Saliency-Aware Nonparametric Foreground Annotation Based on Weakly Labeled Data

Xiaochun Cao, *Senior Member, IEEE*, Changqing Zhang, Huazhu Fu, Xiaojie Guo, *Member, IEEE*, and Qi Tian, *Senior Member, IEEE*

***Abstract*— In this paper, we focus on annotating the foreground of an image. More precisely, we predict both image-level labels (category labels) and object-level labels (locations) for objects within a target image in a unified framework. Traditional learning-based image annotation approaches are cumbersome, because they need to establish complex mathematical models and be frequently updated as the scale of training data varies considerably. Thus, we advocate the nonparametric method, which has shown potential in numerous applications and turned out to be attractive thanks to its advantages, i.e., lightweight training load and scalability. In particular, we exploit the salient object windows to describe images, which is beneficial to image retrieval and, thus, the subsequent image-level annotation and localization tasks. Our method, namely, saliency-aware nonparametric foreground annotation, is practical to alleviate the full label requirement of training data, and effectively addresses the problem of foreground annotation. The proposed method only relies on retrieval results from the image database, while pretrained object detectors are no longer necessary. Experimental results on the challenging PASCAL VOC 2007 and PASCAL VOC 2008 demonstrate the advance of our method.**

***Index Terms*— Foreground annotation, nonparametric, saliency aware, weakly labeled.**

I. Introduction

THE GOAL of image annotation is to predict categories for the objects appearing in the target image, while object localization aims at predicting locations for these objects. In this paper, we concentrate on the problem of foreground annotation, which combines the above two goals. Recently, there has been a significant progress in fully supervised recognition. Usually, the discriminative or generative models are trained on some specific training data. Research efforts along this line allow immediate application of a variety of sophisticated machine learning techniques to learn a series of optimal classifiers or detectors from the training data. However, they have to predefine a few categories and train their models in advance. In most cases, the training process must be repeated if new training samples or new categories are added. To some extent, these approaches are closed-universe ones [1], which are difficult to adapt to new instances or categories.

To break out of the closed universe, plenty of nonparametric methods [1]–[4] have been proposed. Typically, they first retrieve the most similar images from a large database and then transfer the desired information from the retrieved images to the query, instead of training models on specific data in advance. We call these methods using large database data-driven nonparametric methods. Although the data-driven nonparametric method currently offers the most promise for image annotation in large-scale, dynamic data sets, one main problem holds it back from being widely applied. That is the dependence on fully labeled training image data, which is expensive to obtain. The manual annotation of objects in large image sets is tedious and unreliable. Alternatively, it is much easier to obtain sufficient weakly labeled data from only image-level annotations, i.e., category labels. In this paper, we are interested in predicting image-level labels and object locations using weakly labeled images in a data-driven nonparametric approach. The training data are regarded weakly labeled for localization task, when training images only contain the image-level labels of objects of interest.

Furthermore, most existing nonparametric methods usually employ global image descriptors [1]–[3] to promote the image retrieval performance. As a result, they can obtain a retrieval set with a similar scene to the target image. However, the object-level information is not addressed in these methods. Recently, to consider the location information, several techniques have been proposed. Russakovsky *et al.* [5] design an approach called object-centric spatial pooling (OCP). The OCP makes use of the object location to represent foreground–background features. Liu *et al.* [6] use a feature mining procedure to discover the input-image specific, salient and descriptive features, called label-specific features,

Manuscript received September 27, 2014; revised June 17, 2015 and September 10, 2015; accepted September 28, 2015. Date of publication October 26, 2015; date of current version May 16, 2016. This work was supported in part by the National High-Tech Research and Development Program, China, under Grant 2014BAK11B03, in part by the National Natural Science Foundation of China under Grant 61422213 and Grant 61332012, in part by the National Basic Research Program of China under Grant 2013CB329305, in part by the 100 Talents Programme through the Chinese Academy of Sciences, and in part by the Strategic Priority Research Program through the Chinese Academy of Sciences under Grant XDA06010701. *(Corresponding author: Changqing Zhang.)*

X. Cao is with the School of Computer Science and Technology, Tianjin University, Tianjin 300072, China, and also with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China (e-mail: caoxiaochun@iie.ac.cn).

C. Zhang is with the School of Computer Science and Technology, Tianjin University, Tianjin 300072, China (e-mail: cq.max.zhang@gmail.com).

H. Fu is with the Ocular Imaging Department, Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore 138668 (e-mail: huazhufu@gmail.com).

X. Guo is with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China (e-mail: xj.max.guo@gmail.com).

Q. Tian is with the Department of Computer Science, 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.2015.2488637

2162-237X © 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

TABLE I
OVERVIEW OF NONPARAMETRIC METHODS FOR IMAGE ANNOTATION

Methods	Matching	Input	Label of training data	Output label
[2], [8]	Global	Image	Full label	Region level
[1], [3], [9]	Global + Region	Image	Full label	Region level
[6]	Region	Image + Keywords	Weak label	Region level
[10]	Global	Image	Weak label	Image level
[11]	Global	Image + Keywords	Weak label	Image level
[12]	Global + Geotag	Image + Geotag	Full label	Image level
[13]	Object	Image + Keywords	Full label	Region level
Our method	Global + Object	Image	Weak label	Window level

for each label. These methods consider the object-level information. For the weakly supervised setting, object localization is much more difficult for lacking the locations of objects in training images. In our saliency-aware model, we incorporate the salient object detection technique [7], which can automatically generate object windows. Then, the detected salient object windows are utilized to describe images, and thus, the retrieved images, windows, and the inferred image labels contribute to the final object localization.

The contributions of this paper are highlighted as follows.

1) We propose a saliency-aware nonparametric foreground annotation (SANFA) approach based on weakly labeled training data. Unlike those approaches requiring object-level or pixel-level labels, the images in our database are only weakly labeled to indicate what kinds of objects are contained.
2) Our method focuses on the foreground annotation that simultaneously predicts the image-level foreground annotation and provides a window-level label for each tagged object.
3) We exploit the saliency detection technique to connect three related tasks (i.e., image retrieval, image annotation, and object localization) within a unified framework.

II. RELATED WORK

1) Image Annotation: The goal of the conventional automatic image annotation is to assign a target image, and several relevant keywords that reflect its visual content. In our method, the component of image-level annotation has the similar purpose. The existing methods usually define a parametric [14], [15] or nonparametric [16], [17] model to capture the relationship between image features and keywords. In general, nonparametric methods usually assume different nonparametric density representations of the joint word-image space and achieve robust performance. However, in practice, the complexity of the kernel density representation limits its applicability to large-scale data. Our method comes into the domain of nonparametric. However, our approach provides a meaningful foreground label using the objectness cue without estimating kernel density.

2) Object Localization: Most existing object localization approaches are based on machine learning techniques [18]–[22], which usually localize objects with a sliding window. The classifiers have to be evaluated over a large set of candidate windows. Hence, they are computationally expensive. The mostly related method [21] tries to select one window per image containing instances of a given object category. The candidate windows are generated according to objectness scores [23], so the computation is significantly reduced. Our method also transforms the localization problem to the window selection one, but differs in the following aspects. First, object categories are not given as priors but predicted by our system. Second, our method can handle the target images in an online manner. It is not necessary to feed a bunch of target images to the system. Finally, our method is nonparametric rather than learning based.

3) Generic Object Detection: Many works [24]–[26] in the field of recognition only rely on low-level cues. These methods would be improved with considering that the object-level information is helpful to recognition. Kuettel *et al.* [13] verify that learning to properly localize the objects holds great promise for boosting the classification accuracy. In addition, some methods [7], [23], [27]–[31] propose to detect salient objects automatically, without any prior knowledge about their categories, shapes, or sizes. We should note the following aspects. First, saliency is based on human vision system to discovery salient region using low-level bottom-up features. It produces a pixel-level probability map without object concept. Objectness not only contains the saliency role, but also has the object completeness cue, which is helpful to generate the meaningful bounding box for object detecting. Second, unlike some other saliency object extraction methods [29], [31], [32] that aim to segment the salient object from an image in pixel level, the methods used in our framework generate windows that tend to contain objects.

4) Nonparametric Approach: Nonparametric approaches have shown the potential in a wide spectrum of applications. The method introduced in [10] uses the image-level labeled data. However, it only annotates images with keywords. Some other methods can provide the pixel-level annotation, but require the data with more precise information, typically region labels [1], [2] or geographical labels [12]. The requirement of fully labeled data limits the utilization of rich weakly labeled images. The method in [6] parses a single image with image-level label annotations into localized semantic regions by enforcing the auxiliary knowledge on raw outputs of the Web image search. Although the data used in this paper is weakly labeled, the target image has to be annotated with image-level labels. Our method exploits the content-based cue by object windows. It is similar to the method in [13], but our method only depends on the image level instead of the region-level labeled data. A more detailed overview of nonparametric methods is given in Table I.

