Jeon J., et al. (2003)] viewed the annotation process as analogous to the cross-lingual retrieval problem and used a Cross Media Relevance Model to perform both image annotation and ranked retrieval. The experimental results have shown that the performance of this model on the same dataset was considerably better than the models proposed by [Mori Y., et al. (1999)] and [Duygulu P., et al. (2002)]. The essential idea is that of finding the training images which are similar to the test image and propagate their annotations to the test image. CMRM does not assume any form of joint probability distribution on the visual features and textual features so that it does not have a training stage to estimate model parameters. For this reason, CMRM is much more efficient in implementation than the above mentioned parametric models. There are other models like Correlation LDA proposed by [Blei and Jordan (2003)] that extends the Latent Dirichlet Allocation model to words and images. This model is estimated using Expectation-Maximization algorithm and assumes that a Dirichlet distribution can be used to generate a mixture of latent factors.

II. PROPOSED SYSTEM

Focusing on visual query forms, many content-based image retrieval (CBIR) methods and techniques have been proposed in recent years, but they have several drawbacks. On the one hand, for methods based on query by example, a query image is often absent. On the other hand, query by sketch approaches are too complex for common users and a visual content interpretation of a user image concept is difficult. Therefore, image search using keywords is presently the most widely used approach. Content based indexing of images is more difficult than for textual documents because they do not contain units like words. Image search is based on using annotations and semantic tags that are associated with images. However, annotations are entered by users and their manual creation for a large quantity of images is very time-consuming with often subjective results. Therefore, for than a decade, automatic image annotation has been a most challenging task.

A. Method

Automatic image annotation methods require a quality training image dataset, from which annotations for target images are obtained. At present, the main problem with these methods is their low effectiveness and scalability if a large-scale training dataset is used. Current methods use only global image features for search.

1) We proposed a method to obtain annotations for target images, which is based on a novel combination of local and global features during search stage. We are able to ensure the robustness and generalization needed by complex queries and significantly eliminate irrelevant results. In our method, in analogy with text documents, the global features represent words extracted from paragraphs of a document with the highest frequency of occurrence and the local features represent key words extracted from the entire document. We are able to identify objects directly in target images and for each obtained annotation we estimate the probability of its relevance.

2) During search, we retrieve similar images containing the correct keywords for a given target image. For example, we prioritize images where extracted objects of interest from the target images are dominant as it is more likely that words associated with the images describe the objects.

3) We place great emphasis on performance and have thus tailored our method to use large-scale image training datasets. To cope with the huge number of extracted features, we have designed disk-based sensitive hashing for indexing and clustering descriptors. As show in Figure.1

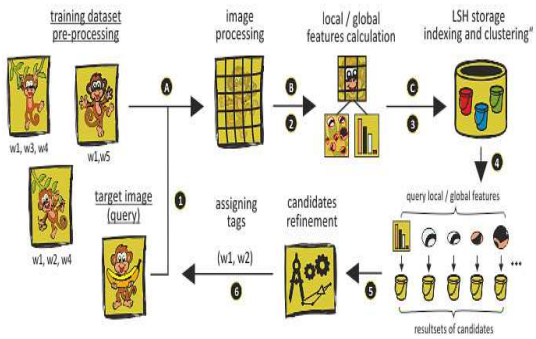

Figure 1: Scheme of my method for automatic image annotation.

III. SYSTEM DESIGN

B. The segmentation algorithm

For image segmentation we have used an our original and efficient segmentation algorithm [Burdescu D., et al. (2009)] based on color and geometric features of an image. The efficiency of this algorithm concerns two main aspects:

a) Minimizing the running time – a hexagonal structure based on the image pixels is constructed and used in color and syntactic based segmentation

b) Using a method for segmentation of color images based on spanning trees and both color and syntactic features of regions.

A similar approach is used in [Felzenszwalb and Huttenlocher (2004)] where image segmentation is produced by creating a forest of minimum spanning trees of the connected components of the associated weighted graph of the image.

