Image Topic Discovery with Saliency Detection

Zhidong Li[1 2]
Zhidong.Li@nicta.com.au
Yang Wang[1 2]
Yang.Wang@nicta.com.au
Jing Chen[2]
Jng.Chen@ieee.org
Jie Xu[1 2]
Jxu@nicta.com.au
John Larid[1 2]
John.Larid@nicta.com.au

[1] The National ICT, AUSTRALIA

[2] The School of Computer Science and Engineering, University of the New South Wales, AUSTRALIA

Abstract

This work proposes a biologically inspired approach to integrate latent topic model with saliency detection. Firstly, a saliency detection algorithm is presented to discriminate salient objects from background parts in the image. A hierarchical latent topic model is proposed to discover image topics by combining subtopics of both salient objects and background parts. We test the algorithm on public image datasets for saliency detection and image categorization. The experimental results show that the proposed approach robustly detects salient objects and categorizes image data, and it outperforms state-of-the-art methods for both saliency detection and unsupervised topic modelling.

1. Introduction

Image topic discovery is a challenging task in computer vision, which is important for content understanding, image retrieval, and event detection. Early work only used low-level features to present an image and measure the similarity between images [1]. In recent years, image categorization by latent topic discovery models [2, 3], which are originally proposed for text classification, has gained considerable attention. Vision researchers [4, 5, 6, 7, 8] also proposed bag-of-words (BoW) approach, in which "visual words" are taken as atomic units to present an image, for visual latent topic discovery. Bosch et al. [9] investigated different formations of visual words over a number of features and showed that the dense sampled visual words over SIFT [10] feature gives the best performance. Li et al. [11] proposed a topic discovery model based on Latent Dirichlet Allocation (LDA) to semantically categorize the natural scene images. Sivic et al. [12] presented unsupervised image categorization by both probabilistic Latent Semantic Analysis (pLSA) and LDA. In addition, spatial information in images can also be employed for topic discovery. For example, Liu and Chen [8] used the location of patches within the same topic to improve image categorization. Sivic et al. [13] used visual word pairs to indicate spatial relationship

BMVC 2010 doi:10.5244/C.24.33

between the visual words. Wang and Grimson [14] used regular grids to group neighbouring pixels of the same object. Cao and Li [5] incorporated spatial coherency of visual words within over-segmented homogenous image regions. Furthermore, Sudderth et al. [6] proposed to combine scene, objects and their parts to discover the image topic. Given annotated image parts, Li and Li [4] proposed a model of event classification by integrating object and scene information. The model was further extended in [15] to simultaneously perform segmentation, annotation and classification.

On the other hand, instead of reading word by word in a document, human is usually attracted by salient objects in the scene before noticing the other parts. In practice, a person can rapidly categorize the images into different classes based on the salient information captured with the prior knowledge [16] or merely through low-level visual attention. Inspired by the psychophysics study of human vision system, the ability of detecting salient objects has been implemented with various computational models. Based on the salient point detection method [17], Walther and Koch [18] introduced a biologically plausible model to detect the salient proto-objects, which can be used to describe the existing prominent object in the salient region. Harel et al. [19] proposed a graph based saliency detection approach. Liu et al. [20] proposed a method to detect the salient object at pixel level, object level, and global level. The method requires supervised learning to determine the weighting of the three levels. Hou and Zhang [21] proposed a fast detection approach based on the spectral residual (SR) of Fourier Transform of an image. The residual information is calculated from the difference between the original image signal and a smooth one in the log amplitude spectrum. Furthermore, Guo et al. [22] suggested that the phase spectrum of Fourier transform alone (PFT) is enough to obtain the salient map. In this paper, we introduce a biologically inspired approach that combines latent topic model with saliency detection to categorize image dataset. First, based on PFT approach and Grab-Cut [23], a saliency detection algorithm is proposed to discriminate salient objects from background parts in the image. Second, a hierarchical generative model is presented to discover image topics by considering subtopics of both salient objects and background parts in the image.