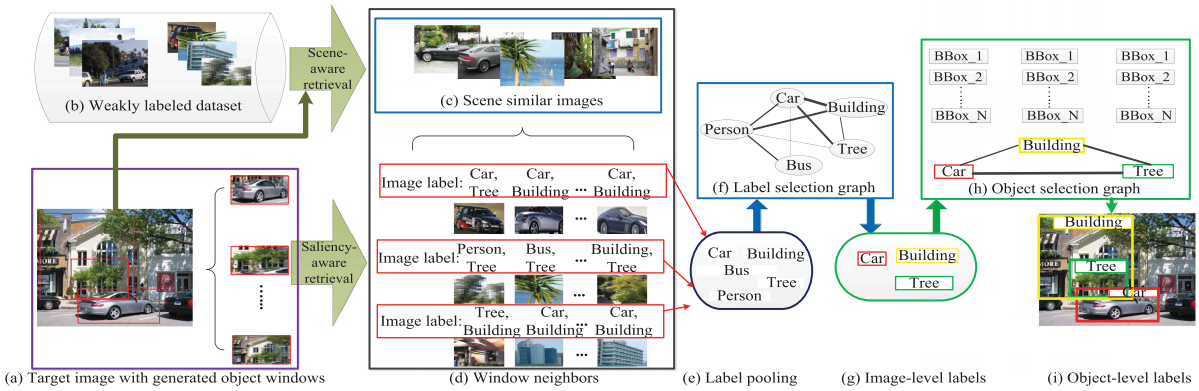

Fig. 1. Framework of our method. Given a (a) target image as query with generated windows, from the (b) weakly labeled database, we retrieve its (c) K_I scene similar images. We call this step scene-aware retrieval. For each window in the target image, we find its (d) K_W nearest neighboring windows from those scene similar images. Then, we compute the per label likelihood score to generate the (e) label pool. Based on the label pool, we construct the (f) label-selection graph and get the (g) image-level annotation. Finally, we construct the (h) window-selection graph to obtain (i) object-level annotation.

5) Weakly Supervised Localization: Weakly supervised learning [33], [34] is important due to the plentiful auxiliary data and its effectiveness. Many weakly supervised localization methods have been proposed; however, most of them [34], [35] are applied to simple data sets, such as Caltech04 [36] or Weizmann horses [37]. A few attempts [22], [38]–[40] have been made to learn models for a larger number of categories on more challenging data sets. The methods [34], [38] use segmentations for weakly supervised object localization. However, the segmentation suffers from a high commutation cost. Therefore, they conduct the method on relatively small data sets. The method [22] utilizes deformable parts model (DPM), which is trained without any detailed object level or ROI annotation. It is suitable for the problem of weakly supervised object localization as well. The model drift detection method [38] detects and stops the iterative learning when the detector starts to drift away from the objects of interest. Note that, all of these methods are learning based.

III. Framework of Our Work

Fig. 1 shows our framework. The target image does not have any annotated information, say neither image-level labels nor object locations. The reference images are with only image-level tags standing for the categories of objects within the images. Given an image, we predict both the image-level and object-level annotations, namely, image tags and object locations. The procedure of our method is as follows.

1) A collection of object windows are sampled on the target image [Fig. 1(a)]. The same offline processing is also carried out on the weakly labeled training data set [Fig. 1(b)].
2) The content-based retrieval is performed in a coarse to fine way. First, a relatively small set of K_I images with similar global scene is constructed by searching on the database [Fig. 1(c)]. The windows corresponding to these images are pooled. For each window in the target image, its K_W nearest neighbors are obtained from the pooled windows [Fig. 1(d)]. Based on these neighboring windows, the likelihoods with respect to each category are calculated [Fig. 1(e)]. Then, we obtain the image-level annotation based on a label-selection graph [Fig. 1(f) and (g)].
3) With the image-level labels, we construct a window-selection graph with each predicted label acting as a node [Fig. 1(h)]. Finally, in Fig. 1(i), the object-level annotation results are determined. Note that, our database is an open universe, and hence, the new images can be freely added into data set in an online manner.

Our workable framework exploits several basic observations as follows.

1) Images with similar scenes tend to contain similar foregrounds. This motivates us to perform the global scene matching.
2) Object window detection can discover the most foregrounds in an image, which guarantees the foreground retrieval performance.
3) Objects of the same category are likely to have consistent or similar appearance. This is a well-known principle, and thus, label transfer is reasonable.
4) A window has a higher probability objectness score if it contains an foreground object, and hence, we make use of the objectness score to measure the foreground window. Based on these observations, we perform image foreground annotation based on retrieving images in both the image level and object level.

A. Hierarchical Retrieval

To provide more accurate retrieval results, we simultaneously consider global scene-level and object-level cues. First, our system performs the scene-level matching with global image descriptors, followed by the object-level matching with the object window. For clarification, we list the definitions of the main notations used in this paper in Table II.

1) Scene-Aware Retrieval: Similar to the existing nonparametric methods [8], [41], [42], our goal is to find a set of related images from the database, each of which is desired to be similar to the query. Typically, it can alleviate the uncertainty and cut the computational cost. The relatively small retrieval set, denoted by $\mathbb{I}_{\text{retr}}$, will serve as the source

TABLE II
NOTATIONS AND DEFINITIONS

Notation	Definition
$\mathbb{I}$	The set of all training images
$\mathbb{I}_{\text{retr}}$	The retrieval set from $\mathbb{I}$ with global scene matching
N_D	The number of windows sampled on each training image
N_T	The number of windows sampled on the target image
K_I	The number of nearest neighboring images for the target image
K_W	The number of nearest neighboring windows for each window of the target image
$\text{bb}(I)$	The function which returns the window set sampled in image I
$\text{lab}(I)$	The function which returns the labels for image I

of being matched by the object windows. Moreover, a good retrieval set should consist of images with the similar type of scene to the target image, and more importantly, as well as with similar objects. In this step, we follow the setting in [1], which employs global image features, including spatial pyramid [43], GIST [44], tiny image [45], and color histogram. We increasingly sort the images in database according to the Euclidean distance for the query with respect to each global image descriptor. Finally, we take the minimum of perfeature ranks to combine the final top nearest neighbors, which achieves better results than averaging the ranks.

2) Saliency-Aware Retrieval: With the retrieval results obtained from global scene matching, the rerank process is based on salient object window matching or foreground matching. For each training image, as well as the target image, a number of windows that probably contain objects of unknown categories are generated through [7]. Such windows are generic for they are not associated with any category label. Each window has an objectness score to indicate the probability of containing an object. Any existing salient object window generation method could be incorporated into our framework. We employed the method [7] in our framework, since it achieves much better performance than other methods [23], [30], and the advantage is more obvious, especially when we select a small number of the top salient objectness windows, as shown in Fig. 2. Though there are also some region proposal methods [46], [47] that can give pixel-level object candidates, most of them are computationally expensive.

We rerank the retrieval set $\mathbb{I}_{\text{retr}}$ by object-level matching. In particular, we sample N_D and N_T windows for each the training image and the target image, respectively. As a result, we have two sets of windows for the training image and the target image. First, for each window of the target image, we retrieve its K_W nearest neighboring windows from the pooled window set sampled on the retrieved images. Afterward, $K_W \times N_T$ windows are pooled. According to the retrieval result, the object-level similarity between the target image I_t and the retrieved image I_i is defined as

$$\text{sim}_I(I_t, I_i) = \sum_{j=1}^{N_T} \sum_{k=1}^{K_W} h\left(w_j, w_j^k\right) \cdot \delta\left(w_j^k, I_i\right) \tag{1}$$

where the function $h(w_j, w_j^k)$ measures the contribution to the similarity of image pair I_t and I_i by the window w_j^k. The operator $\delta(w_j^k, I_i)$ acts as a filter, which indicates that whether

(a) (b) (c)

Fig. 2. Comparison of different salient object window generation methods. We select the top five salient object windows for each method. (a) [7]. (b) [23]. (c) [30].

the window w_j^k is contained in image I_i. It is defined as

$$\delta\left(w_j^k, I_i\right) = \begin{cases} 1, & \text{if } w_j^k \in \text{bb}(I_i) \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The function $\text{bb}(I_i)$ denotes the window set sampled in the image I_i. We define the function $h(\cdot, \cdot)$ according to the objectness score and the appearance similarity

$$h\left(w_j, w_j^k\right) = s_j \cdot s_j^k \cdot \text{sim}_B\left(w_j, w_j^k\right) \tag{3}$$

where w_j is the jth window of the target image and w_j^k is the kth neighboring window of w_j. The corresponding objectness scores of w_j and w_j^k are s_j and s_j^k, respectively. Therefore, the contribution to the image pair similarity of a window pair is determined by two factors, i.e., the objectness score and the appearance similarity. We employ Gaussian kernel to convert distance to similarity as follows:

$$\text{sim}_B(w_1, w_2) = \exp(-\lambda ||\mathbf{x}_1 - \mathbf{x}_2||^2) \tag{4}$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are the descriptor vectors corresponding to the windows w_1 and w_2, respectively, and λ is a tunable parameter. According to (1), we can see that a higher $\text{sim}_I(I_t, I_i)$ score reflects that the image I_i is more similar to the target image I_t in object level. The hierarchical retrieval ensures that the neighboring images are similar to the target image both in scene and object levels.