Fig.2. is presented the hexagonal structure used by the segmentation algorithm:

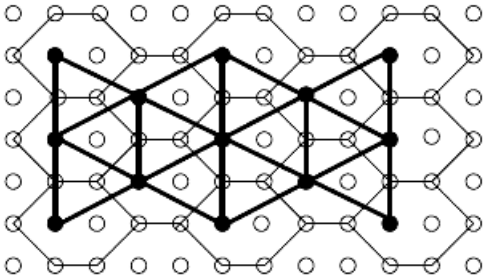

Figure 2: The grid-graph constructed on the hexagonal structure of an image

A particularity of this approach is the basic usage of the hexagonal structure instead of color pixels. In this way the hexagonal structure can be represented as a grid-graph G = (V, E) where each hexagon h in the structure has a corresponding vertex v□V , as presented in Fig.1. Each hexagon has six neighbors and each neighborhood connection is represented by an edge in the set E of the graph. To each hexagon two important attributes are associated: the dominant color and the coordinates of the gravity center. For determining these attributes were used eight pixels: the six pixels of the hexagon frontier, and two interior pixels of the hexagon.

Image segmentation is realized in two distinct steps:

(1) A pre-segmentation step – only color information is used to determine an initial segmentation. A color based region model is used to obtain a forest of maximum spanning trees based on a modified form of the Kruskal's algorithm. For each region of the input image it is obtained a maximal spanning tree. The evidence for a boundary between two adjacent regions is based on the difference between the internal contrast and the external contrast between the regions.

(2) A syntactic-based segmentation – color and geometric properties of regions are used. It is used a new graph which has a vertex for each connected component determined by the color-based segmentation algorithm. The region model contains in addition some geometric properties of regions such as the area of the region and the region boundary. A forest of minimum spanning trees is obtained using a modified form of the Boruvka's algorithm. Each minimum spanning tree represents a region determined by the segmentation algorithm.

C. The Annotation Process

Details about the annotation process are presented below.

1. The dataset

We have used for our experiments the segmented and annotated SAIAPR TC-12 [Segmented and Annotated IAPR TC-12 dataset], [Escalante H.J., et al. (2010)] benchmark which is an extension of the IAPR TC-12 [IAPR TC-12 Benchmark] collection for the evaluation of automatic image annotation methods and for studying Automated annotation of natural images using an extended annotation model 7 their impact on multimedia information retrieval. IAPR TC-12 was used to evaluate content-based image retrieval and multimedia image retrieval methods [Clough P., et al. (2006)], [Grubinger M., et al. (2007)]. SAIAPR TC-12 benchmark contains the pictures from the IAPR TC-12 collection plus: segmentation masks and segmented images for the 20,000 pictures, region-level annotations according an annotation hierarchy, region-level annotations according an annotation hierarchy, spatial relationships information. Each image was manually segmented using a Matlab tool named Interactive Segmentation and Annotation Tool (ISATOOL), that allows the interactive segmentation of objects by drawing points around the desired object, while splices are used to join the marked points, which also produces fairly accurate segmentation with much lower segmentation effort. Each region has associated a segmentation mask and a label from a predefined vocabulary of 275 labels. This vocabulary is organized according to a hierarchy of concepts having six main branches: Humans, Animals, Food, Landscape-Nature, Manmade and Other. For each pair of regions the following relationships have been calculated in every image: adjacent, disjoint, beside, X-aligned, above, below and Y-aligned. The following features have been extracted from each region: area, boundary/area, width and height of the region, average and standard deviation in x and y, convexity, average, standard deviation and skewness in two color spaces: RGB and CIE-Lab.

2. The annotation model based on an object oriented approach

The Cross Media Relevance Model is a non-parametric model for image annotation that assigns words to the entire image and not to specific blobs – clusters of image regions, because the blob vocabulary can give rise to many errors. Some principles defined for the relevance models [Lavrenko V., et al. (2001)], [Lavrenko V., et al.

