Semantically Modeling of Object and Context for Categorization

Chunjie Zhang, Jian Cheng, and Qi Tian, *Fellow, IEEE*

***Abstract*— Object-centric-based categorization methods have been proven more effective than hard partitions of images (e.g., spatial pyramid matching). However, how to determine the locations of objects is still an open problem. Besides, modeling of context areas is often mixed with the background. Moreover, the semantic information is often ignored by these methods that only use visual representations for classification. In this paper, we propose an object categorization method by semantically modeling the object and context information (SOC). We first select a number of candidate regions with high confidence scores and semantically represent these regions by measuring correlations of each region with prelearned classifiers (e.g., local feature-based classifiers and deep convolutional-neural-network-based classifiers). These regions are clustered for object selections. The other selected areas are then viewed as context areas. We treat other areas beyond the object and context areas within one image as the background. The visually and semantically represented objects and contexts are then used along with the background area for object representations and categorizations. Experimental results on several public data sets well demonstrate the effectiveness of the proposed object categorization method by semantically modeling the object and context information.**

***Index Terms*— Context modeling, object categorization, object modeling, semantic representation.**

I. Introduction

AUTOMATICALLY categorizing an image based on its content plays an important role for various visual applications [1]–[3]. The state-of-the-art methods are object-centric-based [4]–[7] which go one step beyond hard partitions of images [8]–[11]. The object-centric-based scheme tries to first locate the positions of objects and then model the objects and background accordingly. The categorization accuracy can be greatly improved by aligning the objects.

Although it is very effective, the object-centric-based scheme still has three drawbacks. First, exact localizations of objects are still very hard due to varied visual appearances and occlusions. The use of human labor with bounding box annotation can alleviate this problem. However, it costs a lot of human labor and is time consuming. Besides, different people may provide varied bounding boxes. The shortage of training data hinders automatic object detection. Moreover, there are too many objects to be efficiently detected. Recently, the objectness proposal technique [12]–[15] becomes popular. It tries to generate a number of rectangular regions instead of identifying the exact locations of objects. We can harvest the useful information from these proposal regions.

Second, visual information is often used for object-centric-based modeling. However, its performance is often hindered by visual polysemy and interclass variations. It is more effective to combine the semantic representations for classification. The semantic representation method uses the human interpretable features to represent images. Human-annotated labels can be viewed as a special case of the semantic representations. The semantic-based strategies often make use of training samples [16]–[20] or information beyond the training data [21]–[25] on the image level. However, noisy information may be introduced, because there are often multiple objects of different classes on one image. It would be more effective to model the object areas. Besides, spatially nearby areas may contain the same object and should be jointly considered. The spatial correlations and semantic representations of image areas should be jointly modeled for efficient representations.

Third, the context and background information play different roles for object categorization. To category different objects, we need to extract the context and background information accordingly [14]. However, many methods ignore this problem by modeling the context and background jointly. We use context to refer to image regions that are adjacent to the object and also have semantic correlations with the categorization task. However, the extracted context areas are often very large, while the background areas are relatively small. There are no background areas for many images. The background regions are less semantically correlated with or even hinder the categorization task. Accurately modeling both context and background is very important for efficient categorization.

Manuscript received September 26, 2017; revised January 16, 2018 and April 21, 2018; accepted July 3, 2018. Date of publication August 7, 2018; date of current version March 18, 2019. This work was supported in part by the Scientific Research Key Program of Beijing Municipal Commission of Education under Grant KZ201610005012 and in part by the National Science Foundation of China under Grant 61429201. The work of Q. Tian was supported in part by ARO under Grant W911NF-15-1-0290 and in part by the Faculty Research Gift Awards through the NEC Laboratories of America and Blippar. *(Corresponding author: Chunjie Zhang.)*

C. Zhang is with the Research Center for Brain-Inspired Intelligence, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: chunjie.zhang@ia.ac.cn).

J. Cheng is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with the University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: jcheng@nlpr.ia.ac.cn).

Q. Tian is with the Department of Computer Sciences, The University of Texas at San Antonio, San Antonio, TX 78249 USA (e-mail: qitian@cs.utsa.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2018.2856096

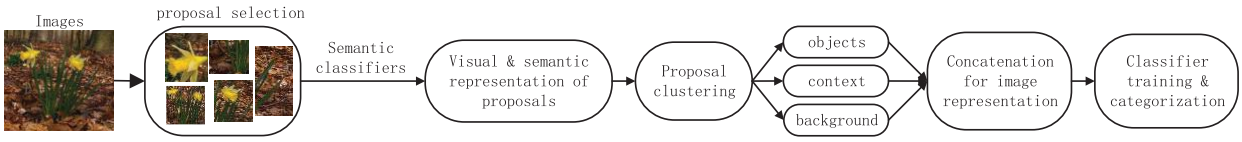

Fig. 1. Flowchart of the proposed object categorization method by semantically modeling the object and context information.

To solve the above-mentioned problems, in this paper, we propose a novel object categorization method by semantically modeling the object and context information (SOC). First, we leverage the objectness proposal technique to select the candidate regions with high confidence scores. For each candidate region, we use the visual representation to predict its semantic correlations using prelearned classifiers to semantically represent candidate regions. Instead of directly using all the candidate regions for categorization, we treat object and context areas separately. To extract object areas, we cluster candidate regions by jointly considering the content similarities and the spatial information of these regions. In this way, we can ensure that each cluster concentrates on one particular object with its surrounding information. For each cluster, we view the area of the corresponding cluster center as the object. The other areas beyond the objects are viewed as the context areas. We only use candidate regions with high confidence scores to avoid introducing too much noise. Besides, the semantic representation also helps to exclude some noisy regions. For each image, we view the other area beyond the object and context areas as the background. For each image, we concatenate the visually and semantically represented objects and contexts along with the background area into a vector as the final representation of this image. We then train linear support vector machine (SVM) classifiers to predict the categories of images. Experimental results on several public image data sets prove the effectiveness of the proposed SOC method for object categorization. Fig. 1 shows the flowchart of the proposed object categorization method by semantically modeling the object and context.

The deep neural networks can also learn semantic representations from images. However, it is different from the semantic representation used in this paper in three aspects. First, the semantic representations used in this paper are explicitly learned vectors, while the semantic representations learned in deep neural networks are matrixes implicitly learned from various layers which require further processing. Second, the semantic classifiers used in this paper can be learned with images of other sources and can also be learned with training images at the class level for each image data set. We can use various prelearned classifiers (e.g., local-feature-based classifiers [1], [8], [11], deep neural-network-based classifiers [27]–[29], [74]–[76], classifiers learned from the Internet [21], [23], [25], attribute-based classifiers [20], [22], and human-annotated labels [24], [25]) instead of only using the information learned from the deep neural networks. Third, the proposed semantic representations are used to represent objectness proposals/image regions, while the deep neural-network-based semantic representations often target the whole image.

In this paper, the object and context areas are represented both visually and semantically. This helps to bridge the semantic gap to some extent. Besides, jointly modeling the object, context, and background finely also helps to align images for effective similarity measurements. The main contributions of the proposed method lie in two aspects.

1) First, instead of only using visual representations, the proposed SOC method makes use of semantic representations which can be obtained using various prelearned classifiers. SOC can be combined with various state-of-the-art image classification methods.
2) Second, the object and context areas are selected by using candidate regions with high confidence scores. This helps to avoid introducing too much noisy information and finally improves the categorization accuracies.