B. Image-Level Annotation

The image-level annotation is based on image retrieval results, which aims to predict the categories of the objects

contained in the target image. The task is similar to the traditional automatic image annotation. However, we expect to find foreground categories instead of image-level categories. Based on the retrieval feedback, the label-selection graph is constructed to obtain the image-level annotation result.

1) Label-Selection Graph Construction: To get the image-level annotation, we introduce a label-selection graph based on the undirected graphical model (UGM). Let $\mathbb{L} = \{L_1, L_2, \ldots, L_M\}$ denote a set of M labels, and $\mathbf{l} = (l_1, l_2, \ldots, l_M) \in \{0, 1\}^M$ is a discrete state vector with the state l_i corresponding to the label L_i. Each variable l_i of $\mathbf{l}$ acts as a node in the label-selection graph. If the corresponding label is selected, then $l_i = 1$, otherwise $l_i = 0$. The configuration of $\mathbf{l}$ for the target image I_t is computed with a maximum *a posteriori* (MAP) estimation using the following conditional log-likelihood:

$$\log P(\mathbf{l}|I_t) = \sum_{i=1}^{M} \psi_u(l_i \mid I_t) + \sum_{i<j} \psi_p(l_i, l_j \mid I_t) - Z(I_t) \quad (5)$$

where Z is the normalization term independent to $\mathbf{l}$. To measure the probability of the target image containing an object of the specific label L_i, we define a minus unary potential function $\psi_u(\cdot)$ as

$$\psi_u(l_i \mid I_t) = \begin{cases} \dfrac{u(L_i \mid I_t) \cdot \eta(L_i)}{\max\left(\{u(L_j \mid I_t) \cdot \eta(L_j)\}_{j=1}^{M}\right)}, & \text{if } l_i = 1 \\ 1 - \dfrac{u(L_i \mid I_t) \cdot \eta(L_i)}{\max\left(\{u(L_j \mid I_t) \cdot \eta(L_j)\}_{j=1}^{M}\right)}, & \text{otherwise} \end{cases} \quad (6)$$

where $\eta(\cdot)$ is the label confidence function [a detailed explanation will be given later in (12)] and $\max(\{u(L_j \mid I_t) \cdot \eta(L_j)\}_{j=1}^{M})$ is a normalization term for the minus unary potential. The term $u(L_i \mid I_t)$ reflects the contribution of all the neighboring windows to the label L_i, which is defined as

$$u(L_i \mid I_t) = \sum_{j=1}^{N_T} \sum_{k=1}^{K_W} h\left(w_j, w_j^k\right) \cdot \delta\left(w_j^k, L_i\right) \quad (7)$$

$$\delta\left(w_j^k, L_i\right) = \begin{cases} 1, & \text{if } w_j^k \in \text{bb}(I) \text{ and } L_i \in \text{lab}(I) \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

where the term $\delta(w_j^k, L_i)$ indicates whether label L_i is present in the image I (from the retrieval set), which contains the window w_j^k. The term $\text{lab}(I)$ denotes the labels of image I. For the purpose of smoothness, a simple label compatibility function is given based on the Potts model $p(l_i, l_j) = [l_i = l_j]$. It favors the case that high cooccurrence labels are assigned to the same state. It is defined as

$$\psi_p(l_i, l_j) = (1 - p(l_i, l_j)) \cdot \varphi(l_i, l_j). \quad (9)$$

The term $\varphi(l_i, l_j)$ measures the cooccurrence frequency between labels L_i and L_j, which is defined as

$$\varphi(l_i, l_j) = \frac{\sum_{I_k \in \mathbb{I}_{\text{retr}}} \text{v}(L_s, I_k)}{|\mathbb{I}_{\text{retr}}|} \quad (10)$$

where $L_s = \{L_i, L_j\}$ and $\text{v}(\cdot, \cdot)$ is defined as

$$\text{v}(L_s, I_k) = \begin{cases} 1, & \text{if } L_s \subset \text{lab}(I_k) \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

We can now transform the task of image annotation to the framework of standard undirected graph model through maximizing the log-likelihood $\log P(\mathbf{l}|I_t)$. Note that, compared with the existing methods, both the unary and pairwise terms are calculated in a nonparametric approach.

2) Label Confidence Level: Actually, the data set is biased between the presence and the absence of tags. For instance, "person" is frequently involved in images together with other tags. Conversely, some tags may seldomly appear together with other tags. In real cases, the cooccurrence degrees of different tags vary considerably, which influences the performance of the weakly labeled training data setting. The frequently cooccurring tags tend to have higher minus unary potential, as the score may be contributed by the matching of the other categories contained in images. To alleviate the influence, we define the label confidence level function $\eta(\cdot)$, which is inversely proportional to the label cooccurrence degree

$$\eta(L_i) = \frac{\sum_{m=1}^{M} P(L_i, m) \cdot C(L_i, m)}{\sum_{m=1}^{M} C(L_i, m)} \quad (12)$$

where $C(L_i, m)$ denotes the number of images in the retrieval set with the label L_i and containing m labels in total. The function $P(L_i, m)$ measures the contribution to minus unary potential $\psi_u(l_i)$ of an image containing L_i with m labels. In our experiments, we define $P(L_i, m) = 1/m$, which indicates that the contribution corresponding to the label L_i is inversely proportional to the image label number. We normalize the original minus unary potential based on $\eta(L_i)$, as shown in (6).

Note that, the retrieval set is refined by the label-constraint retrieval. In particular, we obtain the label scores (minus unary potential) using the original retrieval set $\mathbb{I}_{\text{retr}}$. Then, C (e.g., $C = M/2$) labels with top minus unary potentials are selected. Afterward, we retrieve a more accurate image set from the training database that excludes the images without labels belonging to the top C labels. Then, the label-constraint retrieval set is obtained. We call this method SANFA with label-constraint (SANFA_{LC}).

C. Object Localization

After obtaining the image-level label, we make further efforts to predict locations for objects. In general, the task of object localization is much more challenging than image-level annotation, even with a large amount of annotated images available for training detectors. Most existing approaches are based on object detectors trained beforehand, which usually suffer from several issues as follows.

1) The performance depends on the models and training data.
2) It is not easy to extend the existing models to new categories. That is to say, if there is a new class, they need to train a new detector.

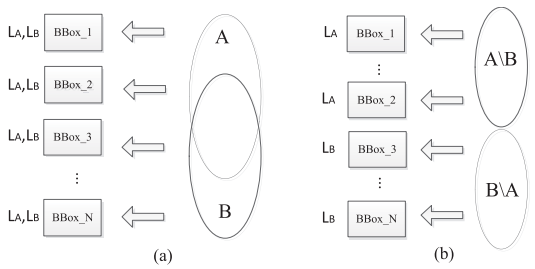

Fig. 3. (a) Overlapping sets. (b) Label-constraint sets. There are two sets $\mathbb{I}_A$ and $\mathbb{I}_B$, standing for two image subsets that contain labels L_A and L_B, respectively. With the label-constraint set, the more accurate information can be transferred to the object windows and, thus, the target image.

3) They often suffer from high computation cost due to sliding window way, although the efficient subwindow search [48] speeds up this significantly.

To address these limitations, we introduce the nonparametric approach and saliency detection technique. We construct a fully connected undirected graph, called window-selection graph, which aims to select one window for each predicted category label. We accomplish this by optimizing an energy function that is defined globally over the predicted labels. Ideally, the optimal configuration is obtained when all selected windows contain an object instance accurately and with correct category labels.