Related Work: [21, 22] proposed to detect salient objects based on the spectral residual and the phase of Fourier transform. However, the methods can only detect small objects or boundary areas since they focus on the high frequency components of an image. In addition, [4, 6] proposed models of event classification by integrating object and scene information in an image. However, these supervised methods requires either manual labelling of training images or annotated web image databases [15, 24]. [5] also proposed a spatially coherent latent topic model for both object segmentation and image categorization with unsupervised learning. In this work, we propose an unsupervised approach to integrate latent topic model with saliency detection. A hierarchical model is proposed to learn image topics by combining subtopics of both salient objects and background parts.

The paper is organized as following. Section 2 presents our algorithm of saliency detection and introduces the hierarchical topic model together with its learning algorithm. Experimental results are discussed in section 3. The paper is then concluded in section 4.

2. Image Discovery Model with Saliency Detection

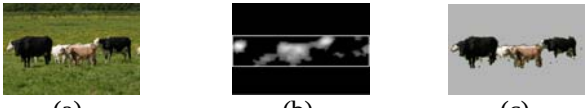
2.1. Saliency Detection

Given an image I, we first compute its saliency map SM by Fourier transform F of the image and inverse Fourier transform F^{-1} [22]:

$$SM = g(I) * \left\| F^{-1}[e^{i \cdot P(F(I))}] \right\| \quad (1)$$

where $g(\cdot)$ is a 2D Gaussian filter with kernel size $\sigma = 8$ and $P(\cdot)$ is the image phase spectrum. The regions with saliency values higher than a threshold (0.75 times the maximum saliency value in the saliency map) are detected as salient objects [22], and the other areas become background parts.

However, the result of the PFT based saliency detection captures the high frequency parts of an image, so that only small parts or boundary areas of salient objects are detected (see Figure 2b). The detected salient regions usually cannot represent the entire object. Walther and Koch [18] also pointed out that how to spread the salient activation over the salient objects was still an open issue. To enhance the result of saliency detection, we handle the saliency activation spreading by Grab-Cut [23]. Grab-Cut is a supervised approach to segment objects, which requires the user to input the rectangle area of foreground. To automatically detect the salient objects, we set the rectangle area as the bounding box of the detected salient regions (see figure 2b) instead of manual input. Figure 2c shows the detection result by the proposed salient detection algorithm.



(a) (b) (c)

Figure 1: (a) One example image. (b) Result of saliency detection based on PFT approach. (c) Salient object detection by the proposed method.

2.2. Topic Discovery

To integrate latent topic discovery with saliency detection, a hierarchical generative model (see Figure 2) is built. In this work, the model performs image topic discovery by combining subtopics of both salient objects and background parts in the image.

2.2.1. Model Description

Given a dataset of M images $\{I_1, I_2, ..., I_M\}$, we use dense uniform sampling patches [4, 9] to obtain the visual words in each image. For each patch, a SIFT feature vector of 128 dimensions are used to describe it. Then the patches are separated into two parts after saliency detection. Two codebooks, one of V_s visual words $W^s = \{w_1^s, w_2^s, ..., w_{V_s}^s\}$ from salient objects and the other of V_b visual words $W^b = \{w_1^b, w_2^b, ..., w_{V_b}^b\}$ from background regions are formed through the K-means algorithm. For each image I_m, there are N_m visual words observed. Here $w = (w_{11}, w_{12}, ..., w_{MN_M})$ represents all the observed words in the image set, and $w_{mn} \in W^s \cup W^b$ is the n^{th} observed word in the m^{th} image.

1. There are T different latent topics $\{t_1, t_2, ..., t_T\}$ in the image dataset. Each topic t_i is generated from a multinomial distribution with parameter θ_i, which is sampled from a

Dirichlet prior with hyper-parameter α.

2. In each image I_m, the N_m observed words consist of N_m^s saliency words and N_m^b background words. Each background word w_j^b is associated with one background subtopic z_k^b. Meanwhile each saliency word w_j^s is associated with a saliency subtopic z_k^s. There are respectively T_s saliency subtopics $Z^s = \{z_1^s, z_2^s, ..., z_{T_s}^s\}$ and T_b background subtopics $Z^b = \{z_1^b, z_2^b, ..., z_{T_s}^b\}$. Here $z = (z_{11}, z_{12}, ..., z_{MN_M})$ represent the subtopics for all the observed words w in the image dataset, and $z_{mn} \in Z^s \bigcup Z^b$ is the subtopic for word w_{mn}.

