Joint Inference of Objects and Scenes with Efficient Learning of Text-Object-Scene Relations

Botao Wang, Dahua Lin, Hongkai Xiong, *Senior Member, IEEE,* Y. F. Zheng, *Fellow, IEEE*

***Abstract*—The rapid growth of web images presents new challenges as well as opportunities to the task of image understanding. Conventional approaches rely heavily on fine-grained annotations, such as bounding boxes and semantic segmentations, which are not available for web-scale images. In general, images over the Internet are accompanied with descriptive texts, which are relevant to their contents. To bridge the gap between textual and visual analysis for image understanding, this paper presents an algorithm to learn the relations between scenes, objects and texts with the help of image-level annotations. In particular, the relation between the texts and objects is modeled as the matching probability between the nouns and the object classes, which can be solved via a constrained bipartite matching problem. On the other hand, the relations between the scenes and objects/texts are modeled as the conditional distributions of their co-occurrence. Built upon the learned cross-domain relations, an integrated model brings together scenes, objects and texts for joint image understanding, including scene classification, object classification and localization, and the prediction of object cardinalities. The proposed cross-domain learning algorithm and the integrated model elevate the performance of image understanding for web images in the context of textual descriptions. Experimental results show that the proposed algorithm significantly outperforms conventional methods in various computer vision tasks.**

***Index Terms*—Scene classification, object classification, object localization, conditional random field.**

I. Introduction

With the explosive growth of images over the Internet, it has never been more desirable for intelligent vision systems that can automatically extract semantics from images, including how they are composed (image segmentation), what scenarios they describe (scene categorization), what objects they embrace (object classification) and where they are (object localization). In general, to answer these questions is non-trivial due to the variations of objects in the visual appearance and the complex interactions among them. Thanks to the continual efforts of the researchers over the past few decades, remarkable progress has been made in several fundamental tasks of computer vision, including object detection [1], scene categorization [2] and image retrieval [3]. Yet, image understanding, as the ultimate goal of computer vision, remains a challenging task [4], and the performance of the state-of-the-art algorithms is still inferior to human intelligence.

B. Wang and H. Xiong are with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. Email: {botaowang, xionghongkai}@sjtu.edu.cn.

D. Lin is with the Department of Information Engineering, The Chinese University of Hong Kong, Email: dhlin@ie.cuhk.edu.hk.

Y. F. Zheng is with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210, USA. Email: zheng@ece.osu.edu.

The work was supported in part by the NSFC, under grants 61425011, U1201255, 61271218, 61529101, 61472234, 61271211, and Shu Guang project (13SG13).

As we move towards web-scale image understanding, the problems become even more difficult. The vast diversity of web images presents a substantial challenge to the vision community – we can no longer rely on images with fine-grained annotations, e.g., bounding boxes and semantic segmentations, to train the models, because it is too lavish to provide detailed annotations for millions of images in thousands of classes. Naturally, challenges come with opportunities. An important distinction of web images as opposed to those in traditional datasets is that they are often associated with descriptive texts, such as captions, keywords and tags, which are highly relevant to the content of the images. As illustrated in Fig. 1, the words *woman* and *bicycle* in the caption reveal a few aspects about the image, even if the image is not visible itself. For example, the image contains at least two object classes: PERSON and BICYCLE, and it may be taken in the street or sports field. In this paper, we aim at bridging the gap between the textual descriptions and the visual analysis of the images for joint image understanding.

Significant efforts have been made to explore effective ways to integrate visual and textual analysis [5]–[12]. Li et al. [13] learned a generative model from the images and tags for joint scene classification, image segmentation and image annotations. Farhadi et al. [5] introduced the "meaning space" as an intermediate representation between images and texts. Fidler et al. [14] advanced a holistic scene understanding framework that jointly reasons about semantic segmentation, the presence of objects and their spatial extent. All of these methods would require strong supervision to extract object classes from the textual descriptions in order to obtain proper representations. To reduce the enormous effort of manual annotation for the tremendous number of images over the Internet, we explore an efficient approach for establishing the link between the textual descriptions and visual concepts.

Some approaches [14, 15] exploit the spatial relations of objects indicated by the prepositions (e.g., *near*, *behind*) for image understanding. Although the prepositions are claimed to be effective for object recognition, they suffer from two limitations in practice. First, the spatial relations defined by prepositions have to be learned with the bounding boxes of objects, which are so immoderately difficult to obtain. Second, the spatial relations from the textual descriptions are actually very scarce in reality. In most cases, people might feel redundant to describe the spatial relations of objects, which are so obvious to the viewers since they can perceive the images. Statistical analysis on public dataset also supports this idea. For example, in the UIUC dataset [5], the mean frequency of effective prepositions of *near*, *in*, *on* and *in front of* is about 0.02 per caption. Compared to the prepositions, Part of Speech (POS) tags are far more robust and abundant in the

