AUTOMATIC IMAGE REGION ANNOTATION THROUGH SEGMENTATION BASED VISUAL SEMANTIC ANALYSIS AND DISCRIMINATIVE CLASSIFICATION

Jing Zhang[⋆]*, Yongwei Gao*[⋆]*, Shengwei Feng*[⋆]*, Yubo Yuan*[⋆]*, Chin-Hui Lee*[†]

[⋆] Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai, China
[†]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA
jingzhang@ecust.edu.cn chl@ece.gatech.edu

ABSTRACT

We propose a new framework for automatic image annotation (AIA) of regions through segmentation based semantic analysis and discriminative classification. Given a test image, it is first segmented by a proposed texture-enhanced JSEG algorithm. Then these regions are represented by an extended bag-of-words model in which a feature vector, based on a visual lexicon with its vocabulary consisting of a visual word or a co-occurrence of multiple visual words, is constructed to represent the region content. Finally a concept classifier learned by a maximal figure-of-merit algorithm is used to predict the region labels. These models are discriminatively trained from image regions with multiple associations between regions and concepts. Experiments on a subset of the Corel 5K data set illustrate that our proposed approach to region AIA achieves more accurate annotation results than some sate-of-the-art algorithms.

***Index Terms*—** image segmentation, texture enhanced JSEG, region annotation, automatic image annotation, MFoM.

1. INTRODUCTION

Automatic image annotation (AIA) is a fundamental problem in computer vision, which is a core technology of image understanding and retrieval [1]. There are many AIA methods and frameworks have been proposed and achieved good performance in image understanding and semantic detection [2], but most of them annotate the images as a whole and haven't considered the semantics of regions.

For achieving more accurate image semantic information, every semantic region of images needs to be analyzed and annotated. Region based annotation methods focus on every independent semantic area, which makes the visual features more accurate to present a certain semantic concept. Compared with global based image annotation, region based methods can achieve more precise and detailed semantic information and become the research hotspot of image understanding in recent years [3][11].

Image segmentation is the basis of region based image annotation methods, which is very important for image content representation and further image content analysis. Accurate image segmentation makes every region having independent semantic information, which is the foundation of region content representation and modeling. Though many image segmentation algorithms have been proposed and used widely in image annotation and retrieval [21]-[25], such as JSEG [21] and NCUT [22]. We still couldn't find one algorithm that can deal with semantic segmentation well.

This research was conducted while visiting Georgia Tech. It has been supported by the National Nature Science Foundation of China (Grant 61402174 and 61370174).

Representation of the image/region content is a fundamental step in AIA, which influences the results to a great extent. Image content representation based on BoW model is proposed by Li et al. [5], and then it is widely used in image and video retrieval fields [6]-[8], such as object recognition [6], image categorization [4] and near duplicate detection [7]. The obvious benefit of BoW model is that we can explore the modeling of syntactic and semantic relationship among the visual symbols.

Based on the image content representation, various statistical image annotation models [2] have been utilized to obtain the relationship between visual features and semantic labels. One of them transforms multi-label annotation problem into several single-label classification problems, such as support vector machine (SVM) [12]-[14] and artificial neural network (ANN) [9][10]. Another one uses specific algorithms to process multi-label data directly, such as cross-media relevance model (CMRM) [11], continuous-space relevance model (CRM) [15], Markov random field (MRF) [16] and conditional random fields (CRF) [17]-[19].

We first propose a textual-enhanced JSEG algorithm, which improves the segmentation results by fusing the texture and color class maps. We then propose a new point-line-region (PLR) model to reduce the over segmentation problem. We also enhance the BoW model to represent the image region content and study the semantic relationship between visual words by learning ensemble visual lexicons. Finally a multi-class maximal figure-of-merit (MC MFoM) approach [27] is used to construct model for region label prediction.

2. TEXTURE-ENHANCED JSEG ALGORITHM

JSEG is one of the popular region-based segmentation algorithms widely used in image annotation and retrieval [21] in which the color information is fully considered. However it is not effective to separate the regions with the same color distribution. Here, an improved texture-enhanced JSEG (TJSEG) algorithm is proposed. To better distinguish the regions with different texture distributions, we combine the texture and color class maps. Then a point-line-region (PLR) model is proposed to reduce over-segmentation.

