Fuzzy Based Contextual Cueing for Region Level Annotation

Sheng-hua Zhong, Yan Liu, Yang Liu, Fu-lai Chung
The Hong Kong Polytechnic University

ABSTRACT

This paper investigates the challenging issue of assigning given image-level annotations to precise regions on natural images. We propose a novel label to region assignment (LRA) technique called Fuzzy-based Contextual-cueing Label Propagation (FCLP) with four parts: First, an image is over-segmented into a set of atomic patches and the local visual information of color features and texture features are extracted. Second, fuzzy representation and fuzzy reasoning are used to model contextual cueing information, especially for the imprecise position information and ambiguous spatial topological relationships. Third, labels are propagated inter images in visual space and intra images in contextual cueing space. Finally, the fuzzy C-means clustering based on K-nearest neighbor (KNN-FCM) is utilized to segment the images into semantic regions and associate with corresponding annotations. Experiments on the public datasets demonstrate the effectiveness of the proposed technique.

Categories and Subject Descriptors

I.4.9 [**Computing Methodologies**]: Image Processing and Computer Vision—*Applications*

General Terms

Algorithm, Performance, Experimentations

Keywords

Label to region assignment, Contextual cueing, Fuzzy Theory.

1. INTRODUCTION

Label to region assignment (LRA) is defined as the assignment of the given image-level annotations to the precise regions within the image automatically. For example, Figure 1(a) is an image with three image-level annotations of water, cow, and grass. The aim of LRA is to segment the image to several regions and associate the annotations with the corresponding semantic regions as shown in Figure 1(b). LRA techniques could replace the tedious manual method of making region-level annotations, so it would be helpful in achieving reliable and visible content-based image retrieval [1].

J. Li et al. proposed a uniform framework of LRA for images in sports domain [2]. They used WordNet to refine the image-level annotations generated from noisy tags and Flickr images as the training data to improve the recognition accuracy of regions. The system showed impressive and stable performance to segment and assign regions for the images of badminton, bocce, croquet, polo, rock climbing, rowing, sailing, snowboarding, and etc. X. Li explored bi-layer sparse coding and label propagation techniques for LRA to natural images, which is the challenging issue in multimedia content analysis [1]. The basic idea is that the regions with the common annotation are more likely to have similar local features, even if these regions are in different images. Their methods showed distinguished performance improvement for images with multiple objects or in a complex background.

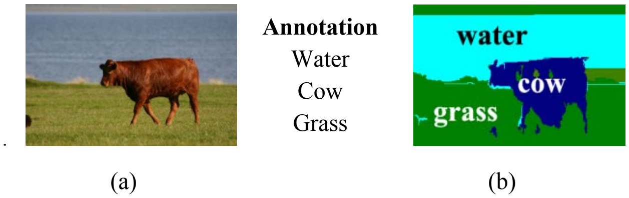


Figure 1. Example of label to region assignment. (a) An image with given image-level annotations (b) The label to region assignment result.

However, visual similarity doesn't work for all the cases in image understanding. Figure 2(a) shows an ordinary image of nature view. Similar with [1], we use Scale-invariant feature transform (SIFT) descriptors as the local features of the uniform sampled data points in Figure 2(b). The values of the local features from different data points are very similar. Figure 2(c)-(f) compares SIFT descriptors from four random selected data points, two from sky and two from sea. Obviously, the difference between (d) and (e) is even smaller than the difference between (d) and (c). Actually, this problem is not caused by SIFT descriptor. Only considering visual features, such as color and texture, the sky and the sea are similar. One interesting observation is that human can distinguish the sea and the sky easily. Human have seen similar views or pictures in real life, so that they have formed the prior knowledge that the sky is generally above the sea. Such kind of prior knowledge was formally defined by psychologists in 1998 as contextual cueing, the manner in which human brains gather information by incidentally learned associations between spatial configurations and target locations [3].

In this paper, we try to provide more semantic understandings of the natural images with the aid of contextual cueing. Different with contextual information, for example, synchronized or unsynchronized logs and text associated with multimedia data, which has been widely used to understand web videos [4], contextual cueing is seldom studied by multimedia society. One possible reason is that the classical bivalent sets theory causes serious semantics loss in describing contextual cueing, such as imprecise position information and ambiguities in spatial topologic. To address this difficulty, we utilize fuzzy representation and fuzzy logic, corresponding to the "degree of truth", to model the relationships of the semantic regions within an image. We describe the contextual cueing with fuzzy theory in section 2, and based on this idea, we propose a novel LRA technique called Fuzzy-based Contextual-cueing Label Propagation (FCLP) in section 3. Section 4 provides performance comparison and demonstration on two public datasets. The paper is closed with conclusion.

ICIMCS'10, December 30-31, 2010, Harbin, China.