(2002)] are applied by this model to automatically annotate images and for ranked retrieval. Relevance models were introduced to perform a query expansion in a more formal manner. Given a training set of images with annotations this model allows predicting the probability of generating a word given the blobs in an image. A test image I is annotated by estimating the joint probability of a keyword w and a set of blobs Σ P(w,b1 ,...,bm)= J T P(J)P(w,b1 ,...., bm | J) (1) For the annotation process the following assumptions are made:

a) it is given a collection C of un-annotated images
b) each image I from C to can be represented by a discrete set of blobs: I = {b1...bm}
c) there exists a training collection T, of annotated images, where each image J from T has a dual representation in terms of both words and blobs: J = {b1 ....bm;w1...wn }
d) P(J) is kept uniform over all images in T
e) the number of blobs m and words in each image (m and n) may be different from image to image.
f) no underlying one to one correspondence is assumed between the set of blobs and the set of words; it is assumed that the set of blobs is related to the set of words. P(w,b1 ,...,bm | J) represents the joint probability of keyword w and the set of blobs {b1...bm} conditioned on training image J.
An intuitive interpretation of this probability is how likely w co-occurs with individual blobs given that we have observed an annotated image J.
In CMRM it is assumed that, given image J, the events of observing a particular keyword w and any of the blobs {b1...bm} are mutually independent, so that the joint probability can be factorized into individual conditional probabilities. This means that) P(w,b1,...,bm|J) can be written as:

$$P(w, b_1, \ldots, b_m \mid J) = P(w \mid J) \prod_{i=1}^{m} P(b_i \mid J) \quad (1)$$

$$P(w \mid J) = (1 - \alpha_J) \frac{\#(w, J)}{|J|} + \alpha_J \frac{\#(w, T)}{|T|} \quad (2)$$

$$P(b \mid J) = (1 - \beta_J) \frac{\#(b, J)}{|J|} + \beta_J \frac{\#(b, T)}{|T|} \quad (3)$$

where:
(1) P(w|J) , P(w|J) denote the probabilities of selecting the word w, the blob b from the model of the image J.
(2) #(w, J) denotes the actual number of times the word w occurs in the caption of image J.
(3) #(w, T) is the total number of times w occurs in all captions in the training set T .
(4) #(b, J) reflects the actual number of times some region of the image J is labeled with blob b.
(5) #(b, T) is the cumulative number of occurrences of blob b in the training set.
(6) |J| stands for the count of all words and blobs occurring in image J.
(7) |T| denotes the total size of the training set.
(8) The prior probabilities P(J) can be kept uniform over all images in T

The smoothing parameters α and β determine the degree of interpolation between the maximum likelihood estimates and the background probabilities for the words and the blobs respectively. The values determined after experiments for the Cross Media Relevance Model were α = 0.1 and β = 0.9. Starting from the principles of the CMRM model we have obtained an object oriented model using the classes presented in table 1 and the mapping presented in table 2
Table1. The classes used by the object oriented model

Table 2. The mapping used between the CMRM model and the object oriented model

CMRM Model	**Object Oriented Model**
P(W|J)	public double PWJ(Concept w,Image J, IobjectContainer db, int cardT)
P(b|J)	Public double PBJ(Blob b, Image J, IobjectContainer db, int cardT)
$P(w,b1_{,..}b_m\|J)$	Public double PWBsJ(Concept w, List<Blob> blobs, Image J, IobjectContainer db, int cardT)
$P(w,b_{1,..}b_m)$	public double PWBs(Concept w, List<Blob> blobs, List<Image> T, IobectContainer db, int cardT)

For that object oriented model we have made some changes in order to improve the results of the annotation

Classes	**Members**	**Member's Type**
Image	PictureName	String
	Regions	List<Region>
Region	Index	Int
	AssignedBlob	Blob
	AssignedConcepts	Concept
	Features-VectorItem	Features-Vector
	MatrixFilePath	String
Blob	Index	Int
	AverageFeatures Vector	FeaturesVector
Features-Vector	Features	List<double>
Concept	Name	String
	OriginalIndex	Int
Regions-Relationship	RegionA	Region
	RegionB	Region
	RelationshipMode	String
Hierachical-Relationship	ParentConcept	Concept
	ChildConcept	Concept

process obtained using the initial version. In [Jeon J., et al. (2003)] it was mentioned that for the CMRM model the experimental results have shown a mean precision value equal with 0.33 and a mean recall value equal with 0.37. We considered that these values could be further improved. In order to achieve this target some changes were involved having as a result a modified model. The experimental results will show better values for mean precision and mean recall.