2.1. Texture-color class map

Due to that texture feature is not considered in JSEG, the regions with similar color and different texture will not be segmented well. We proposed a texture-enhanced JSEG algorithm, in which the original color class map is replaced by a Texture-Color (TC) class map. TC class map is produced by fusing color class map and texture class map, in which color and texture features are considered simultaneously and makes up the shortage of JSEG.

For achieving texture class map, the Gabor texture features of

the original images are extracted and clustered by Kmeans. Then the texture map is constructed by giving every pixel a certain class label. Next, we will produce a new texture-color class map by fusing color class map and texture class map.

Suppose that there are M classes in color class map and N classes in texture class map, and there are at most $M * N$ classes in TC map. We record the pixels with the same Color map label and the same Texture map label as a new TC class label. There are too many classes in TC map, and some classes only have a few pixels. Hence we combine some classes with few pixels into nearby classes by their color and texture similarity.

We merge the similar regions based on the color class map and Gabor feature image. Two regions whose size is less than a specific size, will be merged due to the color class map if the Gabor texture features of them are similar. Hence in the TC class map, the regions with similar color and different texture will not be merged. On the other hand, the regions with similar texture and slightly different color will be merged in terms of its surrounding color map.

After obtaining the TC class maps, J value is calculated. We add the texture feature to construct the TC class map, which makes the segmentation more reasonable, and decreased the over segmentation in some degree by reducing the number of seed points. However the segmentation results are still unsatisfactory. For further resolving the problem of over segmentation, we propose a new PLR model (point-line-region model), which adopts the method of removing the dividing line between two regions to merge them.

2.2. Point-line-region model

Over segmentation is often caused by many wrong boundaries in the segmented image. In the proposed PLR model, there are three different stages, including 'point', 'line' and 'region', to remove some incorrect boundaries and merge over-segmentation regions. In our preliminary experiments, we found that the similarity between two sides of a boundary is important to judge a correct segmentation.

In the 'point' algorithm, we remove boundaries by the similarity between points as illustrated in Fig. 1 (I). Suppose L_i is a dividing line, we choose three points including two endpoints and midpoint to judge the similarity of the regions on both sides. Points A, B and C are chosen and the RGB values of the three points on each side of the picked points are counted. For example, points 1-6 on the two sides of line L_i are chosen. The RGB distance d_p of the two horizontal sides at Points A and C is obtained by Eq. (1), or by Eq. (2) for two vertical sides at Point B:

$$d_p = \sum_{i=1}^{3}\sum_{j=3}^{3} Euc(P_i(x+j,y), P_i(x-j,y)) \quad (1)$$

$$d_p = \sum_{i=1}^{3}\sum_{j=3}^{3} Euc(P_i(x,y+j), P_i(x,y-j)) \quad (2)$$

in which, $P_i(x,y)$ is the RGB values of the i-th chosen pixel in the boundary, and $Euc(P_i(x,y), P_j(x,y))$ is the Euclidean distance of the RGB values between pixels P_i and and two regions on its sides will be merged if the distance is below a threshold.

Similarly we propose the line algorithm to evaluate the similarity between two regions near the dividing line and removing the segmentation line if the two bands are similar as shown in Fig. 1 (II), there are two pixel bands A and B on each side of the line L. Then the HSVH of each band area is extracted. The distance d_l of the two band areas for each boundary is obtained by:

$$d_l = Euc(HSVH_{b_A}, HSVH_{b_B}) \quad (3)$$

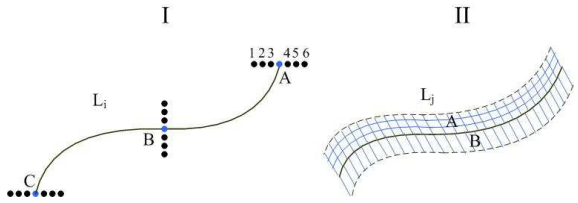

Fig. 1: Point and line algorithms

in which, $HSVH_{b_A}$ is the HSVH of pixel band A, and the Euc is Euclidean distance. If the "line" similarity is less than a threshold we set, the boundary will be removed.