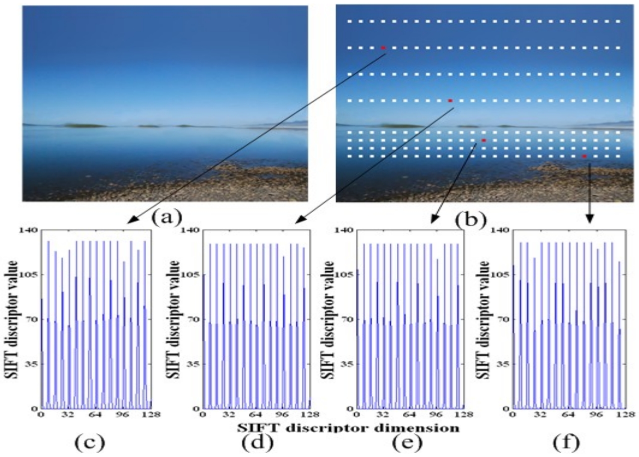


Figure 2. Example of the difficulty to distinguish sky and sea based on visual feature. (a) The original image. (b) The original image with 200 data points. (c) – (f) 128 local features of four random points selected from sky and sea are shown.

2. CONTEXTUAL CUEING WITH FUZZY THEORY

Contextual cueing is a concept in psychology that refers to the manner in which the human brain gathers information from visual elements and their surroundings. Generally, the information is acquired incidentally from past experiences of regularities of the visual world, and gradually formed the knowledge about spatial invariants. Five types of spatial invariants are thought to be important in contextual cueing [5]:

- Probability: the likelihood that certain objects will be present in a scene
- Co-occurrence: the likelihood that certain objects will be present together
- Size: the familiar relative size of objects
- Position: the typical positions of some objects in some scenes
- Spatial topological relationship: left of, right of, above, below, surround, inside, and etc.

The spatial invariants can guide the visual attention, speed the visual elements search, and help the object recognition. Figure 3 is an example to illustrate how contextual cueing works to resolve ambiguity for the recognition. We cannot distinguish which the object is in Figure 3(a), a cup or a hat, because they have very similar appearances. But in Figure 3 (b), it is easy to recognize the target object as a hat when a related object (head) appears below it. And in Figure 3 (c), the surrounding dishware disambiguates the identity of the object effectively.

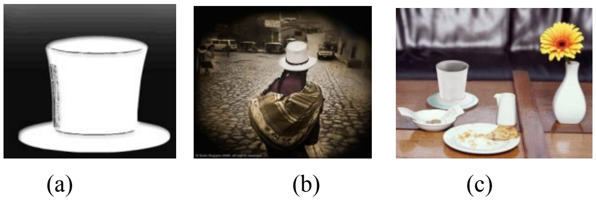


Figure 3. Example of contextual cueing in object recognition. (a) An ambiguous object: hat or cup? (b) A hat on the head (c) A cup surrounding dishware.

Based on these psychology theories and evidences, this paper intends to integrate the contextual cueing in LRA to provide human-like understanding of the image. Two issues should be addressed. The first issue is that how to model the acquired knowledge of contextual cueing, i.e., spatial invariants. The second issue is how to model the formation of contextual cueing, i.e., the learning process of knowledge acquirement.

For the modeling of spatial invariants, the probability and co-occurrence information are available with regard to the problem of LRA, because the image-level annotations are known in advance. In [1], object size has been considered. Some studied from related area, known as simultaneous object recognition and image segmentation, have demonstrated obvious performance improvement even if very simple position information [6] and spatial topological relationship [7] are utilized. However, little work has been conducted to explore spatial invariants for LRA problem of natural images, mainly due to the difficulty of modeling imprecise position information and ambiguous spatial topological relationships for the images with multiple objects and complex backgrounds. To address this problem, we describe the position information using fuzzy representation and model the topological relationships using fuzzy logic. Different from the probability-based theory that measures the likelihood an even occurs, fuzzy theory measures the degree that an event occurs. In Figures 4(a) and (b), is the sky above the building in the image? Yes, we are totally sure because the images have existed. But obviously, the spatial relationship between the sky and the building in Figure 4(a) is different with the one in Figure 4(b). How to describe this kind of difference? In our paper, we use fuzzy membership to quantize the degree of truth for these spatial invariants. Moreover, fuzzy reasoning is used to imitate human's learning process of contextual cueing.

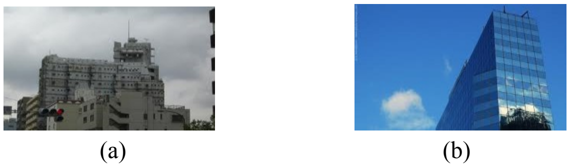


Figure 4. Example of imprecise position information and ambiguous spatial topological relationships in natural images.

3. FUZZY BASED CONTEXTUAL CUEING LABEL PROPAGATION

This paper proposes a novel Fuzzy-based Contextual-cueing Label Propagation (FCLP) technique for label to region assignment to natural images. As shown in Figure 5, an image is first over-segmented into atomic patches. Then visual features and contextual cueing features are extracted from each atomic patch. Two kinds of visual features, color and texture, are coded based on two Bag-of-Words (BOW) codebooks. Labels are propagated inter images using Bi-Layer sparse coding. Fuzzy position and fuzzy spatial topological relationship are used as contextual cueing features. We acquired the knowledge of spatial invariants by learning from web images using fuzzy reasoning. According to the acquired fuzzy membership of positions and spatial topologic, labels are propagated intra images to model the relationships of the semantic regions within an image. Finally, the post processing utilized fuzzy C-means clustering based on K-nearest neighbor (KNN-FCM) to assign the given image-level annotations to the corresponding regions.