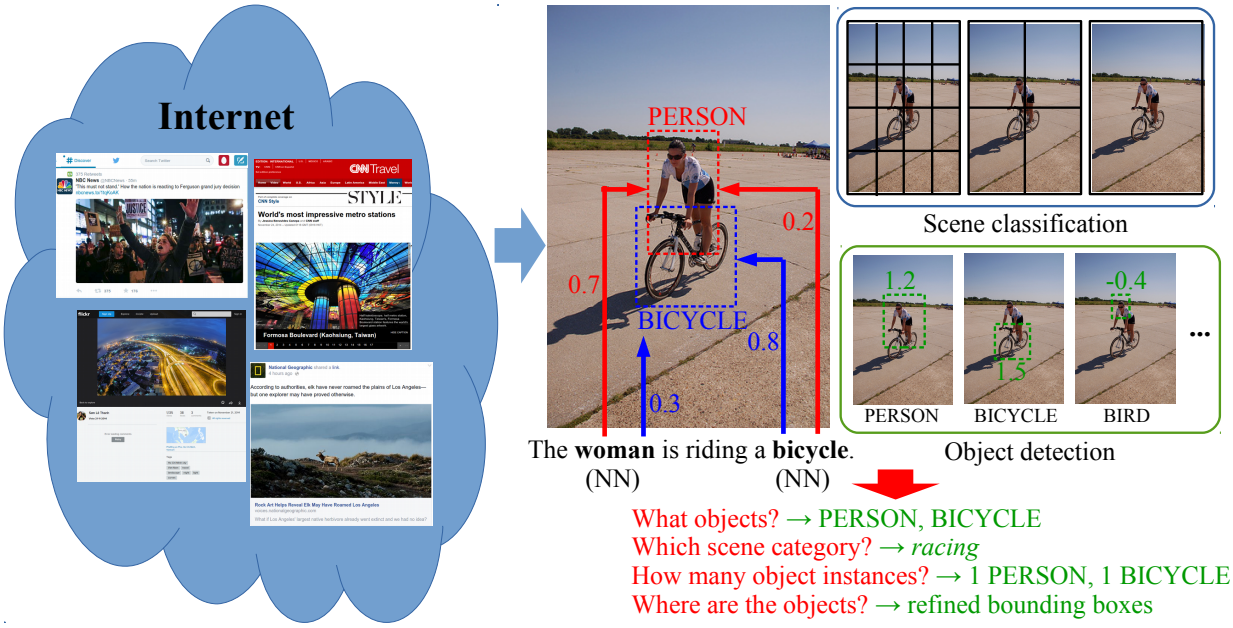


Fig. 1. Massive amount of images over the Internet are associated with textual descriptions, which are highly relevant to the content. The proposed algorithm learns the cross-domain relations between scenes, objects, and texts with image-level annotations, and jointly performs scene classification, object classification, prediction of object cardinalities, and object localization.

textual descriptions. In consequence, they serve as the main source of our textual analysis.

Motivated by these observations, we aim at developing an algorithm to learn the relations across different domains, namely, *scenes*, *objects* and *texts*, for image understanding. In particular, we do not demand fine-grained annotations, such as bounding boxes, to learn those cross-domain relations. Instead, only image-level annotations will be involved, which are much easier to reach, especially for web images. Moreover, an integrated model will be designed, which utilizes those cross-domain relations to perform multiple image inference tasks, including scene classification, object classification (i.e., the existence of object classes), prediction of object cardinalities (i.e., the number of object instances), and object localization.

The contribution of this paper is two-fold. First, we propose an algorithm to learn the cross-domain relations between scenes, objects and texts. To be specific, the relation between the texts and the objects is formulated as the matching probability of the nouns and object classes, which can be derived by solving a set of instance-level constrained bipartite matching problems. Furthermore, the relations of scenes and objects/texts are represented by the relative frequency of occurrence of the objects/nouns in different scene classes. The proposed algorithm needs only image-level annotations, including the scene labels and the object cardinalities, to learn such relations. Therefore, it makes possible to harness the sheer wealth of web images without requiring enormous amount of manual annotations.

Second, an integrated model for image understanding is developed, which incorporates visual and textual features to jointly predict: (1) the scene classes of images, (2) the object classes in images, (3) the cardinalities of object classes, and (4) the locations of object instances. Various off-the-shelf object detectors and scene classifiers can be utilized to provide hypotheses of objects and scenes to the proposed model. Together with the textual descriptions of images, a conditional random field graphical model is established. In particular, two types of unary potentials and three types of binary potentials are elaborately designed to measure the compatibility of scenes, objects and texts with regard to the learned cross-domain relations. Remarkable improvements can be achieved by the proposed model in various computer vision tasks with relatively low computational complexity.

To distinguish the objects and the nouns, we use words in upper case to represent object classes, and words in italics to represent nouns. For example, PEOPLE is an object class, while *people* is a noun in the textual description. The remainder of the paper is organized as follows: Section II reviews the related work. Section III describes the algorithm of learning cross-domain relations. Section IV presents the integrated model for joint image understanding. Section V provides experimental results. Finally, Section VI concludes the paper.

II. Related Work

There is quite a little work using text for image understanding, which generally fall in three categories: feature-based methods, topic-based methods and CRF-based methods.

Feature-based approaches extract features, respectively, from the image and the associated text, which are combined in either feature-level, metric-level, or classifier-level. Li et al. [16] concatenated the textual and visual features to train an SVM for image classification. Wang et al. [17] put forward to learn three SVMs: one over textual features, one over image features, and one combining the scores from the former two. Cheng et al. [3] adopted textual feature and visual feature

in a sequential way, and learned a concept-dependent fusion strategy to combine them. Guillaumin et al. [18] integrated textual and visual features with multiple kernel learning. Some approaches introduce a common space to bridge the textual and visual features. Iyengar et al. [19] recommended a retrieval system where visual and textual features are linked via an intermediate concept layer. Farhadi et al. [5] advocated a common representation called *"meaning space"*, to which images and captions are both mapped. Yeh et al. [20] developed a domain adaptation framework for cross-domain recognition, where the canonical correlation analysis is adopted to derive a joint feature space for associating cross-domain data.

Topic models are also used to model texts and images. In the seminal work by Barnard et al. [21], LDA models [22], formulated upon images and text respectively, are coupled to form a joint distribution. Later, Blei et al. [23] suggested *Corr-LDA* to capture the correlation between images and corresponding captions, assuming one-to-one correspondence between visual and textual topics. Putthividhy et al. [24] brought an extension that is able to exploit correlations across multiple topics. Recently, more sophisticated multi-modal topic models using MRF [25], neural networks [26] and relational models [27], have been studied. Overall, both feature-based and topic-based methods generally consider the textual descriptions as a whole, and do not exploit the links of finer granularity to connect texts and objects, which is a key aspect in our approach.

Graphical models, which are increasingly popular, have been investigated in joint image-text analysis. For image retrieval, Gao et al. [28] represented the image features and the textual features with a bag-of-visual-words and a bag-of-textual-words, respectively. A hypergraph is generated to model the relevance of images where the edges are the shared visual and textual words. Fidler et al. [14] incorporated the nouns, the cardinality of nouns, and four prepositions to construct a conditional random field for image segmentation, object detection and scene classification. However, their approach manually specifies the correspondence between nouns and object classes, while the proposed method *learns* such correspondence automatically. Moreover, their approach requires bounding boxes to learn the spatial relations implied by the propositions, while the proposed method demands only image-level annotations.

Weakly supervised object detection methods have also been proposed, which require no bounding boxes but image-level labels that indicate the presence of object classes. Cinbis et al. [29] designed a multi-fold multiple instance learning approach that iteratively trains the detector and infers the object locations in the images. Bilen et al. [30] collected a set of exemplars, which best describe the training data, by enforcing "soft" similarity between each possible location in the image. Shi et al. [31] modeled the object classes and image backgrounds together in a single Bayesian latent topic model, which can be learned by a mixture of weakly labeled and unlabeled data. Although the aforementioned methods are able to learn the object models in the absence of bounding boxes, the performance of the weakly supervised object detection methods is often inferior to their supervised counterparts [1, 32].