After processing by the point and line algorithms, most of the wrong boundaries have been removed. Then we use the proposed region algorithm as the last step with scalable color descriptor (SCD) to be abstracted to represent every region, and the Euclidean distance d_r is used to compare the similarity between two regions:

$$d_r = Euc(SCD_{r_i}, SCD_{r_j}) \quad (4)$$

where SCD_{r_i} is SCD of region i.

Due to the PLR model, over segmentation can be reduced substantially. Experiments on the Corel 5K dataset illustrated that our texture-enhanced JSEG algorithm can achieve more accurate segmentation results than JSEG and NCUT, and deal with the over segmentation problems well. In Fig. 2, there are several comparative results with the three segmentation algorithms. Obviously, the results of texture-enhanced JSEG method are always superior.

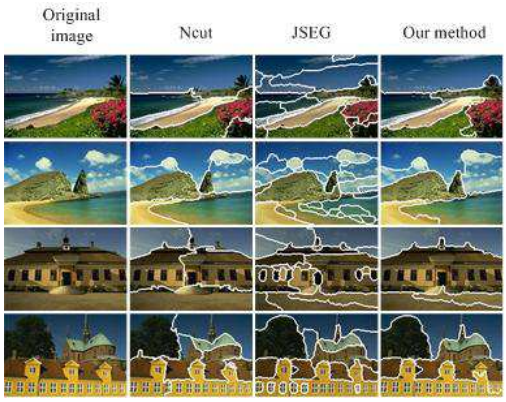

Fig. 2: The examples of segmentation results

3. IMAGE REPRESENTATION BY VISUAL CODEBOOK

To obtain accurate image annotation and retrieval results, representation of the image content is the most fundamental step and influences the results to a great extent. The aim of image representation is to express the image content by visual features, which is desired to associate a given image with several concepts describing the objects and their spatial relations in an image. There are many image content representation methods have been proposed during the last decades, and Bag of Words (BoW) as the famous one is widely used in image annotation and retrieval in recent years. Next we will introduce an extended BoW based image content representation method, which represents the region of image by a set of visual words considering the contextual and spatial dependency.

Every image is divided to several regions with different semantic concepts after image segmentation. Our aim is to represent these regions by visual features which implied the contextual and spatial dependency. We extend BoW model by several aspects. The first one is that we learn a collection of visual terms, each of them may be a single token or a combination of tokens. Secondly, we learn an ensemble of visual lexicons from the different sets of low-level visual features, each describing a partial content of an region. Thirdly, we learn co-occurrence statistics of visual terms to reflect the contextual and spatial relationship.

Different from the conventional BoW model, we segment every region by blocks of 4×4 pixels. Due to the irregular shape of the region, we use the rule that for each grid G which has $M_0 \times M_0$ pixels, if most of the pixels belong to a certain region, the grid is labeled according to that region.

Preliminary experiments prove that color and texture features are more effective for image annotation and retrieval. We chose three low-level visual features, including HSV color histogram (HSVH), color moments (CM) and Gabor texture as the features of region content representation. We extract HSVH, CM and Gabor from every grid and group each of them into a vector, X_{ij}, for the block located at the i-th row and j-th column. All vectors from training images are quantized as a codebook index.

The visual words in a lexicon consist of not only the tokens but also patterns inferred from the token relations (e.g., location, spatial, n-gram, etc.). Here, we construct the visual words by combining unigram and bigram patterns to represent the content of regions. Every block centered at X_{22} in a 3×3 block group has eight neighbouring blocks with its direction-specific bigram patterns obtained from its neighboring blocks. These bigrams are treated as distinctive patterns. When we express a region by extended BoW model, we use not only the patterns of unigram but also the bigrams. After a region of an image is tokenized and the occurrence statistics of visual lexicons are tabulated, a feature vector is extracted for content representation by the techniques in [26].