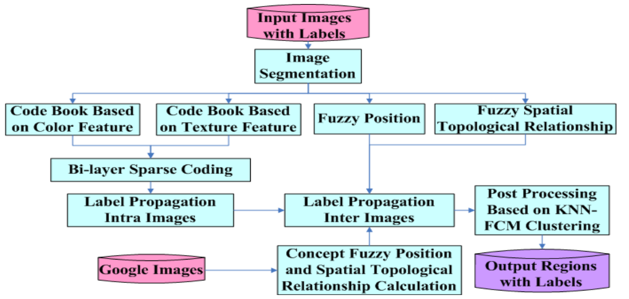

Input Images with Labels
Image Segmentation
Code Book Based on Color Feature
Code Book Based on Texture Feature
Fuzzy Position
Fuzzy Spatial Topological Relationship
Bi-layer Sparse Coding
Label Propagation Intra Images
Label Propagation Inter Images
Post Processing Based on KNN-FCM Clustering
Google Images
Concept Fuzzy Position and Spatial Topological Relationship Calculation
Output Regions with Labels

Figure 5. Sketch of Fuzzy-based Contextual-cueing Label Propagation technique.

3.1 Image Representation

In our technique, two different kinds of features are utilized to present the image, visual features and contextual cueing features. As pre-processing to FCLP, we advocate segmenting images into multiple segmentations. Like [1]-[2], to ensure the segmented patch involve within an object/concept, we start with a modified version of an over-segmentation algorithm [8].

3.1.1 Visual Features

After the over-segmentation step, the feature representation is obtained for those atomic patches. Each atomic patch is described by using Bag-of-Words (BOW) features generated by color (in *Lab* space) and texture features (SIFT descriptor). Within each region, a number of interest points are detected by using the scale invariant saliency (SIFT) detector. In some small patches, SIFT detector cannot detect any interest points. In this case, N_s points are randomly picked from each patch as the chosen points. A codebooks includes two parts are obtained for the chosen points and the region appearance by unsupervised k-means clustering. One is obtained based on the SIFT descriptor of chosen points. Another is obtained based on the LAB color of chosen points and the average LAB value of whole region. The visual feature of an atomic patch x_{ij} in image $\mathrm{x}_i, i=1,...\mathrm{N}$ could be denoted as an *m*-dimensional descriptor feature $x_{i,j} \in \mathbb{R}^m, j=1,2,...,n_i$, where N is the number of image dataset. n_i is the number of patches in image x_i.

3.1.2 Contextual Cueing Features

In our technique, two kinds of contextual cueing features are utilized in semantic space, fuzzy position and spatial topological relationship. To represent these features in each atomic patch, the location of the patch need to be firstly studied. Previous work differed in studying the location of one representative point or complete points set in every patch. Based on the balance of the representational ability and the computational complexity, we choose the center of gravity and the contour points as the typical points to represent each patch.

a) Fuzzy Position

Fuzzy position in contextual cueing is used to represent the typical positions of some objects in images. According to the conclusion in [9], one object category is likely to be within a horizontal section of the image. Therefore, we choose to use the vertical location "top," "middle," and "bottom," to characterize the position. The fuzzy membership of position is defined in Figure 6(a) using a commonly used triangular function, where I_h is the height of the image. The fuzzy membership of position is calculated on the typical points. The average fuzzy value of these points is defined as the fuzzy position of patch *j* denoted as $R^P_{Position}(j)=\{\mu^P_{top}(j),\mu^P_{middle}(j),\mu^P_{bottom}(j)\}$. To distinguish with the fuzzy membership for concept defined later, Letter "*P*" used as a superscript to indicate it is the fuzzy membership for patch.

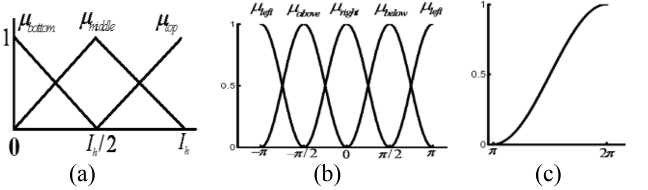


Figure 6. (a) Fuzzy membership for position "top," "middle," "bottom." (b) Fuzzy membership for the spatial relationships in four directions. (c) Fuzzy membership for "surround."

b) Fuzzy Spatial Topological Relationship

Existing work studies defines primitive spatial topological relations involving direction and distance information by using such terms as "right of,", "left of,", "below,", "far below,", "above,", "far above," "surround," and "inside.". The spatial topological relations are influenced by the direction and distance between two patches.