The improvements of the proposed SOC method over object, context and background (OCB) [14] lie in two aspects. First, SOC goes one step further by using both visual and semantic representations of images for classification. Since the backgrounds are cluttered with various objects which do not belong to any classes to be classified, we only use the semantic representations to model the object and context information. Second, SOC extracts the object and context information more properly than OCB. This is because SOC only makes use of objectness proposals with relatively larger responses, while OCB uses other objectness proposals with less confidence. The OCB introduces more noisy information than the SOC when modeling the context information.

The rest of this paper is organized as follows. We give the related work in Section II. In Section III, we give the details of the proposed object categorization method by semantically modeling the object and context information. The experimental results and analysis on six public image data sets are given in Section IV. Finally, we conclude in Section V.

II. Related Work

Object categorization could help to improve the accuracy of many visual applications, such as image classification [1], image captioning [2], and retrieval [3]. Earlier works used hard partition strategies [1], [8]–[11]. As a typical hard partition strategy, the spatial pyramid matching (SPM) [1] was widely used. van Gemert *et al.* [8] softly assigned local features with several visual words, while Zhang *et al.* [9] tried to generate a number of codebooks instead of one codebook for representations. Yang *et al.* [11] combined sparse coding with the hard partition strategy and improved the performances. However, the hard partition strategy has divided the images using predefined rules without considering the content information of objects.

To improve the accuracy of object categorization, object-centric-based [4]–[7] methods became popular. Russakovsky *et al.* [4] tried to separate the object and background information and represented them separately. Chai *et al.* [5] leveraged the segmentation strategy to locate the objects. However, this strategy was time consuming and easily being degraded by inaccurate segmentation results. Chen *et al.* [6] combined the detection and classification task into a unified framework. Yang *et al.* [7] gradually located objects with convolutional neural networks (CNNs) to improve the performances. The object-centric-based strategy went beyond hard partition of images and concentrated on the objects to be classified.

Although it was effective, accurately detecting the objects was still very hard due to variations and large numbers of object classes. To alleviate this problem, objectness proposal techniques [12]–[15] became popular. Cheng *et al.* [12] proposed a fast objectness estimation algorithm with binary representation, while Zitnick and Dollar [13] proposed the edge box algorithm for objectness proposal selection. Zhang *et al.* [14] used these detected regions for joint object, context, and background modeling and improved the classification performances. Wei *et al.* [15] also used the proposals to solve the multilabel image classification problem. However, all of these methods have relied on visual information only which was far from satisfactory for reliable classification.

Semantic representation was also used by researchers with training samples [16]–[20] or information beyond the training data [21]–[25]. Rasiwasia and Vasconcelos [16] proposed a holistic context model to semantically represent images for classification. An exemplar classifier-based weak semantic space was constructed by Zhang *et al.* [18] and then further extended into subsemantic spaces [19]. Torresani *et al.* [20] used classesmes for object recognition. One problem with these methods was the lack of training samples. To alleviate this problem, the usage of information from other sources became another choice. Yang *et al.* [21] used Web images for video indexing with the sample-specific loss. Farhadi *et al.* [22] tried to semantically describe objects by attributes which costed human labor for annotation. Li *et al.* [23] tried to construct the object bank using the Web images with high-level semantic representation. Tang *et al.* [24] alleviated human labor for image tagging, while Russell *et al.* [25] leveraged the Web resources for semantically annotating the images.

In recent years, more and more effective visual representation methods had been proposed, e.g., Fisher vector [26] and CNN [27]. Motivated by these methods, many works [28]–[35] had been conducted which greatly improved the classification accuracy. Chatfield *et al.* [28] evaluated the details of different algorithms. Donahue *et al.* [29] proposed a deep convolutional network for generic visual recognition and greatly improved the performances. Liu *et al.* [30] proposed an elastic net hypergraph learning scheme, while Zhang *et al.* [31] fused multiple cues with deep network for event recognition. Gong *et al.* [32] tried to replace max pooling with multisclae orderless pooling in the deep neural network, while He *et al.* [33] used spatial pyramid pooling. Oquab *et al.* [34] transferred the midlevel representations with CNNs to solve the lacking of training sample problem. Wu and Ji [35] constrained the deep transfer learning process for particular applications. Many works had been done for various visual applications [36]–[67] which greatly improved the performances. Varma and Ray [44] learned the discriminative power invariance of different features. Yuan and Yan [45] minimized the sparse reconstruction error for classification. Xie *et al.* [46] used the binratio similarity, while Angelova and Zhu [48] combined the detection and segmentation for classification. Zhang *et al.* [49] used the nonnegative sparse coding and then extended with correlation constraints [50]. Both the visual and semantic correlations [51]–[55] were used with good performances. The deep convolutional network [56] extended the local-feature-based methods [57]–[59] with improved performances. Lin *et al.* [60] treated different regions discriminatively, while Razavian *et al.* [61] used CNNs. Zhou *et al.* [62] learned the deep features, while Wu and Rehg [63] designed a visual descriptor. Sadeghi and Tappen [64] targeted the scene recognition problem with latent regions. Zhang *et al.* [65], [67] used the semantic representations. However, object, context, and background were mixed together. Researchers had also explored various objectness proposal techniques for detection and classification [14], [68]–[70]. Girshick *et al.* [68] proposed novel regions with CNN features for object detection. It could also be used for the classification task. The algorithm was further extended with fast and faster region-based convolutional neural networks [69], [70] and greatly improved the performances. Qu *et al.* [71] proposed a novel joint hierarchical structure learning strategy to classify images with good improvements. Xie *et al.* [72] proposed a deformation-guided kernel regression algorithm to remove a turbulence effect.

III. Semantically Modeling of Object and Context for Object Categorization

In this section, we give the details of the proposed object categorization method by semantically modeling the object and context information (SOC).

A. Semantical Representation of Objectness Proposals

To avoid directly locating objects, we use the objectness proposal technique [12] to first extract a number of rectangular areas. The proposed method improves over [68] in two aspects. First, we use both visual and semantic representations to cope with variations of images, while [68] only uses the visual features. Second, the proposed SOC method models images with object, context, and background, while [68] uses the detected proposals for classification directly. SOC jointly combines a number of objectness proposals which is more robust to noisy detection and more discriminative than [68]. One selected area may contain an object with high probability. We then try to extract useful information from these proposal areas. Many methods try to harvest the discriminative information with visual information only. However, due to the polysemy problem of visual features, visual-based methods cannot fully separate different objects from the noisy information. We can alleviate this problem by semantically

representing these proposals. A noisy region may be misclassified by the objectness proposal algorithm. However, if we measure its semantic correlations with a number of semantics, the responses would be able to tell the differences.

We sort the object proposals using their confidence scores in the descending order. We then choose the first N proposals. Formally, let $\boldsymbol{V} = [\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_N}]$ be the visual features of N selected proposal areas with their confidence scores sorted in the descending order as $\boldsymbol{CS} = [cs_1, cs_2, \ldots, cs_N]$, where cs_1 is the highest score and cs_N is the lowest score, $\boldsymbol{v_n} \in \mathbb{R}^{M\times 1}, n = 1, \ldots, N$. $\boldsymbol{L} = [\boldsymbol{l_1}, \boldsymbol{l_2}, \ldots, \boldsymbol{l_N}]$ are the corresponding locations of proposals. We can use various visual representation schemes (e.g., Fisher vector and CNN) to improve the discriminative power. The semantic representation of $\boldsymbol{v_n}$ can be obtained by measuring its semantic correlations with the prelearned semantic classifiers $f_p(*)$, $p = 1, \ldots, P$, where P is the number of semantics (or image classes). The semantic classifiers can be learned with images of other sources and can also be learned with training images at the class level for each image data set. We use linear SVM classifiers as the semantic classifiers in this paper. Specially, we train one classifier to separate the images of the pth class from other images of different classes. Since the trained classifier tries to separate images of the same class from the other classes, if we apply it to predict the category of one test image or image area, the output would bear some semantic similarities between the test image and the particular class. In this way, we call this strategy semantic-based, as shown in [16]–[19]. Both generative classifiers and discriminative classifiers can be used for semantic classifier learning. If $\boldsymbol{sr_n}$ is the semantic representation of the nth proposal, the pth element of $\boldsymbol{sr_n}$ can then be calculated as $sr_{n,p} = f_p(\boldsymbol{v_n})$ with $\boldsymbol{sr_n} \in \mathbb{R}^{P\times 1}$ as

$$\boldsymbol{sr_n} = [f_1(\boldsymbol{v_n}); \ldots; f_P(\boldsymbol{v_n})]. \tag{1}$$

The advantages of the semantical-based representations lie in two aspects. First, instead of only using visual features, we can harvest the useful information from training images for more discriminative representations. Second, we can also get rid of some noisy proposals. This can help to increase the accuracy of subsequent object modeling process which eventually improves the categorization performances.