1) Label Constrained Retrieval: We select a window for each predicted label L_i. Typically, the selected window should match with the retrieval windows that contain objects of the category L_i. However, for the weakly labeled data setting, it remains challenging to determine the category of the object, since an image may be labeled with multiple labels. Therefore, to alleviate the uncertainty, we refine the image retrieval set by enforcing the label constraint, as shown in Fig. 3. The label constrained images in set $\mathbb{I}_{\mathrm{LC}}$ are not the single-labeled images. They may have multiple labels. However, for each selected image, only one label from the predicted image-level label set can be contained. Accordingly, we denote the label constrained image set as

$$\mathbb{I}_{\mathrm{LC}} = \bigcup_{L_A \in \mathbb{L}_P} \left(\mathbb{I}_{L_A} - \bigcup_{L_B \in \mathbb{L}_P} \left(\mathbb{I}_{L_A} \bigcap \mathbb{I}_{L_B} \right) \right) \tag{13}$$

where $\mathbb{L}_P$ is the predicted label set. Note that, the label-constrained retrieval here is different from the case in image-level annotation step, although they both aim to reduce the uncertainty.

2) Window-Selection Graph Construction: Similar to the label-selection graph, we employ the UGM to model the window-selection graph. However, there are some main differences between the label-selection graph and the window-selection graph as follows. First, the candidate label set is the predicted labels $\mathbb{L}_P = \{L_1, L_2, \ldots, L_P\}$, instead of the entire label set $\mathbb{L}$ as in the label-selection graph. Typically, $\mathbb{L}_P$ is a relatively small subset of $\mathbb{L}$ with corresponding state vector $\mathbf{l}_P$, and each variable l_i of vector $\mathbf{l}_P$ acts as a node of the window-selection graph. Second, the state space turns out to be $\mathbb{W} = \{w_1, \ldots, w_{N_T}\}$ instead of a binary. It indicates which window is selected for a predicted label. The configuration of $\mathbf{l}_P$ for the target image is obtained with an MAP estimation of the following conditional log-likelihood:

$$\log P(\mathbf{l}_P | I_t) = \sum_{i=1}^{P} \phi_u(l_i | I_t) + \sum_{i<j} \phi_p(l_i, l_j | I_t) - Z(I_t) \tag{14}$$

where $Z(I_t)$ is independent to $\mathbf{l}_P$. The minus unary potential function ϕ_u encodes the log-likelihood of the variable l_i taking some state, namely, how likely one window contains an object of the category L_i. The definition of ϕ_u is

$$\phi_u(l_i = w_j | I_t) = s_j \cdot \sum_{k=1}^{K_W} s_j^k \cdot \mathrm{sim}_B\left(w_j, w_j^k\right) \cdot \delta\left(w_j^k, L_i\right) \tag{15}$$

where the minus pairwise potential function ϕ_p favors the lowly overlapping windows

$$\phi_p(l_i = w_{l_i}, l_j = w_{l_j} | I_t) = -\frac{\mathrm{area}\left(w_{l_i} \bigcap w_{l_j}\right)}{\mathrm{area}\left(w_{l_i} \bigcup w_{l_j}\right)} \tag{16}$$

where the term area $(\cdot)$ denotes the area of a part of an image. The penalty, constructed from the measure of area overlap used in the VOC challenges [49], is employed in our method. Compared with other possible object localization metrics [50], this formulation is simple and has several favorable properties, e.g., invariance to scale and translation.

Note that, for both the image-level and object-level annotations, our model connects the nodes (labels) of the target image, rather than other elements within the image like the typical cases of conditional random fields (CRFs) (e.g., pixels in segmentation [51], body parts in human pose estimation [52], and windows from a group of images [21]). In this paper, the graph is small, so we can effectively optimize the problem in the framework of undirected graph model by using exact inferences (e.g., graphCut [53]).

IV. Experiments

We conduct our experiments on three related image-processing tasks, including image retrieval, image annotation, and object localization to investigate the proposed method.

A. Experimental Settings

The experiments are conducted on PASCAL VOC 2007 and PASCAL VOC 2008 [49]. Each data set includes 20 categories. The first one contains 9963 images in total, and in our experiments, we adopt the standard train (5011 images)/test (4952 images) split. For PASCAL VOC 2008, the original standard training data contains some images from the PASCAL 2007 test data, and we eliminate these images and use the left (3040) images as training data. We also use the test data in PASCAL VOC 2007 as our test data in PASCAL VOC 2008, because there are no new datasets with the object location groundtruth. We compare our method with several related methods on PASCAL VOC 2007 but not on PASCAL VOC 2008 for the localization task, because there are seldom results reported on this data set. Similar to [1], we use the spatial pyramid (three levels, SIFT dictionary of size 200)

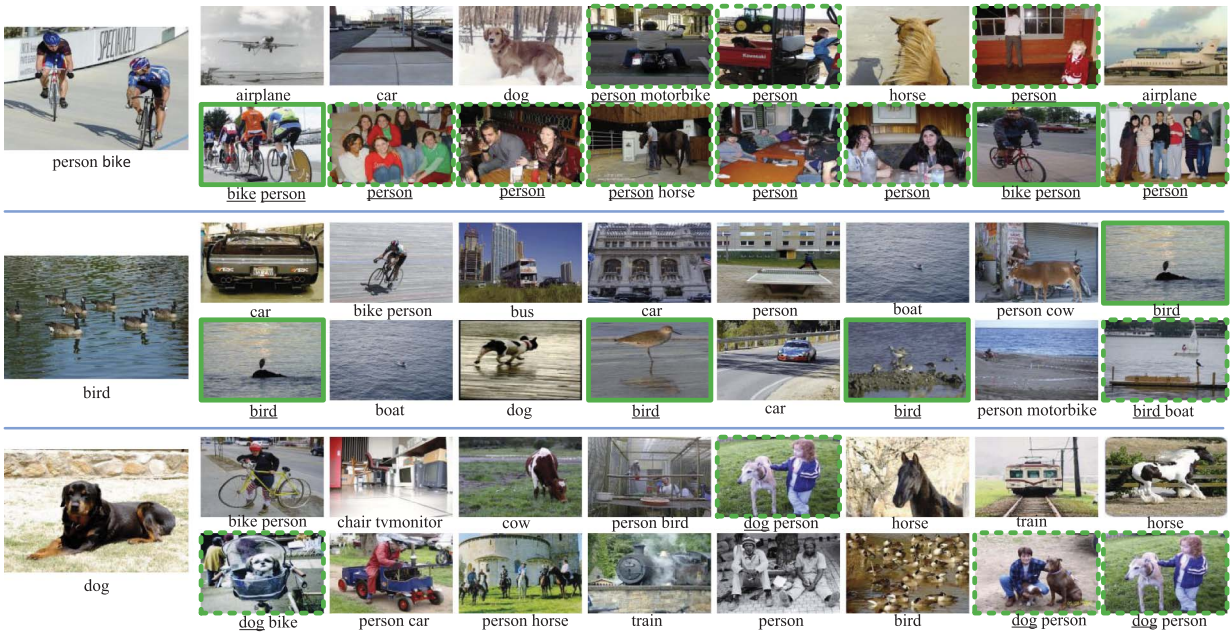

Fig. 4. Scene-aware and saliency-aware retrieval results. The images of the first column are query images. The odd rows are scene-aware retrieval results, and the even rows are the results of saliency-aware retrieval. We list the top eight neighboring images for each method. The images highlighted in dashed rectangles are weakly related to the queries, while those in thick solid rectangles indicate the strong relationship.

in our experiments to describe the salient object windows, and the dimensionality is 4200. Our experiments involve four parameters. We set $N_D = 10$, $N_T = 30$, and $K_I = K_W = 100$ throughout our experiments. For each image in database, we sample a few accurate object windows, while for the target image, a larger number of windows may be helpful for taking care of the less salient object. Therefore, we suggest that N_T is set to be larger than N_D.