For image x_i, we first calculated the angle θ made by the line passing through the typical points' pairs $p_i(j_1,j_2)$ belonging to patch x_{ij_1} and x_{ij_2} ($j_1=1,...n_i, j_2=1,...n_i, j_1\neq j_2$). Then the dominant angle θ_d between two patches is determined according to the angle histograms of contour points in two patches. Based on the dominant angle θ_d, the membership functions of four spatial topological relationships between x_{ij_1} and x_{ij_2} are defined in Figure 6(b) according to [10]. The membership function of the fuzzy set 'surround' is given by (1) and Figure 6(c), where θ_r is the range of the angle θ. Considered the distance of the center of gravity g_{j_1}, g_{j_2}, the membership function of "far above" is given by (2). The fuzzy membership of three pairs of opposite relation is denoted as (3).

$$\mu_{surround}(\theta_r)=\begin{cases}\cos^2(\theta_r/2) \text{ if } \pi\leq\theta_r\leq 2\pi \\ 0 \quad \text{otherwise}\end{cases} \tag{1}$$

$$\mu_{far\ above}(j_1,j_2)=\begin{cases}1\wedge(1.5\times\|g_{j1}-g_{j2}\|_2/I_h) & \text{if } \mu_{above}(j_1,j_2)>0 \\ 0 & \text{otherwise}\end{cases} \tag{2}$$

$$\mu_{left}(j_1,j_2)=\mu_{right}(j_2,j_1),\ \mu_{surround}(j_1,j_2)=\mu_{inside}(j_2,j_1),\ \mu_{surround}(j_1,j_2)=\mu_{inside}(j_2,j_1) \tag{3}$$

The fuzzy spatial topological relation between patch j_1 and j_2 in image x_i is defined as: $R^{PP}_{Spatial}(j_1,j_2)=\{\mu^{PP}_{ri}(j_1,j_2)/ri,\mu^{PP}_{le}(j_1,j_2)/le,$ $\mu^{PP}_{be}(j_1,j_2)/be,\mu^{PP}_{fb}(j_1,j_2)/fb,\mu^{PP}_{ab}(j_1,j_2)/ab,\mu^{PP}_{fa}(j_1,j_2)/fa,\mu^{PP}_{su}(j_1,j_2)/su,\mu^{PP}_{in}(j_1,j_2)/in\}$ $\in F(P\times P)\text{of}P\times P\text{=}\{(\mathrm{x}_{ij},\mathrm{x}_{ij})|\mathrm{x}_{ij}\in P\wedge\mathrm{x}_{ij}\in P\}$, where the abbreviated name of each spatial topological relation is written.

3.2 Label Propagation Inter Images

As pre-processing to label propagation inter images, Bi-Layer sparse coding from [1] is utilized to construct a linear combination between the patches with the same annotation in visual feature space. After the Bi-Layer sparse coding, the linear combination coefficient $\hat{\alpha}_{i_1,j_1,i_2,j_2}$ was obtained, which denoted as linear combination relationship of the patch j_1 from image x_{i_1} to the patch j_2 from image x_{i_2} in feature space.

Label propagation inter images is based on the linear combination coefficients obtained from Bi-Layer sparse coding. The label is propagated from the candidate region to the selected patches of the remaining images and vice versa. Supposed that $\mathrm{z}_i \in \mathbb{R}^{\mathrm{n}_k}$ indicates the annotation vector, n_k is the total number of image annotations. The binary element $\mathrm{z}_i(k)$ takes 1 if the *i*th image contains the *k*th annotation and 0 otherwise. Firstly, the patch-level annotation vector $\{z_{i,j}\}$, $z_{i,j} \in \mathbb{R}^{n_k}$ is initialized with annotation vector $z_i \in \mathbb{R}^{n_k}$ of image x_i . Then for every candidate patch j_1 from image x_{i_1} , if $\hat{\alpha}_{i_1,j_1,i_2,j_2} > 0$, z_{i_2,j_2} and z_{i_1,j_1} are updated by (4), where $i_1, i_2 = 1,\ldots\mathrm{N}$; $j_1, j_2 = 1,\ldots n_i$. And β_{i_2,j_2} is the weight coefficient calculated according to the size of the j_1th atomic patch and normalized by the image size of the i_1th image.

$$z_{i_2,j_2} = z_{i_2,j_2} + \hat{\alpha}_{i_1,j_1,i_2,j_2}\ , \quad z_{i_1,j_1} = z_{i_1,j_1} + \sum_{i_2=1}^{N}\sum_{j_2=1}^{n_i}(\hat{\alpha}_{i_1,j_1,i_2,j_2} \times \hat{\beta}_{i_2,j_2}) \tag{4}$$

After label propagation inter images, the patch-level annotation vector $\{z_{i,j}\}$ is updated. Based on $\{z_{i,j}\}$, we define $w^0_{i,j,k} = z_{i,j,k} / \bigvee_{k'=1}^{n_k}(z_{i,j,k'})$, which is the initial membership value of the patch x_{ij} to label k .

