International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.9, No.2 (2016), pp.247-270
http://dx.doi.org/10.14257/ijsip.2016.9.2.22

Image Holistic Scene Understanding Based on Global Contextual Features and Bayesian Topic Model

Lin Li

Institute of Intelligent Computing and Information Technology, Chengdu Normal University, Chengdu, 611130, China
lilin200909@gmail.com

Abstract

Image holistic scene understanding based on global contextual features and Bayesian topic model is proposed. The model integrates three basic subtasks: the scene classification, image annotation and semantic segmentation. The model takes full advantage of global feature information in two aspects. On the one side, the performance of image scene classification and image annotation are boosted by incorporating image global contextual features; On the other side, the performance of image semantic segmentation is also boosted by new superpixel region segmentation method and new superpixel regions and patch feature representation. 1) For image scene classification and image annotation: (1) We improve the feature engineering methods by using the PHOW proposed by Vedaldi [1]; (2) Furthermore, global contextual features are learned by semantic features. 2) For semantic segmentation: (1) We improve the super-pixel segmentation method by using UCM in the literature [2]; (2)We proposed new feature representation for super-pixel region and patches by incorporating DSIFT, texton filter banks, RGB color, HOG, LBP and location features. The experiments testify that model performance has raised on all three sub-tasks.

Keywords: *Image holistic scene understanding, Scene understanding, Global contextual features, Bayesian topic model, Probabilistic graphical models*

1. Introduction

We live in a world filled with contextual information, we are always embedded in some contextual information associated with understanding when we identify a particular object [3]. The scientific community has confirmed the existence of such contextual relationship, the current computer vision technology is also being gradually trying to simulate the human cognitive model in various forms of image understanding and vision tasks.

The directed graph model is very convenient to integrate this kind of dependency or related information. There are many very successful models[4, 5, 6, 7, 8, 9, 10, 11]. However, there are still a lot of problems we need to solve for this kind of image understanding model.

First, there are many problems not clear for human brain's cognitive integrality[3], such as the organization of contextual information in the brain, the contribution of contextual information to object recognition, the how of some scenarios of context information storing in the brain, and so on. Furthermore, we need more in-depth studies of computer, biology, medicine and other aspects.

Secondly, although there are many advantages for the graph models to simulate reality in logical reasoning and context information relations, but these advantages are only relative, there are a lot of inaccurate or inappropriate places. Some scenes may require special research consistent with their own characteristics in the directed graph models.

Copyright ⓒ 2016 SERSC

Third, with the development of computer technology, the research will face new opportunities and challenges. For example, formerly a large data processing is very difficult, but now we can solve some of these bottlenecks in the original research question by more high-performance computing platform. Of course, we also need to do related research to deal with the specific situation.

Finally, the directed graph model is a traditional topic, it has many mature principles. However, the application of image understanding based on the directed graph model is relatively new. Especially the research in terms of human cognition from a holistic point of view is still on the startup stage. So it is necessary to explore more models or systems in line with the overall human cognition as close as possible, with the combination of the traditional theory and image holistic scene understanding.

2. Related Work

Contextual information is useful for image understanding, Hoiem and Efros [4] study the contextual characteristics of 3D scenes, propose a model for placing local object detection in the context of the overall 3D scene by modeling the interdependence of objects, surface orientations, and camera viewpoint. The model has two advantages: (1) subtle relationships (such as the object size related to the viewpoint) can be easily represented; and (2) additions and extensions to the model are easy (the direct method requires complete retraining whenever anything changes). To add a new object to this model, one only needs to train a detector for the object and supply the distribution of the object's height in the 3D scene. But the model has a number of basic assumptions and limits: all objects are on the same ground and perpendicular to the ground, camera inclination is small, and within reasonable limits; Camera-dithering is zero or an image can be calibrated; The argument of the camera itself is standard, and so on. So we can see that the actual application conditions of this model are quite harsh.

In order to overcome these limitations, Hoiem and Efros [12] propose another integrated 3D scene understanding system with estimates of surface orientations, occlusion boundaries, objects, camera viewpoint, and relative depth. The model is not the camera projection in the two-dimensional plane, but the context of the relationship between the visual elements of a real 3D scene. The basic feature factors that the model considered are surface characteristics of the image itself, image occlusion and depth estimation, object and camera angle. The model also takes into account the context of interactive features such as the interaction between surface features and object, interactions between surface features and occlusion, interactions between objects and occlusion. Because this method takes into account of various factors, so the recognition performance boosts greatly.

Sudderth *et. al.,* [5] propose hierarchical probabilistic model for the detection and recognition of objects in cluttered, natural scenes. The model is based on a set of parts which describe the expected appearance and position, in an object centered coordinate frame, of features detected by a low-level interest operator. Each object class has its own distribution over these parts, which are shared between objects. The model learns the parameters via a Gibbs sampler which uses the graphical model's structure to analytically average over many parameters. The model integrates two sub-tasks of the detection and identification.

Cao and L. Fei-Fei[13] propose a model named Spatial-LTM(Spatially coherent Latent Topic Model) for simultaneously object segmentation and scene classification. A major drawback of the Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA) models is the assumption that each patch in the image is independently generated by its corresponding latent topic. While such representation provides an efficient computational method, it lacks the power to describe the visually coherent images and scenes. Spatial-LTM has the following advantages: (1) Spatial-LTM provides a unified representation for spatially coherent bag of words topic models; (2) Spatial-LTM can simultaneously segment and classify objects, even in the case of occlusion and multiple instances; and (3) Spatial-LTM can be trained either unsupervised or supervised, as well as when partial object labels are provided.

Tu et al[6] propose a Bayesian framework for parsing images into their constituent visual patterns. The parsing algorithm optimizes the posterior probability and outputs a scenic representation in a "parsing graph", in a spirit similar to parsing sentences in speech and natural language. This computational framework integrates two popular inference approaches – generative (top-down) methods and discriminative (bottom-up) methods. The former formulates the posterior probability in terms of generative models for images defined by likelihood functions and priors. The latter computes discriminative probabilities based on a sequence (cascade) of bottom-up tests/filters.

The model on natural images of complex city scenes shows that image segmentation can be improved by allowing object specific knowledge to disambiguate low-level segmentation cues, and conversely object detection can be improved by using generic visual patterns to explain away shadows and occlusions.

Li Fei-Fei and Perona[7] proposed a framework for considering two different layers of image-related information, and enhance the robustness of image classification. Each conditional random field (CRF) of model layer can capture the condition of the field to observe any interaction. This method takes into account the short-range interactions such as the pixels smooth and long-range interactions such as mutual configuration between objects and regions. The common method can be extended to different areas such as a pixel dimension to the context object recognition. To further enhance the integration of the model, Li Fei-Fei and L.-J. Li[14] integrate semantic information through a combination of the scene and object classification, semantic information. However the training sample purity for this model is strict, the model need strictly divided and labeled images.

Literature [11] is the one of the most comprehensive models for understanding the holistic scene understanding based on Bayesian graph. The model combines three tasks of scene classification, image annotation and semantic segmentation successfully at the same time. However, the model has shortcomings to affect the further performance improvements: 1) Lack of contextual information to reflect the image features. This part of the information is helpful for model classification and annotation; 2) The model is insufficient to mine local features such as super-pixel and image patches features.

We propose a holistic scene model based on global contextual features and Bayesian topic model to explore the contextual information and local image feature characteristics in order to enhance the overall image scene understanding performance.

3. Framework of Image Holistic Scene Understanding Based on Global Contextual Features and Bayesian Topic Model

As shown in Figure 1, image understanding can be hierarchical and top-down. First, the image can be understood as equestrian from the scene level.

Secondly, the extracting image feature information can form the semantic feature space π, and semantic feature space can be further learned to get the contextual feature space ς.

Third, there are two aspects of the visual image information: 1) the class annotation information such as image contains the sky, trees, horses, and saddle. 2) Visual information comprises of: (1) image block information in figure 2. (2) image can be segmented into different regions, which contains the super-pixel segmentation information.