B. Object Area Selection

After semantically representing the proposal areas, we harvest the useful information from them for categorization. We follow the object-centric-based strategy and try to select the object areas. This is achieved by first clustering these proposal areas and then extracting the objects. We combine the visual and semantic representations of proposals for joint representations.

Formally, for the nth proposal, let $\boldsymbol{x_n} = [\boldsymbol{v_n}; \boldsymbol{sr_n}]$ be the concatenated representation. The similarity between the mth proposal and the nth proposal can then be calculated as

$$\hat{s}_{m,n} = \exp^{-\|\boldsymbol{x_m} - \boldsymbol{x_n}\|^2/\sigma} \tag{2}$$

where σ is the scaling parameter. In this way, we can jointly combine the discriminative power of visual and semantic representations for similarity comparison.

Instead of only using content information, the locations of different proposals should also be considered. In this paper, we use the Intersection over Union (IoU) scores [12], [14] to measure the location similarity of two proposal regions. The IoU score of the mth proposal and the nth proposal can be calculated as

$$\widetilde{s}_{m,n} = \frac{|\boldsymbol{l_m} \bigcap \boldsymbol{l_n}|}{|\boldsymbol{l_m} \bigcup \boldsymbol{l_n}|} \tag{3}$$

where $\bigcap$ and $\bigcup$ represent the intersection operation and union operation, respectively.

The overall similarity $s_{m,n}$ between the mth proposal and the nth proposal is obtained by combining the content similarity $\hat{s}_{m,n}$ and location neighborhood $\widetilde{s}_{m,n}$ as

$$s_{m,n} = \hat{s}_{m,n} + \alpha \times \widetilde{s}_{m,n} \tag{4}$$

where α is the parameter for balancing the relative influences of $\hat{s}_{m,n}$ and $\widetilde{s}_{m,n}$. Setting α to zero equals to only using the content information without considering the spatial locations of proposals.

We use $s_{m,n}$ to select the object areas. Instead of using k-means clustering as [14] did, we use k-medodis clustering algorithm instead. The reasons why we use k-medodis instead of k-means clustering lie in two aspects. First, k-means is more easier to be disturbed by outliers. Second, k-means uses the average of the samples within one cluster as the cluster center, while k-medodis uses one sample with the least summed distances. For our proposal clustering problem, k-medodis is more appropriate for selecting the objects. Besides, [14] used the overlapped area of the proposals within one cluster for object selection. This strategy often extracts the object with a very small size. However, by using the k-medodis clustering, we can use the cluster center for object selection directly which is more appropriate with proper coverage.

We cluster the N objectness proposals into $\widetilde{K}$ clusters using k-medodis clustering. We sort the clusters by the number of proposals assigned to this cluster. Only the first K clusters ($K < \widetilde{K}$) with the largest number of proposals are used. In this way, we can get rid of the noisy proposals for better modeling of objects. For the kth cluster, if $\hat{\boldsymbol{l}}_{\boldsymbol{k}}$ is the location of the corresponding cluster center, we then choose this region as the object area $\boldsymbol{A_k} = \hat{\boldsymbol{l}}_{\boldsymbol{k}}$. For each image, K object areas are selected. To represent the selected objects, we concatenate the corresponding visual and semantic representations in the fixed spatial order (left and right and top and down) as

$$\boldsymbol{A_o} = [\boldsymbol{A_1}; \ldots; \boldsymbol{A_K}]. \tag{5}$$

After selecting the objects, we can locate the surrounding context and background areas.

C. Context and Background Area Selection

Context information has been proven very useful for reliable categorization of different objects. To make use of the context information, we need to first determine the corresponding

locations. Although the usage of more objectness proposals is a plausible way to solve this problem, it suffers from contamination of noisy proposals as proposals with relatively lower confidence scores bear little object correspondences. If we use too many proposals, the context and background areas would be unbalanced. In fact, the proposals with high confidence scores already have enough information for reliably modeling the context information.

To extract context areas, we use the proposals for each cluster by extracting the object areas. For the kth cluster, let $\overline{\boldsymbol{l}}_1, \ldots, \overline{\boldsymbol{l}}_{P_k}$ be the locations of P_k proposals that are assigned to this cluster. We use the covered area of all the P_k proposals for context area selection by excluding the object areas. Formally, the context areas can be selected as

$$\boldsymbol{A_c} = \left(\overline{\boldsymbol{l}}_1 \bigcup \overline{\boldsymbol{l}}_2 \bigcup \ldots \bigcup \overline{\boldsymbol{l}}_{P_k}\right) - \boldsymbol{A_o}. \tag{6}$$

The $\bigcup$ operation is used to include the context information as much as possible. Since the selected proposals are of high confidence scores with the noisy proposals removed, the union areas contain the objects and context information. The context areas can be selected by subtracting the objects accordingly.

After the object and context areas are chosen, we can view the other areas within an image as the background. If $\boldsymbol{I}$ is an image, then the background area can be selected as

$$\begin{aligned} \boldsymbol{A_b} &= \boldsymbol{I} - \left(\overline{\boldsymbol{l}}_1 \bigcup \overline{\boldsymbol{l}}_2 \bigcup \ldots \bigcup \overline{\boldsymbol{l}}_{P_k}\right) \\ &= \boldsymbol{I} - \boldsymbol{A_o} - \boldsymbol{A_c}. \end{aligned} \tag{7}$$

We then combine the visual and semantic representations of objects and contexts with the background representation for joint representation. Formally, if $\overline{\boldsymbol{v}}_{o,i}, \overline{\boldsymbol{sr}}_{o,i}, i = 1, \ldots, K$ are the visual and semantic representations of the objects, $\overline{\boldsymbol{v}}_{c,i}, \overline{\boldsymbol{sr}}_{c,i}, i = 1, \ldots, K$ are the visual and semantic representations of the context areas, and $\overline{\boldsymbol{v}}_b$ is the visual representation of the background, then the final image representation $\boldsymbol{h}$ can be obtained as $\boldsymbol{h} = [\overline{\boldsymbol{v}}_{o,1}; \overline{\boldsymbol{sr}}_{o,1}; \ldots; \overline{\boldsymbol{v}}_{o,K}, \overline{\boldsymbol{sr}}_{o,K}; \overline{\boldsymbol{v}}_{c,1}; \overline{\boldsymbol{sr}}_{c,1}; \ldots; \overline{\boldsymbol{v}}_{c,K}; \overline{\boldsymbol{sr}}_{c,K}; \overline{\boldsymbol{v}}_b]$. We use both the visual and semantic representations for object and context representations in order to improve the discriminative power and avoid the influences of misclassified proposals. However, for the background, we only use the visual information instead. This is because the background areas contain various visual information; if we represent them semantically, the corresponding representation would have little discriminative power. After obtaining the concatenated representations of images, we train linear SVM classifiers to predict the categories of images.

