Hindawi Publishing Corporation
Advances in Multimedia
Volume 2014, Article ID 857682, 7 pages
http://dx.doi.org/10.1155/2014/857682

Research Article

Automatic Image Tagging Model Based on Multigrid Image Segmentation and Object Recognition

Woogyoung Jun,[1] Yillbyung Lee,[1] and Byoung-Min Jun[2]

[1]*Department of Computer Science, Yonsei University, 50 Yonsei-ro, Seodaemun, Seoul 120-749, Republic of Korea*
[2]*Department of Computer Engineering, Chungbuk National University, Cheongju 361-763, Republic of Korea*

Correspondence should be addressed to Woogyoung Jun; woogyoung@yonsei.ac.kr

Received 17 May 2014; Accepted 7 December 2014; Published 22 December 2014

Academic Editor: Balakrishnan Prabhakaran

Since rapid growth of Internet technologies and mobile devices, multimedia data such as images and videos are explosively growing on the Internet. Managing large scale multimedia data with correct tags and annotations is very important task. Incorrect tags and annotations make it hard to manage multimedia data. Accurate tags and annotation ease management of multimedia data and give high quality retrieve results. Fully manual image tagging which is tagged by user will be most accurate tags when the user tags correct information. Nevertheless, most of users do not make effort on task of tagging. Therefore, we suffer from lots of noisy tags. Best solution for accurate image tagging is to tag image automatically. Robust automatic image tagging models are proposed by many researchers and it is still most interesting research field these days. Since there are still lots of limitations in automatic image tagging models, we propose efficient automatic image tagging model using multigrid based image segmentation and feature extraction method. Our model can improve the object descriptions of images and image regions. Our method is tested with Corel dataset and the result showed that our model performance is efficient and effective compared to other models.

1. Introduction

Nowadays, we are always online. Desktop computers, laptop computers, and even smartphones are connected online anytime and anywhere. It is very easy to share multimedia data with our mobile devices and explosive growth of social network services such as Facebook, Flickr, and Twitter helps with tremendous growth of multimedia data on the Internet. To manage these multimedia data, reliable tag and annotation information should be improved. How to manage such large scale of multimedia is the most famous topic these days. Well-tagged image is effective for management and retrieval. We focus on automatic image tagging model using image segmentation and feature extraction. Since an image presents multiple objects on single image, we mainly focus on how to extract multiple objects successfully. We find out image segmentation technique and propose a multigrid based image segmentation method. Sometimes an image may contain single object but most of user created contents contain multiple objects in image (Figure 1). Therefore, extracting visual features from whole image has limitation for tagging or annotating an image. Feng et al. [1] also proposed grid based method which is more effective than the basic image segmentation models. But it still has limitation for multiobject problem in segmented region. Therefore, we propose a multigrid image segmentation method which is able to extract features of multiobjects presented in an image. Experimental results showed that our model presented efficient, effective, and most accurate image tagging results compared to other models.

We present related researches in Section 2 and propose our multigrid image segmentation model in Section 3. In Section 4, we present our novel automatic image tagging model based on multigrid image segmentation method. We present experimental results in Section 5. Finally, we reach to conclusions and feature works in Section 6.

2. Related Researches

Typically there are three types of image tagging models. Those are automatic, manual, and half-automatic models. Manual tagging is most accurate and reliable for image tags. Nevertheless, it takes tremendous cost to tag image manually.

(a) (b)

Figure 1: An example of images: an image contains single object (a) and an image contains multiobjects (b).

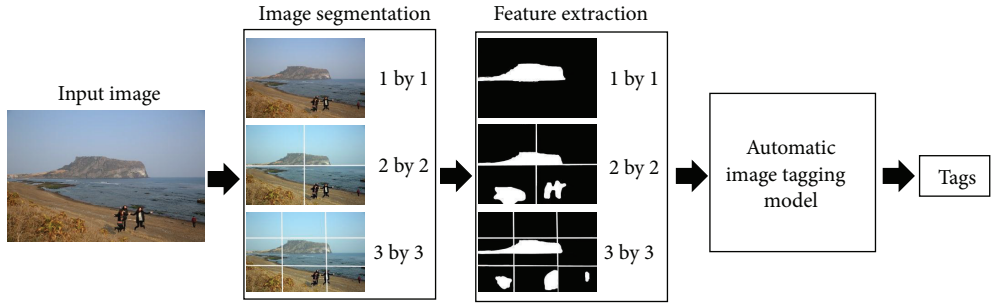

Figure 2: Overview of our grid based image segmentation, feature extraction, and automatic image tagging model.