We extract the low-level visual feature from every block, and cluster them into N clusters. Then a visual lexicon, A=A_1, A_2, ..., A_N, with N visual terms is constructed. For traditional BoW model, every region will be represented as a vector of N dimensions, V=(v_1, v_2, ..., v_N), each component being the statistics of the visual term occurred in the region. In our extended BoW model, we combined unigram and bigram patterns in the codebook, which makes the number of tokens become N+N*N. Hence every region in our method is represented as a vector V with N+N*N dimensions. For more accurately representing the visual contextual and spatial dependency, we can extend the codebook with n-gram patterns. However as you can see, the tokens of codebook will increase apace and the vector V will have a high dimension, which will influence the accuracy of region annotation. During many experiments on different patterns, we choose the one which combines unigram and bigram patterns to construct codebook and represent the regions by visual words.

4. IMAGE REGION ANNOTATION ALGORITHM

We use an MC MFoM approach [27] to build the concept models for image region annotation. It learns multi-category classifiers by optimizing a metric-oriented objective function. Therefore it is more robust and better than the popular SVM and CRF classifiers, especially for learning in the case with sparse training.

All the training images are segmented by texture-enhanced JSEG and every region is assigned a concept manually. A training set is defined as $T = \{(V, l)|V \in R^D, l \in C\}$, where (V, l) is a training sample, V is a D-dimensional feature extracted from a region as discussed in Section 3. l is the manually assigned label of corresponding region. Usually one label corresponds to one region. $C = \{C_j, 1 \leq j \leq K\}$ is denoted as the predefined label set, in which K is the total number of labels and C_j is the j-th label. We will learn a discriminant function for the j-th keyword with the parameter set Λ_j and $g_j(V; \Lambda_j)$,.

According to the multiple-label decision rule in Eq.(5), region V are assigned multiple relevant labels in the evaluation stage.

$$\begin{cases} Accept\ V \in C_j \quad if\ g_j(V;\Lambda_j) - g_j^-\left(V;\Lambda^-\right) > 0 \\ Reject\ V \in C_j, Otherwise \end{cases} 1 \leq j \leq N \quad (5)$$

where $g_j^-(V; \Lambda^-)$ is named as class anti-discriminant function for the j-th keyword, which is defined as in Eq.(6).

$$g_j^-(V; \Lambda^-) = \log \left[\frac{1}{|C_j^-|} \sum_{i \in C_j^-} \exp\left(g_i(V; \Lambda_i)\right)^\eta \right]^{\frac{1}{\eta}} \quad (6)$$

where C_j^- is a subset containing the most competitive keyword models against C_j, $|C_j^-|$ is its cardinality, Λ^- is the parameter set for all competitive keyword models, and η is a positive constant. Eq.(6) measures the score as a geometric average of scores among all competing categories, which works as a negative model for the j-th label.

In the learning stage of MC MFoM, we estimate the parameter set $\Lambda = \{\Lambda_j, 1 \leq j \leq N\}$ by optimizing a metric-oriented objective function. A one-dimensional class misclassification function $d_j(V; \Lambda)$ is used to smooth the discrete decision rule in Eq.(5),

$$\mathrm{d_j(V;\Lambda) = -g_j(V;\Lambda_j) + g_j(V;\Lambda_j^-)}, \quad (7)$$

where $d_j(V; \Lambda) < 0$ when a correct decision is made. Otherwise, $d_j(V; \Lambda) \geq 0$. It works similar to Eq.(5). Since Eq.(7) is valued from -∞ to +∞, for the keyword C_j, a class loss function $k_j(V; \Lambda)$ is defined for normalization as Eq.(8).

$$k_j(\mathrm{V};\Lambda) = \frac{1}{1+\mathrm{e}^{-\alpha(\mathrm{d_j(V;\Lambda)}+\beta)}} \quad , \quad (8)$$

where α is a positive constant controlling the size of the learning window and learning rate, and β is a constant measuring the offset of $d_j(V; \Lambda)$ from 0. They are determined by amounts of experiments. Eq.(8) simulates the error count made by the j-th region model for a given sample V. With the above definitions, most commonly used metrics, precision, recall and F_1, are approximated over training set, T by Eq.(9)-(11).