III. Weakly-supervised Learning of Cross-domain Relations

The relations between scenes, objects and texts are modeled in two levels. Specifically, on the one hand, the relation between the objects and texts is represented by the matching probability of the object classes and the nouns. On the other hand, the relations between the scenes and objects/texts are represented by the frequency of occurrence of the objects/nouns in different scene classes. In the following, we describe how to obtain those cross-domain relations with image-level annotations in detail.

To begin with, some notations will be introduced. A sample is comprised of an image I and a textual description, which is summarized by the cardinalities of nouns $\boldsymbol{b} = (b_1, \cdots, b_N)$. Here, N is the number of noun classes, and b_u is the cardinality of the u-th noun class. If a noun is not in the textual description, its cardinality is zero; otherwise, it is a positive integer.

As in [14], two types of cardinalities of nouns are extracted from the textual descriptions by the Stanford parser [33]: the exact ones and the uncertain ones. The exact cardinalities can be obtained by retrieving the nouns either in singular form or with numerical modifiers. For example, we can extract 3 *kids* (noun with numerical modifier) and 1 *river* (noun in singular form) from the sentence *"three kids are swimming in the river"*. In the other case, the uncertain cardinalities are characterized by nouns in plural form without numerical modifiers. For example, *"some cars"* implies that the number of *cars* is at least 2.

In this way, the observation of a sample can be denoted by $(I, \boldsymbol{b})$. In the training set, the annotations of the samples will be provided, including the scene labels $\{s_t\}_{t=1}^K$ and the cardinalities of the object classes $\{\boldsymbol{c}_t\}_{t=1}^K$, where K is the number of samples in the training set. $s_t \in \{1, 2, \cdots, S\}$ is the scene label of the t-th sample, where S is the number of scene classes. $\boldsymbol{c}_t = (c_1^t, \cdots, c_M^t)$ is the vector of object cardinalities of the t-th sample, where M is the number of object classes, and c_v^t is the cardinality of the v-th object class in the t-th sample.

A. Relation Between Texts and Objects

The relation between the noun classes and the object classes is modeled with their matching probability $P(v|u) : 1 \leq v \leq M, 1 \leq u \leq N$, which depicts the semantic relevance of the nouns and the object classes. For example, $P(\text{PERSON}|\textit{boy})$ is supposed to be large, and $P(\text{CAR}|\textit{bird})$ is supposed to be small. To derive such matching probability, the learning algorithm is composed of two steps: (1) resolve the instance-level matching of the nouns and objects for each training sample by solving a constrained bipartite matching problem, and (2) estimate the conditional matching probabilities $P(v|u)$ based on the resulting correspondence.

The instance-level correspondence of the nouns and the objects of a sample is represented by a matching matrix $X \in \{0, 1\}^{N \times M}$. $X(u, v) = 1$ if the u-th noun class is matched to the v-th object class for this sample; otherwise, $X(u, v) = 0$. Given the cardinalities of nouns and objects of

the sample, namely, $\boldsymbol{b} = (b_1, \cdots, b_N)$ and $\boldsymbol{c} = (c_1, \cdots, c_M)$, X can be obtained by solving the following constrained optimization problem:

$$\hat{X} = \underset{X}{\operatorname{argmax}} \sum_{u=1}^{N}\sum_{v=1}^{M} b_u X(u,v) P_0(v|u),$$
$$\text{s.t.} \begin{cases} \sum_{u=1}^{N} b_u X(u,v) \le c_v, \forall\ 1 \le v \le M, \\ \sum_{v=1}^{M} X(u,v) \le \mathbf{1}(b_u > 0), \forall\ 1 \le u \le N, \end{cases} \quad (1)$$

where $\mathbf{1}$(statement) is a boolean function, which is equals to 1 if the statement is true, and 0 otherwise. $P_0(v|u)$ is the initial matching probability of the nouns and the objects, which can be estimated from $\{\boldsymbol{b}_t\}_{t=1}^K$ and $\{\boldsymbol{c}_t\}_{t=1}^K$ by

$$P_0(v|u) = \frac{\sum_{t=1}^K \mathbf{1}(b_u^t = 1)\mathbf{1}(c_v^t = 1)}{\sum_{t=1}^K \mathbf{1}(b_u^t = 1)}, \quad (2)$$

Obviously, P_0 is the relative frequency of co-occurrence of the nouns and objects in the training set. Although P_0 is a coarse estimate of the relevance of the nouns and objects, empirically it is good enough as a prior guide to the matching problem, because highly relevant objects and nouns are very likely to co-occur in the samples.

The goal of Eq. (1) is to match as many pairs of nouns and objects as possible, and two constraints are enforced to the optimization problem. The first constraint ensures that the total cardinalities of the nouns that are mapped to an object class must not exceed cardinality of that object class. For example, the textual description is *"two boys and a girl are playing in the garden"*, and the annotation indicates that there are 3 instances of PERSON in the image. In this case, this constraint requires that no more than 3 instances from 2 *boys*, 1 *girl* and 1 *garden* can be mapped to the object PERSON.

The second constraint ensures that a noun can be mapped to one object class at most. In reality, the mapping between the nouns to the object classes is many-to-many. In other words, multiple nouns can refer to the same object class, and one noun can also refer to multiple object classes. It is well recognized that people tend to be more specific when they describe images, so that they rarely use one noun to represent multiple object classes. For example, in the UIUC dataset [5], there are only three nouns that can represent multiple object classes, i.e., *furniture* (DININGTABLE, CHAIR, SOFA), *vehicle* (BUS, CAR) and *animal* (BIRD, CAT, DOG, COW, HORSE, SHEEP), and their frequencies of occurrence are extremely low (0.0072, 0.0034 and 0.0026 per caption, respectively). Without loss of generality, a many-to-one mapping from the nouns to the objects is established here. Note that if a noun does not exist in the text (i.e., $\mathbf{1}(b > 0) = 0$), it will not be mapped to any object class.

Eq. (1) is a binary integer linear programming problem, which can be solved efficiently. One simple way to solve it approximately is to relax the problem into a real-valued linear programming problem and then threshold the resulting solution.