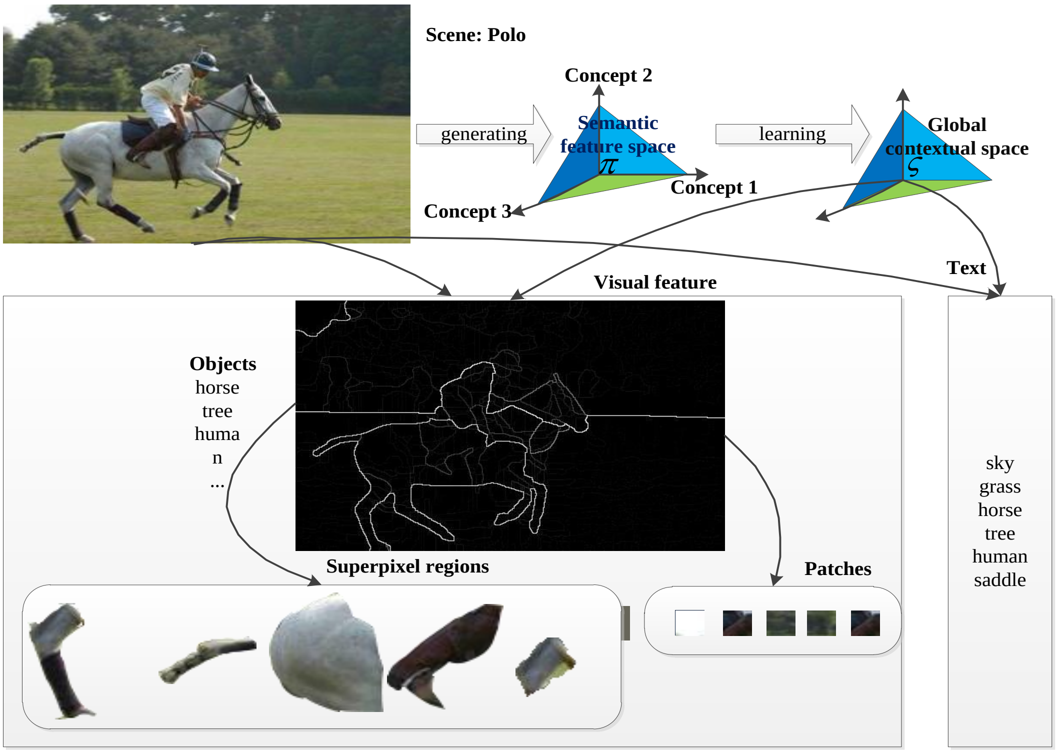


Figure 1. A Diagram of Hierarchical Scene Understanding

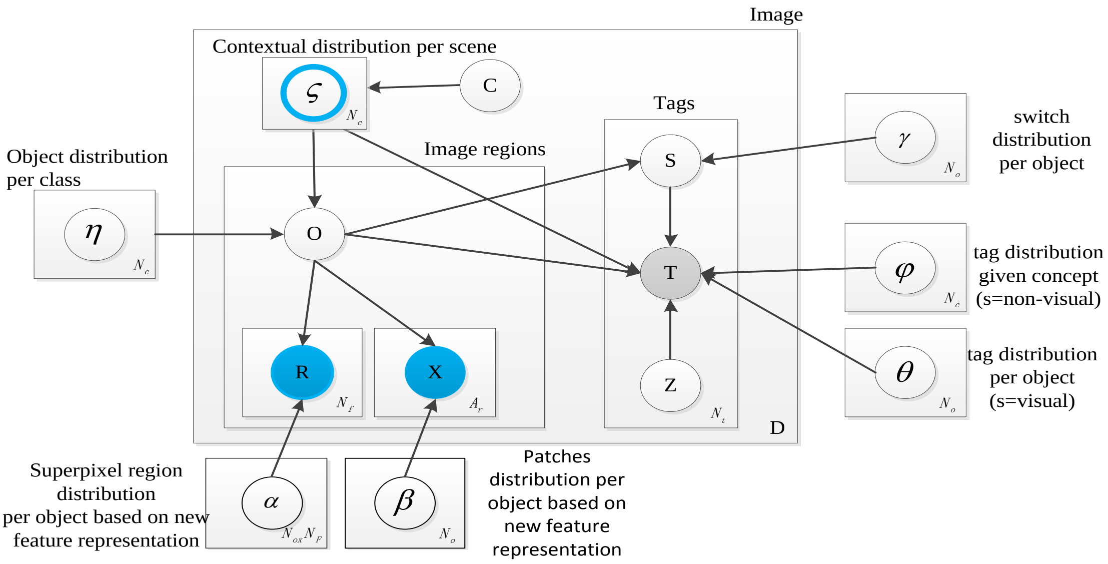


Image
Contextual distribution per scene
ς
N_c
C
Tags
Image regions
Object distribution per class
η
N_c
O
S
T
Z
N_t
R
N_r
X
A_r
D
γ
N_o
switch distribution per object
φ
N_c
tag distribution given concept (s=non-visual)
θ
N_o
tag distribution per object (s=visual)
α
$N_{ox}N_F$
β
N_o
Superpixel region distribution per object based on new feature representation
Patches distribution per object based on new feature representation

Figure 2. Holistic Scene Understanding based on Global Contextual Features and Bayesian Topic Model

Figure 2 is based on the global context information and Bayesian topic model. Comparing with the literature [11], the main difference in our model is shown in the blue zone mark: 1) Adding the contextual feature space ς in order to eliminate the semantic ambiguity; 2) improved regional image information representation in R and X as shown in Figure 2.

Copyright ⓒ 2016 SERSC

3.1. The Generative Model

First, for a given image $d \in D$, we generate semantic feature space by the extracted dense PHOW(Pyramid Histogram Of visual Word) features.

Secondly, we get the context of the feature space ς by learning the semantic feature space.

Third, our model is different with the literature [11], our superpixel edge detection and segmentation is hierarchical image segmentation method[2] proposed by Arbelaez, etc., as shown in Figure 3: The first line (a) is the the original images, the second line (b) is the UCM (Ultrametric Contour Map) image, the third line (c) is an super-pixel segmentation area based on UCM.

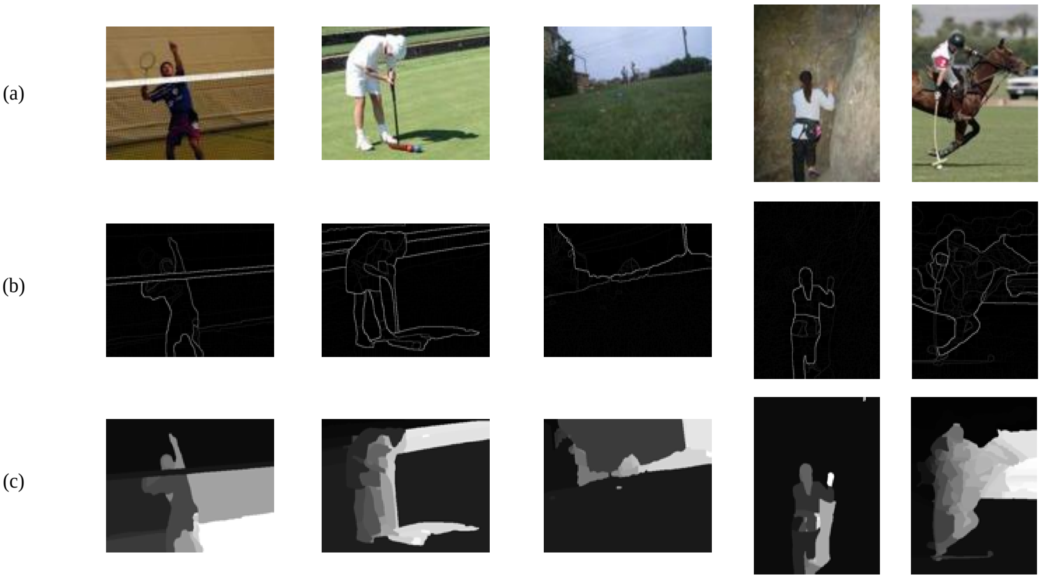


Figure 3. A Diagram of UCM Segmentation

Fourth, representation for each super pixel area and image patches is also different with the literature [11], a new feature extracting method is described in Section 5.3.

Fifth, the variable T represents stage, whichn the training stage which is visible, in order to generate an image and the corresponding annotations,fromlass scene C is sampled by a given prior distribution. Now, given a scenario and the corresponding visual information, we are ready to generate the texture component.

3.2. Generating the Visual Component

Scene contextual space C is a multivariate normal distribution, which is the contextual space got by learning images semantic space.

First, for N_r image area, an object is obtained by sampling of the object under known condi,ns with distribution of $O \propto Multi(\eta_c)$, the appearance of the image is also sampled as the same way:

1) For each $i \in F$, F is regional feature classes described in section 5.3.3, which has shape texture filtering, RGB (Red, Green, Blue) color, HOG (Histogram of Oriented Gradient), LBP (Local Binary Patterns) and location features. Exterior features is sampled through global distribution $R_i \propto Multi(\alpha_i \mid O)$, here each object and each class of feature regional has unique super parameter α_i.

2) The image patches are sampled through the distribution of $X \propto Multi(\beta | O)$, which forms A_r sets.

3.3. Generating the Tag Component

Meanwhile, the regional index variable Z is sampled from a uniform distribution, Z is used to account for the different numbers of tags and regions in this image, as suggested by Blei and Jordan[15].

As mentioned above, the switch variable S allows tags T to correspond to either visually relevant (*i.e.,* the objects) or visually irrelevant (*i.e.,* more abstract information) parts of the scene. This is formulated by allowing tags T to be drawn from either the distribution governed by object O or the one controlled by scene class C. These ideas are summarized in the following generative procedure. For each of the N_t image tags:

1, The index variable sampling: $Z \propto Unif(N_i)$. Z is responsible for connecting the image area and annotation.

1, The switching variable sampling: $S \propto Binomial(\gamma_{oz})$. (a) If $S = non - viusal$, sample a tage: $T \propto Mult(\varphi_c)$. (b) If $S = viusal$, sample a tag $T \propto Mult(\theta_{OZ})$.