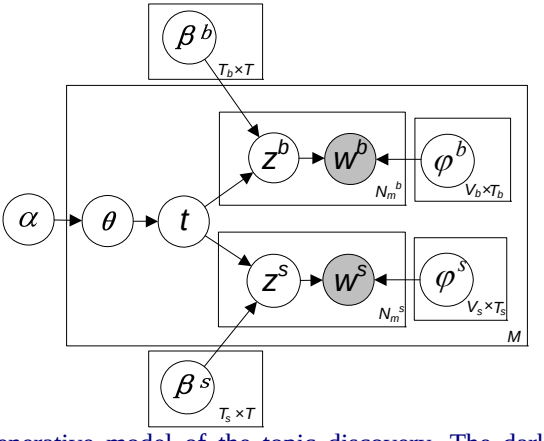


Figure 2: Proposed generative model of the topic discovery. The dark nodes represent observed data and the light nodes represent latent variables and parameters (see the text for details).

3. The saliency subtopics are generated according to the parameter matrix $\beta^s = (\beta_{k,i}^s)$, where each element $\beta_{k,i}^s = p(z_k^s \mid t_i)$. On the other hand, the saliency words are generated according to the parameter matrix $\varphi^s = (\varphi_{j,k}^s)$ with $\varphi_{j,k}^s = p(w_j^s \mid z_k^s)$.
4. The background subtopics are generated according to the parameter matrix $\beta^b = (\beta_{k,i}^b)$, where each element $\beta_{k,i}^b = p(z_k^b \mid t_i)$. The background words are generated according to the parameter matrix $\varphi^b = (\varphi_{j,k}^b)$ with $\varphi_{j,k}^b = p(w_j^b \mid z_k^b)$.

2.2.2. Model Learning

The learning process is to estimate the parameters of the generative model, and then infer the latent variables using the estimated parameters. Given the proposed model in figure 2, the model parameters can be estimated by maximizing the log likelihood of the observed data. For example, EM algorithm could be applied for likelihood maximization [3, 7]. However, in general it is intractable to compute according to [2]. In this work, we estimate the model parameters by iterating the following two steps. First, given the topic label estimation, the parameters β^b, β^s, φ^b and φ^s are optimized and the subtopics are inferred; then the assignment of topic labels are optimally approximated given the inferred subtopics. These two steps are performed alternatively.

In the first step, given the estimated subtopics z, the posterior probability of latent topics $t = (t_{11}, t_{12}, ..., t_{MN_M})$ becomes:

$$p(t \mid z) = \frac{p(t,z)}{\sum_{t} p(t,z)}, \tag{2}$$

where

$$p(t,z) = \int d\theta Dir(\theta \mid \alpha) \prod_{m,n} p(t_{mn} \mid \theta) p(z_{mn} \mid t_{mn}). \tag{3}$$

Without prior information, each of the subtopic z are randomly initialized. t_{mn} is the topic label for visual word w_{mn}. Suggested by Griffiths and Steyvers in [25], the topic inference for each word can be optimally approximated by Collapsed Gibbs Sampling. The topic label can be approximated as

$$p(t_{mn} = t_i \mid t_{-mn}, z, \alpha) \propto \begin{cases} \dfrac{n^{t_i}_{-mn,z_{mn}}}{\sum_k n^{t_i}_{-mn,z^b_k}} \cdot \dfrac{n^m_{-mn,t_i} + \alpha_{t_i}}{\sum_{i'} n^m_{-mn,t_{i'}} + \alpha_{t_{i'}}}, & if\ w_{mn} \in W^b \\ \dfrac{n^{t_i}_{-mn,z_{mn}}}{\sum_k n^{t_i}_{-mn,z^s_k}} \cdot \dfrac{n^m_{-mn,t_i} + \alpha_{t_i}}{\sum_{i'} n^m_{-mn,t_{i'}} + \alpha_{t_{i'}}}, & otherwise \end{cases}, \tag{4}$$

where t_{-mn} is the set of topic labels for all the words except w_{mn}, $n^t_{-mn,z}$ is the number of words in the image dataset with subtopic z assigned to topic t excluding w_{mn}, and $n^m_{-mn,t}$ is the number of words in the m^{th} image assigned to topic t excluding w_{mn}. The sampling process will optimally assign a topic label for each word in same subtopic,