3.3 Label Propagation Intra Images

By utilizing fuzzy logic reasoning, labels are propagated intra images in semantic space based on the similarity compare of contextual cueing features between the patch and concept defined from common knowledge.

In order to obtain the images from a common knowledge base for learning the contextual cueing spatial invariants, we firstly query Google Image using the annotation and pair of annotation available in the MSRC and COREL databases. Then we calculated the common fuzzy position of different objects and the fuzzy spatial topological relationship between different objects. The fuzzy membership for the common knowledge is calculated in the same way as the fuzzy membership for patches defined in 3.12. Fuzzy position of concept $R^C_{Position}(k) = \{\mu^C_{top}(k), \mu^C_{middle}(k), \mu^C_{bottom}(k)\}$ is constructed to represent the common position of concept k . Fuzzy spatial topological relationship of concept $R^{CC}_{Spatial}(k_1,k_2) = \{\mu^{CC}_{ri}(k_1,k_2)/ri, \mu^{CC}_{le}(k_1,k_2)/le, \mu^{CC}_{be}(k_1,k_2)/be, \mu^{CC}_{fb}(k_1,k_2)/fb,$ $\mu^{CC}_{ab}(k_1,k_2)/a, \mu^{CC}_{fa}(k_1,k_2)/fa, \mu^{CC}_{su}(k_1,k_2)/su, \mu^{CC}_{in}(k_1,k_2)/in\} \in F(C \times C)$ is constructed to represent the common spatial relationship between concept k_1 and k_2. Here, concept is proposed to distinguish the annotation in our fuzzy based contextual cueing model. The statistical results of the fuzzy membership is shown in Figure 7(a) and Figure 7(b). Higher the variable is, whiter the color is shown.

With the aid of knowledge about the common position and spatial topological relationships of every concept, we use fuzzy logic reasoning to compare the similarity of those two contextual cueing relationships between $R^P_{Position}, R^{PP}_{Spatial}$, $R^C_{Position}$, $R^{CC}_{Spatial}$. Algorithm 1 summarizes the label propagation inter images procedure. In our algorithm, the fuzzy membership w_{i_1,j_1,k_1} of the patch x_{ij} to label k_1 is updated by the similarity measure $S^{PC}_{Position}(j_1,k_1)$ and $S^{PC}_{Spatial}(j_1,k_1)$, which refers to the position and spatial topological relationships between patches and concepts.

```
Algorithm 1: Label Propagation Intra Image
Input: Initial membership value $w^0_{i_1,j_1,k_1}$; Fuzzy membershipvariable $R^P_{Position}, R^{PP}_{Spatial}, R^C_{Position}, R^{CC}_{Spatial}$; Iteration number $N_{max}$; the tolerance factor $\varepsilon_{\min}$; Experience parameter $\lambda$.
Output: Final fuzzy label membership vector $w^{out}_{i_1,j_1,k_1}$
for $n = 1,\ldots, N_{max}$ do
  for $j_1 = 1,\ldots, \mathrm{n}_i$ do
   for $k_1 = 1,\ldots, N_c$ do
    for $j_2 = 1,\ldots, \mathrm{n}_i$ do
    for $k_2 = 1,\ldots, N_c$ do
      $R^{PP}_{Spatial}(j_1,j_2) \circ R^{CC}_{Spatial}(k_1,k_2) = \vee[R^{PP}_{Spatial}(j_1,j_2) \wedge R^{CC}_{Spatial}(j_1,j_2)]$
      $R^P_{Position}(j_1) \circ R^C_{Position}(k_1) = \vee[R^P_{Position}(j_1) \wedge R^C_{Position}(k_1)]$
      $S^{PC}_{Position}(j_1,k_1) = [R^P_{Position}(j_1) \circ R^C_{Position}(k_1)] / \vee[R^P_{Position}(j_1) \vee R^P_{Position}(k_1)]$
      $S^{PC}_{Spatial}(j_1,j_2,k_1,k_2) = [R^{PP}_{Spatial}(j_1,j_2) \circ R^{CC}_{Spatial}(k_1,k_2)] / \vee[R^{PP}_{Spatial}(j_1,j_2) \vee R^{CC}_{Spatial}(k_1,k_2)]$
    end for
    end for
$w^n_{i_1,j_1,k_1} = \lambda \times w^{n-1}_{i_1,j_1,k_1} \times S^{PC}_{Position}(j_1,k_1) + (1-\lambda) \times w^{n-1}_{i_1,j_1,k_1} \times \frac{1}{n_{i_1}-1} \times \sum_{j_2=1, j_2 \neq j_1}^{n_{i_1}} \{\bigwedge_{k_2=1, k_2 \neq k_1}^{n_k} [w^{n-1}_{i_1,j_2,k_2} \times S^{PC}_{Spatial}(j_1,j_2,k_1,k_2)]\}$
    $w^n_{i_1,j_1,k_1} = w^n_{i_1,j_1,k_1} / \bigvee_{k_1=1}^{n_k} w^n_{i_1,j_1,k_1}$
    if $\sum_{j_1=1}^{n_i}\sum_{k_1=1}^{N_c} \| w^n_{i_1,j_1,k_1} - w^{n-1}_{i_1,j_1,k_1} \|_F < N_c \times \mathrm{n}_i \times \varepsilon_{\min}$
     $w^{out}_{i_1,j_1,k_1} = w^n_{i_1,j_1,k_1}$; break;
    end if
   end for
  end for
end for
$w^{out}_{i_1,j_1,k_1} = w^n_{i_1,j_1,k_1}$
```