4. Composite Model

Combining all the generation process, the resulting joint distribution of the scene class C, contextual distribution ς, object O, the regional R, the joint probability of images, annotation small X, implicit variables T, and Z of distribution becomes:

$$p(C,\varsigma,O,R,X,S,T,Z | \eta,\alpha,\beta,\gamma,\theta,\varphi) = p(C) \times p(\varsigma | C) \times (\prod_{n=1}^{N_r} p(O_n | \eta,\varsigma))$$

$$\times \prod_{n=1}^{N_r}((\prod_{i=1}^{N_F} p(R_{ni} | \mathrm{O}_\mathrm{n}, \alpha_i)) \times \prod_{r=1}^{A_r} p(X_{nr} | O_n, \beta)) \qquad (1)$$

$$\times \prod_{m=1}^{N_t} p(Z_m | N_r) p(S_m | O_{Z_m}, \gamma) p(T_m | O_{Z_m}, S_m, \theta, \varsigma, \varphi)$$

The Model integrates the three tasks in an integrated framework: classification, annotation and segmentation. The equation 1 shows the impact on the scene classification and segmentation tasks.

We get $p(c)$ from learning the scene context C, and generate a hierarchical representation by coupling ς with its objects and region. The overall recognition performance is enhanced through three layers joint modeling. We define a unique distribution $p(O | \varsigma)$ for each object with the context ς of each scene. In addition, the scene C distribution also affects the right contextual marked distribution $p(T | \varsigma)$. This scene class influence serves as a top-down contextual facilitation of the object recognition and annotation tasks.

In addition, the model has a unique feature of simultaneous segmentation, annotation and object identification based on texture and visual models. The model performance is superior to the general BoW (Bag of Words) model proposed by Cao and Li Fei-Fei[13] which contains basic information and images global small area patches.

5. Feature Engineering of Global Contextual Model

Inspired by the literature [16], our study focuses on feature integration issues in the holistic image contextual scene understanding based on Bayesian topic model.

5.1. Semantic Feature Extraction

According to Bayesian decision criteria for classification or annotation tasks, we only need the largest posterior probability. Given an image, we define a posterior probability vector as: $\pi=\left(\pi_1,...,\pi_L\right)^T$, where $\pi_c=P(c|I), c=1,...,L$ is a collection of semantic feature description. π is semantic multivariate normal distribution (Semantic Multinomial, SMN), forming the corresponding semantic space. So, we establish the mapping from the image to the abstract semantic space, π_c □is quite different from the image space χ, with a clear semantic information. The semantic features inherits PHOW many advantages such as: invariation to configuration of the scene, low computational complexity, high-level abstract expression. The surface features such as corners, edge direction, spectrum and other semantic features are widely used in image classification or annotation by calculating their adjacent relationship in semantic space to classify or match. The generation process of semantic feature representation is as shown in Figure 4.

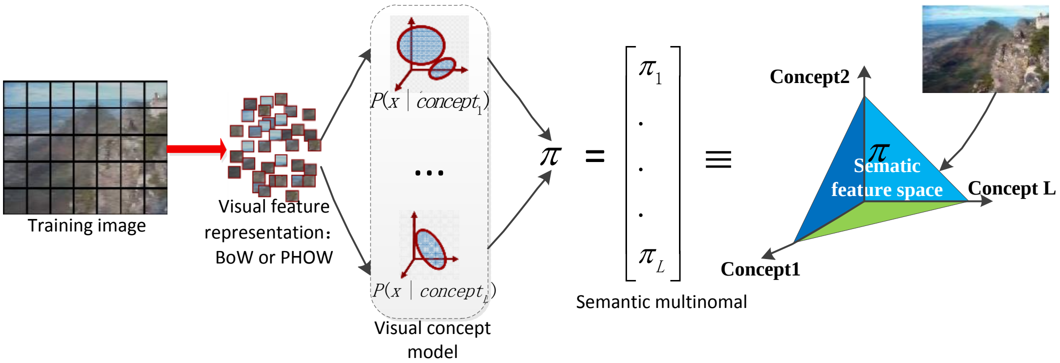


Figure 4. The Generating Process of Semantic Feature Representation

But there are some semantic feature expression issues: A major problem is that the semantic features based on semantic description of the image appearance have two aspects ambiguity: 1) context-free information may have a similar appearance semantic description such as clouds and smoke; 2) semantic feature descriptor can explain the coexistence context, but cannot reveal the contextual dependencies. There are two issues of ambiguity co-occurrence. One is contextual co-occurrence such as the interdependence of image patches got by segmentation can't be explained, the other is ambiguity co-occurrences, for examples: image patches cannot be correctly explained, this may have an ambiguous interpretation with only semantic feature information, we aren't quite sure the fine differences between the street scene and the campus scene.

5.2. Global Contextual Feature

A possible choice to solve the semantic ambiguity is explicitly modeling the contextual dependence by imposing restrictions on physical feature representation, such as star-shaped structure modeling[17] or object-relational modeling[18]. However, this approach will significantly increase the complexity, reduce the invariance of feature representation, and sacrifice the model generalization. A more robust approach is to retain the basic visual features on a higher level of abstraction to represent image features, so small ambiguity between the image patches can be easily detected. The generating process of global contextual feature representation is shown in Figure 5.

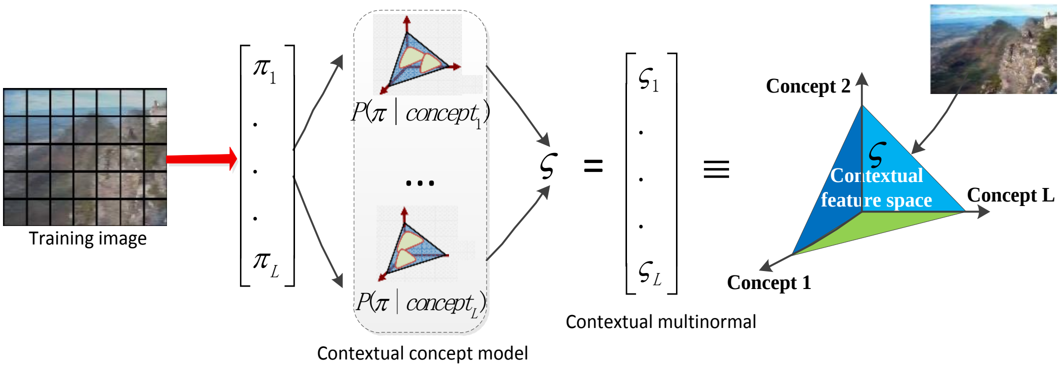


Figure 5. The Generating Process of Global Contextual Feature Representation

5.2.1. From Semantic Feature to Contextual Feature: The basic idea is that the same scene should have similar ambiguous coexistence. Although there may be coexistence between image patches in the street scene and bedroom scenes, but not all of the street scenes and bedroom scene are the same. Typically, ambiguous coexistence is a coincidence, otherwise it becomes a contextual coexistence. Therefore, the coexistence is impossible to detect from a single image, but through all the combined detection of semantic feature sets of scene image.

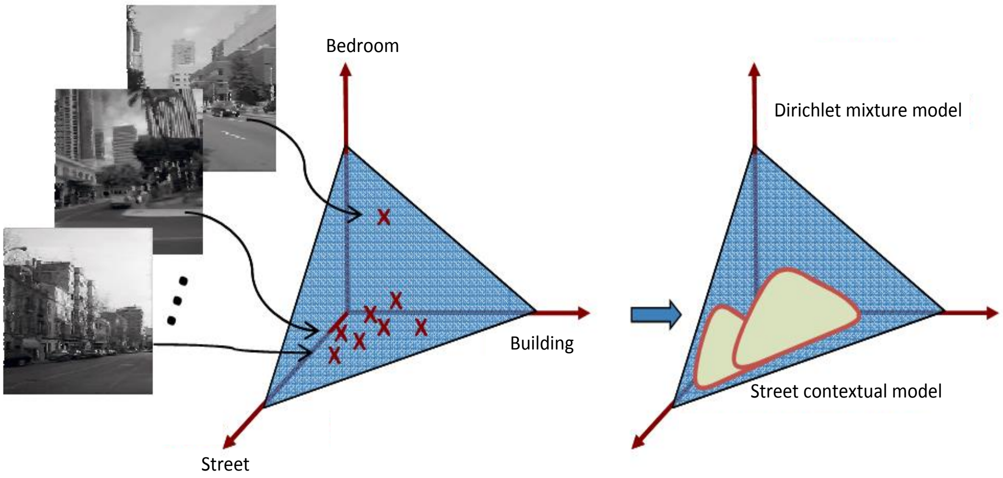


Figure 6. The learning process of contextual model

As shown in Figure 6, by adding a layer of semantic features we can get the contextual features. The scene C is modeling by the distribution probability of the training images in training data set D_c. We define this kind of SMN c distribution as contextual model. If D_c is large enough, and the model is dominated by the stable feature characters in the scene c. Therefore, the model will give a higher probability to those semantic space areas which is dominated by the contextual coexistence. At the same time, it will give low probability to those with ambiguous coexistence region.