D. Object Categorization

Suppose we have Q training images with representations and labels as $(\boldsymbol{h_q}, y_q), q = 1, \ldots, Q$ of C classes. We learn the classifiers in the one-vs-all way. For the cth class, $c = 1, \ldots, C$, we try to learn linear classifier as

$$\hat{y}_{q,c} = \boldsymbol{w}_c^T \times \boldsymbol{h}_q + b_c \tag{8}$$

where $\boldsymbol{w}_c$ and b_c are the parameters for the cth class.

To learn the classifier, we try to minimize the summed loss over training images with L_2 constraint as

$$[\boldsymbol{w}_c, b_c] = \underset{[\boldsymbol{w}_c, b_c]}{\operatorname{argmin}} \sum_{q=1}^{Q} \ell(\hat{y}_{q,c}, y_q) + \lambda\left(\|\boldsymbol{w}_c\|^2 + b_c^2\right) \tag{9}$$

where $\ell(*, *)$ is the loss function. λ is the parameter for controlling the relative influences of summed loss and the constraint. We use the hinge loss $\ell(\hat{y}_{q,c}, y_q) = \max(1 - \hat{y}_{q,c} \times y_q, 0)$ in this paper.

After the classifiers are learned, we can predict object's categories using (8) by assigning an image to the class which has the largest response with the corresponding classifier. Algorithm 1 gives the procedures of the proposed object categorization method by semantically modeling the object and context information.

Algorithm 1 Procedures of the Proposed Object Categorization Method by Semantically Modeling the Object and Context Information

Input:

Training images, selected proposals with sorted order, prelearned classifiers $f_p(*)$, σ, α, cluster number K, testing images.

Output:

The predicted categories;

1: Semantically represent the selected proposals using Eq. 1.
2: Clustering the proposals using Eq. 4 with k-medodis clustering.
3: Select the object areas using Eq. 5.
4: Extract the context and background areas using Eq. 6 and Eq. 7 respectively.
5: Learn the classifiers for image class prediction by optimizing over Problem 9.
6: Predict the categories of images using Eq. 8.
7: **return** The predicted categories of images.

The computational complexity of detecting objectness proposals is $O(Q)$, where Q is the number of training images. The main complexity cost of object, context, and background selection lies in the clustering process with (4) whose computational complexity is of $O(N\widetilde{K}t)$, where N is the number of proposals with relatively high scores, $\widetilde{K}$ is the cluster number, and t is the number of iterations. Usually, $\widetilde{K}$ and t are much smaller than N. The computational complexity of training classifiers in (8) is $O(QC)$. Hence, the overall computational complexity of the proposed SOC method is $O(QC+NKt)$. After the classifiers are learned, we can predict its class using (8). For a testing image, we need less than 1 s for prediction when tested on a desktop computer using Intel Core i7-6700@3.4 GHz and 16-GB RAM.

IV. Experiments

To evaluate the effectiveness of the proposed object categorization method by semantically modeling the object and context information, we conduct object categorization experiments on the Flower-17 data set [37], the Flower-102 data set [38],

TABLE I
CLASSIFICATION RATE COMPARISON ON THE FLOWER-17 DATA SET

Algorithm	Performance
LR-GCC [9]	91.52± 1.24
OCB-FV [14]	97.84± 0.78
ObjectBank [23]	80.90
Nilsback [37]	71.76± 1.76
Varma [44]	82.55± 0.34
KMTJSRC-CG [45]	88.90± 2.30
mTDP [46]	94.89± 0.90
SOC-FV	98.23± 0.75
SOC	98.76± 0.64

the Caltech-101 data set [39], the Caltech-256 data set [40], the MIT-Indoor data set [41], and the PASCAL VOC 2007 data set [42]. We directly compare with the performances reported by other researchers using the same experimental setup for fair comparisons. We use the features extracted using the CNNs [27] for visual representations and train the classifiers for semantic modeling. The pretrained classifiers are learned from training images of the corresponding image data set. These classifiers are trained to separate images of different classes apart. In other words, the pretrained classifiers are data-set-dependent. We use a linear SVM classifier in this paper. Image completion [43] is used to cope with nonrectangular regions. A classification rate is used for quantitative comparison on the five data sets except on the PASCAL VOC 2007 data set. For the PASCAL VOC 2007 data set, the average precision (AP) is used for performance evaluation.

A. Flower-17 Data Set

The Flower-17 data set has a total of 1360 images with 80 images for each class. There are 17 classes of images (*Buttercup, Colts foot, Daffodil, Daisy, Dandelion, Fritillary, Iris, Pansy, Sunflower, Windflower, Snowdrop, Lily valley, Bluebell, Crocus, Tigerlily, Tulip*, and *Cowslip*). We follow the experimental setup as [37] and use 40/20/20 images for train/validate/test, respectively. We repeat the selection process for 10 times. The performance comparisons with other methods are given in Table I.

We can have three conclusions from Table I. First, compared with hard-partition-based methods [9], [37], [44], [45], the usage of object-centric representation is more effective to align objects for efficient similarity measurements. Besides, when compared with the other object-based representation method [14], the usage of semantic-based representation can increase the discriminative power. In this way, the proposed SOC is able to improve over [14]. Moreover, SOC is also able to improve over task-specific-based method [46]. The experimental results on the Flower-17 data set show the effectiveness of the proposed method.

To show the per class accuracy, we also give the boxplot of the performance of the proposed SOC method on each class of the Flower-17 data set in Fig. 2. The proposed SOC method is able to achieve more than 96% accuracy for the hardest class. Most of the flowers are centered on the images, which can be efficiently clustered and represented by the proposed method.

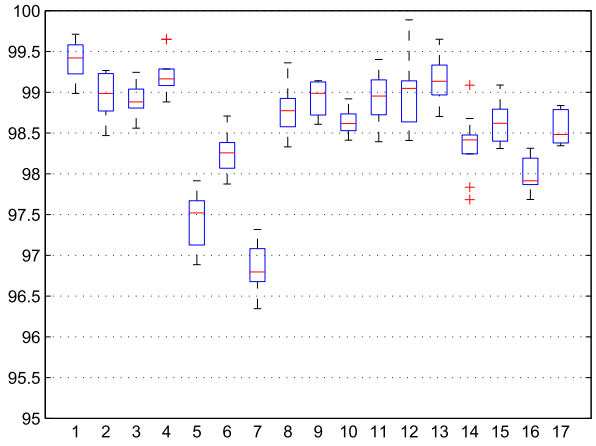

Fig. 2. Boxplot of per class performances of SOC on the Flower-17 data set. The numbers from 1 to 17 in the horizontal row represent Buttercup, Colts foot, Daffodil, Daisy, Dandelion, Fritillary, Iris, Pansy, Sunflower, Windflower, Snowdrop, Lily valley, Bluebell, Crocus, Tigerlily, Tulip, and Cowslip, respectively.

TABLE II
MEAN CLASSIFICATION RATE COMPARISONS ON THE FLOWER-102 DATA SET

Methods	Classification rate
TriCoS [5]	85.2
LR-GCC [9]	75.7
OCB-FV [14]	91.3
Nilsback [38]	72.8
KMTJSRC-CG [45]	74.1
Hu [47]	86.8
Det+Seg [48]	80.7
SOC-FV	92.6
SOC	94.5

B. Flower-102 Data Set

This data set is an extended data set of the Flower-17 data set with 102 classes of 8189 flower images; 10/10/rest images are used for train/validate/test, respectively, as in [38]. We give the classification rate comparisons on the Flower-102 data set in Table II.