$$t_{mn} = \arg\max_{t_i} p(t_{mn} = t_i \mid t_{-mn}, z, \alpha). \tag{5}$$

In the second step, given the estimated latent topics t, let $n^{t_i}_{w^b_j}$ be the number of occurrences of the background word w^b_j in the latent topic t_i, and all of them constitute the co-occurrence matrix X^b=($n^{t_i}_{w^b_j}$), The log likelihood of X^b can be written as

$$L(X^b \mid \beta^b, \varphi^b) = \sum_i \sum_j n^{t_i}_{w^b_j} \log p(t_i, w^b_j), \tag{6}$$

where

$$p(t_i, w^b_j) = p(t_i) \sum_k p(z^b_k \mid t_i) p(w^b_j \mid z^b_k). \tag{7}$$

Thus the estimation of the parameter matrix $\beta^b = (p(z^b_k \mid t_i))$ and $\varphi^b = (p(w^b_j \mid z^b_k))$ can be obtained by maximizing the log likelihood using the expectation-maximization (EM) algorithm. Similar to [7], the E step is equivalent to computing the posterior probability $p(z^b_k \mid t_i, w^b_j)$ given the estimated $p(z^b_k \mid t_i)$ and $p(w^b_j \mid z^b_k)$,

$$p(z^b_k \mid t_i, w^b_j) = \frac{p(z^b_k \mid t_i) p(w^b_j \mid z^b_k)}{\sum_{k'} p(z^b_{k'} \mid t_i) p(w^b_j \mid z^b_{k'})}, \tag{8}$$

the parameters $p(z_k^b \mid t_i)$ and $p(w_j^b \mid z_k^b)$ are randomly initialized. Meanwhile the M step corresponds to the following updates, given the computed $p(z_k^b \mid t_i, w_j^b)$ from the E step.

$$p(w_j^b \mid z_k^b) = \frac{\sum_i n_{w_j^b}^{t_i} p(z_k^b \mid t_i, w_j^b)}{\sum_{j'} \sum_i n_{w_{j'}^b}^{t_i} p(z_k^b \mid t_i, w_{j'}^b)}, \tag{9}$$

and

$$p(z_k^b \mid t_i) = \frac{\sum_j n_{w_j^b}^{t_i} p(z_k^b \mid t_i, w_j^b)}{\sum_{k'} \sum_j n_{w_j^b}^{t_i} p(z_{k'}^b \mid t_i, w_j^b)}, \tag{10}$$

Similarly, the EM algorithm can be employed to estimate the parameters $\beta^s = (p(z_k^s \mid t_i))$ and $\varphi^s = (p(w_j^s \mid z_k^s))$. In the E step, given the estimated $p(z_k^s \mid t_i)$ and $p(w_j^s \mid z_k^s)$,

$$p(z_k^s \mid t_i, w_j^s) = \frac{p(z_k^s \mid t_i,) p(w_j^s \mid z_k^s)}{\sum_{k'} p(z_{k'}^s \mid t_i,) p(w_j^s \mid z_{k'}^s)}. \tag{11}$$

In the M step, given the computed $p(z_k^b \mid t_i, w_j^b)$ from the E step,

$$p(w_j^s \mid z_k^s) = \frac{\sum_i n_{w_j^s}^{t_i} p(z_k^s \mid t_i, w_j^s)}{\sum_{j'} \sum_i n_{w_{j'}^s}^{t_i} p(z_k^s \mid t_i, w_{j'}^s)}, \tag{12}$$

$$p(z_k^s \mid t_i) = \frac{\sum_j n_{w_j^s}^{t_i} p(z_k^s \mid t_i, w_j^s)}{\sum_{k'} \sum_j n_{w_j^s}^{t_i} p(z_{k'}^s \mid t_i, w_j^s)}. \tag{13}$$

Then for each visual word w_{mn}, the corresponding subtopic is estimated as

$$z_{mn} = \begin{cases} \arg\max_{z_k^b} p(z_k^b \mid t_{mn}, w_{mn}), \text{if } w_{mn} \text{ is in background} \\ \arg\max_{z_k^s} p(z_k^s \mid t_{mn}, w_{mn}), \text{otherwise} \end{cases}. \tag{14}$$

The topic of each image I_m is estimated as:

$$t_{I_m} = \arg\max_{t_i} n_{I_m}^{t_i}, \tag{15}$$

where $n_{I_m}^t$ is the number of words assigned with topic t in an image I_m.