For example, streets and buildings often coexist, the contextual model gives a higher probability to both scenarios for the coexistence reason. On the other hand, streets and bedrooms are rarely coexist, then will be assigned a low probability. Thus, the representation of the image through the posterior probability will emphasize the contextual coexistence, a compress coincidence coexistence. Here, high-level abstraction of posterior probability is called contextual features, the probability vector associated each image is called CMN (Contextual Multinomial), which forms a contextual space.

Copyright ⓒ 2016 SERSC

5.2.2. Contextual Feature Model

The CCM (Contextual Concept Model) is obtained by learning the semantic space S. The random variable C is defined in terms of a scene indexes set K, $c \in 1,..K$. Here assumed scene vocabulary K and visual spatial scenes vocabulary L have equivalent meaning namely: $K = L$. This assumption means that the contextual model can explain the object relationship in scenes.

Since scene S itself is a probability simplex, c is the scene of a Dirichlet distribution mixture in the scene S. Then we have this:

$$P(\pi \mid c, \wedge^c) = \sum_k \beta_k^c Dir(\pi, \alpha_k^c) \tag{2}$$

Here, the parameter $\wedge^c = \left\{\beta_k^c, \alpha_k^c\right\}$, β_k is the probability density function is $\left(\sum_k \beta_k^c = 1\right)$, $Dir(\pi,\alpha)$ is the dirichlet distribution with parameter $\alpha = \left(\alpha_1,...,\alpha_L\right)$.

$$Dir(\pi;\alpha) = \frac{\Gamma(\sum_{i=1}^{L} \alpha_i)}{\prod_{i=1}^{L} \Gamma\left(\pi_i\right)} \prod_{i=1}^{L} \left(\pi_i\right)^{\alpha_i - 1} \tag{3}$$

Where, $\Gamma(\cdot)$ is the gamma function, Parameter $\wedge^y$ is got by training in the semantic space π of all the images among D_c.

5.2.3. Contextual Model Learning

It needs to note that this structure is generic, so any posterior probability vector to produce the appearance of the identification system may be used to study the proposed model. In fact, the contextual model can even use non-clear appearance modeling such discrimination classifier.

$$\pi^* = \arg\min_{\pi} \frac{1}{N} \sum_{n=1}^{N} KL(\pi \,\|\, \pi^n) \quad s.t \quad \sum_{i=1}^{L} \pi_i = 1 \tag{4}$$

We can prove that the above equation can be transformed into the following equation:

$$\pi_i^* = \frac{\exp \frac{1}{N} \sum_n \log \pi_i^n}{\sum_i \exp \frac{1}{N} \sum_n \log \pi_i^n} \tag{5}$$

It can be further simplified to:

$$\pi_i^* = \frac{\exp\left\{\frac{1}{n} \sum_n \log P(x_n \mid i)\right\}}{\sum_j \exp\left\{\frac{1}{n} \sum_n \log P(x^n \mid i)\right\}} \tag{6}$$

So the image SMN is:

$$\pi^* = \arg\max_{\pi} P(\pi \mid I) \tag{7}$$

However, this optimization is no direct solution, usually solved with approximate reasoning approach including Laplace, variational approximation or sampling methods, this study is the variational method:

$$\pi_i^* = \frac{\gamma_i - 1}{\sum_j \gamma_j - 1} \tag{8}$$

Here, γ_i is solved by the following iterations:

$$\gamma_i^* = \sum_n \phi_{ni} + \alpha_i \tag{9}$$

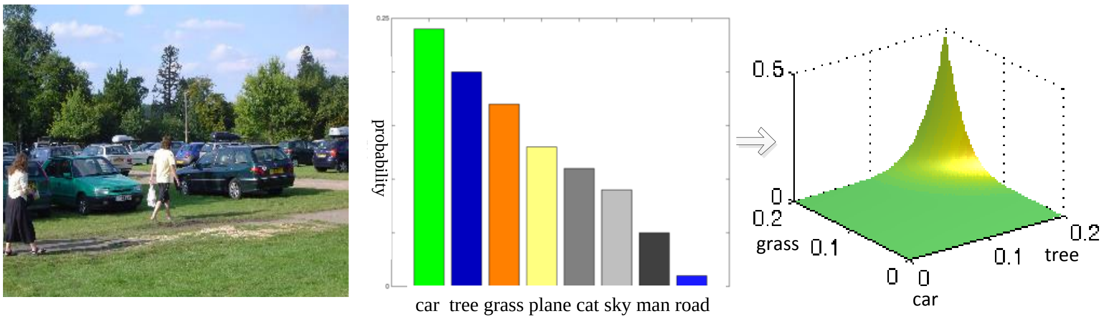
$$\phi_{ni}^* \propto P(x_n \mid w_n = 1) e^{\psi(\gamma_i) - \psi(\sum_j \gamma_j)} \tag{10}$$

Here, α_i is a priori $P(\pi)$, and compatible with uniformity class priori assumptions, it generally set to 1. $\psi(\cdot)$ is the second gamma function (Digamma function), ϕ_{ni} and γ_j are the parameters of variational distribution.

5.2.4. Contextual Feature Space

The contextual model $P(\pi \mid c)$ based on semantic space S has a similar role as the surface model based on visual space X with the $P(x \mid c)$. According to the Bayes rule, the formula is

$$P(c \mid \pi) = \frac{P(\pi \mid c) P(c)}{P(\pi)} \tag{11}$$



(a) (b) (c)

Figure 7. From Semantic Feature Space to Contextual Feature Space

By retaining all the posterior probability $\varsigma_c = P(c \mid \pi)$ based on contextual scene, we can design new semantic space. We use a CMN vector $\varsigma = (\varsigma_1, \ldots, \varsigma_L)^T$ as the contextual multivariate normal distribution of the image I, which exists in a new probabilistic simplex. Thus, we establish a contextual representation from the image I to CMN ς mapping.

The CMN generating process is shown in Figure 7. Among the Figure 7, (a) is the original image, (b) is based on the original image SMN feature space formed by the feature extracting representation, (c) is the CMN learned from SMN feature space.

5.3. New Feature Representation for Superpixel Regions and Patches

It's obviously inadequate for the superpixel area representation in the original literature [11] indicated by R in Figure 2, which doesn't well describe the overall characteristics of the region. So, in order to improve the performance of semantic segmentation, we propose the following two measures: 1) use UCM method to better preserve the regional segmentation and object edges; 2) use a similar practical method of literature [19, 20], the pixel area of the global super-pixel features include:

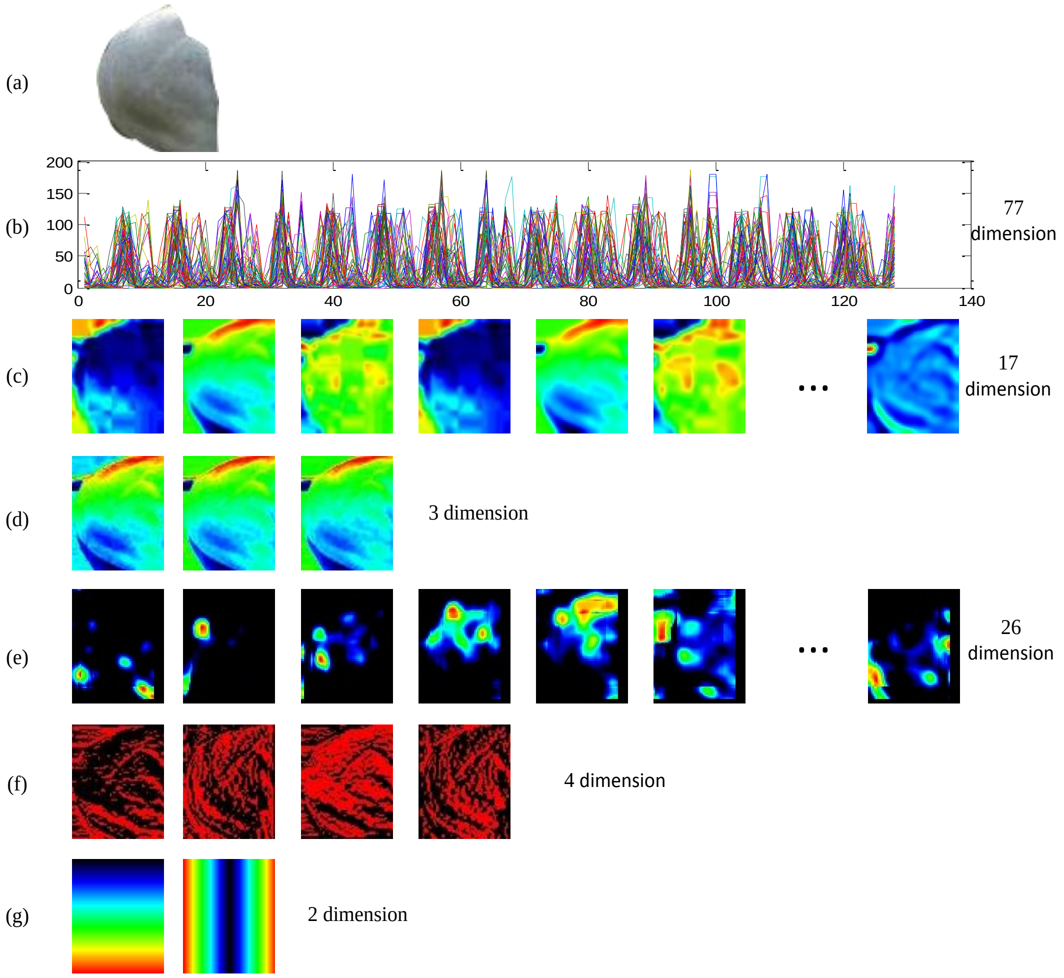


Figure 8. New Feature Representation of Superpixel Regions and Patches

1) DSIFT features as shown in Figure 8 (b) has a total of 77 dimensions. DSIFT features can reflect a variety of invariant characteristics of the image.