We can have similar conclusions from Table II as in the Flower-17 data set. On one hand, by representing objects directly instead of hard partition, we can model the images more efficiently. On the other hand, the selection of objectness proposals is also more effective than the segmentation-based schemes [5], [48]. Besides, by combining the semantic-based representation, we can improve over the visual-based method [14]. As there are more classes of images with similar visual appearances in the Flower-102 data set than the Flower-17 data set, the relative improvements are also larger because of the consideration of semantics and the more efficient modeling of object and context information.

C. Caltech-101 Data Set

This data set has a total of 9144 images of 101 classes. There are 31–800 images for different classes. We randomly select 15/30 training images and use the rest images for testing. We repeat this random selection process for 10 times to get

TABLE III
MEAN CLASSIFICATION RATE COMPARISONS ON THE CALTECH-101 DATA SET

Methods	15 images	30 images
SPM [1]	56.40	64.40 ± 0.80
KC [8]	−	64.14 ± 1.18
ScSPM [11]	67.00 ± 0.45	73.20 ± 0.54
OCB-CNN [14]	84.51 ± 0.84	95.05 ± 0.92
DECAF [29]	−	84.77 ± 0.70
KMTJSRC-CG [45]	65.00 ± 0.70	−
mTDP [46]	−	93.68 ± 0.50
VGG [56]	−	92.70 ± 0.50
NBNN [57]	65.00 ± 1.14	70.40
SOC-FV	74.83 ± 0.52	77.59 ± 0.66
SOC	86.13 ± 0.95	96.87 ± 0.83

reliable results. Table III gives the categorization performance comparisons with other baseline methods.

Compared with the traditional SPM-based representation strategy, directly modeling objects is very useful for reliable representations. Hence, SOC improves over [1], [8], [11] dramatically. Besides, SOC also outperforms NBNN [57] which treats visual information independently without training classifiers. This indicates that the location information also plays an important role for efficient categorization. Moreover, the usage of deeper network for image modeling is more effective than that of shallow models. Hence, CNN-based methods are able to outperform the local-feature-based methods. By modeling the semantic representations along with the visual information, SOC can improve the performances over CNN-based methods [29], [56].

D. Caltech-256 Data Set

The Caltech-256 data set has 29 780 images of 256 classes; 15, 30, 45, and 60 images per class are randomly selected for classifier training. The other images are used for testing. We repeat the random selection process for 10 times and give the performance comparisons with other baseline methods in Table IV. We also give the performance of the proposed method when combined with sparse-coding-based visual representations. Note that other more discriminative representation methods can also be used to get the semantic representations.

We can see from Table IV that the proposed SOC method is able to improve categorization performances on the Caltech-256 data set. Especially, by extracting CNN-based visual features instead of local features, we can get more reliable semantic classifiers for modeling. Since images of the Caltech-256 data set have larger visual variations than that of the Caltech-101 data set, the semantic-based representations are more useful to separate different proposals and can get more representative object and context areas. Besides, it helps to get rid of noisy proposals. Moreover, compared with other semantic-based methods [19], [20], [23], the proposed SOC method represents objects more accurately. The use of more discriminative visual features also improves the performances.

We do use the bounding box annotation for objectness proposal detection. Some baseline methods (e.g., TriCoS [5], OCB [14], and Det + Seg [48]) have also used the bounding box information for detection and classification. Other methods use information from other sources for classification. For example, ObjectBank [23] harvests the semantic representations of images from the Internet, and Classesmes [20] also uses the trained categories selected from an ontology of visual concepts from other sources. The proposed SOC method improves over these baseline methods even when auxiliary information is used.

E. MIT-Indoor Data Set

This data set has 15 620 images of 67 classes. We follow the data split setup as in [41]. Especially, 80 images per class are used for training, while the other images are used for testing. The performance comparison is given in Table V.

As this data set consists of images of indoor scenes, the interclass variation is relatively larger. Hence, only using visual information cannot achieve satisfactory results. Besides, since objects are cluttered, simply dividing images with a predefined strategy also introduces noisy information. Hence, the object-centric-based modeling scheme is more appropriate whose effectiveness can be seen from Table V. However, to generate semantically discriminative representations, we need to harvest information from training images. Automatically collected images from the Internet are contaminated with noise. This is why SOC is able to outperform ObjectBank [23] dramatically. Moreover, the proposed SOC is able to outperform many CNN-based methods [62], [63] by incorporating semantical information and directly modeling objects. The results on the MIT-Indoor data sets again show the effectiveness of the proposed SOC method.

F. PASCAL VOC 2007 Data Set

There are twenty classes (*aeroplane, bicycle, boat, bottle, bus, bird, car, cat, cow, chair, dining table, dog, horse, person, sheep, motorbike, train, potted plant, soft*, and *tv/monitor*) of more than 10 000 images with the predefined train/validate/test splits. We first learn the classifiers and select the parameters with the train and validate split. These images are then merged to retrain the classifier with the selected parameters. We apply learned classifiers for performance evaluations on the test split. Table VI gives the AP performances on the PASCAL VOC 2007 data set. We also give the performances of the proposed method (SOC using VGG network) when combined with VGG [56].

The PASCAL VOC 2007 data set is very difficult to categorize, because objects are cluttered and occluded. Besides, different objects have varied sizes, which also increases the difficulty of accurate recognition. Moreover, modeling of nonrigid objects is more difficult than that of rigid objects. The object-centric-based strategy can cope with these problems to some extent. We use the objectness proposals instead of locating objects directly to avoid introducing extra noise. Since objects in this data set often appear together, the usage of semantic representation can help to model them better. The semantic representation of one particular proposal with two or more objects has positive responses of the corresponding semantics. In this way, we are able to improve

TABLE IV
PERFORMANCE COMPARISONS ON THE CALTECH-256 DATA SET

Methods	15 images	30 images	45 images	60 images
KC [8]	–	27.17 ± 0.46	–	–
LR-GCC [9]	39.21 ± 0.48	45.87 ± 0.41	–	–
SPM [11]	23.34 ± 0.42	29.51 ± 0.52	–	–
ScSPM [11]	27.73 ± 0.51	34.02 ± 0.35	37.46 ± 0.55	40.14 ± 0.91
OCB-SC [14]	40.56 ± 0.53	49.87 ± 0.50	54.18 ± 0.49	57.60 ± 0.51
OCB-FV [14]	44.03 ± 0.46	53.15 ± 0.44	57.84 ± 0.40	59.03 ± 0.45
S^3R [19]	37.85 ± 0.48	43.52 ± 0.44	46.86 ± 0.63	–
classesmes [20]	–	36.00	–	–
ObjectBank [23]	–	39.00	–	–
FV[26]	38.5	47.4	52.1	54.80 ± 0.40
SPM [40]	–	34.10	–	–
VGG [56]	–	–	–	86.20 ± 0.30
NBNN [57]	35.20	42.80	–	–
LScSPM[58]	30.00 ± 0.14	35.74 ± 0.10	38.54 ± 0.36	40.32 ± 0.32
SOC-SC	49.85 ± 0.57	57.18 ± 0.49	61.12 ± 0.55	63.84 ± 0.43
SOC-FV	54.62 ± 0.71	61.85 ± 0.63	65.87 ± 0.56	66.17 ± 0.68
SOC	73.45 ± 0.78	82.61 ± 0.64	85.34 ± 0.58	87.85 ± 0.55