B. Evaluation Metrics

Area under ROC curve (AUC) and ranking loss are employed to evaluate the performance of image annotation. The larger the AUC score is, the better the performance is. While for ranking loss, smaller score indicates better performance. The definition of $\text{AUC}_{\text{macro}}$ is defined as

$$\begin{aligned}\text{AUC}_{\text{macro}} &= \frac{1}{q}\sum_{j=1}^{q}\text{AUC}_j \\ &= \frac{1}{q}\sum_{j=1}^{q}\frac{\left|\{(\mathbf{x}', \mathbf{x}'')|f(\mathbf{x}', y_j) \geq f(\mathbf{x}'', y_j), (\mathbf{x}', \mathbf{x}'') \in \mathcal{Z}_j \times \bar{\mathcal{Z}}_j\}\right|}{|\mathcal{Z}_j||\bar{\mathcal{Z}}_j|}\end{aligned} \tag{17}$$

where $f(\mathbf{x}, y)$ is a real-valued function giving the confidence of $y \in \mathcal{Y}$ being the proper label of the test instance $\mathbf{x}$. $\mathbf{x}'$ and $\mathbf{x}''$ are two different instances. q is the cardinality of the label set. $\mathcal{Z}_j$ and $\bar{\mathcal{Z}}_j$ correspond to the set of test instances with and without label y_j, respectively. The $\text{AUC}_{\text{micro}}$ is defined as (18), shown at the bottom of this page, where y' and y'' correspond to the relevant and irrelevant labels, and $\mathcal{S}^+$ and $\mathcal{S}^-$ correspond to the set of relevant and irrelevant instance-label pairs. The definition of ranking loss is

$$\begin{aligned}\text{Ranking loss} = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{|Y_i||\bar{Y}_i|}|\{(y', y'')|f(\mathbf{x}_i, y') \leq f(\mathbf{x}_i, y''), \\ (y', y'' \in Y_i \times \bar{Y}_i)\}|\end{aligned} \tag{19}$$

where Y_i is the label set associated with instance $\mathbf{x}_i$ and p is the number of test instances. The detailed definitions of these metrics can be referred to [54].

C. Results of Image Retrieval

First, we test our method on the image retrieval task, which is strongly related to the image-level annotation and the object localization. The example results of the scene-aware retrieval and the saliency-aware retrieval are given in Fig. 4. The images in the first column are queries. Each query is associated with two rows of the retrieval results. The odd rows are the retrieval results using only global features, and the even rows are the results with saliency-aware retrieval. According to the results, the method using global features can find neighboring images with the similar scene, but often fails to find images containing objects of the same category. In other words, the scene-aware method usually concentrates on the scene-level similarity. Instead, our method concerns both the scene-level and object-level cues. In particular, as shown in the first group (the first

$$\text{AUC}_{\text{micro}} = \frac{|(\mathbf{x}', \mathbf{x}'', y', y'')|f(\mathbf{x}', y') \geq f(\mathbf{x}'', y''), (\mathbf{x}', y') \in \mathcal{S}^+, (\mathbf{x}'', y'') \in \mathcal{S}^-|}{|\mathcal{S}^+||\mathcal{S}^-|} \tag{18}$$

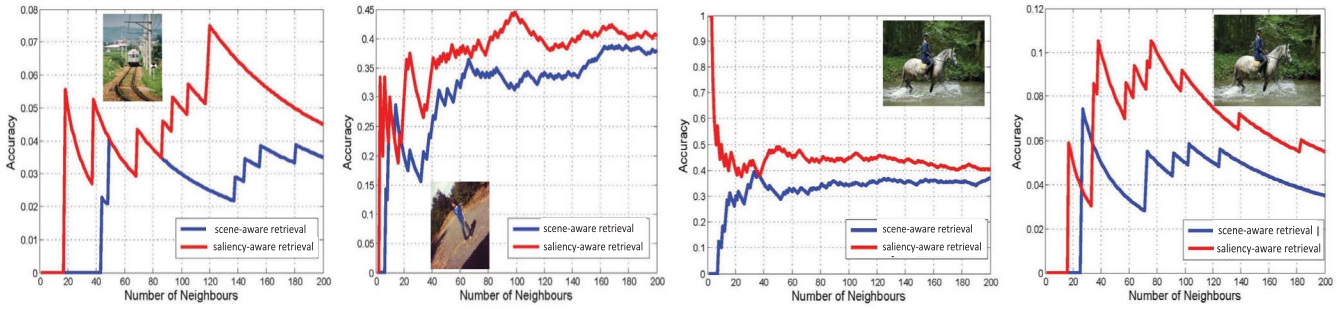

Fig. 5. Retrieval samples. We compare the retrieval accuracy with respect to the number of retrieved neighbors. (a) Category label: train. (b) Category label: person. (c) Category label: person. (d) Category label: horse.

and second rows) of Fig. 4, six out of eight (75%) images contain "person" but no "bike," we call them weakly related ones, since only a part of categories are matched. Two out of eight (25%) images are strongly related, which contain both "person" and "bike." However, with the global scene matching in the first row, only three (37.5%) weakly related image are found in the top eight neighbors. The second and third groups further confirm that our method using the saliency-aware retrieval can find more accurate neighboring images that contain objects of the same category as the target image. This also ensures that foreground has the priority to be annotated over background in the step of image-level annotation.

Fig. 5 gives the accuracy curves with respect to the number of neighboring images of the scene-aware model and the proposed saliency-aware model for three example queries (two images containing single object and one image containing multiple objects). The trend of the curves for other queries is consistent with that of the curves for this case. Our approach improves the accuracy of the scene-aware model on both the single-object images [Fig. 5(a) and (b)] and the multiobject images [Fig. 5(c) and (d)]. The overall performance will be further verified by the image-level annotation in Section IV-B.

D. Results of Image-Level Annotation

We first compare the performance in terms of the traditional image annotation measurements, including the standard criteria of AUC and ranking loss. For the baseline, we use the k-nearest neighbor (k-NN) method with global features, including GIST, Color Hist, and SPM. We also compare our method to the learning-based methods TagProp [14], TRAM [55], and FastTag [56]. For all these method, we select the top k ($k = 100$) nearest neighboring images for voting. The results are reported in Tables III and IV. The criteria labeled with $\uparrow$ indicate the higher the better, while the ones labeled with $\downarrow$ indicate the lower the better. The bold numbers and the italic numbers indicate that the corresponding method outperforms the baseline and all the compared methods, respectively. Taking the PASCAL VOC 2007 for example, the proposed method outperforms the baseline on all criteria.

The mean average precision for the object recognition task is reported in Tables V and VI. We also compare the results with several related methods. According to the results, there is an obvious advantage in incorporating object location compared with k-NN and TagProp, which can also be regarded as data-driven methods. The methods labeled with stars are nonparametric ones, and the others are learning-based methods. We observe that SANFA_{LC} outperforms not only the other nonparametric techniques but also the parametric technique [40]. The result is mildly surprising and shows the power of utilizing the saliency cue. Though SANFA does not outperform the k-NN in the image-level annotation stage, SANFA outperforms the k-NN in localization stage, since the correctly labeled objects are more salient with SANFA than k-NN, which makes them easy to be correctly localized.

TABLE III
IMAGE-LEVEL ANNOTATION PERFORMANCE ON PASCAL 2007. P: PARAMETRIC METHOD. NP: NONPARAMETRIC METHOD

	Method	$\text{AUC}_{\text{micro}}\uparrow$	$\text{AUC}_{\text{macro}}\uparrow$	Ranking Loss$\downarrow$
P	TAGPROP [14]	***0.8545***	0.7997	0.1997
	TRAM[55]	0.8280	***0.8385***	***0.1716***
	FastTag [56]	0.8066	0.8043	0.1963
NP	k-NN	0.8451	0.7822	0.2423
	SANFA	0.8446	0.7812	0.2423
	SANFA_{LC}	**0.8462**	**0.8059**	**0.1993**

TABLE IV
IMAGE-LEVEL ANNOTATION PERFORMANCE ON PASCAL 2008. P: PARAMETRIC METHOD. NP: NONPARAMETRIC METHOD

	Method	$\text{AUC}_{\text{macro}}\uparrow$	$\text{AUC}_{\text{macro}}\uparrow$	Ranking Loss$\downarrow$
P	TAGPROP[14]	0.8397	0.7833	0.2163
	TRAM[55]	0.8211	0.8133	0.1868
	FastTag[56]	***0.8426***	***0.8291***	***0.1714***
NP	k-NN	0.8223	0.7659	0.2596
	SANFA	0.7900	0.7584	0.2654
	SANFA_{LC}	**0.8310**	**0.7731**	**0.2490**

E. Results of Object Localization

Localization performance is measured with the percentage of target images in which an instance is correctly localized (**CorLoc**). The correctness of localization is measured according to the PASCAL criterion [49] (window-intersection-over-union ≥ 0.50). We evaluate the localization performance of the proposed method SANFA by compared with some