Half-automatic models such as Google Image Labeler are good way to tag image pretty accurate, but it also has limitation that users have to spend time in playing game and it might cause suffering from noise tags. Therefore, fully automatic image tagging model is most interesting research field in these days despite of lower performance compared to manual and half-automatic models. Many researchers make effort to increase accuracy.

Learning based automatic image tagging models are most recent research interests. Starting keyword based methods, semantic keyword methods are proposed. Most recently, for more effective image tagging, setting up the relationship between textual features and visual features is currently the main topic. Jeon et al. [2] and Yang et al. [3] proposed cross media model which tags images with joint probabilities of semantic information and visual features. They used discrete features to tag images and it can lose helpful visual information. Carneiro et al. [4] proposed SML model which is semisupervised learning model which is not suitable for image segmentation. Wang et al. [5] combined global and local regions and, to improve tagging performance, they used contextual features. Lindstaedt et al. [6] proposed visual folksonomy based automatic image tagging especially for fruits and vegetables. In addition, Manh and Lee [7] focus on small object segmentation based on visual saliency in natural images and Divya et al. [8], Santosh and Shyam [9], and Patil and Kokare [10] demonstrated that image segmentation and automatic image tagging models help with semantic image retrieval.

Unlike these algorithms, our model focused on efficient and effective multigrid based image segmentation model and object recognition. And we propose an image tagging model based on our multigrid image segmentation method (Figure 2). Our proposal for multigrid based image segmentation and object recognition is shown in next section and then we propose efficient automatic image tagging model.

3. Image Segmentation Model

Most of image region segmentation depends on surrounding contrast. Cheng et al. [11] proposed global contrast-based region detection algorithm. For image segmentation, Felzenszwalb and Huttenlocher [12] proposed graph-based image segmentation method, and Xiong et al. [13] proposed hierarchical deformable model for face detection. And color contrast for each image region is being calculated. In this paper, we calculate weight for each region for image region segmentation.

Let $D_r(r_1, r_2)$ be the distance between image regions r_1 and r_2; then $D_r(r_1, r_2)$ can be calculated as follows:

$$D_r(r_1, r_2) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} f(c_{1,i}) f(c_{2,j}) D(c_{1,i}, c_{2,j}), \quad (1)$$

where $f(c_{1,i})$ is probability of color $c_{1,i}$ on image region r_1 and $f(c_{2,j})$ is probability of color $c_{2,j}$ on image region r_2. Distance between $c_{1,i}$ on region r_1 and $c_{2,i}$ on region r_2 $D(c_{1,i}, c_{2,j})$ is distance between the two pixels. It takes long time to calculate whole color distance because Lab color space is 255^3.

Therefore, we used histogram based compress Lab color space. Therefore, we can recalculate $D_r(r_1, r_2)$ as follows:

$$D_r(r_1, r_2) = \sum_{m=1}^{n_B} \sum_{n=1}^{n_B} f(b_m) f(b_n) D(b_m, b_n), \quad (2)$$

where b_m is mth bin of color $c_{1,i}$ in region r_1 and b_n is nth bin of color $c_{2,j}$ in region r_2. n_B is number of histogram bins. Now we can calculate $f(b_*)$ as follows:

$$f(b_*) = \frac{\#(b_*)}{N(r_k)}, \quad (3)$$

where $\#(b_*)$ is number of bin colors b_* and $N(r_k)$ denotes number of pixels in region $r_k \cdot f(b_*)$. If the pixel of certain color appears many times, it means that it is main color of certain region. If we calculate directly $f(b_*)$ within (3), then similar color may be assigned to another bin and especially it can be noise when the region is small. To overcome such problem, we redefine (3) as follows:

$$f'(b_*) = \frac{\sum_{i=1}^{l} (Z - D(b_*, b_i)) \#(b_*)}{[(l-1)Z] N(r_k)}, \quad (4)$$

where l is the number of similar colors with b_* in histogram. $D(b_*, b_i)$ is distance between b_* and the ith similar color of b_*. Z and $(l-1)Z$ are normalized factor and $Z - D(b_*, b_i)$ is linear transformed weight. Now we calculate weight for certain region by comparing with other regions. We calculate region importance r_h as follows:

$$I(r_h) = \sum_{r_h \neq r_i} w(r_i) D_r(r_h, r_i). \quad (5)$$

$D_r(r_h, r_i)$ can be calculated with (2). $w(r_i)$ denotes number of pixels in region r_i and it can be weight of region r_i as well. Since (5) does not concern spatial relationship, we recalculate (5) with spatial relationship as follows:

$$I(r_h) = \sum_{r_h \neq r_i} \exp\left(-\frac{D_s(r_h, r_i)}{\sigma_s^2}\right) w(r_i) D_r(r_h, r_i), \quad (6)$$

where $D_s(r_h, r_i)$ denotes spatial distance between regions r_h and r_i. $D_s(r_h, r_i)$ is calculated using Euclidean distance measure. σ_s is used to control spatial weight.