TABLE V
MEAN CLASSIFICATION RATE COMPARISONS ON THE MIT-INDOOR DATA SET

Methods	Classification rate
SPM [1]	34.40
OCB-SC [14]	52.60
OCB-CNN [14]	75.84
S^3R [19]	40.10
ObjectBank [23]	37.60
Gong [32]	68.90
Quattoni [41]	26.50
mTDP [46]	75.61
Doersch [59]	66.87
Lin [60]	68.50
Razavian [61]	69.00
Zhou [62]	70.08
CENTRIST [63]	36.90
LPR-LIN [64]	44.84
SOC-FV	58.37
SOC	79.63

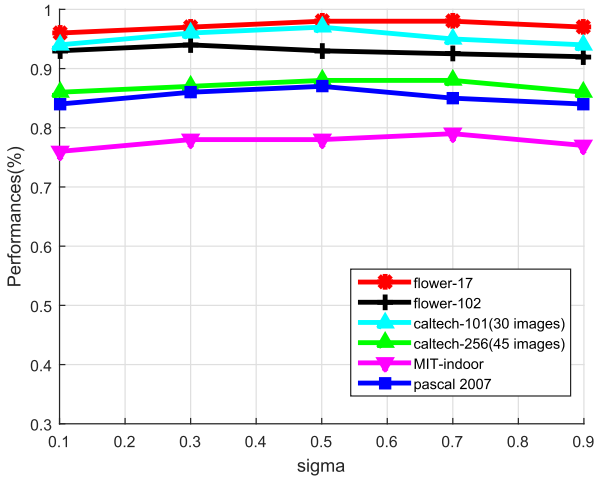

Fig. 3. Influences of σ on the Flower-17 data set, the Flower-102 data set, the Caltech-101 data set (30 images), the Caltech-256 data set (45 images), the MIT-Indoor data set, and the PASCAL VOC 2007 data set.

over OCB [14]. Besides, by leveraging the more efficient CNN-based visual features, we can further improve the performances by semantically representing the objects and contexts for categorization. Moreover, SOC improves over OCB [14] with larger improvements on nonrigid objects than on rigid objects. This is because we can get more representative object and context areas for modeling over [14]. By representing images more discriminatively using [56] than [28], we can further improve the classification performances.

G. Influences of Parameters

The parameters σ and α play important roles for object categorization. To quantitatively evaluate their influences, we plot the performance influences of σ and α in Figs. 3 and 4, respectively.

We can see from Fig. 3 that the influence of σ is relatively stable for different data sets. We believe this is because the joint usage of visual and semantic representations is robust to variations. As long as the relative similarities of different proposals are preserved, we can get discriminative object and context areas. In other words, the final SOC-based representation is not very sensitive to σ, which shows the robustness of the proposed SOC method.

We can see from Fig. 4 that the influence of α is larger than that of σ. The spatial information and the content information play different roles for reliable object area selections. For one particular data set, if objects often appear at different places, a smaller α value is more appropriate than a larger one and vice versa. However, the usage of spatial and content information does help to improve the categorization performances.

The number of selected proposals N also affects the categorization performance. We give its influences on the six data sets in Fig. 5. The performances increase with proposal numbers at first. However, if we use too many proposals, the performance will decrease. This is because it incorporates more noisy information with the increment of N. Besides, we need relatively more proposals for the data sets with larger interclass variations than other data sets. The influences of

TABLE VI
PERFORMANCE COMPARISONS ON THE PASCAL VOC 07 DATA SET

object class	Best07[42]	FV[26]	DeCAF[29]	CNN[28]	PRE[34]	SPM[4]	OCP[4]	OCB-FV[14]	OCB-CNN[14]	VGG[56]	SOC-FV	SOC	SOC-VGG
airplane	77.5	80.0	87.4	95.3	88.5	72.5	74.2	82.1	96.2	–	85.5	97.3	99.5
bicycle	63.6	67.4	79.3	90.4	81.5	56.3	63.1	69.3	91.7	–	71.7	95.2	98.1
bird	56.1	51.9	84.1	92.5	87.9	49.5	45.1	54.4	93.3	–	57.5	96.7	97.8
boat	71.9	70.9	78.4	89.6	82.0	63.5	65.9	72.6	90.1	–	74.8	94.8	97.5
bottle	33.1	30.8	42.3	54.4	47.5	22.4	29.5	34.2	56.8	–	38.6	68.1	75.4
bus	60.6	72.2	73.7	81.9	75.5	60.1	64.7	73.8	82.3	–	76.1	82.5	88.9
car	78.0	79.9	83.7	91.5	90.1	76.4	79.2	81.7	91.9	–	83.9	94.6	95.2
cat	58.8	61.4	83.7	91.9	87.2	57.5	61.4	64.5	92.4	–	66.2	95.4	96.0
chair	53.5	56.0	54.3	64.1	61.6	51.9	51.0	58.3	66.2	–	61.9	74.9	79.6
cow	42.6	49.6	61.9	76.3	75.7	42.2	45.0	53.6	78.5	–	56.3	81.4	87.3
table	54.9	58.4	70.2	74.9	67.3	48.9	54.8	60.7	76.3	–	65.3	82.3	89.6
dog	45.8	44.8	79.5	89.7	85.5	38.1	45.4	47.2	91.1	–	49.8	92.8	94.1
horse	77.5	78.8	85.3	92.2	83.5	75.1	76.3	81.4	94.7	–	82.2	96.6	97.2
motorbike	64.0	70.8	77.2	86.9	80.0	62.8	67.1	72.3	89.2	–	76.5	91.5	93.5
person	85.9	85.0	90.5	95.2	95.6	82.9	84.4	86.9	95.6	–	89.1	96.9	97.6
plant	36.3	31.7	51.1	60.7	60.8	20.5	21.8	35.4	62.4	–	39.5	70.3	76.7
sheep	44.7	51.0	73.8	82.9	76.8	38.1	44.3	53.8	84.1	–	58.8	85.1	89.4
sofa	50.9	56.4	57.0	68.0	58.0	46.0	48.8	58.7	70.4	–	61.5	75.6	80.2
train	79.2	80.2	86.4	95.5	90.4	71.7	70.7	82.6	96.2	–	84.1	97.5	97.8
tv	53.2	57.5	68.0	74.4	77.9	50.5	51.7	61.1	77.8	–	62.3	80.9	84.8
mAP	59.4	61.7	73.4	82.4	77.7	54.3	57.2	64.2	83.9	89.3	66.5	87.5	90.8

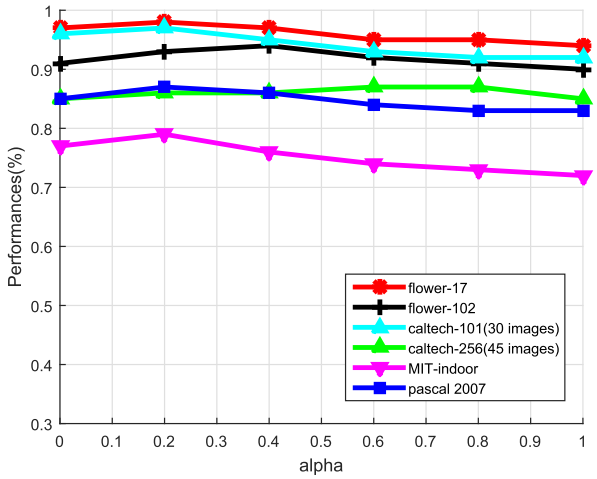

Fig. 4. Influences of α on the Flower-17 data set, the Flower-102 data set, the Caltech-101 data set (30 images), the Caltech-256 data set (45 images), the MIT-Indoor data set, and the PASCAL VOC 2007 data set.