Furthermore, we consider a weaker form of annotation, where only the existence of object classes is available, and the object cardinalities are unknown. Such annotations are also abundant over the Internet, such as tags and keywords. The existence of object classes is denoted by a binary vector $\boldsymbol{z} = (z_1, \cdots, z_M)$, where $z_v = 1$ if the v-th object class exists in the image, and $z_v = 0$ otherwise. As the cardinalities of objects are not present, the nouns will be encoded in similar fashion by $\boldsymbol{q} = (q_1, \cdots, q_N)$, where $q_u = 1$ if the u-th noun class exists in the text, and $q_u = 0$ otherwise. With $\boldsymbol{z}$ and $\boldsymbol{q}$, the instance-level matching of nouns and objects can also be obtained by solving a slightly different version of Eq. (1):

$$\hat{X} = \underset{X}{\operatorname{argmax}} \sum_{u=1}^{N}\sum_{v=1}^{M} X(u,v) P_0(v|u),$$
$$\text{s.t.} \begin{cases} \sum_{u=1}^{N} X(u,v) \le z_v, \forall\ 1 \le v \le M, \\ \sum_{v=1}^{M} X(u,v) \le q_u, \forall\ 1 \le u \le N. \end{cases} \quad (3)$$

In this case, the first constraint requires that an object class, if present, can be mapped to no more than one noun class.

Once the matching matrices $\{\hat{X}_t\}_{t=1}^K$ are computed for all training samples, the matching probability between the nouns and the objects can be re-estimated by

$$P(v|u) = \frac{\sum_{t=1}^K b_u^t \hat{X}_t(u,v)}{\sum_{t=1}^K b_u^t}, \quad (4)$$

where the denominator is the number of matched instances of the u-th noun class, and the numerator is the number of instances of the u-th noun class that are mapped to the v-th object class in the training set. Later, experiments in Section V will show that the matching probability obtained by Eq. (4) is very accurate even without the cardinalities, while better results can be achieved using the cardinalities.

B. Relations Between Scenes-Objects and Scenes-Texts

The relations between scenes-objects and scenes-texts are defined in similar fashion, i.e., the frequency of occurrence of the object/noun classes in different scene classes.

To be specific, given the scene labels $\{s_t\}_{t=1}^K$ and the object cardinalities $\{\boldsymbol{c}_t\}_{t=1}^K$ in the training set, the frequency of object class $v \in \{1, \cdots, M\}$ in scene $s \in \{1, \cdots, S\}$ is defined as

$$F_O(s,v) = \frac{\sum_{t=1}^K \mathbf{1}(c_v^t > 0)\mathbf{1}(s_t = s)}{\sum_{t=1}^K \mathbf{1}(s_t = s)}. \quad (5)$$

Here, the denominator is the number of samples in scene class s, and the numerator is the number of samples in scene class s containing object v.

Likewise, given the scene labels $\{s_t\}_{t=1}^K$ and the noun cardinalities $\{\boldsymbol{b}_t\}_{t=1}^K$ in the training set, the frequency of noun $u \in \{1, \cdots, N\}$ in the textual description of scene $s \in \{1, \cdots, S\}$ is

$$F_N(s,u) = \frac{\sum_{t=1}^K \mathbf{1}(b_u^t > 0)\mathbf{1}(s_t = s)}{\sum_{t=1}^K \mathbf{1}(s_t = s)}. \quad (6)$$

Again, the denominator is the number of samples in scene class s, and the numerator is the number of samples in scene class s with noun u in the textual description.

IV. Integrated Model for Joint Image Analysis

With the cross-domain relations learned by the proposed method, an integrated model brings together the scenes, objects and texts to jointly predict: (1) the scene classes of images, (2) the object classes in the images, (3) the cardinalities of object classes, and (4) the locations of object instances.

As illustrated in Fig. 2, the proposed model comprises three types of vertices from different domains, namely, *the scene vertex*, *the object vertices*, and *the text vertices*. Specifically, the model contains only one scene vertex with a random variable $s \in \{1, \cdots, S\}$, which encodes the scene class of the image. Moreover, there are M object vertices in the model, which encode the existence of the M object classes. The v-th object vertex is associated with a random variable $z_v \in \{0, 1\}$, denoting the presence ($z_v = 1$) or absence ($z_v = 0$) of the v-th object class. Consequently, the presence of object classes can be indicated by $\boldsymbol{z} = (z_1, \cdots, z_M)$. Finally, there are N text vertices in the model, which encode the existence of N distinctive nouns in the textual descriptions. Similar to the object vertices, the u-th text vertex is associated with a random variable $q_u \in \{0, 1\}$, denoting the presence ($q_u = 1$) or absence ($q_u = 0$) of the u-th noun in the textual description. Likewise, $\boldsymbol{q} = (q_1, \cdots, q_N)$ is the indicator of the presence of the noun classes.

To measure the probability of the labeling of the vertices in the model, two unary potentials and three binary potentials are carefully designed. The unary potentials include the *scene potential* f_S and the *object potentials* f_O. They evaluate the labeling of the scene vertex and the object vertices based on the visual appearance of the image. To establish the relations between scenes, objects and nouns, three types of edges are introduced to link the vertices from different domains to form a unified graphical model. Each type of edge is associated with a distinct binary potential function, which measures the compatibility of the labeling of the two vertices based on the learned cross-domain relations. In total, there are three types of binary potentials, namely, the *text-object potentials* f_{TO}, the *scene-object potentials* f_{SO}, and the *scene-text potentials* f_{ST}.

A. Unary Potentials

Defined over the scene vertex, the scene potential measures the likelihood with which the image belongs to each scene class based on the predictive scores of a set of off-the-shelf scene classifiers. In particular, the off-the-shelf scene classifiers take the image as input and produce a set of classification scores $\{D_s\}_{s=1}^{S}$, where $D_s \in \mathbb{R}$ is the predictive score of the s-th scene class. Various image classification algorithms, such as SPM [2] and Object Bank [34], can be validated in the proposed model. In turn, the scene potential f_S over the scene vertex s is defined as

$$f_S(s) = D_s, \quad 1 \leq s \leq S. \tag{7}$$

Obviously, f_S favors the scene class with the high predictive score.

A distinct object potential is defined for each object vertex to measure the likelihood that the image contains the object class. First, a set of off-the-shelf object detectors are used to generate object hypotheses, e.g., bounding boxes, from the image. For example, DPM [1] and exemplar SVM [35] perform the object classification over the sliding windows in a greedy manner and eliminate redundant detections via non-maximum suppression. In particular, Barinova et al. [36] developed a probabilistic framework based on Hough transform, which permits detection of multiple objects without invoking non-maximum suppression heuristics.

Each object hypothesis is associated with a classification score, indicating the confidence of the hypothesis, Naturally, the largest classification score of an object class is chosen as the predictive score of the existence of the object class in the image. Finally, the object potential f_O over the object vertex z_v is defined as

$$f_O(z_v) = z_v(d_v - L_v), \quad \forall\ 1 \leq v \leq M. \tag{8}$$

Here, L_v is introduced as the bias of the v-th object class to calibrate the predictive scores, since the distribution of the predictive scores varies across classes.