2) The shape of the texture features of the filter as shown in Figure 8 (c) is obtained from the group of 17-dimensional convolution filter. Here, the dimension of the filter consists of 9 scales Gaussian CIELab color space of the three channels, four-dimensional Gaussian derivative (x and y directions having two dimensions), and 4 dimensional Laplacian of Gaussian, and a total of 17 dimensions.

3) RGB color features as shown in Figure 8 (d) has values in RGB color space of the three channels, a total of three dimensions.

4) HOG features as shown in Figure 8 (e) has a total of 26 dimensions of the HOG feature.

5) LBP features as shown in Figure 8 (f) has the total four dimensions of LBP features.

6) Position feature, as shown in Figure 8 (g) is used to represent the current pixel height and centerline position, which has two dimensions.

7) We also use the feature representation as discussed before for the small image patches indicated by X in Figure 2.

Copyright ⓒ 2016 SERSC

According to the literature [19, 20] and the results of our experiments show that the new composite feature representation as mentioned before is helpful for semantic segmentation.

6. Model Learning

The model learning is same as the literature [11]'s CBS (Collapsed Gibbs Sampling) proposed by Neal[21]. The parameters meaning of o , R, X, S, T, Z , O_{dn} , R_{dn} , X_{dn} , S_A , T_B , Z_A , $\overline{O}_{dn}$ $\overline{R}_{dn}$, $\overline{X}_{dn}$, $\overline{S}_A$, $\overline{T}_B$, $\overline{Z}_A$, $n_{co,-dn}$, π_o , N_o and $n_{co',-dn}$ is same as in literature [11]. ς_d is contextual distribution. Due to Markov property of variable O , we can eliminate these variables and integrate out parameters $\eta, \alpha, \beta, \gamma, \theta, \varphi$. Then, the posterior over the object O_{dn} can be described as:

$$\begin{aligned} & p(O_{dn} = o \mid \overline{O}_{dn}, \varsigma_d, R, X, S, T, Z) \\ & \qquad \propto p(O_{dn} = o \mid \overline{O}_{dn}, \varsigma_d) \times \qquad (1) \\ & \qquad p(R_{dn} \mid \overline{R}_{dn}, O) \times p(X_{dn} \mid \overline{X}_{dn}, O) \times p(Z_A \mid N_r) \times \quad (2) \\ & \qquad p(S_A \mid O, Z, \overline{S}_A) \times p(T_B \mid O, Z, S, \overline{T}_B) \qquad (3) \end{aligned} \tag{12}$$

Using standard Dirichlet integral formulation, we obtain the first element of this product:

$$p(O_{dn} \mid \overline{O}_{dn}, \varsigma_d = c) = \frac{n_{co,-dn} + \varsigma_c + \pi_o}{\sum_{o'} n_{co',-dn} + N_c \varsigma_c + N_o \pi_o} \tag{13}$$

Where, N_c is the total numbers of different classes.

The second and third item in equation 1 can be learned by the same way.

7. Model Inference

Our model inference is also similar to literature [11]'s model reasoning.

7.1. Image Classification

The purpose of classification is judgment the scene class of the unknown image by calculating of the implicit object variable. We use the visual potential model (such as visual potentials, regional and small surface information) to calculate the contextual probability of each scenario, and then select the maximum probability class for the final scene class.

$$\begin{aligned} p(\varsigma \mid R_d, X_d) &= \frac{p(\varsigma, R_d, X_d)}{p(R_d, X_d)} \\ &\propto \prod_{N_r} \sum_{O} p(R \mid O) p(X \mid O) p(O \mid \varsigma) \end{aligned} \tag{14}$$

Finally, choose $c = \arg\max p(\varsigma \mid R_d, X_d)$ for the final classification classes.

7.2. Image Annotation

The image annotation results derive from image semantic segmentation, image segmentation is based on the probability of each class of objects, and we take the classes of object classes in image as image annotation classes.

7.2. Image Segmentation

Semantic segmentation infers the accurately the position of each pixel of an object in an image. This can be got by integrating all the scene object classes:

$$\begin{aligned} p(O \mid R, X) &= \sum_C p(O, \varsigma, R, X) \propto \sum_C p(O, C, R, X, \varsigma) \\ &= \sum_C p(O \mid \varsigma) p(R \mid O) p(X \mid O) p(\varsigma \mid C) p(C) \end{aligned} \tag{15}$$

You can see the impact of the top-down object segmentation classes $p(C)$ not only affected by the scene, but also by visual features on the bottom.

8. Experimental Design

We tested our new model on two data sets.

8.1. Data Sets

Table 1. The Statistics of Scene Classification, Annotation and Segmentation on Msrc-v2 Dataset

Type of segmentation and annotation	Type of scene classification
22 types: Building, Grass, Tree, Cow, Sheep, Sky, Aeroplane, Water, Face, Car, Bicycle, Flower, Sign, Bird, Book, Chair, Road, Cat, Dog, Body, Boat, Background	21 types: Sign, Bird, Dog, Cat, Bicycle, Tree, Water, Sheep, Person, Building, Cow, Chair, Aeroplane, Grass, City, Flower, Book, Boat, Nature, Car, Face

8.1.1. UIUC sports data set

The UIUC sports data set[14] contains eight classes: badminton, bocce, croquet, polo, rock climbing, Boating (Rowing), sailing, snowboarding. Each class contains 800 images, a total of 6400 images, 200 images were randomly selected for each class as testing, the rest for the training.

8.1.2. Msrc-v2 data set

The Msrc-v2 data set[22] is currently used for testing the semantic segmentation and classification. The original database consists of 591 images, of which the statistics of scene classification, semantic annotation statistics is as shown in Table 1. The number of training set is 335 images, and 256 images for the testing. The image annotation class is the first 7 classes, a total of 22 classes (including the background).

8.2. Experimental setup

We test our model on the UIUC and Msrc-v2 dataset. The testing hardware environment is the CPU of Intel P6100, 2.00 GHZ and memory of 6 GB RAM. Our development and testing platform are Ubuntu12.04 operating system, with Matlab2013b and gc ++ development.

1) Super-pixel segmentation: super-pixel region segmentation methods is changed from the literature [23] approach to UCM, threshold of super-pixel segmentation is set to 0.1.

2) For CMN training

(1) Our method is different from the literature [16] that the basic feature of the image is color PHOW descriptors extracted by SIFT scale dense set to 7, DSIFT step size set to 5, and Color set to opponent.

(2) The mixing ingredients is set 45. According to the literature [16] validation, there are very small gains for the accuracy when the ingredients is greater than 40. To avoid singularity, setting a variation value is set to 0.01 for minimum, and the maximum is 100. The similar measure for SMN is conducted by KL divergence.

3) The default model super parameters is : $\alpha = 1, \beta = 1$

4) The default topic class is set to 20.

9. Experimental Results and Analysis

9.1. Image Scene Classification

On the UIUC and Msrc-v2 dataset, scene classification accuracy comparison of each class is shown in Table 2 and Table 3. Figure 9 and Figure 10 is a visual comparison of the confusion matrix of our model and literature [11] on two data sets.

Table 2. Classification Results on Dataset UIUC

Method	badminton	bocce	croquet	polo	rock climbing	rowing	sailing	snow boarding	Average
Literature [11]	67	41	68	56	56	35	57	54	54.3
Our model	69	58	77	60	74	51	60	59	63.5

Table 3. Classification Results on Dataset Msrc-v2 (%)

Method	sign	bird	dog	cat	bicycle	tree	water	sheep	person	building	cow
Literature [11]	62	67	58	70	85	85	67	94	42	46	90
Our model	69	73	67	70	85	85	73	94	42	46	95

From table 2, 3 and Figure 9, 10 , we can see that:

(1) For each classification accuracy, our proposed model were significantly better than the literature [11].

(2) The average accuracy increases from 54.3% to 63.5%.