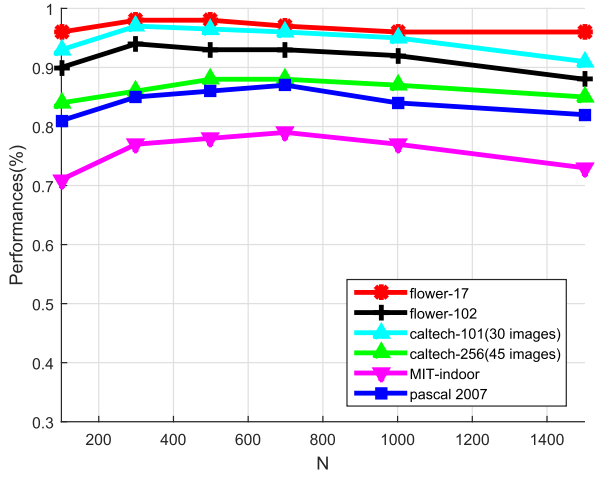

Fig. 5. Influences of the number of proposals (N) on the Flower-17 data set, the Flower-102 data set, the Caltech-101 data set (30 images), the Caltech-256 data set (45 images), the MIT-Indoor data set, and the PASCAL VOC 2007 data set.

the cluster numbers K are also given in Fig. 6. We can see that the performances are stable as long as K is not set to a small value. Besides, we need a larger K value for the data sets with larger variations.

H. ILSVRC 2012 Data Set

Finally, we test the proposed SOC method on the ILSVRC 2012 data set [73] which has 1000 classes of images. This data set contains more images than the other data sets used in this paper. We follow the same experimental setup as in [74] and use the 1.28 million images to train the model and the 100k images for performance evaluation. We combine the proposed SOC method with the ResNet-101 [74] by using it as the initial visual representations. The top-1 and top-5 error rates are used for quantitative evaluation of the performances. The error rate comparisons are given in Table VII. We can see from Table VII that the proposed SOC method improves over VGG and ResNet when combined with them for visual and semantic representations. The semantic representations can make use of various visual features for more discriminative representations of images. Besides, the improvement of SOC-VGG over VGG is larger than that of SOC-ResNet-101 over ResNet-101. It becomes harder to improve the performances when a relatively lower error rate has been achieved.

The improvements of the proposed method vary on different data sets with different baseline methods for two reasons. First, the difficulties of different data sets are varied. The PASCAL VOC 2007 data set and the ILSVRC 2012 data set are more difficult to classify than the Flower and Caltech data sets. Second, the performances of different baseline methods also vary from each other. It is relatively harder to make improvements over well-performed baseline methods

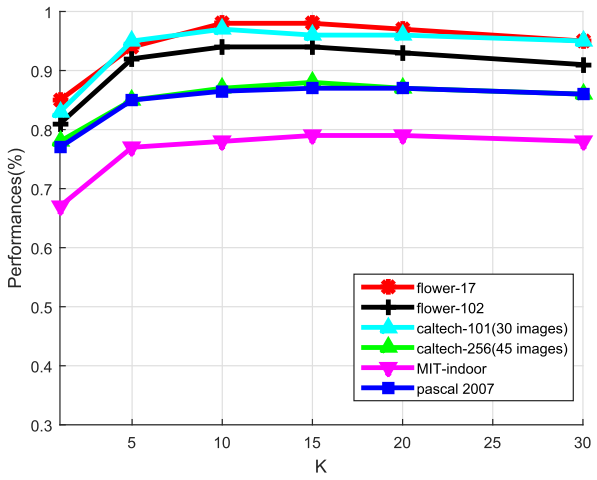

Fig. 6. Influences of the number of cluster centers (K) on the Flower-17 data set, the Flower-102 data set, the Caltech-101 data set (30 images), the Caltech-256 data set (45 images), the MIT-Indoor data set, and the PASCAL VOC 2007 data set.

TABLE VII
ERROR RATE COMPARISONS ON THE ILSVRC 2012 DATA SET

Methods	top-1	top-5
DeCAF [29]	-	19.2
Zeiler & Fergus [75]	37.5	16.1
VGG [56]	24.4	7.1
GoogLeNet [76]	-	6.7
ResNet-101 [74]	19.9	4.6
SOC-VGG	23.1	6.5
SOC-ResNet-101	19.3	4.4

(e.g., ResNet GoogLeNet and VGG). However, the proposed method does improve over these methods. This proves the effectiveness of the proposed methods.

V. CONCLUSION

In this paper, we proposed a novel object categorization method by semantically representing the object and context information. Objectness proposals were extracted and represented both visually and semantically. The semantic representations were obtained using prelearned classifiers. The prelearned classifiers can be trained using training images or information from other sources. Proposals with high confidence scores were used to avoid noise contamination. The selected proposals were then clustered. For each cluster, we used cluster center and other proposals assigned to this cluster for object and context area selection. The other areas beyond the object and context were regarded as the background. We concatenated the object and context along with the background information for representations and categorizations. Experimental results on several public image data sets well demonstrated the effectiveness of the proposed SOC method.

REFERENCES

[1] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 2169–2178.

[2] J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: Fully convolutional localization networks for dense captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4565–4574.

[3] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell, "Natural language object retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4555–4564.

[4] O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei, "Object-centric spatial pooling for image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 1–15.

[5] Y. Chai, E. Rahtu, V. S. Lempitsky, L. Van Gool, and A. Zisserman, "TriCoS: A Tri-level class-discriminative co-segmentation method for image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 794–807.

[6] Q. Chen, Z. Song, J. Dong, Z. Huang, Y. Hua, and S. Yan, "Contextualizing object detection and classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 1, pp. 13–27, Jan. 2015.

[7] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "CRAFT objects from images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 6043–6051.

[8] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283, Jul. 2010.

[9] C. Zhang, C. Liang, L. Li, J. Liu, Q. Huang, and Q. Tian, "Fine-grained image classification via low-rank sparse coding with general and class-specific codebooks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 7, pp. 1550–1559, Jul. 2017.

[10] C. Zhang, Q. Huang, and Q. Tian, "Contextual exemplar classifier-based image representation for classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 8, pp. 1691–1699, Aug. 2017.

[11] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1794–1801.

[12] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3286–3293.

[13] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.

[14] C. Zhang, G. Zhu, C. Liang, Y. Zhang, Q. Huang, and Q. Tian, "Image class prediction by joint object, context, and background modeling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 2, pp. 428–438, Feb. 2018.

[15] Y. Wei *et al.*, "HCP: A flexible CNN framework for multi-label image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1901–1907, Jun. 2015.

[16] N. Rasiwasia and N. Vasconcelos, "Holistic context models for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 902–917, May 2012.

[17] N. Rasiwasia and N. Vasconcelos, "Scene classification with low-dimensional semantic spaces and weak supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Anchorage, AK, USA, Jun. 2008, pp. 1–6.

[18] C. Zhang, J. Liu, Q. Tian, C. Liang, and Q. Huang, "Beyond visual features: A weak semantic image representation using exemplar classifiers for classification," *Neurocomputing*, vol. 120, pp. 318–324, Nov. 2013.

[19] C. Zhang *et al.*, "Object categorization in sub-semantic space," *Neurocomputing*, vol. 142, pp. 248–255, Oct. 2014.

[20] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *Proc. Eur. Conf. Comput. Vis.*, Crete, Greece, 2010, pp. 776–789.

[21] Y. Yang, Z.-J. Zha, Y. Gao, X. Zhu, and T.-S. Chua, "Exploiting Web images for semantic video indexing via robust sample-specific loss," *IEEE Trans. Multimedia*, vol. 16, no. 6, pp. 1677–1689, Oct. 2014.

[22] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 1778–1785.

[23] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1378–1386.