3.4 Post Processing Based on KNN-FCM Clustering

In the post processing part, we use KNN-FCM [11] clustering to segment the images into semantic regions and associate with corresponding annotations. Firstly, the initial cluster centers are computed by KNN. Then FCM clustering algorithms are utilized to generate cluster $F_{k'}$ $k' = 1,\ldots K_i$ with the center $c_{ik'}$, where K_i is the number of annotations in image x_i . In the end, those patches with in a same cluster are merged to form a semantic region. The final region-level label is set as the one with the largest value in the label vector according (5).

$$l_{ij_1k_1} = \begin{cases} 1 \ \text{ if } \max\limits_{1 \le k' \le m_i}(c_{ik'}) = c_{ik_1}, \mathrm{x}_{ij} \in F_{k'} \\ 0 \qquad\qquad\qquad\qquad \text{otherwise} \end{cases} \tag{5}$$

4. EXPERIMENTS AND RESULTS

To demonstrate performance of our proposed technique, we conduct two experiments on two public datasets, MSRC and COREL Stock Photo CDS. The quantitative label-to-region assignment accuracy measures as the percentage of pixels with agreement between the assigned label and ground truth. Two kinds of techniques are compared with the proposed FCLP. One is a series of Binary Support Vector Machine (BSVM) based algorithms with different values for the parameter of maximal patch size, namely, SVM1: 150 pixels, SVM2: 200 pixels, SVM3: 400 pixels, and SVM4: 600 pixels. The BSVM is implemented based on the lib-SVM library and the Gaussian Radial Basis Function kernel is used by setting the kernel parameter as 1. The other is two latest LRA techniques of label propagation with one-layer sparse coding and bi-layer sparse coding [1]. For the proposed techniques, the parameter λ, $T_{\max}$ and $\varepsilon_{\min}$, actually shows stable performance under different values. In our experiments, we set $\lambda = 0.7$, $\varepsilon_{\min} = 0.1$, $N_{\max} = 50$, $N_s = 5$, and the dimension of the BOW feature vector m = 628, including 500 dimensions and 128 dimensions for SIFT and LAB color descriptors respectively.

4.1 MSRC Dataset

MSRC dataset contains 591 images with ground-truth of image-level annotations and region-level annotations. Similar with previous work on this dataset [1], we remove the images with only single annotation or infrequent annotation. This gives rise to 380 images with totally 18 categories: building, grass, tree, cow, boat, sheep, sky, mountain, aeroplane, water, bird, book, road, car, flower, cat, sign, and dog. Table 1 shows the accuracy comparison of a series of SVM-based algorithms, one-layer and bi-layer LRA, and our proposed technique FCLP. FCLP performs much better than all other algorithms. The detailed comparison results are illustrated in Figure 9.

Table 1. Label-to-region assignment accuracy comparison.

Dataset	SVM1	SVM2	SVM3	SVM4	One-Layer	Bi-Layer	FCLP
MSRC	0.24	0.22	0.27	0.25	0.54	0.65	**0.72**

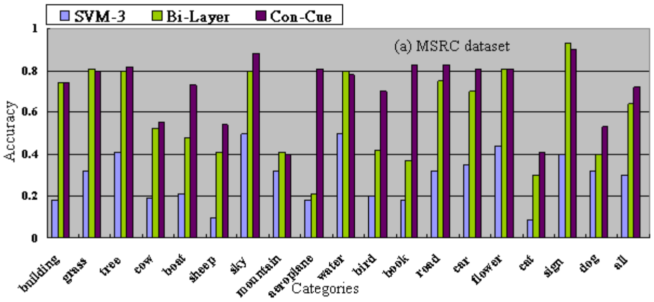


Figure 9. Detailed LRA accuracies for MSRC dataset. The horizontal axis shows the name of each label/annotation and the vertical axis represents the accuracy of LRA.

Figure 10 demonstrates some interesting observations of the performance comparison between Bi-layer technique and FCLP technique. Figure 10 (a) is the image with given image-level annotations of sky, building, tree and road. Since sky and road have very similar SIFT features, Bi-Layer technique assigns the road annotation to the sky region. Moreover, this error influences the further region segmentation and assignment as shown in Figure 10 (b). With the aid of fuzzy position and fuzzy spatial topological memberships shown in Figure 7, FCLP technique recognizes the sky region correctly. Another example is shown in Figure 10 (d), which contains sky, road, tree and car. Bi-Layer technique assigns the road annotation to the sky region again. More importantly, bi-Layer technique is less effective for handling the categories for foreground objects [1]. As shown in Figure 10(e), the car hasn't been segmented and recognized correctly because of the small size. FCLP also considers the sizes of the objects, but as shown in Figure 10 (g), the car is correctly assigned. The key is how to use contextual cueing appropriately in image understanding. As mentioned in section 2, five types of spatial invariants, such as size and position, are thought to be important in contextual cueing, so if only one or two spatial invariants are utilized, the image analysis result may over-emphasize certain aspect of contextual cueing. Moreover, in this case, fuzzy theory successfully demonstrates its effectiveness in modeling human's understandings of the visual world.