TABLE V

IMAGE-LEVEL ANNOTATION PERFORMANCE ON PASCAL 2007 IN TERMS OF AP. P: PARAMETRIC METHOD. NP: NONPARAMETRIC METHOD

Method		plane	bike	bird	boat	bottle	bus	car	cat	chair	cow
P	TAGPROP [14]	0.5981	0.1730	0.2509	0.4182	0.1069	0.1959	0.5386	0.2002	0.2930	0.1249
	TRAM [55]	0.5746	0.1511	0.1655	0.3386	0.0990	0.1736	0.5064	0.1813	0.2449	0.0835
	FastTag [56]	0.5701	0.1932	0.2320	0.3716	0.1219	0.2625	0.4884	0.2009	0.2856	0.1097
	Nguyen [40]	0.3070	0.1650	0.2300	0.1490	0.0490	0.2960	0.2650	0.3530	0.0720	0.2340
NP	k-NN*	0.5058	0.1069	0.2184	0.3456	0.0884	0.1338	0.4307	0.170	0.2840	0.1233
	SANFA*	0.5136	0.0992	0.2146	0.3073	0.0866	0.1285	0.4176	0.1659	0.2683	01047
	SANFA$_{LC}$*	0.5385	0.1273	0.2376	0.3528	0.1031	0.2313	0.4635	0.1864	0.2979	0.1187

Method	dingtable	dog	horse	mtbike	person	ptplant	sheep	sofa	train	tv	**AVG**
TAGPROP [14]	0.2355	0.2241	0.5514	0.3647	0.6468	0.0973	0.1411	0.1830	0.3911	0.1912	***0.2964***
TRAM [55]	0.1815	0.2034	0.4881	0.2603	0.6060	0.0775	0.0956	0.1404	0.3842	0.1544	0.2555
FastTag [56]	0.1794	0.2171	0.4349	0.2680	0.6042	0.0980	0.1449	0.1452	0.3882	0.1542	0.2735
Nguyen [40]	0.2050	0.3210	0.2440	0.3310	0.1720	0.1220	0.2080	0.2880	0.4060	0.0700	0.2240
k-NN*	0.2054	0.2135	0.3233	0.2573	0.4216	0.0962	0.1397	0.1793	0.3145	0.1554	0.2353
SANFA*	0.1911	0.2070	0.3002	0.2329	0.4187	0.0946	0.1258	0.1686	0.2861	0.1450	0.2231
SANFA$_{LC}$*	0.1714	0.2334	0.3823	0.2738	0.5751	0.1217	0.1367	0.1643	0.3407	0.1931	**0.2629**

TABLE VI

IMAGE-LEVEL ANNOTATION PERFORMANCE ON PASCAL 2008 IN TERMS OF AP. P: PARAMETRIC METHOD. NP: NONPARAMETRIC METHOD

Method		plane	bike	bird	boat	bottle	bus	car	cat	chair	cow
P	TAGPROP [14]	0.5870	0.0940	0.1774	0.4044	0.1249	0.1839	0.4580	0.2270	0.2725	0.0938
	TRAM [55]	0.5703	0.0915	0.1627	0.2766	0.0774	0.1677	0.4536	0.1985	0.2106	0.0609
	FastTag [56]	0.5866	0.1549	0.2387	0.3620	0.1145	0.2424	0.4089	0.2161	0.2598	0.1253
NP	k-NN*	0.5649	0.0800	0.1677	0.3929	0.0927	0.1956	0.3976	0.1866	0.2590	0.0991
	SANFA*	0.5268	0.0789	0.1722	0.2538	0.0969	0.1624	0.3728	0.1692	0.2023	0.0967
	SANFA$_{LC}$*	0.5838	0.0965	0.2160	0.3422	0.1083	0.1866	0.4356	0.1914	0.2768	0.1212

Method	dingtable	dog	horse	mtbike	person	ptplant	sheep	sofa	train	tv	**AVG**
TAGPROP [14]	0.2149	0.2194	0.2005	0.1764	0.6267	0.0962	0.1222	0.1341	0.3370	0.2313	0.2489
TRAM [55]	0.1503	0.1661	0.1086	0.1474	0.5552	0.0749	0.0623	0.1141	0.3110	0.1586	0.2059
FastTag[56]	0.1916	0.2047	0.2478	0.2309	0.5889	0.0868	0.1142	0.1295	0.3831	0.1671	***0.2527***
k-NN*	0.1999	0.2009	0.1387	0.1440	0.4141	0.0790	0.1390	0.1617	0.3179	0.1917	0.2212
SANFA*	0.1561	0.2066	0.1169	0.1576	0.4118	0.0798	0.1064	0.1080	0.2566	0.1561	0.1944
SANFA$_{LC}$*	0.2036	0.2254	0.2298	0.2301	0.4226	0.1019	0.1373	0.1732	0.3152	0.1759	**0.2387**

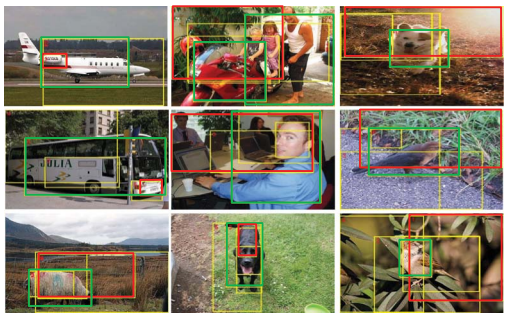

Fig. 6. Top five salient object windows detected by method [7]. Red rectangles: the highest scored salient object windows. Green rectangles: our localization results.

weakly supervised the localization methods. The performance is measured on the entire PASCAL 2007 test set. The detailed results are shown in Table VII. The comparisons are learning based and even more complex (i.e., requiring segmentation or DPM based), while SANFA is free of training and segmentation. For all comparisons, we select the first object window proposal in each image as the annotation of the object of interest. For the method SALIENT-ONLY [7], the most salient object window is selected as the object localization result.

Fig. 6 investigates the reason why our method can outperform SALIENCY-ONLY. The most salient object windows may not be the best localizations, and incorporating the retrieved reference windows is necessary and helpful. The first two images of the first row correspond to the first images of Figs. 7 and 8, respectively. They are typical example images on which we correctly detected good candidate windows (in green).

The results in terms of each category are shown in Tables VIII and IX. Note that the performance of TagProp [14] is relatively bad although it achieves a promising result in the image-level annotation. The main reason is that these methods do not focus on salient objects while matching. Hence, although they can tag the image with more accurate labels, the tagged nonsalient objects are more difficult to localize. The results corresponding to the methods labeled with stars are localization results with ground-truth labels, which is the same setting as the most existing localization methods.

We qualitatively evaluate the localization ability of our method. First, we localize objects with the predicted labels acting as input. Second, for difficult ones, ground-truth labels are given, and then we predict one window for each label. In Fig. 7, the images in the left rectangle are successfully annotated, with both correctly predicted labels and windows for objects. The results in the right rectangle are wrong image-level annotation results. The examples in Fig. 8 (solid-line box) are some images that are difficult to obtain fully correct

TABLE VII

OBJECT LOCALIZATION COMPARISON ON PASCAL VOC 2007 IN TERMS OF CORRECT LOCALIZATION ON POSITIVE TRAINING IMAGES (CorLoc). P: PARAMETRIC METHOD. NP: NONPARAMETRIC METHOD

Method		Property	AVG
P	Nguyen[40]	training needed, jointly learning of localization and classification	22.40
	MIML [57]	training needed, segmentation needed, multi-label multi-instance framework	23.60
	Siva and Xiang [39]	training needed, negative mining	28.90
	Mode-drift[38]	training needed, iterative learning of a detector	30.4
NP	SALIENT-ONLY[7]	no training, saliency-aware	8.43
	WS-SANFA	no training, saliency-aware, data-driven	23.54