Now we propose method to segment image into multigrid to recognize objects. We segment images based on multigrid based image segmentation method. And then we extract object feature as already mentioned. Finally, we extract visual feature from each segmented image. Since we extract visual features from multigrid segmented images, we can extract most objects in image. In this paper, we segment images into 3 steps. In first step, we extract feature from entitled image. In second step, we extract features from 2 by 2 grid segmented images. In third step, we extract features from 3 by 3 grid segmented image. The number of steps increases more than 3 by 3 grid; then the well-extracted number of objects (Figure 3) and accuracy (Figure 4) decrease.

Number of well-extracted objects and their accuracy show best result on step 3. That is, smaller objects in images

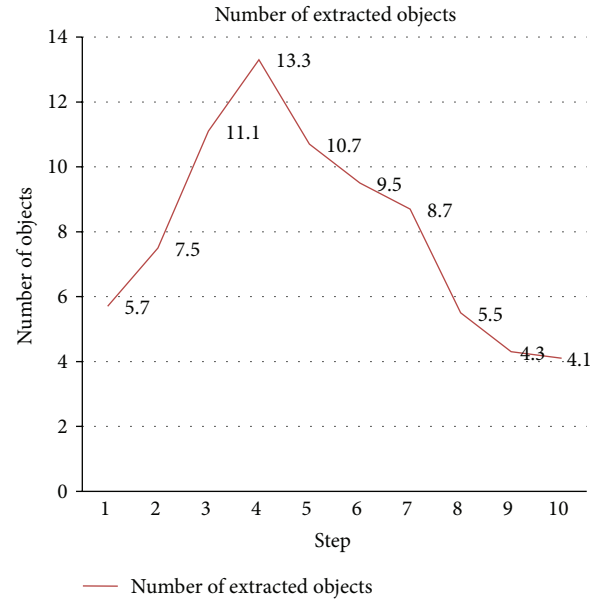

Figure 3: Number of well-extracted objects for segmentation steps.

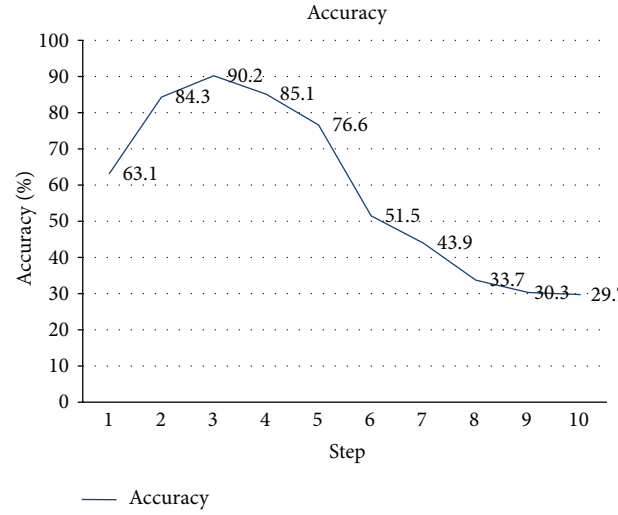

Figure 4: Accuracy of well-extracted objects for segmentation steps.

are not that important in that image and more important object can be segmented into other regions which means important object feature can be lost.

We can see an object extracted from segmented images (Figure 5). When we extract feature from entitled image, only one object is extracted. When we extract features from 2 by 2 segmented images, we could recognize more detailed objects in images. Meanwhile, we could recognize more detailed objects in 3 by 3 segmented images.