If the predictive score is large enough (i.e., $d_v > L_v$), the object potential implies that the object class exists in the image (i.e., $f_O(z_v = 1) > f_O(z_v = 0)$). Otherwise it favors that the object class does not exist in the image (i.e., $f_O(z_v = 1) < f_O(z_v = 0)$). Thus, f_O is a reasonable estimator of the existence of the object classes based on image features, whose discriminative capability depends on the off-the-shelf object detectors.

B. Binary Potentials

The binary potentials reflect the relations between the vertices from two different domains. Thus, three types of binary potentials are defined in the model: the text-object potentials f_{TO}, the scene-object potentials f_{SO} and the scene-text potentials f_{ST}. It is worth mentioning that the binary potentials are defined with the learned cross-domain relations as described in Section III.

The text-object potential, linking a text vertex and an object vertex, captures the compatibility of the existence of the noun and the object. Specifically, the binary potential between the u-th text vertex and the v-th object vertex is defined as

$$f_{TO}(q_u, z_v) = \max(P(v|u) - T, 0) q_u z_v, \tag{9}$$

where $P(v|u)$ is the matching probability of the u-th noun class and the v-th object class, and T is a threshold. The underlying rationale is that if the relevance between a noun and an object is sufficiently large, then the existence of the noun in the text strongly indicates the presence of the corresponding object in the image. On the other hand, if the relevance between the noun and the object is small, f_{TO} is always zero regardless of q_u and z_v, meaning that the labeling of one vertex will not affect the labeling of the other one because they are not relevant.

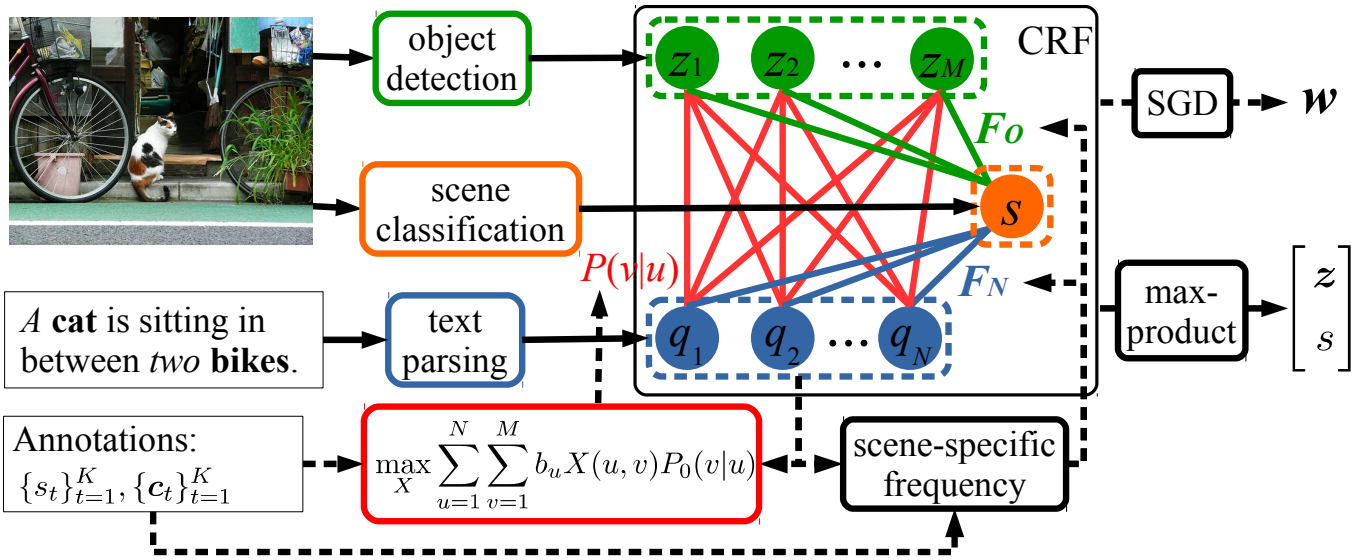


Fig. 2. The framework of the proposed model. A set of off-the-shelf object detectors and scene classifiers provide hypotheses of the objects and the scenes to the underlying model. Together with the learned cross-domain relations among scenes, objects and texts, a conditional random field is constructed to jointly predict the scene classes, presence of objects, the cardinalities of objects, and the locations of objects. The solid path shows the training procedure and the dashed path shows the testing procedure.

The scene-object potential f_{SO} measures the likelihood that an object class exists in different scenes. Specifically, the binary potential between the scene vertex and the v-th object vertex is defined as

$$f_{SO}(s, z_v) = F_O(s, v) z_v + (1 - F_O(s, v))(1 - z_v), \quad (10)$$

where $F_O(s, v)$ is the frequency of occurrence of object v in the scene class s, which is defined in Eq. (5). As the scene label and the object indicator are both to be inferred, f_{SO} favors the most likely combination of the existence of the object class and the scene class.

The scene-text potential captures the dependency of the nouns on scene classes. Thus, the binary potential between the scene vertex and the u-th text vertex is defined as

$$f_{ST}(s, q_u) = F_N(s, u) q_u + (1 - F_N(s, u))(1 - q_u), \quad (11)$$

where $F_N(s, u)$ is the frequency of occurrence of noun u in scene s, which is defined in Eq. (6). Eq. (11) demonstrates that if noun u exists in the textual description, the scene-text potential favors the scene class in which the noun is most likely to appear.

C. Joint Inference of Objects and Scenes

With the unary potentials and the binary potentials, the joint probability of the labeling of the object vertices $\boldsymbol{z}$ and the scene vertex s conditioned on the image I and the text $\boldsymbol{q}$ can be factorized in terms of the graphical model:

$$\begin{aligned} p(\boldsymbol{z}, s|I, \boldsymbol{q}) = \frac{1}{Z(I, \boldsymbol{q})} \exp \Big[& w_1 f_S(s) + w_2 \sum_{v=1}^M f_O(z_v) \\ & + w_3 \sum_{v=1}^M f_{SO}(s, z_v) + w_4 \sum_{u=1}^N f_{ST}(s, q_u) \\ & + w_5 \sum_{u=1}^N \sum_{v=1}^M f_{TO}(q_u, z_v) \Big]. \end{aligned} \quad (12)$$

Here, $\boldsymbol{w} = [w_1, w_2, w_3, w_4, w_5]$ are the weights that balance the contribution of different potentials, and $Z(I, \boldsymbol{q})$ is the normalizing parameter.