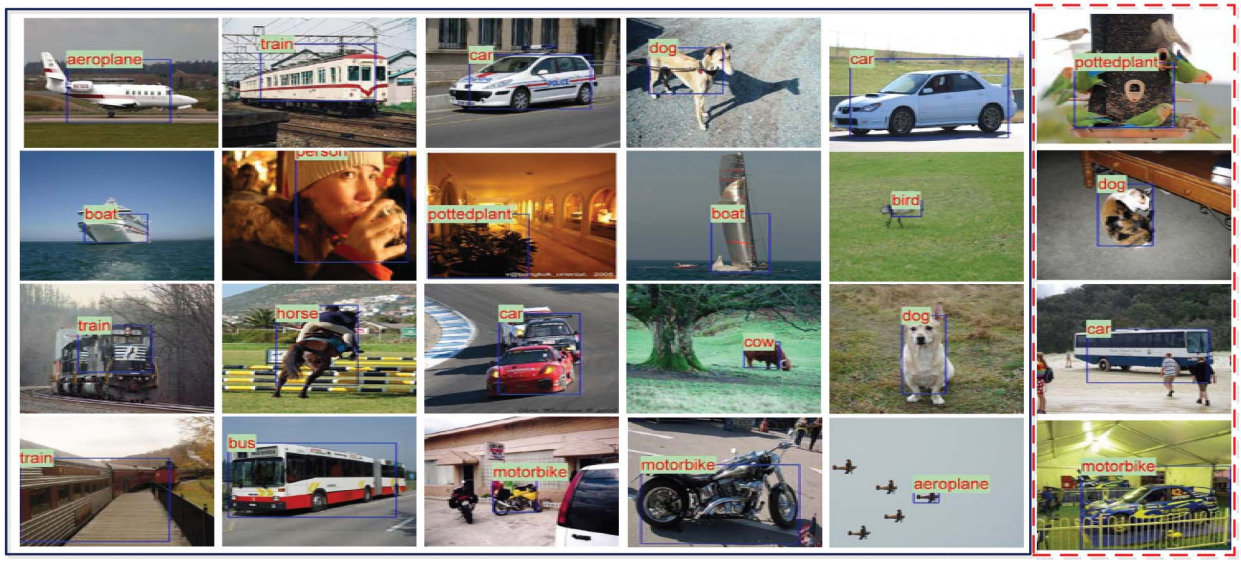

Fig. 7. Object localization with predicted labels. The images in the left five columns are correct detections, and those in the last column are wrong detections.

TABLE VIII

OBJECT LOCALIZATION COMPARISON ON PASCAL VOC 2007 IN TERMS OF CORRECT LOCALIZATION ON POSITIVE TRAINING IMAGES (CorLoc)

Method		plane	bike	bird	boat	bottle	bus	car	cat	chair	cow
Image	kNN	12.00	0.00	0.00	0.00	0.00	0.00	4.58	0.00	0.00	0.00
	TagProp [14]	14.67	0.00	0.42	1.56	0.00	0.00	8.21	0.54	0.00	0.00
	FastTag [56]	13.17	19.60	1.04	3.41	2.92	0.00	3.74	1.51	0.18	0.00
	SANFA	8.67	1.74	0.42	1.56	0.00	8.47	17.37	1.08	0.44	2.06
	SANFA$_{LC}$	7.33	2.33	2.50	0.78	0.63	27.12	12.98	1.62	0.00	6.19
Image+Tags	Mode-drift*[38]	42.4	46.5	18.2	8.8	2.9	40.9	73.2	44.8	5.4	30.5
	SANFA$_{GT}$*	26.00	20.35	17.08	14.84	3.14	47.46	41.60	35.51	6.55	25.77

Method	dingtable	dog	horse	mtbike	person	ptplant	sheep	sofa	train	tv	**AVG**
kNN	0.00	0.40	0.00	0.00	9.58	0.00	0.00	0.00	0.00	0.00	1.33
TagProp [14]	0.00	1.60	0.00	0.00	9.38	0.00	0.00	0.00	1.72	0.00	1.90
FastTag [56]	0.00	0.00	0.00	0.00	10.09	0.00	4.08	0.00	2.70	0.00	3.15
SANFA	17.50	5.60	26.26	27.06	2.37	0.64	0.00	2.86	7.76	1.15	6.65
SANFA$_{LC}$	7.50	9.20	12.12	35.88	0.10	3.21	0.00	0.95	10.34	0.57	***7.07***
Mode-drift*[38]	19.0	34.0	48.8	65.3	8.2	9.4	16.7	32.3	54.8	5.5	**30.40**
SANFA$_{GT}$*	25.00	32.80	35.86	41.18	9.58	7.69	9.46	28.57	25.86	18.39	**23.54**

image-level labels. However, given correct image-level labels, our method can localize the objects correctly. For the third row, although the top five windows with highest objectness scores are all centered on the person in the image, our approach can select the correct window for the dog with the help of the appearance similarity. The images in Fig. 8 (dashed-line box) are some example images, in which the objects are difficult to localize.

(a) (b) (c) (d) (e)

Fig. 8. Object localization with ground-truth labels. (a) Target image. (b) All windows sampled on the target image. (c) Top five windows with highest objectness scores. (d) Localization results with our method. (e) Ground-truth localization.

TABLE IX

OBJECT LOCALIZATION COMPARISON ON PASCAL VOC 2008 IN TERMS OF CORRECT LOCALIZATION ON POSITIVE TRAINING IMAGES (CorLoc)

Method		plane	bike	bird	boat	bottle	bus	car	cat	chair	cow
Image	kNN	13.33	0.00	0.00	0.00	0.00	0.00	4.58	0.00	0.00	0.00
	TagProp [14]	11.22	0.00	0.35	0.00	0.00	0.00	0.26	0.54	0.00	0.00
	FastTag [56]	12.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	SANFA	9.33	3.49	0.83	0.00	0.63	1.69	23.28	0.54	0.00	0.00
	SANFA$_{LC}$	8.00	0.58	2.08	0.00	0.00	14.41	4.39	2.16	0.44	4.12
Image+Tags	SALIENT-ONLY*[7]	9.33	6.98	5.42	3.91	0.00	14.41	13.55	9.73	1.75	5.15
	SANFA$_{GT}$*	27.80	19.60	19.38	16.48	2.92	38.25	27.61	32.83	5.50	20.47

Method	dingtable	dog	horse	mtbike	person	ptplant	sheep	sofa	train	tv	**AVG**
kNN	0.00	1.20	0.00	0.00	9.87	0.00	0.00	0.00	0.00	0.00	1.22
TagProp [14]	0.00	1.15	0.00	0.00	11.78	0.00	0.00	0.00	0.00	0.00	1.26
FastTag [56]	0.00	0.00	0.00	0.00	11.68	1.350	0.00	0.00	0.00	0.00	1.54
SANFA	15.00	10.08	3.03	4.71	0.39	1.92	0.00	0.00	2.59	0.57	3.94
SANFA$_{GT}$*	5.00	8.80	0.51	25.29	0.00	3.21	1.35	0.00	8.62	3.45	***4.62***
SALIENT-ONLY*[7]	5.00	8.40	7.07	12.35	0.89	3.21	21.62	5.71	21.55	12.64	8.43
SANFA$_{GT}$*	15.79	31.87	34.41	36.05	11.68	7.09	13.27	15.77	25.10	14.90	**20.84**

V. CONCLUSION

We have devised a workable framework for image foreground annotation to predict category labels and, more importantly, windows for objects. The task becomes more challenging when only weak image-level labels are provided. By making use of the saliency objectness cue, our method executes the matching in the object level and, thus, avoids the risk of retrieving only scene similar images that ignore

the similarity of objects. The saliency cue strengthens the awareness of the foreground and, thus, improves the object recognition and localization performance. The quantitative and qualitative experimental results have verified that the proposed method archives the state-of-the-art performance. In the future, we will focus on incorporating cosaliency detection techniques [58], [59] to boost the annotation.

REFERENCES

[1] J. Tighe and S. Lazebnik, "Superparsing," *Int. J. Comput. Vis.*, vol. 101, no. 2, pp. 329–349, Jan. 2013.

[2] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing via label transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2368–2382, Dec. 2011.

[3] G. Singh and J. Košecká, "Nonparametric scene parsing with adaptive feature relevance and semantic context," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 3151–3157.

[4] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

[5] O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei, "Object-centric spatial pooling for image classification," in *Proc. 12th Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 1–15.

[6] X. Liu, S. Yan, J. Luo, J. Tang, Z. Huango, and H. Jin, "Nonparametric label-to-region by search," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3320–3327.

[7] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun, "Salient object detection by composition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 1028–1035.

[8] B. Russell, A. Torralba, C. Liu, R. Fergus, and W. T. Freeman, "Object recognition by scene alignment," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2007, pp. 1241–1248.

[9] D. Eigen and R. Fergus, "Nonparametric image parsing using adaptive neighbor sets," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2799–2806.

[10] X. Li, L. Chen, L. Zhang, F. Lin, and W.-Y. Ma, "Image annotation by large-scale content-based image retrieval," in *Proc. 14th Annu. ACM Int. Conf. Multimedia (MM)*, 2006, pp. 607–610.

[11] X.-J. Wang, L. Zhang, X. Li, and W.-Y. Ma, "Annotating images by mining image search results," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1919–1932, Nov. 2008.

[12] J. Kleban, E. Moxley, J. Xu, and B. S. Manjunath, "Global annotation on georeferenced photographs," in *Proc. ACM Int. Conf. Image Video Retr. (CIVR)*, 2009, pp. 1–8.

[13] D. Kuettel, M. Guillaumin, and V. Ferrari, "Segmentation propagation in ImageNet," in *Proc. 12th Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 459–473.