4. Automatic Image Tagging Model

In this section, we introduce our automatic image tagging model. We combine with our multiscale segmented images

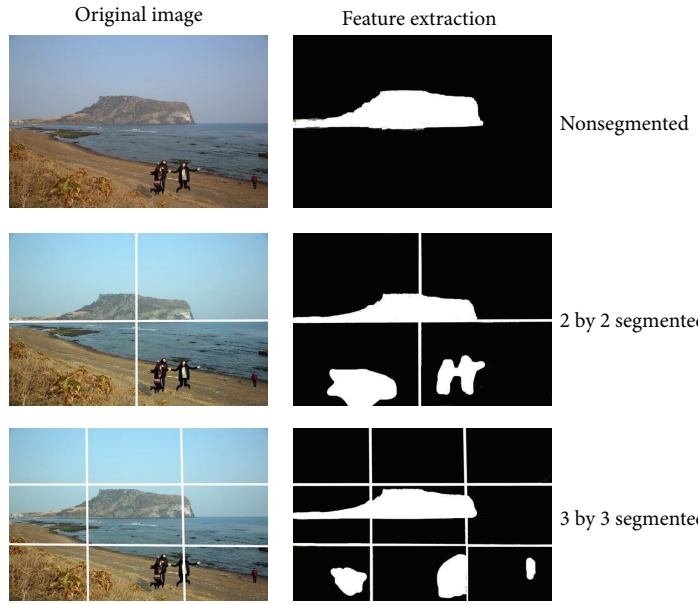

Figure 5: An illustration of objects extracted from segmented region.

introduced in Section 3. Visual features extracted from each region are single object of segmented regions.

For all input image I, I is segmented into 3 by 3 grid. Let us say N is the number of segmented image regions. We extract d-dimensional feature vector F_i from each region r_i. And we define visual generation probability $P_F(\sim \mid I)$. We used Multiple Bernoulli Distribution [12] to calculate visual general probability. U is unlabeled image and F_U is feature vector of U. W_L is subset of tag label. $P(F_U, W_L)$ is similarity between F_U and W_L. The process of jointly generating F_U and W_L is as follows:

(1) select an image I from training set with $P_r(I)$;

(2) obtain segmented image regions;

(3) for each training image I, $i = 1, \ldots, N$;

(4) generate visual descriptions from ith region by using conditional probability;

(5) for each word v is in tag set;

(6) generate tag set by using Multiple Bernoulli Distribution;

(7) using (7), calculate joint probability of visual description and labels in our model:

$$P\left(F_U, W_L\right) = \sum_{I \in r} \left\{ P_r\left(I\right) \times \prod_{i=1}^{N} P_F\left(F_U^i \mid I\right) \times \prod_{v \in W_L} P_V\left(v \mid I\right) \times \prod_{v / \in W_L} \left(1 - P_V\left(v \mid I\right)\right) \right\}. \tag{7}$$

In (7), $P_r(I)$ is the probability of image I from training set. Since there is no prior knowledge, $P_r(I)$ can be assumed to obey uniform distribution:

$$P_r\left(I\right) = \frac{1}{|r|}, \tag{8}$$

where $|r|$ is the size of training image set.

Probability $P_F(\sim \mid I)$ is used to estimate visual generation probability of regions. Assume F_I is visual features of regions in 3 by 3 segmentation; $P_F(\sim \mid I)$ can be calculated as follows:

$$P_F\left(F_U^i \mid I\right) = \frac{1}{N} \sum_{j=1}^{N} \frac{\exp\left\{-\left(F_U^i - F_I^j\right)^T \sum^{-1} \left(F_U^i - F_I^j\right)\right\}}{\sqrt{2^d \pi^d \left|\sum\right|}}, \tag{9}$$

where N is the number of image regions and d is the dimension of visual features. Equation (9) uses Gaussian kernel function to estimate the visual description F_I^j of each region in image I. Gaussian kernel is determined by covariance matrix $\sum = \mu \cdot I$.

$P_V(v \mid I)$ is vth component of Multiple Bernoulli Distribution. It means that probability of tag set W_L which is generated by training image I. Bayesian estimation is used for each tag label as follows:

$$P_V\left(v \mid I\right) = \frac{\varepsilon \cdot \emptyset_{v,I} + L_v}{\varepsilon + |r|}, \tag{10}$$

where L_v is the number of labels v in training set and $|r|$ is the size of training image set. $\emptyset_{v,I}$ is a binary function (if I contains label then 1, else 0). ε is parameter of weight $\emptyset_{v,I}$.