The modified version concerns the following two aspects of the annotation task that will be taken into account when computing the probabilities:

only the images having regions associated with the clusters identified based on the regions of the new image will be considered Using only the concepts and the images associated with the identified clusters, more accurate values are obtained for the computed probabilities. In the initial version all concepts and images were taken into account. The main drawback of this version was represented by the fact that it was possible to have several concepts that were not relevant at all (or assigned to other clusters than the ones identified) for a given image, but their frequency in the training set was high, so a major contribution to the probability value. Because the probability is calculated as a sum of the contribution of each concept, high probability values were not always accurate.

3. Steps involved by the annotation process

The annotation process contains several steps:

- Obtaining the ontology – the information provided by the dataset is processed by the Importer module. The concepts associated with images and their hierarchical structure is identified. The Ontology creator module is using that information to generate the ontology. The files containing feature values (extracted from image regions) are processed and stored in the database.
- Obtaining the clusters – we have used K-means algorithm to quantize the feature vectors obtained from the training set and to generate blobs. After the quantization, each image in the training set was represented as a set of blobs identifiers. For each blob it is computed a median feature vector and a list of concepts that were assigned to the test images that have that blob in their representation. The clustering process is summarized in Fig.3.

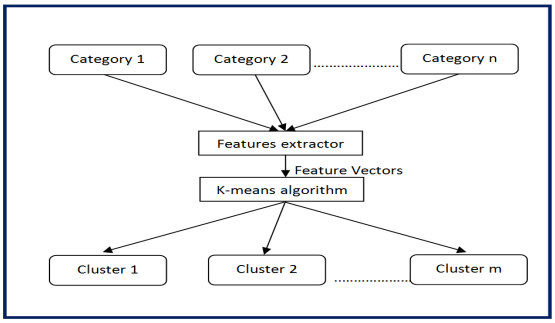

Fig.3. Clustering process

- Image segmentation – the Segmentation module is using the segmentation algorithm described in Section 3 to obtain a list of regions from each new image. In Fig.3. it is presented the list of regions obtained after segmentation together with the annotation result.
- Automated image annotation – this task is performed according with the steps involved by the Annotate Image method presented above. An example is given in Fig.4.

Fig.4. Image annotation

The entire annotation process is summarized in Fig.5.

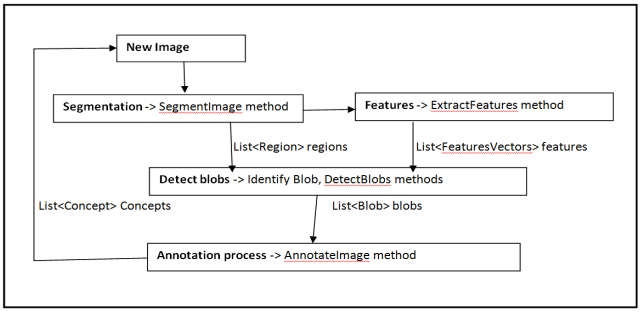

Fig.5. Image annotation process

All tasks involved by this process are implemented in a system having the architecture presented in Fig.6.