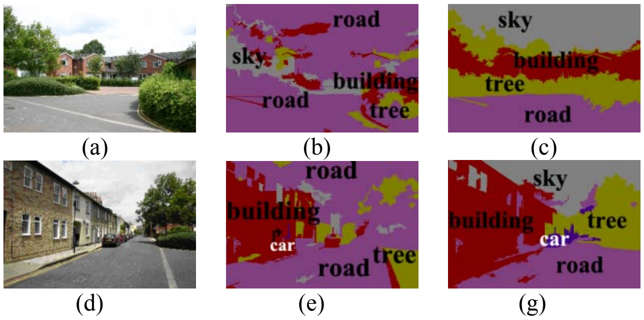


Figure 10. Comparison of LRA results. (a) An image with annotations of sky, building, tree, road (b) Bi-layer result (c) FCLP result (d) An image with annotations of sky, building, tree, car, and road. (e) Bi-layer result (f) FCLP result.

4.2 COREL Dataset

COREL dataset is the most broadly adopted dataset in the community of image retrieval. Similar with previous work in [1], we selected 150 images and manually annotate the ground-truth, which contains the 8 categories: grass, cow, mountain, sky, bear, water, tree, and building. Table 2 shows the accuracy comparison of a series of SVM-based algorithms, one-layer and bi-layer LRA, and our proposed technique FCLP. The detailed comparisons of individual objects are illustrated in Figure 11. Obviously, FCLP achieves best performance under all the cases.

Table 2. Label-to-region assignment accuracy comparison.

Dataset	SVM1	SVM2	SVM3	SVM4	One-Layer	Bi-Layer	FCLP
COREL	0.29	0.32	0.34	0.33	0.51	0.62	**0.70**

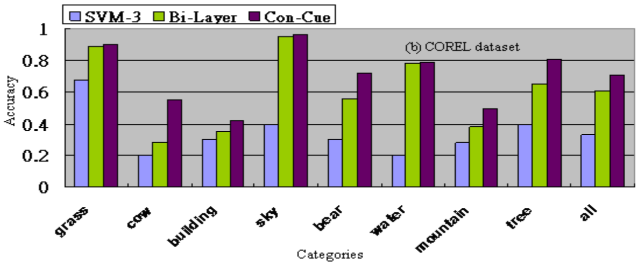


Figure 11. Detailed LRA accuracies for COREL dataset. The horizontal axis shows the name of each label/annotation and the vertical axis represents the accuracy of LRA.

Figure 12 demos some LRA results on COREL dataset, covering all eight categories of regions. Compared with previous work, our proposed technique shows much higher accuracy on the objects with relatively explicit position information or spatial topological relationship with other objects, even their appearance is similar, such as sky, airplane, road, car, and boat. Moreover, all existing techniques suffer from the performance decreasing if more objects appear in the image. But FCLP may benefit from it because more objects may provide more contextual cueing information to help the understanding of the whole image.

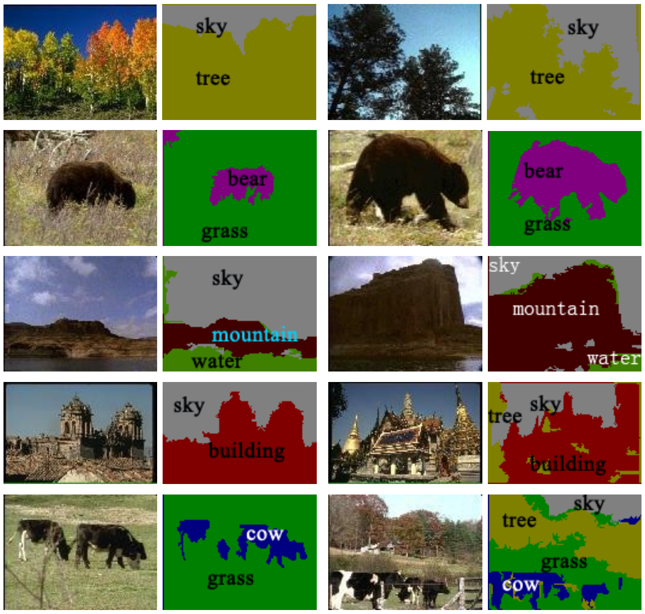


Figure 12. Examples of label-to-region assignment results. Each color in the labeled images denotes one class of localized region.