3. Experiments

We evaluate our techniques with public image datasets for both salient object detection and image categorization. The experiments are conducted with MATLAB implementation on a 2 GHz Pentium 4 machine with 2 GB of RAM. The following parts present the experimental results.

For saliency detection, the algorithm is tested on the MSRC image database [26]. The image database contains 20 categories of totally 591 images. The main attentive objects in these images are manually labelled. We compared our method with the spectral-residual based saliency detection [21] (SR) and frequency-tuned saliency detection [27] (FT). Some sample images and detection results are shown in Figure 3. It can be seen that the spectral information only helps capture high frequency components of the image. The FT method also requires image segmentation. With the help of Grab-Cut, the proposed approach accurately detects the entire salient objects. Table 1 exhibits the average accuracy of saliency detection by different methods. The detection rate is measured by the percentage of pixels in agreement with the ground truth.

	Proposed	FT	SR
Accuracy	73.5%	67.6%	63.2%

Table 1: The overall accuracy of saliency detection.

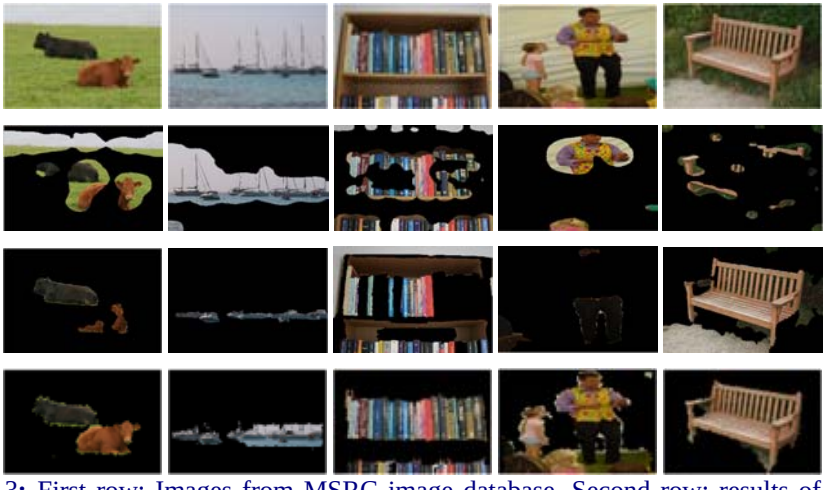
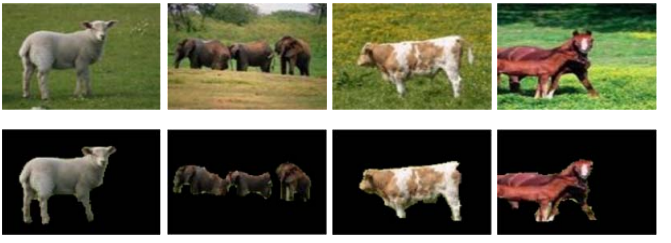


Figure 3: First row: Images from MSRC image database. Second row: results of salient object detection by SR. Third row: results of salient object detection by FT. Fourth row: results of salient object detection by the proposed method.

For image categorization, we simply assume $T = T_b = T_s$ in the experiments. The algorithm is first tested with an image dataset of four object categories based on MSRC image database and Corel image database: horse, cow, sheep, and elephant. Figure 4 shows some example images together with the detected salient objects. We compare our approach with the baseline method of LDA categorization [12]. The overall categorization results of both methods are shown in figure 4. In the experiment, our method consistently

outperforms the LDA based categorization. Since the dataset consists of images with similar background, the information from background may degrade the performance of topic discovery by LDA. However, the problem is alleviated by the proposed approach through discriminating salient objects from background parts in the image.



	Accuracy
Proposed	56.30%
LDA	51.37%

Figure 4: *Left*: Example images and the detected salient objects (sheep, elephant, cow and horse); *Right*: The overall accuracy of object categorization.