2) On Msrc-v2 dataset

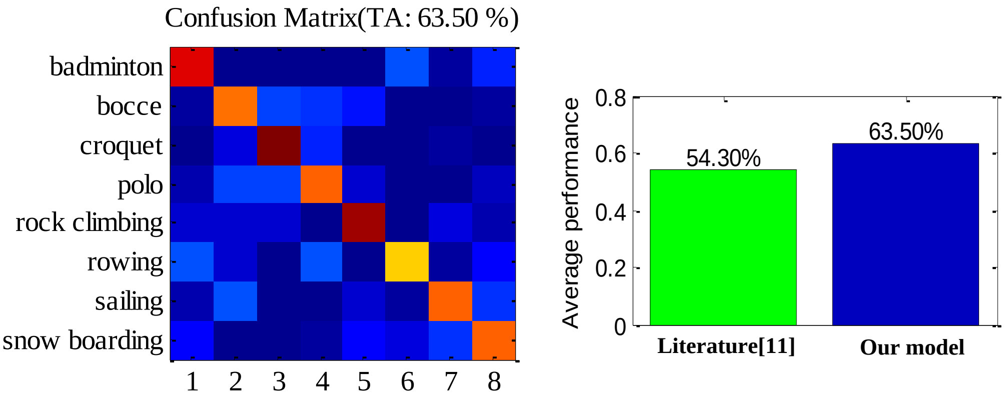
Table 4. Classification Results on Dataset UIUC

Method	chair	aeroplane	grass	city	flowers	books	boat	nature	car	face	Average
Literature [11]	77	92	100	91	86	100	29	75	85	92	76.2
Our model	77	92	100	82	93	100	43	75	85	100	78.9

(1) For each classification accuracy, except the city class that the accuracy declines, and the classes of cat, bike, tree, sheep, person and building that the accuracy remains unchanged, the classification accuracy of the remaining 14 classes is improved.

(2) The average accuracy increases from 67.2% to 78.9%.

The results of these two data sets are verified that the contextual features are helpful to enhance understanding of the overall scene classification performance.



(a)

(b)

Figure 9. Classification Result on Dataset UIUC

9.2. Image annotation

Evaluation criteria of image annotation are similar to the one of image retrieval by using the first N-related object classes, here N is seven. The calculation formula as shown in equation 16.

$$F_\beta = \frac{(1+\beta^2)\,precision \times recall}{\beta^2\,precision + recall} \tag{16}$$

Here F_β measure is a composite indicator that can reflect better performance than single precision or recall criteria.

By comparing the results of the two data sets, we can see that

1) On UIUC dataset

(1) In the first seven classes, the performance of average precision, average recall and F_β of our model is beyond the literature [11] except the human class.

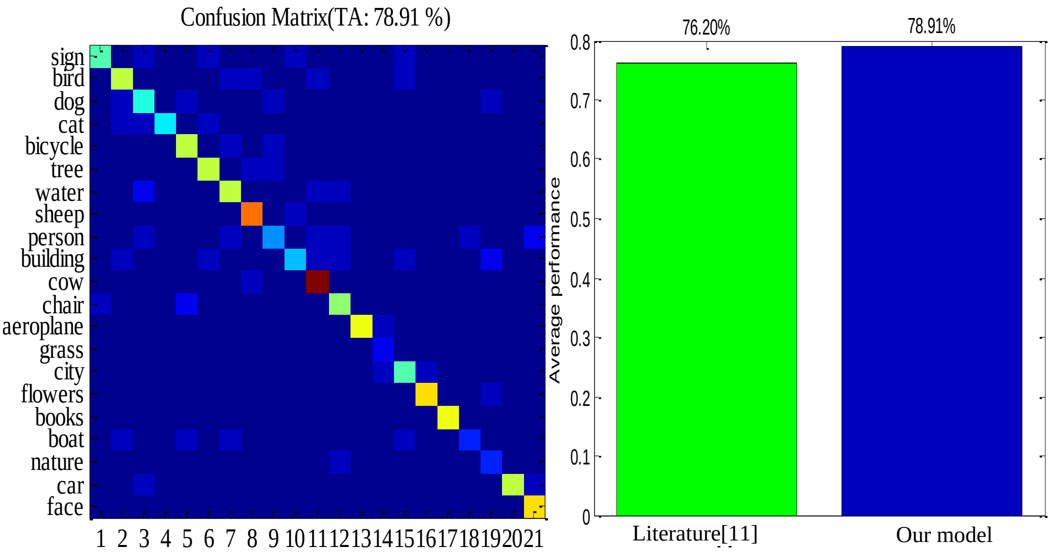
(2) The overall performance increases from 51% to 66%, an increase of 15%.

2) On Msrc-v2 dataset

(1) In the first seven classes, the performance of average precision, average recall and F_β of our model is beyond the literature [11] except the building class.

(2) The overall performance increases from 75% to 79%, an increase of 4%.

These results prove the effectiveness of the proposed model.



(a) (b)

Figure 10. Classification on Dataset Msrc-v2

Table 5. UIUC Image Annotation Results on UIUC Dataset

Object types	Literature [11]			Our model		
	Average precision	Average recall	F_β	Average precision	Average recall	F_β
human	0.85	0.98	0.91	0.56	0.95	0.70
horse	0.17	0.91	0.29	0.45	0.86	0.59
grass	0.33	0.86	0.48	0.66	0.93	0.77
sky	0.44	0.92	0.59	0.65	0.98	0.78
tree	0.38	0.93	0.54	0.69	0.95	0.80
net	0.27	0.85	0.41	0.34	0.90	0.49
sand	0.24	0.46	0.32	0.45	0.56	0.50
Average	0.38	0.84	0.51	0.54	0.88	0.66

Table 6. Image Annotation Results on Msrc-v2 Dataset

Object types	Literature [14]			Our model		
	Average precision	Average recall	F_β	Average precision	Average recall	F_β
Cow	0.45	0.95	0.61	0.55	0.98	0.70
Tree	0.56	0.87	0.68	0.58	0.93	0.71
Face)	0.80	0.98	0.88	0.85	0.98	0.91
Car	0.77	0.89	0.83	0.75	0.95	0.84

Copyright ⓒ 2016 SERSC

Road	0.57	0.56	0.56	0.63	0.70	0.66
Dog	0.86	0.88	0.87	0.88	0.96	0.92
Building	0.68	0.99	0.81	0.74	0.95	0.77
Average	0.67	0.87	0.75	0.65	0.92	0.79

Figure 11 is the annotation examples on two data sets. The bold annotation with blue color is the correct results, tags with red italics is for the wrong annotation, and the black annotation is a neutral. Because of the size of possible classes in UIUC data set is much larger than Msrc-v2 data set, the data listed here are only the probability of the annotation class greater than 0.001 on UIUC dataset, the one greater than 0.01 on the Msrc-v2 dataset.

In Figure 11, column (a) is the results on UIUC dataset, column (b) is the results on the Msrc-v2 dataset. In each column, the image on the left is the right labeled results, and the image on the right is the error labeled results.

9.3. Image Segmentation

In the comparison of image segmentation, our test plan is similar to the literature [22]. From Table 7,8 and Figure 12,13, we can see that:

1) For the global semantic segmentation accuracy, our model has greatly improved performance when compared with the literature [11], increase of accuracy from the original 77.7% to 84.1%, an increase of 6.4%.

2) For each classification accuracy, except the classes of sky, aircraft, roads and boat, our models are superior to or at least same as the literature [11].

3) For the line 2 to 5 in Figure 13, we can clearly see that semantic segmentation result of our the model is better than the literature [11].

4) From the above analysis, we can see that our new feature representation method to improve expression and segmentation is effective.

9.4. Experimental Discussion

Figure 14 is a comparison between SMN and CMN feature space representation, and row (1), (2) images are from the UIUC data set, row (3), (4) images are from Msrc-v2 data set. The column (a) is an image, column (b) is the SMN feature space, the column (c) is the contextual CMN for further transforming from the SMN feature space.

As shown in Figure 14, the interference problem between classes of SMN representation are prominent. However, as shown in the column (c), after using the context of spatial learning, interfering factors are clearly removed. Taking the row (4) as example, the three kinds of semantic graph probability: road, building and car are also discriminating, but interference problem between classes is relatively large, by contextual learning sky probability is decreased from the original 0.05 to 0.005, and for the real scene car class is increased from original 0.24 to 0.71. This significantly enhances the true class and suppresses interference classes, the other lines also have a similar phenomenon, which illustrates the necessity and value of the transformation from SMN to CMN for further image classification and annotation.

10. Conclusions and Future Works

For the holistic image scene understanding based on the directed graph, we propose two methods to enhance holistic scene understanding respectively, one is improving the image classification and annotation performance based on feature engineering and global contextual information, the other is to improve semantic expression segmentation by new image super pixel area and image small feature representation. Experiments show that our new holistic model shows higher overall performance compared with the literature [11] in scene classification, image annotation and semantic segmentation tasks, we summarized as below:

1) Scene Classification: we studies feature fusion problem by incorporating the overall global contextual information. Integrated model of context-based features and Bayesian directed graph is proposed, the model effectively reduces the semantic ambiguity of feature representation. Experiments show that: (1) On the UIUC data set, scene classification performance increases from the original 54.3% to 63.5%, an increase of 9.2%. (2) On Msrc-v2 data sets, performance increases from 76.2% to 78.9%, an increase of 3.7%.