$$FN_\mathrm{j} \approx \sum\nolimits_{V \in T} k_j(V; \Lambda) \cdot 1(V \in C_j), \quad (9)$$

$$FP_\mathrm{j} \approx \sum\nolimits_{V \in T} (1 - k_j(V; \Lambda)) \cdot 1(V \notin C_j), \quad (10)$$

$$TP_\mathrm{j} \approx \sum\nolimits_{V \in T} (1 - k_j(V; \Lambda)) \cdot 1(V \in C_j), \quad (11)$$

where TP_j is the true positive, FP_j is the false positive, and FN_j is the false negative for the j-th keyword. $1(.)$ is an indicator function of any logical expression. In the experiment, the micro-averaging F_1 is our preferred objective function defined as:

$$K(V;\Lambda) = 2\sum\nolimits_{i=1}^{N} TP_i / \left[\sum\nolimits_{i=1}^{N} FP_i + \sum\nolimits_{i=1}^{N} FN_i + 2\sum\nolimits_{i=1}^{N} TP_i\right], \quad (12)$$

solved by a generalized probabilistic descent algorithm [27].

5. EXPERIMENTS

We choose the Corel 5K datasets as experimental data, which is publicly available and diffusely used to evaluate the methods of image annotation and retrieval [20]. Three themes including 'building', 'seaside', 'transportation' are chosen as our experimental image datasets. 300 images are chosen for each theme, of which 200 images are used to train models and 100 images for testing. In the image dataset of building, there are 5 labels: $sky, building, plant, road, water$. And in the image dataset of seaside, there are 8 labels: $sky, building, plant, sand, sea, boat, rock, sun$. In the dataset of transportation, there are 5 labels: $sky, building, plant, car, plane$. All shown experimental results are measured by the averages of precision (mP), recall (mR), and F-measure (mF) over all labels and the number of concepts detected.

In our experiments, we segmented every image by the proposed texture-enhance JSEG algorithm first, then every region is segmented into grids with a grid size of 4×4. After that, a 256-dimensional HSVH vector, a 225-dimensional CM vector and a 60-dimensional Gabor texture vector are extracted from each grid [19]. Then $k-means$ clustering is used to get 64 or 128 symbols of each feature for region tokenization. Every region is represented by a visual words with unigram and bigram patterns of the codebook. A linear classifier is trained using MC MFoM for label prediction, which was illustrated in Section 4.

The experimental results on three different datasets are illustrated in Table 1. From these results, our proposed method is very effective for image region annotation. The average precision and recall on the 'building' dataset are 58.47 and 53.86. For the 'transportation' dataset, the precision and recall are 77.06 and 76.84 respectively, which are very good performance for region annotation.

Table 1: Average precision, recall and F_1 of three concept groups

Dataset	mP	mR	mF
Buildings	58.47%	53.86%	54.25%
Seaside	56.76%	52.21%	53.45%
Tansportation	77.06%	76.84%	76.29%

For achieving more accurate annotation results, We have done many experiments with different parameters, codebooks and features. The first set of experiments is the comparison of different features. There are three features in our experiments, including HSVH, CM and Gabor textrue, in which two of them are about color and one is texture. We illustrated the experimental results in Table 2 with color feature HSVH, texture feature Gabor texture, combination of HSVH and Gabor textrue, and combination of all the three features. It is seen that the results of HSVH is better than Gabor Texture, and fusion of three features is better than all of other features.