To predict the object classes in the image and the scene class of the image, the maximum-a-posteriori (MAP) estimate of Eq. (12) will be computed

$$[s^*, \boldsymbol{z}^*] = \operatorname*{argmax}_{s, \boldsymbol{z}} \; p(s, \boldsymbol{z}|I, \boldsymbol{q}). \quad (13)$$

As the labeling of text vertices can be observed from the textual description, the graphical structure among s and $\boldsymbol{z}$ reduces to a tree. Hence, the optimal solution can be achieved using the max-product algorithm.

D. Cardinalities of Object Classes

Once the object indicator $\boldsymbol{z} = (z_1, \cdots, z_M)$ is inferred by the proposed model, the cardinalities of object classes can be predicted by

$$\hat{c}_v = z_v(\lambda_v^T \hat{c}_v^T + \lambda_v^I \hat{c}_v^I), \quad v = 1, \cdots, M, \quad (14)$$

where $\hat{c}_v^T$ is the cardinality of the v-th object class predicted from the textual description, and $\hat{c}_v^I$ is the cardinality of the v-th object class predicted from the image.

If the proposed model infers that the object class does not exist, its cardinality is zero. Otherwise, the predicted cardinality of the object class is the linear combination of the text-based prediction and the image-based prediction. λ_v^T and λ_v^I are the coefficients of the v-th object class. Next, we describe how to obtain the text-based prediction and the image-based prediction of the object cardinality, as well as the optimal coefficients.

Given the cardinalities of the nouns extracted from the text $\boldsymbol{b} = (b_1, \cdots, b_N)$, the text-based prediction of object cardinality is

$$\hat{c}_v^T = \sum_{u=1}^{N} b_u P(v|u), \quad v = 1, \cdots, M, \tag{15}$$

which is the "expectation" of the cardinalities of the nouns based on their matching probability with the object class. For example, if 2 *boys* and 1 *bike* are extracted from the text, and assuming that the matching probability of PERSON with *boys* and *bike* is 0.9 and 0.1, respectively, the predicted cardinality of PERSON is $2 \times 0.9 + 1 \times 0.1 = 1.9$.

Considering that predicting the object cardinality with only text is not sufficient in case there are objects which are not mentioned in the textual description. The image-based prediction of the object cardinality is devised, which is equal to the number of object instances found by the off-the-shelf object detectors. For example, if the detector of the v-th object class returns 3 instances, typically in the form of bounding boxes, the image-based prediction of the object cardinality is $\hat{c}_v^I = 3$.

Provided with the cardinalities of nouns $\{\boldsymbol{b}_t\}_{t=1}^K$ and the ground truth cardinalities of objects $\{\boldsymbol{c}_t\}_{t=1}^K$ in the training set, the optimal coefficients λ_T and λ_I can be determined by minimizing the mean squared error of the predicted object cardinalities and the ground truth cardinalities:

$$\begin{aligned} \min_{\lambda_v^T, \lambda_v^I} \frac{1}{K} \sum_{t=1}^{K} (\lambda_v^T \hat{C}_v^T + \lambda_v^I \hat{C}_v^I - C_v^t)^2, \\ \text{s.t. } 0 \leq \lambda_v^T, \lambda_v^I \leq 1. \end{aligned} \tag{16}$$

The coefficients are lower bounded by zero and upper bounded by one, because in the extreme case where the objects referred by the text and the detectors are complementary. The true cardinality of the object in the image is the sum of the text-based prediction and the image-based prediction. Note that Eq. (16) is a least squares estimation problem with linear constraints, which can be solved efficiently.

E. Localization of Objects

Once the cardinalities of the object classes are predicted, object detection can be refined to generate more confident object hypotheses. Concretely, the object hypotheses from a certain object detector are sorted in descending order according to their classification scores, so that the first hypothesis is the most confident one. Given the predicted object cardinality $\hat{c}$, the top $\lceil \hat{c} \rceil$ initial hypotheses are chosen as the final object hypotheses, and other hypotheses will be discarded. If the number of initial hypotheses is smaller than $\lceil \hat{c} \rceil$, all of them will be reserved. In this way, a large number of false positive detections can be removed based on the predicted cardinalities of the object classes.

V. Experiments

We mainly evaluate the proposed method on the UIUC dataset [5], and also report its performance on the TSU dataset [13]. As the UIUC dataset, which contains 1000 images from the PASCAL VOC 2008 dataset, is not scene-oriented, we extract a subset of 630 images, whose scene classes can be unambiguously determined. The images are manually categorized into eight well-defined scene classes, namely, *airport*, *dining room*, *farm*, *living room*, *railway*, *racing*, *street*, and *water*. Each image is accompanied with five descriptive captions collected from the human annotators on Amazon Mechanical Turk. In sum, the dataset has $630 \times 5 = 3150$ samples, each of which is an image-caption pair. Table I summarizes the number of samples in the scene classes, and Fig. 3 displays some example images. In addition, the number of instances in the object classes is presented in Table II. Since there are not sufficient instances of BIRD, CAT and DOG, the remaining 17 object classes will be evaluated in the experiments. We randomly select 60% of the samples for training and the rest for testing. Note that if an image is grouped into the training set, all five samples containing this image will be used for training as well, so that the training set and the testing set can be totally disjoint in both images and textual descriptions. It is worth mentioning that all off-the-shelf object detectors are trained upon the PASCAL VOC 2007 dataset, which has no overlap with the UIUC dataset.

TABLE I
SCENE CLASSES AND NUMBER OF SAMPLES

scene	airport	dining room	farm	street
#samples	250	351	580	395
scene	racing	living room	railway	water
#samples	440	365	260	505

TABLE II
NUMBER OF OBJECT INSTANCES IN THE DATASET

PLANE	BICYCLE	BIRD	BOAT	BOTTLE	BUS	CAR
270	240	55	265	195	285	460
CAT	CHAIR	COW	TABLE	DOG	HORSE	MOTOR
0	345	245	285	65	145	280
PERSON	PLANT	SHEEP	SOFA	TRAIN	MONITOR	
1205	135	240	255	260	155	

A. Matching of Nouns and Objects

To evaluate the accuracy of the noun-object matching algorithm, the matching probability $P(v|u)$ is computed with all 5000 samples in the UIUC dataset. Each noun is hard-assigned to the object class of the largest matching probability. We manually labeled the ground truth mapping from the nouns to the object classes. By varying the threshold, a precision-recall curve can be sketched to measure the accuracy of the proposed algorithm. With regard to three conditions:

Fig. 3. Image examples of 8 scene classes in the subset of the UIUC dataset.