[14] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *Proc. IEEE 12th Int. Conf. Comput. Vis. (ICCV)*, Sep./Oct. 2009, pp. 309–316.

[15] C. Wang, D. Blei, and L. Fei-Fei, "Simultaneous image classification and annotation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1903–1910.

[16] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proc. 26th Annu. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2003, pp. 119–126.

[17] V. Lavrenko, R. Manmatha, and J. Jeon, "A model for learning the semantics of pictures," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2003, pp. 553–560.

[18] H. Harzallah, F. Jurie, and C. Schmid, "Combining efficient object localization and image classification," in *Proc. 12th Int. Conf. Comput. Vis. (ICCV)*, Sep./Oct. 2009, pp. 237–244.

[19] M. B. Blaschko and C. H. Lampert, "Learning to localize objects with structured output regression," in *Proc. 10th Eur. Conf. Comput. Vis. (ECCV)*, 2008, pp. 2–15.

[20] C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Efficient subwindow search: A branch and bound framework for object localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2129–2142, Dec. 2009.

[21] T. Deselaers, B. Alexe, and V. Ferrari, "Localizing objects while learning their appearance," in *Proc. 11th Eur. Conf. Comput. Vis. (ECCV)*, 2010, pp. 452–466.

[22] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 1307–1314.

[23] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 73–80.

[24] X. Zhou, K. Yu, T. Zhang, and T. S. Huang, "Image classification using super-vector coding of local image descriptors," in *Proc. 11th Eur. Conf. Comput. Vis. (ECCV)*, 2010, pp. 141–154.

[25] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3360–3367.

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

[27] T. Liu *et al.*, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[28] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 17–24.

[29] J. Han, K. N. Ngan, M. Li, and H.-J. Zhang, "Unsupervised extraction of visual attention objects in color images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 1, pp. 141–145, Jan. 2006.

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

[31] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, "Background prior-based salient object detection via deep reconstruction residual," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 8, pp. 1309–1321, Aug. 2015.

[32] J. Han, S. He, X. Qian, D. Wang, L. Guo, and T. Liu, "An object-oriented visual saliency detection framework based on sparse coding representations," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 12, pp. 2009–2021, Dec. 2013.

[33] F. Zhu and L. Shao, "Weakly-supervised cross-domain dictionary learning for visual recognition," *Int. J. Comput. Vis.*, vol. 109, nos. 1–2, pp. 42–59, Aug. 2014.

[34] C. Galleguillos, B. Babenko, A. Rabinovich, and S. Belongie, "Weakly supervised object localization with stable segmentations," in *Proc. 10th Eur. Conf. Comput. Vis. (ECCV)*, 2008, pp. 193–207.

[35] H. Arora, N. Loeff, D. A. Forsyth, and N. Ahuja, "Unsupervised segmentation of objects using efficient learning," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–7.

[36] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2003, pp. II-264–II-271.

[37] E. Borenstein and S. Ullman, "Learning to segment," in *Proc. 8th Eur. Conf. Comput. Vis. (ECCV)*, 2004, pp. 315–328.

[38] P. Siva and T. Xiang, "Weakly supervised object detector learning with model drift detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 343–350.

[39] P. Siva, C. Russell, and T. Xiang, "In defence of negative mining for annotating weakly labelled data," in *Proc. 12th Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 594–608.

[40] M. H. Nguyen, L. Torresani, F. De la Torre, and C. Rother, "Weakly supervised discriminative localization and classification: A joint learning process," in *Proc. IEEE 12th Int. Conf. Comput. Vis. (ICCV)*, Sep./Oct. 2009, pp. 1925–1932.

[41] J. Hays and A. A. Efros, "IM2GPS: Estimating geographic information from a single image," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.

[42] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing: Label transfer via dense scene alignment," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1972–1979.

[43] 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. (CVPR)*, Jun. 2006, pp. 2169–2178.

[44] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Prog. Brain Res.*, vol. 155, pp. 23–36, Oct. 2006.

[45] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1958–1970, Nov. 2008.
[46] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Apr. 2013.
[47] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 328–335.
[48] C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Beyond sliding windows: Object localization by efficient subwindow search," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.
[49] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.
[50] B. Hemery, H. Laurent, and C. Rosenberger, "Comparative study of metrics for evaluation of object localisation by bounding boxes," in *Proc. 4th Int. Conf. Image Graph. (ICIG)*, Aug. 2007, pp. 459–464.
[51] Y. Zhang and T. Chen, "Efficient inference for fully-connected CRFs with stationarity," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 582–589.
[52] D. Ramanan, "Learning to parse images of articulated bodies," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2006, pp. 1129–1136.
[53] C. Rother, V. Kolmogorov, and A. Blake, "'GrabCut': Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, Aug. 2004.
[54] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.
[55] X. Kong, M. K. Ng, and Z.-H. Zhou, "Transductive multilabel learning via label set propagation," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 3, pp. 704–719, Mar. 2013.
[56] M. Chen, A. Zheng, and K. Weinberger, "Fast image tagging," in *Proc. Int. Conf. on Mach. Learn. (ICML)*, pp. 1274–1282, Jun. 2013.
[57] X. Xue, W. Zhang, J. Zhang, B. Wu, J. Fan, and Y. Lu, "Correlative multi-label multi-instance image annotation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 651–658.
[58] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3766–3788, Oct. 2013.
[59] D. Zhang, J. Han, C. Li, and J. Wang, "Co-saliency detection via looking deep and wide," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2994–3002.

Xiaochun Cao (SM'14) received the B.E. and M.E. degrees from Beihang University, Beijing, China, and the Ph.D. degree from the University of Central Florida, Orlando, FL, USA, all in computer science.

He spent about three years with ObjectVideo Inc., Reston, VA, USA, as a Research Scientist. From 2008 to 2012, he was a Professor with Tianjin University, Tianjin, China. He is currently a Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing. He has authored or co-authored over 100 journal and conference papers.

Prof. Cao is a fellow of the Institution of Engineering and Technology. His dissertation was nominated for the University Level Outstanding Dissertation Award. In 2004 and 2010, he was a recipient of the Piero Zamperoni Best Student Paper Award at the International Conference on Pattern Recognition. He is an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING.

Changqing Zhang received the B.S. and M.S. degrees from the College of Computer Science, Sichuan University, Chengdu, China, in 2005 and 2008, respectively. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Tianjin University, Tianjin, China.

His current research interests include machine learning, data mining, and computer vision.

Huazhu Fu received the B.S. degree in mathematical sciences from Nankai University, Tianjin, China, in 2006, the M.E. degree in mechatronics engineering from the Tianjin University of Technology, Tianjin, in 2010, and the Ph.D. degree in computer science from Tianjin University, Tianjin, in 2013.

He was a Post-Doctoral Research Fellow with Nanyang Technological University, Singapore, for two years. He is currently a Research Scientist with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore. His current research interests include computer vision, image processing, and medical image analysis.

Dr. Fu's Ph.D. dissertation was nominated for the China Computer Federation Outstanding Dissertation in 2014.

Xiaojie Guo (M'13) received the B.E. degree in software engineering from the School of Computer Science and Technology, Wuhan University of Technology, Wuhan, China, in 2008, and the M.S. and Ph.D. degrees in computer science from the School of Computer Science and Technology, Tianjin University, Tianjin, China, in 2010 and 2013, respectively.

He is currently an Assistant Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China.

Dr. Guo was a recipient of the Piero Zamperoni Best Student Paper Award at the International Conference on Pattern Recognition (International Association on Pattern Recognition) in 2010.

Qi Tian (SM'03) received the Ph.D. degree in ECE from the University of Illinois at Urbana–Champaign (UIUC), 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 Professor with the Department of Computer Science, The University of Texas at San Antonio (UTSA). His current research interests include multimedia information retrieval, computer vision, pattern recognition and bioinformatics and published over 310 refereed journal and conference papers. He received the Best Paper Awards in ACM ICMR 2015, PCM 2013, MMM 2013, and ACM ICIMCS 2012, a Top 10% Paper Award in MMSP 2011, a Student Contest Paper Award in ICASSP 2006.

Dr. Tian research projects are funded by ARO, NSF, DHS, Google, FXPAL, NEC, SALSI, Akiira Media Systems, and UTSA, etc. He received the 2010 ACM Service Award and 2014 Research Achievement Award from the College of Science, UTSA. He is the Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA (TMM), the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), *Multimedia System Journal* (MMSJ), and in the Editorial Board of the *Journal of Multimedia* (JMM), and the *Journal of Machine Vision and Applications* (MVA).