Table 2: Average precision, recall and F_1 of features and fusions

Features	mP	mR	mF
HSVH	57.30%	56.50%	56.09%
Gabor Texture	44.01%	44.47%	41.70%
HSV+Gabor Texture	58.63%	55.91%	55.96%
HSVH+Gabor Texture+CM	64.10%	60.98%	61.33%

The second set of experiments is comparison of different codebooks. We compared the experimental results with different number of tokens and different patterns of codebooks, mainly including 64-token codebook with unigram patterns, 64-token codebook with two-dimensional unigram patterns, 128-token codebook with unigram and bigram patterns, and 64-token codebook with unigram and bigram patterns. Codebook with unigram patterns is similar with traditional BoW model, which represents the region as the statistics of visual words. Codebook with unigram and bigram patterns, described in Section 3, adds the contextual and spatial dependency of grid into the patterns, which makes the region content representation more detailed and approaching semantic. Codebook with two-dimensional unigram patterns which combined HSVH feature and Gabor texture feature from the level of patterns, which constructed more new patterns with different color and texture information and makes the region content representation more rich. The experimental results are illustrated in Table 3. As we can see, 64-token codebook with unigram and bigram achieves the best performance than others.

Table 3: Average precision, recall and F_1 of different codebooks

Patterns of Codebook	mP	mR	mF
64-token unigram	55.84%	56.85%	54.81%
64-token two-dimensional unigram	56.58%	53.86%	53.84%
128-token unigram and bigram	54.01%	50.71%	50.41%
64-token unigram and bigram	64.10%	60.98%	61.33%

To indicate the superiority of our method, we compared it with the algorithm proposed in [19]. Since these two algorithms are all region based automatic image annotation methods, we did the experiments on the same datasets with similar conditions, which makes our results more credible and comparable. The comparison of performance achieved by the two algorithms is shown in Table 4.

Table 4: Average precision, recall and F_1 of two algorithms

Algorithms	mP	mR	mF
Our algorithm	57.61%	53.04%	53.85%
MLSIA	45.55%	37.93%	41.44%

To make the annotation results more visual, we illustrated some examples of image segmentation and region annotation results in Fig. 3. Our proposed algorithm can achieve accurate region annotation results close to the ground truth.

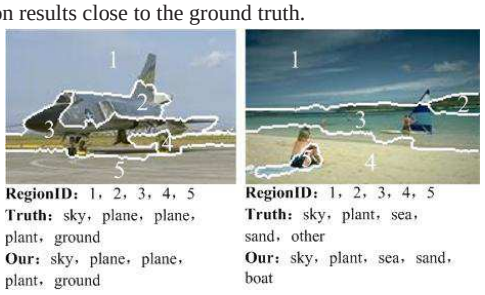

Fig. 3: Some image region annotation results

6. CONCLUSION AND FUTURE WORK

In this paper, we proposed a new framework for automatic image region annotation. An image is first segmented by our proposed texture-enhanced JSEG algorithm, and the co-occurrence statistics of the visual words are used to characterize the content of region. For accurate image region annotation, we use MC MFoM to train discriminative concept models to predict the labels of image region. Experiments on the subset of Corel 5K dataset illustrates that our proposed framework can achieve good performances on image region annotation. In the future, we will experiment on the whole Corel 5K dataset and prove the framework is effective on all kinds of concepts. In addition, we will add semantic position relationship analysis to our framework for even better image region annotation.

7. REFERENCES

[1] F. Kang, "Automatic image annotation," Doctoral Dissertation, Michigan State University, 2007.

[2] D.S. Zhang, Md.M. Islam, G.J. Lu, "A review on automatic image annotation techniques," Pattern Recognation, Vol.45, Issue 1, Pages 346-362, January 2012.

[3] C.B. Yang, M.Dong, J. Hua, "Region-based Image Annotation using Asymmetrical Support Vector Machine-based Multiple-Instance Learning," Proc. of IEEE Conference on Computer Vision and Pattern Recognation, 2006.

[4] H. Zhang, A. Berg, M. Maire, J. Malik, "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol.2, pp.2126-2136, 2006.

[5] F.F. Li, P. Perona, "A Bayesian hierarchical model for learning natural scene categories," IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol.2, pp.524-531, 2005.

[6] J. Sivic, A. Zisserman, "Video google: a text retrieval approach to object matching invideos," in: Proc. of Ninth IEEE International Conference on Computer Vision, vol. 2, pp. 1470-1477, 2003.