[24] J. Tang, Q. Chen, M. Wang, S. Yan, T.-S. Chua, and R. Jain, "Towards optimizing human labeling for interactive image tagging," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 9, no. 4, p. 29, Aug. 2013.

[25] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and Web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 157–173, 2008.

[26] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the Fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, 2013.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 25th Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[28] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. BMVC*, 2014, pp. 1–12.

[29] J. Donahue *et al.*, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. I-647–I-655.

[30] Q. Liu, Y. Sun, C. Wang, T. Liu, and D. Tao. (2016). "Elastic net hypergraph learning for image clustering and semi-supervised classification." [Online]. Available: https://arxiv.org/abs/1603.01096

[31] X. Zhang *et al.*, "Deep fusion of multiple semantic cues for complex event recognition," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1033–1046, Mar. 2016.

[32] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 392–407.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[34] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1717–1724.

[35] Y. Wu and Q. Ji, "Constrained deep transfer feature learning and its applications," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5101–5109.

[36] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, 2006, pp. 801–808.

[37] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 1447–1454.

[38] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. ICCVGIP*, Dec. 2008, pp. 722–729.

[39] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop*, Jun./Jul. 2004, p. 178.

[40] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," CalTech, Pasadena, CA, USA, Tech. Rep. CNS-TR-2007-001, 2007.

[41] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 413–420.

[42] M. Everingham, A. Zisserman, C. Williams, and L. Van Gool, "The PASCAL visual object classes challenge 2007," Pascal Challenge, U.K., Tech. Rep., 2007.

[43] I. Drori, D. Cohen-Or, and H. Yeshurun, "Fragment-based image completion," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 303–312, 2003.

[44] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *Proc. 11th IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[45] X.-T. Yuan and S. Yan, "Visual classification with multitask joint sparse representation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3493–3500.

[46] G.-S. Xie, X.-Y. Zhang, X. Shu, S. Yan, and C.-L. Liu, "Task-driven feature pooling for image classification," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1179–1187.

[47] W. Hu *et al.*, "Bin ratio-based histogram distances and their application to image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 12, pp. 2338–2352, Dec. 2014.

[48] A. Angelova and S. Zhu, "Efficient object detection and segmentation for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 811–818.

[49] C. Zhang, J. Liu, Q. Tian, C. Xu, H. Lu, and S. Ma, "Image classification by non-negative sparse coding, low-rank and sparse decomposition," in *Proc. CVPR*, Jun. 2011, pp. 1673–1680.

[50] C. Zhang, J. Liu, C. Liang, Z. Xue, J. Pang, and Q. Huang, "Image classification by non-negative sparse coding, correlation constrained low-rank and sparse decomposition," *Comput. Vis. Image Understand.*, vol. 123, pp. 14–22, Jun. 2014.

[51] C. Zhang, J. Cheng, and Q. Tian, "Incremental codebook adaptation for visual representation and categorization," *IEEE Trans. Cybern.*, vol. 48, no. 7, pp. 2012–2023, Jul. 2018.

[52] C. Zhang, J. Cheng, and Q. Tian, "Multiview label sharing for visual representations and classifications," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 903–913, Apr. 2018.

[53] C. Zhang, J. Cheng, and Q. Tian, "Structured weak semantic space construction for visual categorization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3442–3451, 2018, doi: 10.1109/TNNLS.2017.2728060.

[54] C. Zhang, J. Cheng, C. Li, and Q. Tian, "Image-specific classification with local and global discriminations," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2017.2748952.

[55] C. Zhang, J. Sang, G. Zhu, and Q. Tian, "Bundled local features for image representation," *IEEE Trans. Circuits Syst. Video Technol.*, to be published, doi: 10.1109/TCSVT.2017.2694060.

[56] K. Simonyan and A. Zisserman. (2015). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[57] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[58] S. Gao, I. W.-H. Tsang, and L.-T. Chia, "Laplacian sparse coding, hypergraph Laplacian sparse coding, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 92–104, Jan. 2013.

[59] C. Doersch, A. Gupta, and A. A. Efros, "Mid-level visual element discovery as discriminative mode seeking," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 494–502.

[60] D. Lin, C. Lu, R. Liao, and J. Jia, "Learning important spatial pooling regions for scene classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3726–3733.

[61] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 512–519.

[62] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 487–495.

[63] J. Wu and J. M. Rehg, "CENTRIST: A visual descriptor for scene categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1489–1501, Aug. 2011.

[64] F. Sadeghi and M. Tappen, "Latent pyramidal regions for recognizing scenes," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 228–241.

[65] C. Zhang, G. Zhu, Q. Huang, and Q. Tian, "Image classification by search with explicitly and implicitly semantic representations," *Inf. Sci.*, vol. 376, pp. 125–135, Jan. 2017.

[66] C. Zhang, J. Cheng, and Q. Tian, "Image-level classification by hierarchical structure learning with visual and semantic similarities," *Inf. Sci.*, vol. 422, pp. 271–281, Jan. 2018.

[67] C. Zhang, J. Cheng, L. Li, C. Li, and Q. Tian, "Object categorization using class-specific representations," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2017.2757497.

[68] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[69] R. Girshick, "Fast R-CNN," in *Proc. ICCV*, Dec. 2015, pp. 1440–1448.

[70] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NIPS*, 2015, pp. 91–99.

[71] Y. Qu *et al.*, "Joint hierarchical category structure learning and large-scale image classification," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4331–4346, Sep. 2017.

[72] Y. Xie, W. Zhang, D. Tao, W. Hu, Y. Qu, and H. Wang, "Removing turbulence effect via hybrid total variation and deformation-guided kernel regression," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4943–4958, Oct. 2016.

[73] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[74] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun. 2016, pp. 770–778.

[75] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. ECCV*, 2014, pp. 818–833.

[76] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. CVPR*, 2015, pp. 1–9.

Chunjie Zhang received the B.E. degree from the Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, China, in 2006, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2011.

He was an Engineer with the Henan Electric Power Research Institute, Henan, China, from 2011 to 2012. He held a post-doctoral position at the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing. He then joined the School of Computer and Control Engineering, University of Chinese Academy of Sciences, as an Assistant Professor. He is currently an Assistant Professor with the Institute of Automation, Chinese Academy of Sciences. His current research interests include image processing, machine learning, pattern recognition, and computer vision.

Jian Cheng received the B.S. and M.S. degrees from Wuhan University, Wuhan, China, in 1998 and 2001, respectively, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2004.

From 2004 to 2006, he held a post-doctoral position at Nokia Research Center, Beijing. He has been with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, since 2006. His current research interests include machine learning methods and their applications for image processing and social network analysis.

Qi Tian (F'15) received the B.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 1992, the M.S. degree in ECE from Drexel University, Philadelphia, PA, USA, in 1996, and the Ph.D. degree in ECE from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, in 2002.

He was a tenured Associate Professor from 2008 to 2012 and a tenure-track Assistant Professor from 2002 to 2008. He is currently a Full Professor with the Department of Computer Science, The University of Texas at San Antonio (UTSA), San Antonio, TX, USA. His research projects are funded by ARO, NSF, DHS, Google, FXPAL, NEC, SALSI, CIAS, Akiira Media Systems, HP, Blippar, and UTSA. He has published over 390 refereed journal and conference papers. His current research interests include multimedia information retrieval, computer vision, pattern recognition, and bioinformatics.

Dr. Tian received the 2010 Google Faculty Award, the 2010 ACM Service Award, the 2014 Research Achievement Award from the College of Science, UTSA, the 2016 UTSA Innovation Award, and the 2017 UTSA President's Distinguished Award for Research Achievement. He is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the *ACM Transactions on Multimedia Computing, Communications, and Applications*, and *Multimedia System Journal*.