5. CONCLUSION

This paper proposed a novel FCLP technique for LRA problem to natural images. We integrate contextual cueing, a concept widely studied in psychological, to improve the semantic understanding of the images. Fuzzy representation and fuzzy reasoning are utilized to describe the contextual cueing knowledge and imitate the contextual cueing process. Moreover, FCLP inherits the merits of label propagation methods, which reduce the training cost by taking advantage of the similarity among the data with common labels. The experiments on two public datasets demonstrate that the proposed technique has shown obvious performance improvement of LRA for the images with multiple objects and complex background.

6. REFERENCES

[1] Xiaobai Liu, Bin Cheng, Shuicheng Yan, Jinhui Tang, Tat Seng Chua, Hai Jin, “Label to Region by Bi-Layer Sparsity Priors”, In *Proceedings of ACM Multimedia*, (Oct. 2009), 115-124.

[2] J. Li, R. Socher, and L. Fei-Fei, “Towards total scene understaning: classification, annotation and segmentation in an automatic framework”, In *CVPR*, 2009.

[3] Chun, M. M. & Jiang, Y.. ‚Contextual cueing: Implicit learning and memory of visual context guides spatial attention”, In *Cognit. Psychol.*, vol. 36, 1998, 28-71.

[4] Y.-G. Jiang, C.-W. Ngo, and S.-F. Chang, “Semantic context transfer across heterogeneous sources for domain adaptive video search", In *ACM Multimedia*, (Oct. 2009), 155-164.

[5] I. Biederman, R. Mezzanotte, and J. Rabinowitz, “Scene perception: detecting and judging objects undergoing relational violations”, In *Cognitve Psychology*, 14(2), 1982, 143–77.

[6] Yuan, J., Li, J., Zhang, B., “Exploiting spatial context constraints for automatic image region annotation”, In *ACMMM*, 2007, 595–604.

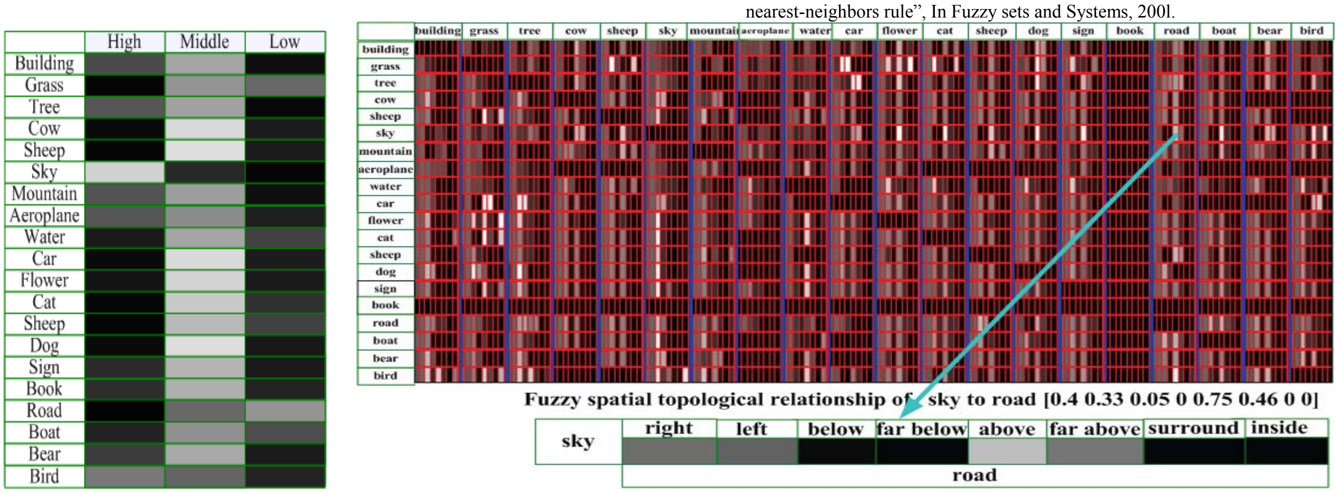
[7] C. Galleguillos, A. Rabinovich and S. Belongie, “Object categorization using co-occurrence, location and appearance”, In *CVPR*, (June. 2008).

[8] P. Felzenszwalb and D. Huttenlocher, “Efficient graph-based imagesegmentation”, In *IJCV*, 59(2), 2004, 167–181.

[9] A. Torralba, A. Oliva, M.S. Castelhano and J.M. Henderson., “Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search”, In *Psychological Review.*, 2006, 766-786.

[10] K. Miyajima and A. Ralescu, “Spatial organization in 2D images”, In *IEEE International Conference on Fuzzy Systems,*1994, 100–105.

[11] N.Zahid, 0.Abouelala, M.Limouri, A．Essaid, “Fuzzy clustering based on K-nearest-neighbors rule”, In Fuzzy sets and Systems, 200l.



(a) (b)

Figure 7. (a) Fuzzy membership of different objects types on Google database for position “top,” “middle,” “bottom.” (b) Fuzzy membership for eight different spatial relationships of different objects types on Google database. Fuzzy memberships are transformed into 0 to 255 and higher the fuzzy membership is, whiter the color is shown.