[7] X. Wu, W.L. Zhao, C.W. Ngo, "Near-duplicate keyframe retrieval with visual keywords and semantic context," Proceedings of the 6th ACM international conference on Image and video retrieval (CIVR '07), pp. 162-169, 2007.

[8] S. Alvarez, M. Vanrell, "Texton theory revisited: A bag-of-words approach to combine textons," Pattern Recognitio, vol.45, pp.4312-4325, 2012.

[9] S.B. Park, J.W. Lee, S.K. Kim, "Content-based image classification using a neural network," Pattern Recognition Letters, vol.25, no.3, pp. 287-300, 2004.

[10] F.D. Frate, F. Pacifici, G. Schiavon, C. Solimini, "Use of neural networks for automatic classification from high-resolution images," IEEE Transactions on Geoscience and Remote Sensing, vol.45, no.4, pp.800-809, 2007.

[11] Y. Wang, T. Mei, S.G. Gong, X.S Hua, "Combining global, regional and contextual features for automatic image annotation," Pattern Recognition, vol.42, no.2, pp.259-266, 2009.

[12] X.J. Qi, Y.T. Han, "Incorporating multiple SVMs for automatic image annotation," Pattern Recognition, vol.40, no.2, pp.728-741, 2007.

[13] D.P. Tao, L.W. Jin, W.F. Liu, X.L. Li, "Hessian Regularized Support Vector Machines for Mobile Image Annotation on the Cloud," IEEE Transactions on Multimedia, vol. 15, no. 4, pp. 833-844, 2013.

[14] J.W. Hu, K.M. Lam, "An efficient two-stage framework for image annotation," Pattern Recognition, vol. 46, no. 3, pp.936-947, 2013.

[15] V. Lavrenko, et al., "A model for learning the semantics of pictures," Proc. of NIPS'03, 2003.

[16] P. Carbonetto, et al., "A statistical model for general contextual object recognition", Proc. of ECCV'04, 2004.

[17] X. He, R.S. Zemel, M.Á. Carreira-Perpiñán, "Multiscale conditional random fields for image labeling," Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 695-702, 2004.

[18] T. Mensink, J. Verbeek, G. Csurka, "Tree-structured CRF Models for Interactive Image Labeling," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 2, pp. 476-489, 2012.

[19] J. Zhang, YX. Zhao, D. Li, ZH. Chen, YB. Yuan. " A Novel Image Annotation Model Based on Content Representation with Multi-layer Segmentation ." Jounal of Neural Computing and Applications, Vol. 26, Issue 6, , pp.1407-1422, June 2015.

[20] P. Duyhulu, et al.,"Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," Proc. of ECCV 2002.

[21] Y. Deng, B. S. Manjunath, and H. Shin, Color image segmentation, Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Vol:2, Fort Collins, CO, 23-25 Jun, 1999.

[22] C Zahn, Graph Theoretical Methods for Detecting and Describing Gestalt Clusters, IEEE Transactions on Computation, Vol:20, pp:68-86, 1971.

[23] S.P. Zhu, X. Xia, Q.R. Zhang, and Belloulata, K., An Image Segmentation Algorithm in Image Processing Based on Threshold Segmentation, The third International IEEE Conference on Signal-Image Technologies and Internet-Based System, pp:673-678, 2007.

[24] A.D.Sappa, Unsupervised contour closure algorithm for range image edge-based segmentation, IEEE Transactions on Image Processing, Volume: 15, Issue: 2, pp:377-384, 2005.

[25] F. Moscheni, S. Bhattacharjee, M. Kunt, Spatio-temporal segmentation based on region merging, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 20, Issue: 9 pp:897-915, 1998.

[26] S. Gao, D.H. Wang and C.H. Lee. "Automatic image annotation through multi-topic text categorization", Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp:14-19 May 2006.

[27] S. Gao, W. Wu, C.H. Lee and T.S. Chua, "A MFoM learning approach to robust multiclass multi-label text categorization," Proc. of International Conference on Machine Learning. 2004.