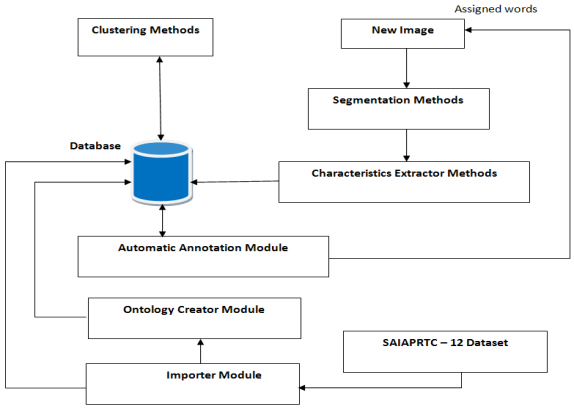

Fig.6. System's architecture

4. Approaches and Implementation

In the work of Makadia et al. [10], they extract 3 color histograms namely, RGB, HSV and LAB and 4 textures namely, Gabor, Haar, GaborQ and HaarQ. These are only basic global colors and texture features. We believe that using these features to represent the image is not enough. We need more higher-level features that could represent image globally at the scene level as well as locally at the Region Of Interest (ROI) level. Human exhibits the exquisite ability at rapidly identifying the gist of the scene of the image. Usually, a human observer of an image at a fraction of second can summarize the essential information about the image such as indoor/outdoor, street, beach, landscape, etc. [3, 13]. Saliency is also a very important point of interest when human observes image because they tend to focus on some important regions or ROIs. Study has shown that the concurrent use of gist of the scene and saliency is a major trait of human vision system [14]. These give reasons for our idea.

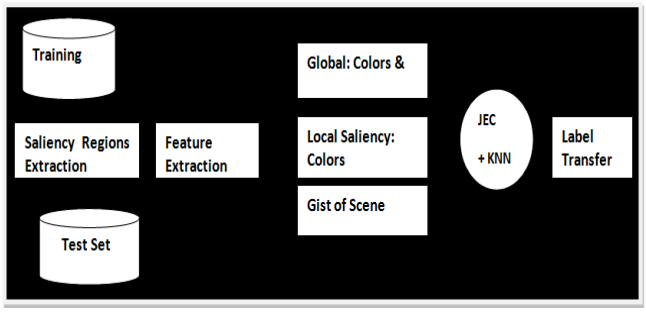

Figure 7: Flow Diagram of the Approach

In this paper, we would like to capture these important features in addition to the basic ones proposed in [10]. The original research on gist of the scene has been reported in [12] with quite a successful rate. For saliency detection, Itti et al.'s work [8] has been the most popular one. However, it is rather complex and computationally expensive. A recent approach introduced by Hou et al. in [5] is simple and gives good performance in real-time computation. Therefore, we choose to implement the later in our work. The outline of our approach is shown in Figure 1. First the features are extracted at image level as well as ROI level. Then we combine the distance of image equally and use K Nearest Neighbor (KNN) method for label transfer.

CONCLUSION

It is widely acknowledged that image annotation is an open and very difficult problem in computer vision. Solving this problem at the human level may, perhaps, require that the problem of scene understanding be solved first. However, identifying objects, events, and activities in a scene is still a topic of intense research with limited success. In the absence of such information, most of the image annotation methods have suggested modeling the joint distribution of keywords and images to learn the association of keywords and low-level image features such as color and texture. Most of these state-of-the-art techniques require elaborate modeling and training efforts. The goal of our work was not to develop a new annotation method but create a family of very simple and intuitive baseline methods for image annotation, which together create a useful annotation evaluation platform. Comparing existing annotation techniques with the proposed baseline methods helps us better understand the utility of the elaborate modeling and training steps employed by the existing techniques. Our proposed baseline methods combine basic distance measures over very simple global color and texture features. K-Nearest Neighbors computed using these combined distances form the basis of our simple greedy label transfer algorithm. Our thorough experimental evaluation reveals that nearest neighbors, even when using the individual basic distances, can outperform a number of existing annotation methods. Furthermore, a simple combination of the basic distances (JEC), or a combination trained on noisy labeled data (Lasso), outperforms the best state-of-the-art methods on three different datasets.

REFERENCES

[1] K. C. Sia, Irwin King, “Relevance Feedback based on Parameter estimation of target distribution” In IEEE International Joint Conference on Neural Networks, pages 1974-1979, 2002

[2] Simon Tong, Edward Chang. “Support vector machine active learning for image retrieval in multimedia” In proceedings of the ninth ACM International Confrence on Multimedia, pages 107-118. 2001.

[3] Dengsheng Zhang, Md. Monirul Islam, Guojun Lu, "A review on automatic image annotation techniques", Pattern Recognition, Volume 45, Issue 1, Pages 346-362, January 2012.