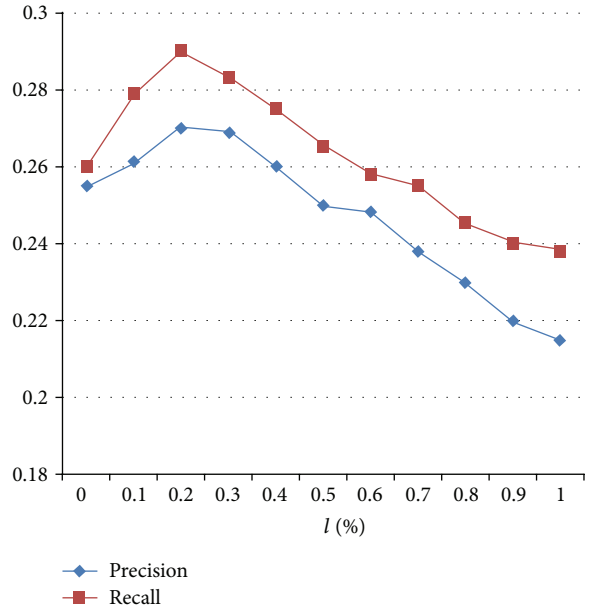

Figure 6: Parameter determination, tag results on l.

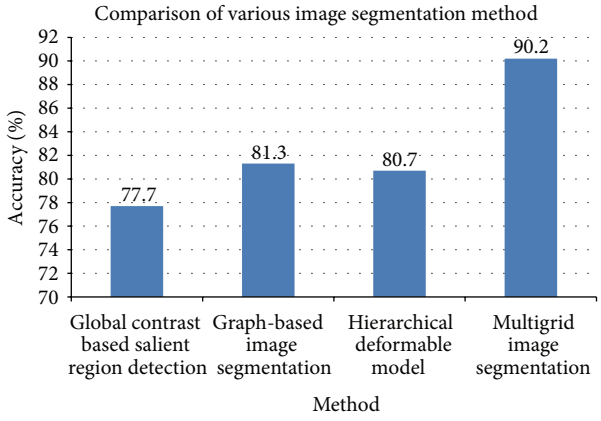

Figure 7: Accuracy comparison result of various image segmentation methods.

5. Experiments

To evaluate our automatic image tagging model based on multigrid image segmentation, we compare our model with other models using Corel dataset. Corel dataset is a popular dataset in automatic image tagging, which includes over 5,000 images. This section focuses on how to construct an effective automatic image tagging model. For the convenience of comparing with other models, we do not use some new visual features. We use the same 30-dimensional features including 9-dimensional RGB color moments, 9-dimensional Lab color moments, and 12-dimensional Gabor texture features. To evaluate other automatic image tagging models, we use precision, recall, and F-measure to evaluate tagging results (Figure 6). In addition, we also count the labels that are correctly tagged at least once, denoted as NZR which reflects the coverage level of annotation words.

Table 1: Experimental results of image segmentation method comparison.

Models	Precision	Recall	F-measure	NZR
Global contrast based salient region detection	0.21	0.12	0.12	57
Graph-based image segmentation	0.26	0.31	0.27	117
Hierarchical deformable model	0.26	0.32	0.28	123
Multigrid image segmentation	0.29	0.32	0.30	151

Table 2: Experimental results.

Models	Precision	Recall	F-measure	NZR
Cross Media Relevance Model	0.10	0.09	0.09	66
Multiple Bernoulli Relevance Model	0.24	0.25	0.24	122
Transductive Multi-Instance Multilabel	0.23	0.27	0.25	130
Supervised Learning Model	0.23	0.29	0.26	137
Our model	0.27	0.29	0.28	144

We need to determine the parameter value of l in (4) according to experiments. l is the number of similar bins in histogram. Horizontal coordinate axis means the ratio of similar color bins in histogram (Figure 4). We can find out optimal parameter value for best annotation results when the ratio of similar color bins is 20%. Precisions and recalls of annotation results would fall if l increases, because higher l will reduce the region contrast, and then feature extracted image regions would be affected to some degrees.

To improve performance of our multigrid image segmentation method, we compared with current methods introduced in Section 2 which are global contrast based salient region detection method [11], graph-based image segmentation [12], and hierarchical deformable model [13]. Experimental result demonstrates accuracy for each method and our method shows best performance compared to other methods (Figure 7).

Meanwhile, we evaluate our model with other image segmentation methods with precision, recall, F-measure, and NZR (Table 1).

Finally, we present our automatic image tagging model performance. We compared our model with some state-of-the-art models, including Cross Media Relevance Model [2, 3], Multiple Bernoulli Relevance Model [1], Transductive Multi-Instance Multilabel [14], and Supervised Learning Model [4]. We can find that our model is very effective and tagging results are better than those state-of-the-art models (Table 2). Our model obtained the highest precision 0.27 which is at least 12% higher than other models. Recall achieves 0.29 which is the same as Supervised Learning Model and the recall is higher than other models obviously.