2) Image annotation: the global performance improvement of contextual feature can also helpful. Experimental results show that: (1) On the UIUC data set, the overall performance of image annotation increases from the original 51.0% to 66.0%, an increase of 15.0%. (2) On the Msrc-v2 data set, the overall performance of image annotation increases from 75.0% to 79.0%, an increase of 4.0%.

3) Semantic segmentation: (1) This paper changes super pixel region method of the literature [23] approach to UCM; (2) We propose the use of ultra-small pixel area and the image description for new feature representation, which constitutes of the texture filtering, RGB color, HOG features, LBP features and location features. Experiments show that the method is effective, the overall performance of the semantic segmentation part from the original 77.7% to 84.1%, an increase of 6.4%.

Copyright ⓒ 2016 SERSC

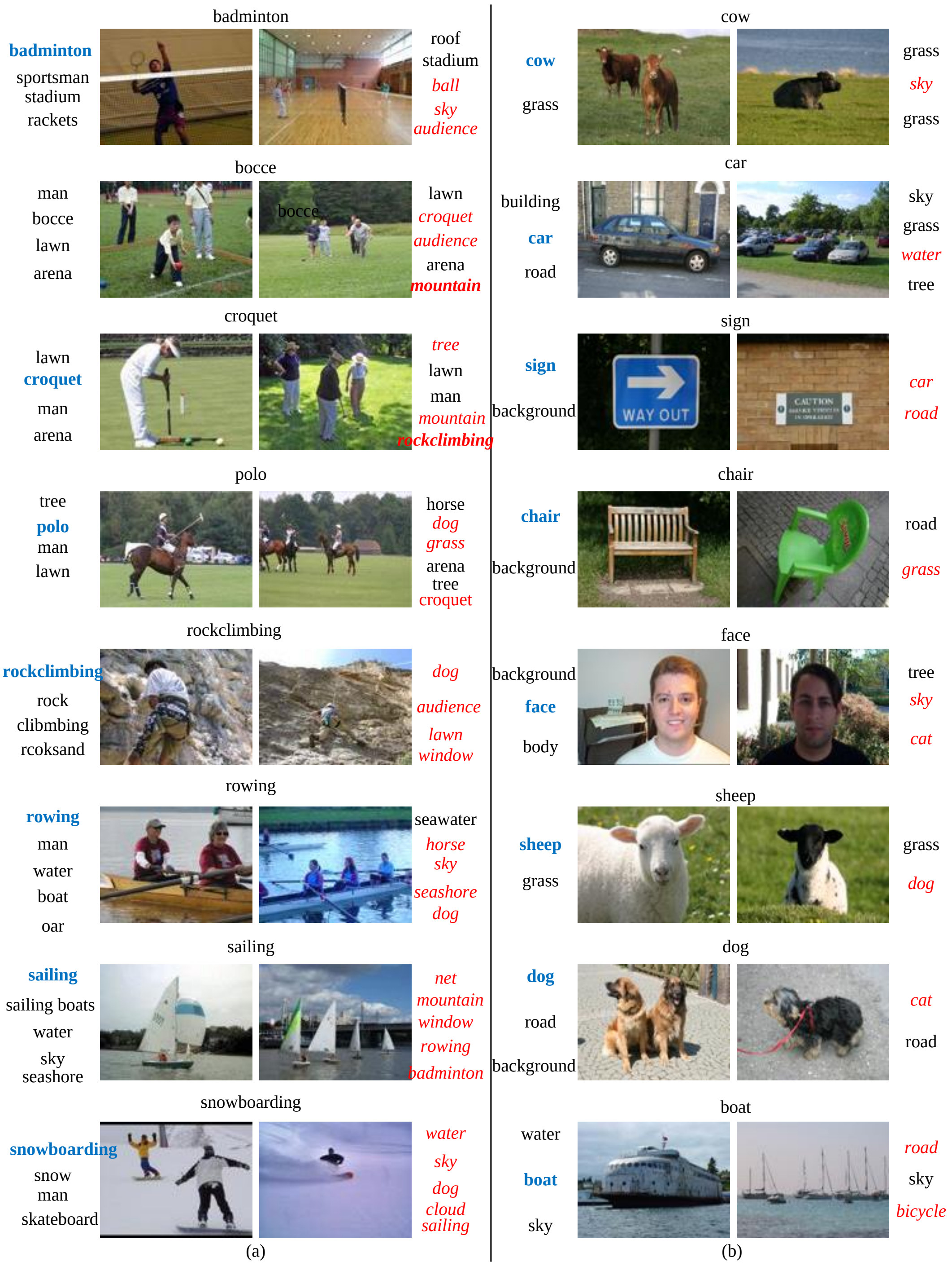
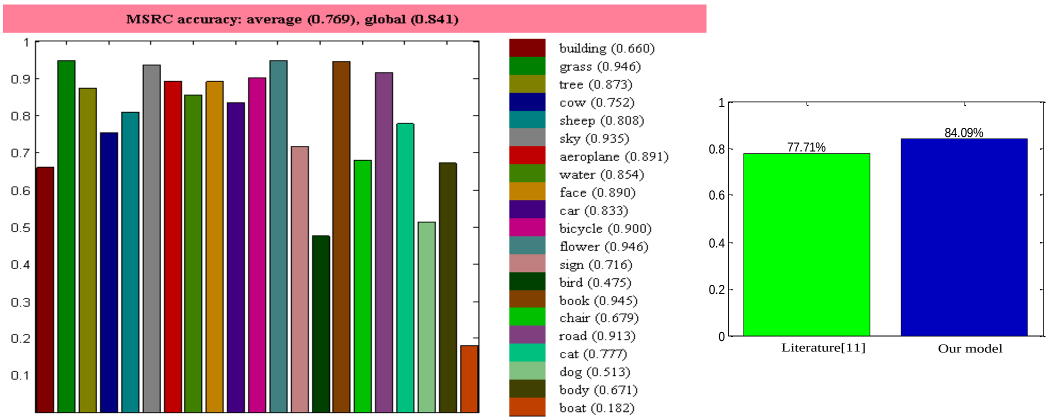
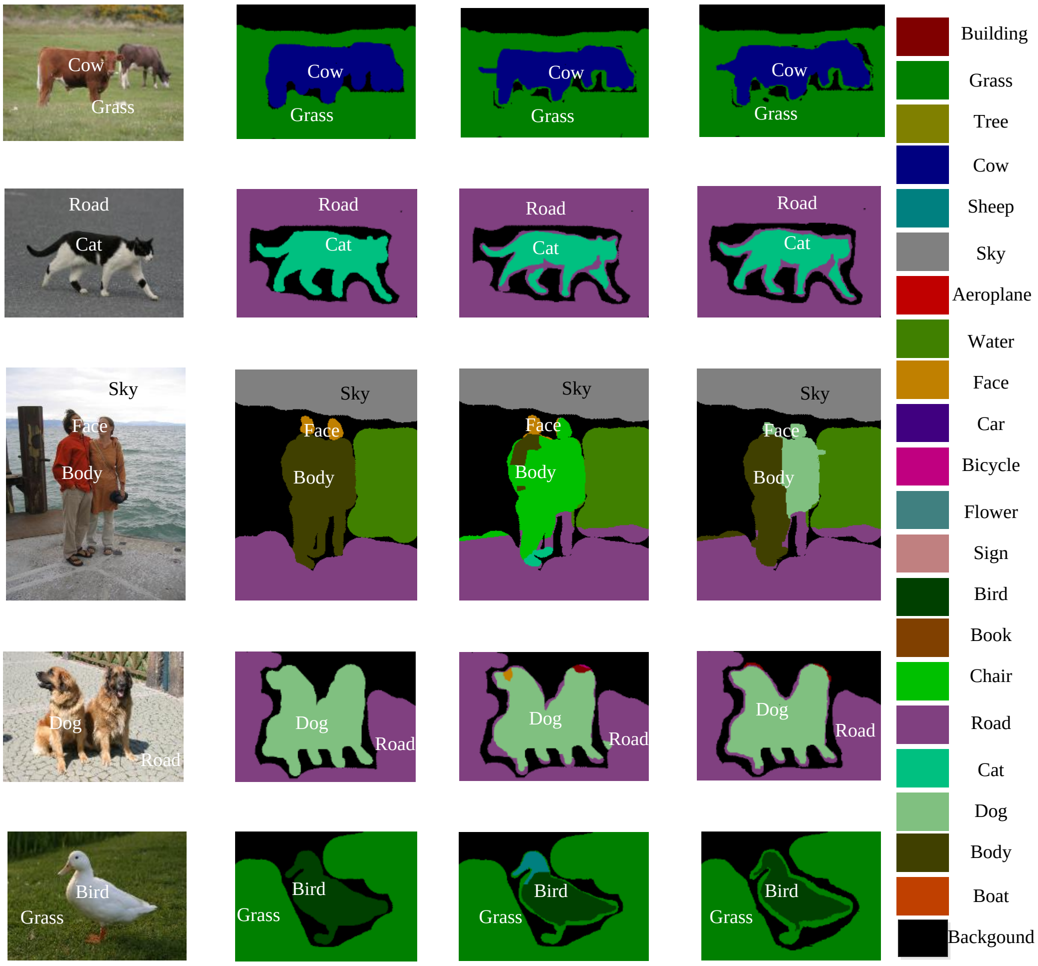