[4] Manjunath, B.S.,Ohm, J.-R., Vasudevan, V.V.; Yamada, A.,2001."Color and texture descriptors," Circuits and Systems for Video Technology, IEEE Transactions on , vol.11, no.6, pp.703,715

[5] Bin Wang; Zhiwei Li; Nenghai Yu; Mingjing Li,2007. "Image Annotation in a Progressive Way," IEEE International Conference on Multimedia and Expo, vol., no., pp.811-814 M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[6] P. Duygulu, K. Barnard, N. de Freitas, D. Forsyth,2002. "Object recognition as machine translation: learning a lexicon for a fixed mage vocabulary", In Seventh European Conference on Computer Vision (ECCV),Vol.4,pp.97-112.

[7] Dhatri Pandya,Prof. Bhumika Shah, “Comparative

Study on Automatic Image Annotation",IJETAE, Volume 4, Issue 3, March2014.

[8] D. Zhang, Md. M. Islam, G. Lu, 2012. "A review on automatic image annotation techniques", Pattern Recognition, vol. 45, no. 1,pp.346–362.

[9] N. Ueda and K. Saito, "Parametric mixture models for multi-labeled text," in *Advances in Neural Information Processing Systems*, vol. 15. MIT Press, Cambridge, MA, 2003, pp. 721–728.

[10] J. Li, J. Z. Wang, "Real-time computerized annotation of pictures", IEEE PAMI 30 (6) (2008) 985-1002.

[11] G. Carneiro, A.B. Chan, P. J. Moreno, N. Vasconcelos, Supervised learning of semantic classes for image annotation and retrieval, IEEE PAMI 29 (3) (2007) 394-410.

[12] Zhiyong Wang, Wan-Chi Siu, Dagan Feng "Image Annotation with Parametric Mixture Model Based Multi-class Multi- labeling". IEEE, 2008. pp. 634-639.

[13] Zhenxing Niu, Gang Hua, Xinbo Gao, Qi Tian, "Semi-supervised Relational Topic Model for Weakly Annotated Image Recognition in Social Media", CVPR-2014 proceeding in IEEExplore.

[14] Yunchao Gong, Yangqing Jia, Thomas K. Leung, "Deep Convolutional Ranking for Multilabel Image Annotation", arXiv: 1312. 4894v2 [cs.CV] 14 Apr 2014

[15] Sean Moran, Victor Lavrenko, "Sparse Kernel Learning for Image Annotation", ACM 978-1-4503-2782-4/14/04

[16] Lamberto Ballanf, Tiberio Uricchiof, Lorenzo Seidenari, and Alberto Del Bimbo, " A Cross-media Model for Automatic Image Annotation", ACM International Conference on Multimedia Retrieval 2014 (ACM ICMR'14)

[17] LEI WANG, LATIFUR KHAN, "Automatic Image Annotation and Retrieval Using Weighted Feature Selection"

[18] Wei Li and Maosong Sun, "Automatic Image Annotation Using Maximum Entropy Model", IJCNLP 2005, LNAI 3651, pp. 34 – 45, 2005

[19] J. Li and J. A. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Transactions on PAMI, 25(10): 175-1088, 2003.

[20] Edward Chang, Kingshy Goh, Gerard Sychay and Gang Wu. CBSA: Content-based soft annotation for multimodal image retrieval using bayes point machines. IEEE Transactions on Circuts and Systems for Video Technology Special Issue on Conceptual and Dynamical Aspects of Multimedia Content Descriptions, 13(1): 26-38, 2003.

[21] Minmin Chen, Alice Zheng, Kilian Q. Weinberger," Fast Image Tagging" CNET 08/2012, Proceedings of the 30 th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

[22] Jiwoon Jeon and R. Manmatha "Using Maximum Entropy for Automatic Image Annotation", Computer Science Department Faculty Publication Series-2004. Paper 136.

[23] Alexei Yavlinsky, Edward Scho_eld, and Stefan Ruger, "Automated Image Annotation Using Global Features and Robust Nonparametric Density Estimation" In CIVR'05