Table 3: Tag results.

Images	Ground truth	Multiple Bernoulli Relevance Model	Our model
	Beach palm people tree	*Cloud* stone ruins pyramid *sky*	**Palm beach** *cloud water sky*
	Blooms cactus flower needle	*Grass* **flower** *petal* water sky	**Flower** *grass petal* **bloom** *sand*
	Flower garden house tree	Water bird nest *sky* grass	**Tree** *sky* **house** *grass* stone
	Bird nest tree	Flower *leaf plant* **tree** sky	**Bird tree** *leaf plant grass*

F-measure of our model achieves 0.28, and it is approximately 8% higher than Supervised Learning Model which obtained the highest F-measure in previous state-of-the-art models. In addition, in criterion of NZR which reflects the coverage of annotation words, our model reaches 144 and it is also the highest in all models.

We compare our model with MBRM model, and the rankings of tagging labels are sorted in descending order with tagging probability (Table 3). If labels are in ground truths, we use bold type. Here, we do not select test images that are perfectly tagged by our model. We can easily find that tagging results of our model show better performance than MBRM model. In addition, we also find that some tag words do not appear in ground truth annotations of the dataset, but some of these words can also describe the contents of images. That is, some correct tags are ignored by users. These labels are in italic type. For example, *clouds*, *water*, and *sky* do not belong to the ground truth in first image, but these labels can be used to describe the contents of first image without question. Besides, some labels in other images also have the similar situations.

6. Conclusions

In this paper, we proposed multigrid image segmentation method. And then we also proposed an automatic image tagging model based on our multigrid image segmentation method. Since segmented image may contain multiple objects, we proposed multigrid image segmentation method. Our model presented high performance compared to other image segmentation methods. With experimental results on automatic image tagging models, our image tagging model showed better performance compared to other state-of-the-art models especially on object feature extraction. To evaluate our proposed multigrid image segmentation method, we compared with other image segmentation methods and to evaluate our automatic image tagging model, we used Corel dataset and compared with other famous models: Cross Media Relevance Model, Multiple Bernoulli Relevance Model, Transductive Multi-Instance Multilabel, and Supervised Learning Model. Our model showed efficient, effective, and accurate performance in all evaluated functions precision, recall, F-measure, and NZR.

Since there are limitations and many works to do, and multimedia data is growing even in this moment, more powerful, reliable, and accurate models must be prevented. With large amount of data being created every moment, we also need to focus on real time automatic annotation model.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] S. L. Feng, R. Manmatha, and V. Lavrenko, "Multiple Bernoulli relevance models for image and video annotation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1002–1009, 2004.

[2] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proceedings of the 26th Annual International ACM SIGIR*, pp. 119–126, 2003.

[3] Y. Yang, Z. Huang, and Z. Ma, "Robust cross-media transfer for visual event detection," in *ACM Multimedia'12*, pp. 1045–1048, 2012.

[4] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 394–410, 2007.

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

[6] S. Lindstaedt, R. Mörzinger, R. Sorschag, V. Pammer, and G. Thallinger, "Automatic image annotation using visual content and folksonomies," *Multimedia Tools and Applications*, vol. 42, no. 1, pp. 97–113, 2009.

[7] H. T. Manh and G. Lee, "Small object segmentation based on visual saliency in natural images," *Journal of Information Processing Systems*, vol. 9, no. 4, pp. 592–601, 2013.

[8] U. J. Divya, K. Hyunseoul, L. Jun, and K. Jee-In, "Fractal based method on hardware acceleration for natural environments," *Journal of Convergence*, vol. 4, no. 3, pp. 6–12, 2013.

[9] K. V. Santosh and K. N. Shyam, "Color directional local quinary patterns for content based indexing and retrieval," *Human-Centric Computing and Information Sciences*, vol. 4, no. 6, 2014.

[10] P. B. Patil and M. B. Kokare, "Interactive semantic image retrieval," *Journal of Information Processing Systems*, vol. 9, no. 3, pp. 349–364, 2013.

[11] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '11)*, pp. 409–416, June 2011.

[12] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[13] Y. Xiong, P. Gang, C. Zhaoquan, and Z. Kehan, "Occluded and low resolution face detection with hierarchical deformable model," *Journal of Convergence*, vol. 4, no. 2, pp. 11–14, 2013.

[14] 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, pp. 661–670, 2010.