1) the initial matching probability, i.e., Eq. (2);
2) the matching probability without the constraint of object cardinalities, i.e., Eq. (3);
3) the matching probability with the constraint of object cardinalities, i.e., Eq. (1).

the precision-recall curves are shown in Fig. 4.

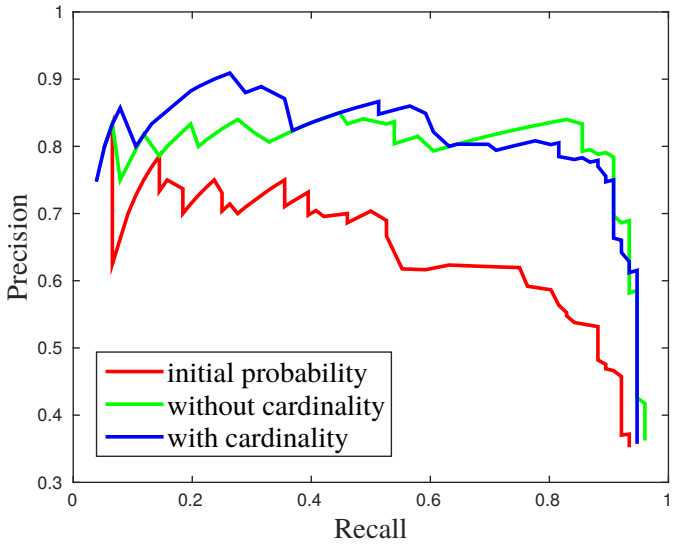


Fig. 4. Precision-recall curves of the noun-object matching.

Upon Fig. 4, the average precision (AP) can be attained for each curve by computing the area under the curve, which measures the overall performance. The AP of the initial matching probability is 0.578. Without the cardinalities of the nouns and the objects, the proposed algorithm increases the AP to 0.731, which obtains a 26.5% improvement over the initial guess. By incorporating the cardinalities, the AP reaches 0.750, which is a 29.7% improvement over the initial guess. On the one hand, the matching probability constrained by the cardinalities is more accurate. On the other hand, the gain brought by the cardinalities is small (2.6%), because most of the samples in the dataset contain only one instance of a specific object or noun, which leads to many identical instance-level matchings with or without cardinalities. Table III lists part of the detailed matching results, which are obtained when the threshold is 0.5. In particular, the last row of the table lists the nouns that are determined to be *unmatched*.

The experimental results clearly demonstrate that the proposed matching algorithm effectively establishes the correspondence between the nouns and the object classes. Taking a close look, even the "wrong" matches, which are not directly related to the matched object classes, are semantically relevant, and will also help reduce the ambiguities arising in object and scene classification.

TABLE III
EXAMPLES OF MAPPING OF NOUNS AND OBJECTS

Object	Correct	Wrong
AEROPLANE	*plane, airplane, jet*	*runway*
BICYCLE	*bicycle, bike*	*cyclist*
BIRD	*bird, hummingbird, seagull, duck*	*beak*
BOAT	*boat, ship, cruise, sailboat*	*shore*
CAR	*car, SUV*	*traffic*
CAT	*cat, kitten*	
COW	*cow, bull, calf*	
DOG	*dog, puppy, pug, chihuahua*	
HORSE	*horse, foal, pony*	*carriage*
MOTORBIKE	*motorbike, scooter, motorcycle*	
PERSON	*man, woman, girl, people, guy,boy*	*family*
SHEEP	*sheep, lamb, goat*	
SOFA	*couch, sofa*	
TRAIN	*train, engine, locomotive*	*railroad*
TVMONITOR	*computer, screen, television, monitor*	
Unmatched	*table, room, front, field, water, road*	*mother*

B. Scene Classification

To evaluate the proposed algorithm in scene classification, we fix the object detector and test various scene classifiers. The popular deformable part-based model (DPM) [1] is utilized as the object detector to compute the unary potentials of the object vertices. Three scene classification algorithms are adopted in the proposed model: spatial pyramid matching (SPM) [2], sparse coding SPM (SC) [37], and locality-constrained linear coding (LLC) [38]. Initially, a bag of C-SIFT descriptors [39] are densely collected for each image. Then, a codebook of 1024 visual words is trained over these descriptors by the k-means algorithm. Consequently, a three-level ($1\times1, 2\times2, 4\times4$) spatial pyramid representation is validated for each image with different encoding schemes depending on the underlying scene classifiers. Finally, a linear SVM classifier is trained for each scene class using the one-*versus*-the-rest scheme.

For each scene classifier, the proposed model is tested under four conditions to evaluate respective effects of scenes, objects, and texts for scene classification.

1) The model consists of only the scene vertex;

2) The model consists of the scene vertex and the object vertices;
3) The model consists of the scene vertex and the text vertices;
4) The full structure of the model is enabled, consisting of the scene vertex, the object vertices and the text vertices.

Fig. 5 compares the mean accuracy of different combinations of scene classifiers and model structures. Table IV displays the class-specific accuracy of each scene class.

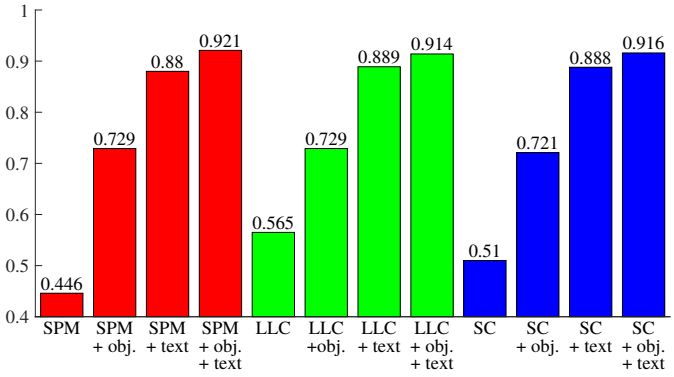


Fig. 5. Mean accuracy of scene classification using SPM, LLC, and SC as the scene classifier.