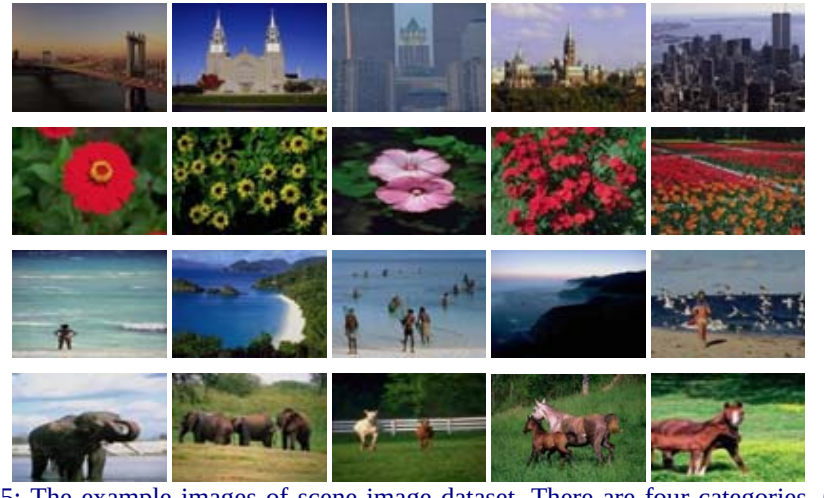
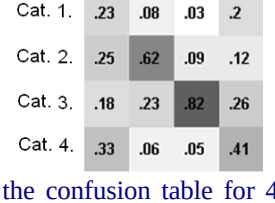


Figure 5: The example images of scene image dataset. There are four categories. Cat. 1: city of New-York and Ottawa (first column); Cat. 2: perennial and us_garden (second column); Cat. 3: beaches and Cal_sea (third column); Cat. 4: horse and elephant (fourth column).



	Accuracy
Proposed	52.25%
Spatial_LTM	48.25%

Figure 6: *Left*: the confusion table for 4-way categorization of the scene image dataset. Rows are the categorised performance while columns present the ground truth classes; *Right*: the overall categorization performance.

In the previous image set, the images labels are assigned according to the foreground objects. However, in practice images with different kinds of objects may belong to the same category and image topics may depend on both/either background and/or foreground

information. To further investigate this aspect with our model, we compose a scene image dataset in which images are labelled according to background and foreground information. Eight classes of images are selected from the Corel image database: horse, elephant, city of New-York, city of Ottawa, beaches, Cal_sea, perennial, us_garden (100 images of each class). The images are then divided into four categories (city of New-York and Ottawa; perennial and us_garden; beaches and Cal_sea; horse and elephant) by merging similar classes. Some example images are shown in figure 5. Our approach is compared with an enhanced latent topic model with spatial constraints (*Spatial-LTM*) [5]. Figure 6 shows the overall performance and the confusion tables by both methods. The proposed approach further improves the categorization accuracy through integrating latent topic model with saliency detection.

Bocce	.18	.02	.03	.04	.12	.25	.21	.16
Badminton	.17	.28	.03	.01	.05	.06	.2	.21
Rowing	.06	.05	.36	.21	.09	.05	.09	.1
Rock climbing	.06	.1	.34	.19	.08	.06	.04	.14
Snowboarding	.06	.05	.04	.05	.24	.19	.25	.13
Croquet	.28	.01	0	.04	.17	.42	.04	.05
Polo	.07	.09	.04	.04	.09	.1	.45	.13
Sailing	.21	.07	.01	.01	.01	.18	.16	.36

	Accuracy
Proposed	30.69%
Spatial_LTM	27.60%

Figure7: *Left*: the confusion table for 8-way categorization of the event image dataset. Rows are the categorized performance while columns present the ground truth classes; *Right*: the overall categorization performance.

We further test our method on the event image database [4], which contains 8 categories of events: badminton, bocce, croquet, polo, rock climbing, rowing, sailing, and snowboarding. Each of these events is defined by the presented athlete and corresponding background. 100 randomly selected images from each event are used to test our algorithm. Event classification is a challenging task since athletes may pose similarly in different games and background scenes are cluttered. Classes such as croquet and bocce are quite similar and hard to be distinguished even by human. According to [15], LDA based categorization with supervised learning achieved 36% overall categorization accuracy. The results of both the proposed approach and aforementioned Spatial-LTM are illustrated in Figure 7 together with the confusion table.