Figure 11. Some Image Annotation Results on Two Data Sets



(a) (b)

Figure 12. Semantic Segmentation Comparison on Dataset Msrc-v2 (2)



(a) Image (b) Ground truth (c) Reference [11] (d) Our model

Figure 13. Semantic Segmentation Comparison on Dataset Msrc-v2 (2)

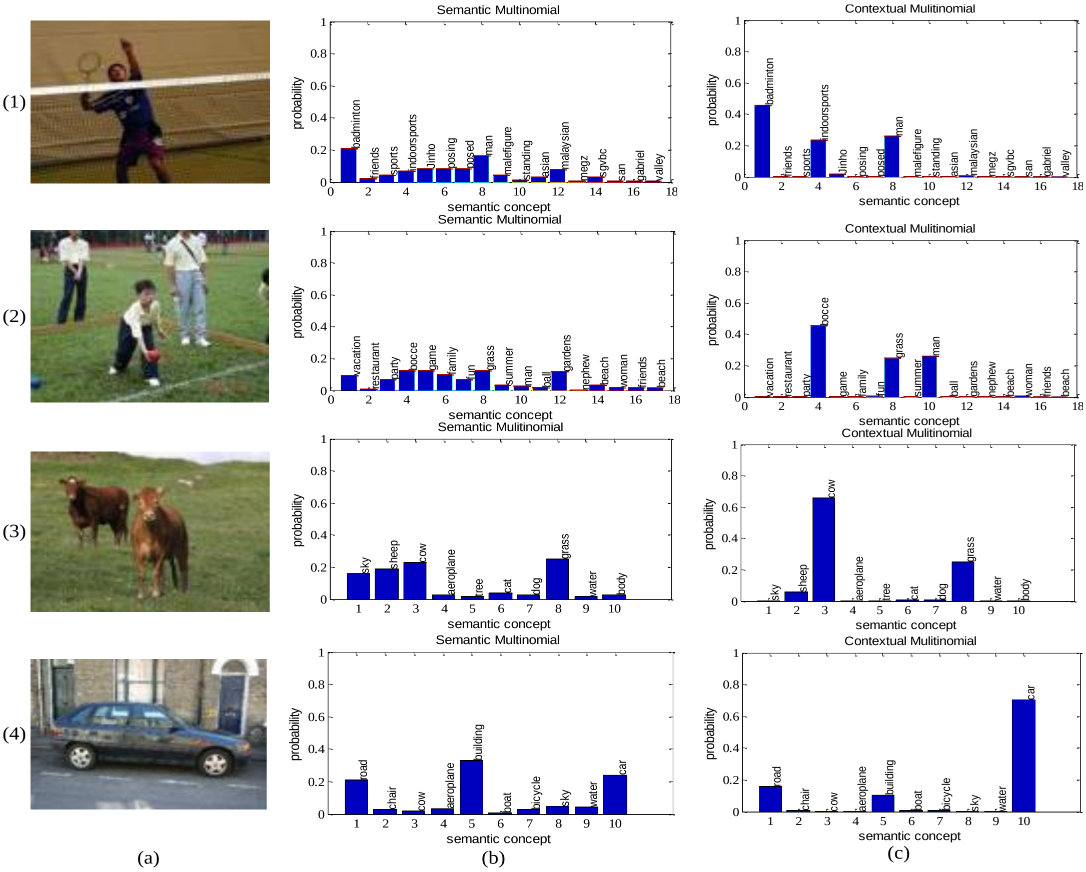
Table 7. Segmentation Results on Dataset Msrc-v2 (%)

Copyright ⓒ 2016 SERSC

Method	buliding	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle
Literature [11]	52	85	78	64	65	95	93	84	86	81	78
Our model	66	95	87	75	81	94	89	85	89	83	90

Table 8. Segmentation Results on Dataset Msrc-v2(%cont.)

Method	flower	sign	bird	book	chair	road	cat	dog	body	boat	average	global
Literature [11]	85	62	30	88	58	98	63	42	63	23	70.2	77.7
Our model	95	72	48	95	68	91	78	51	67	18	76.9	84.1



(a) (b) (c)

Figure 14. Comparison between SMN and CMN

Table 8. Classification Result Comparisons between SMN and CMN on Dataset UIUC

Method	badminton	bocce	croquet	polo	rock climbing	rowing	sailing	snow boarding	Average
Literature[11]	67	41	68	56	56	35	57	54	54.3
SMN	67	48	73	59	65	40	60	55	58.4
Our Model (based on CMN)	69	58	77	60	74	51	60	59	63.5

Acknowledgments

This work was supported in part by the Scientific Research Fund of Sichuan Provincial Education Department of China (No. 15TD0038).

References

[1] A. Vedaldi and A. Zisserman, “Efficient additive kernels via explicit feature maps”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 480, **(2012)**.

[2] P. Arbelaez, M. Maire, C. Fowlkes and J. Malik, “Contour detection and hierarchical image segmentation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 898 **(2011)**.

[3] M. Bar, “Visual objects in context”, Nature Reviews Neuroscience, vol. 5, no. 617, **(2004)**.

[4] D. Hoiem, A. A. Efros and M. Hebert, “Putting objects in perspective”, International Journal of Computer Vision, vol. 80, no. 3, **(2008)**.

[5] E. B. Sudderth, A. Torralba, W. T. Freeman and A. S. Willsky, “Learning hierarchical models of scenes, objects, and parts”, Proceedings of 10th IEEE International Conference on Computer Vision, Beijing, China, **(2005)** October 17-21.

[6] [6] Z. Tu, X. Chen, A. L. Yuille, and S.-C. Zhu. Image parsing: Unifying segmentation, detection, and recognition. International Journal of Computer Vision. 63, 113**(2005)**.

[7] [7] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, **(2005)** September. 21-23; San Diego, CA, United states.

[8] [8] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. Proceedings of 11th IEEE International Conference on Computer vision, **(2007)** October 14-21; Janeiro, Brazil.

[9] [9] L. Fei-Fei and L.-J. Li. What, where and who? telling the story of an image by activity classification, scene recognition and object categorization. Proceedings of 2010 IEEE Conference on Computer Vision and Pattern Recognition. **(2010)** June 21-23; San Francisco, CA, United states.

[10] [10] J. J. Corso. Toward parts-based scene understanding with pixel-support parts-sparse pictorial structures. Pattern Recognition Letters. 34, 762**(2013)**.

[11] [11] L.-J. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, (2009) June 20-25; Miami, FL, United states.

[12] [12] D. Hoiem, A. A. Efros, and M. Hebert. Closing the loop in scene interpretation. Proceedings of 26th IEEE Conference on Computer Vision and Pattern Recognition. **(2008)** June 23-28; Anchorage, AK, United states.

[13] [13] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. Proceedings of 11th IEEE International Conference on Computer Vision.**(2007)** October 14-21;Janeiro, Brazil

[14] [14] L.-J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. Proceedings of 11th IEEE International Conference on Computer Vision. **(2007)** October 14-21;Rio de Janeiro, Brazil.

[15] [15] D. M. Blei and M. I. Jordan. Modeling annotated data. Proceedings of 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.**(2003)** July 28 - August 1; Toronto, Canada

[16] [16] N. Rasiwasia and N. Vasconcelos. Holistic context models for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 34, 902**(2012)**.

[17] [17] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence. 32, 1627**(2010)**.

[18] [18] A. Torralba, K. P. Murphy, and W. T. Freeman. Contextual models for object detection using boosted random fields. in Advances in neural information processing systems. Proceedings of the 18th Advances in Neural Information Processing Systems, **(2004)** December 13-18; Vancouver, BC, Canada

[19] [19] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. Proceedings of 9th European Conference on Computer Vision, **(2006)** May 7-13; Graz, Austria.

[20] [20] S. Gould, T. Gao, and D. Koller. Region-based segmentation and object detection. Proceedings of Advances in Neural Information Processing Systems. **(2009)** December 7-9;Vancouver, BC, Canada

[21] [21] R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. Journal of Computational and Graphical Statistics. 9, 249**(2000)**.

[22] [22]. J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. **(2012)** June 18-20; Providence, RI, United states.

[23] [23]. P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. International Journal of Computer Vision. 59, 167 **(2004)**.

Authors

Lin Li, Ph.D. an associate professor at department of Computer Science and technology, Chengdu Normal University. He is also a system analyser of computer and software technology of P.R. China and member of China Computer Federation. His research interest covers machine learning and its application in image processing and computer vision. He is the corresponding author of this paper.

Copyright ⓒ 2016 SERSC