4. Conclusion

In this work we have proposed an unsupervised approach to integrate latent topic model with saliency detection. In our framework, salient objects are discriminated from background parts through enhanced saliency detection. Then the image topics are discovered by combining information from both salient objects and background parts. The experimental results on public image sets demonstrate the effectiveness of the approach for both saliency detection and image categorization. In the future, we will extend our method to video analysis by incorporating the spatiotemporal information into the model.

5. Acknowledgement

National ICT Australia (NICTA) is funded by the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council.

References

[1] A. Vailaya, M.A.T. Figueiredo, A.K. Jain and H.J. Zhang. Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10(1):117-130, 2001.

[2] D.M. Blei, A.Y. Ng and M.I. Jordan., Latent dirichlet allocation. *Journal of Machine Learning Research*, (3):993-1022, 2003

[3] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. SIGIR*, 1999.

[4] L.J. Li and F.F. Li, What, where and who? classifying events by scene and object recognition. In *Proc. ICCV*, 2007.

[5] L. Cao and F.F. Li. Spatially coherent latent topic model for concurrent object segmentation and classification. In *Proc. ICCV*, 2007.

[6] E.B. Sudderth, A. Torralba, W.T. Freeman and A.S. Willsky, Learning hierarchical models of scenes, objects, and parts. In *Proc. CVPR*, 2007.

[7] J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman and W.T. Freeman, Discovering objects and their location in images. In *Proc. ICCV*, 2005.

[8] D. Liu and T. Chen. Unsupervised image categorization and object localization using topic models and correspondences between images. In *Proc. ICCV,* 2007.

[9] A. Bosch, A. Zisserman and X. Munoz, Scene classification via pLSA, In *Proc. ECCV,* 2006.

[10]D.G. Lowe, Distinctive Image Features from scale-Invariant key-points. *International Journal of Computer Vision*, 2003.

[11]F.F. Li and P. Perona, A bayesian hierarchical model for learning natural scene categories, In *Proc. CVPR,* 2005.

[12]J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman and W.T. Freeman. Discovering object categories in image collections. In *Proc. ICCV,* 2005.

[13]J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman and W.T. Freeman, Discovering objects and their location in images. In *Proc. ICCV*, 2005.

[14]X. Wang and E. Grimson. Spatial Latent Dirichlet Allocation. In *Proc. NIPS*, 2007.

[15]L.J. Li, R. Socher and F.F. Li, Towards total scene understanding: classification, annotation and segmentation in an automatic framework, In *Proc. ICCV*, 2009.

[16]P.V. Marius, F.F. Li and K. Sabine, Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature,* 460:94-97, 2009.

[17]L. Itti, C. Koch and E. Niebur, A model of saliency-based visual attention for rapid scene analysis. *IEEE T-PAMI*, 20(11):1254-1259, 1998.

[18]D. Walther and C. Koch, Modelling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407, 2006.

[19]J. Harel, C. Koch and P. Perona, Graph-Based visual saliency. *Advances in Neural Information Processing Systems*, 2007.

[20]T. Liu, J. Sun, N.N. Zheng, X. Tang and H.Y. Shum. Learning to detect a salient object. In *Proc. CVPR*, 2005.

[21]X. Hou and L. Zhang. Saliency Detection: A spectral residual approach. In *Proc. CVPR*, 2007.

[22]C. Guo, Q. Ma and L. Zhang. Spatio-Temporal saliency detection using phase spectrum of quaternion fourier transform. In *Proc. CVPR*, 2008.

[23]C.R.V. Kolmogorov and A. Blake, Grab-Cut interactive foreground extraction using iterated graph cuts. in *Proc. ACM Siggraph*, 2004.

[24]R. Fergus, F.F. Li, P. Perona and A. Zisserman. Learning object categories from Google's Image Search. In *Proc. CVPR*, 2005.

[25]T. Griffiths and M. Steyvers. Finding scientific topics. in *Proc. National Academy of Sciences*. 2004.

[26]J. Winn, A. Criminisi and T. Minka. Object categorization by learned universal visual dictionary. In *Proc. ICCV*, 2005.

[27]R. Achantay, S. Hemamiz, F. Estraday and S. Süsstrunk, Frequency-Tuned salient region detection. In *Proc. ICCV*, 2009.