Several conclusions can be drawn from this experiment. Remarkably, the incorporation of text can significantly improve the accuracy of image-based scene classification. When only the scene classifiers are validated, the use of text improves the mean accuracy of SPM, LLC and SC by 97%, 57% and 74%, respectively. When the scene classifiers and the object detectors are both enabled, the use of text improves the mean accuracy of SPM, LLC and SC by 26%, 25% and 27%, respectively. It can be derived from the fact that textual descriptions are more robust and informative in characterizing scene classes, while image features are sensitive to illumination, deformation and noise. Also, texts provide a wider range of real-world concepts than objects of interest, which are helpful in determining the scene classes of images. Moreover, object detection also plays a key role in scene classification. Without the text, object detection improves the mean accuracy by 63%, 29% and 41% for SPM, LLC and SC, respectively. After incorporating the text, object detection improves the mean accuracy by 5%, 3% and 3% for SPM, LLC and SC, respectively. It can be understood that object-based image features, e.g., Classemes [40] and Object Bank [34], provide mid-level semantic representations of image contents, which are discriminative for image classification. Although most of the objects in images can be covered by the textual descriptions, those non-salient objects, which are not mentioned in the text, can be discovered by object detectors, resulting in marginal improvements over the joint scene-text inference. Furthermore, the proposed model is robust against different types of scene classifiers, because consistent gains are obtained with SPM, LLC and SC.

Furthermore, we also compare the performance of the proposed method with TSU [13] in scene classification. Similar to the proposed method, TSU also utilizes images and texts (tags) for joint image analysis, including scene classification, image annotation, and image segmentation. Experiments are conducted upon the dataset used in [13], which consists of eight scene classes. Each scene class contains 800 images crawled from Flickr, and 600 images are randomly selected for training and the rest for testing. Since the object labels are not available in the dataset, we use the "scene classifier + text" version of the proposed method for comparison, where SPM is used as the off-the-shelf scene classifier. The result is displayed in Table V. Clearly, the proposed method outperforms TSU in every scene class, and the mean accuracy of the proposed method is 63% higher than TSU.

C. Object Classification

To evaluate the proposed model in object classification, we adopt the classic SPM as the scene classifier, and test three object detectors: the classic deformable part based model (DPM) [1], the exemplar-SVM (ESVM) [35], and the state-of-the-art RCNN [32]. Similar to the scene classification, the proposed model is tested under four conditions to evaluate the impact of objects, texts and scenes for object classification:

1) The model consists of only the object vertices, which measures the performance of the object detector;
2) The model consists of the object vertices and the text vertices, so that the existence of objects are inferred using both visual cues and the textual cues, which are linked by the matching probability of the nouns and the objects;
3) The model consists of the object vertices and the scene vertex. In other words, this configuration evaluates the influence of visual context in predicting the presence of objects.
4) The full version of the proposed model is used, consisting of the object vertices, the scene vertex and the text vertices.

Considering the relation between the nouns and the object classes, a text-only baseline is also tested, which uses only the text vertices to predict the objects.

The class-specific *average precision (AP)* of object classification is displayed in Table VI, and the mean AP over the 17 object classes is illustrated in Fig. 6, where the mean AP of the text-only baseline appears with the dash line.

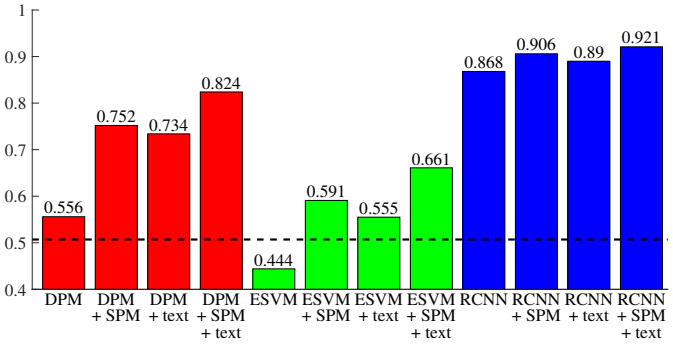


Fig. 6. Mean average precision of object classification using DPM and Exemplar-SVM as the object detector.

Table VI demonstrates that the proposed model gets the highest AP in 16 out of 17 classes when DPM and ESVM

TABLE IV
ACCURACY OF SCENE CLASSIFICATION USING SPM, LLC AND SC AS THE SCENE CLASSIFIERS

scene	airport	dining room	farm	living room	railway	racing	street	water
SPM	.450	.393	.804	.552	.086	.143	.648	.375
SPM + obj.	.750	.679	.891	.724	.714	.619	.849	.469
SPM + text	.880	.821	.991	.876	.857	.876	.869	**.819**
SPM + obj. + text	**.980**	**.871**	**.996**	**.883**	**.943**	**.933**	**.935**	.806
LLC	.600	.357	.761	.759	.429	.190	.673	.531
LLC + obj.	.700	.714	.870	.828	.714	.571	.874	.406
LLC + text	.960	.800	.987	.890	.874	.886	.869	**.825**
LLC + obj. + text	**.970**	**.850**	**.991**	**.890**	**.926**	**.924**	**.930**	.806
SC	.600	.393	.783	.655	.286	.048	.724	.312
SC + obj.	.650	.750	.891	.759	.714	.619	.849	.375
SC + text	.930	.800	.991	.883	.874	.895	.869	**.825**
SC + obj. + text	**.960**	**.893**	**.991**	**.897**	**.920**	**.914**	**.925**	.806

TABLE V
SCENE CLASSIFICATION ACCURACY OF TSU AND THE PROPOSED METHOD

	badminton	bocce	croquet	polo	rockclimbing
TSU	.67	.41	.68	.56	.56
Proposed	**.98**	**.89**	**.90**	**.98**	**.84**
	rowing	sailing	snowboarding		*avg.*
TSU	.35	.57	.54		.54
Proposed	**.84**	**.78**	**.88**		**.88**

are the object detector, and in 14 out of 17 classes when RCNN is the object detector. Fig. 6 shows that the object detectors often obtain poor performance on their own. When scene classification is performed, the mean AP of object classification is improved by 35%, 33% and 4% for DPM, ESVM and RCNN, respectively, because there is a distinct distribution of object classes in each scene class, which, in return, helps determine the objects in the image based on the information about the scene. Moreover, when the texts are used, the mean AP is be improved by 32%, 25% and 3% for DPM, ESVM and RCNN, respectively, because the scene classes can be well characterized by the nouns in the textual description. Eventually, the best performance is obtained by jointly modeling objects, scenes and texts, which improves the performance of the object detectors by 48%, 49% and 6% for DPM, ESVM and RCNN, respectively. Overall, RCNN obtains the best performance among the three object detectors, showing the effectiveness of the deep networks in object classification.

D. Cardinalities of Object Classes

To evaluate the accuracy of the prediction of object cardinalities, SPM is selected to classify the scene classes of images, and the best-performing RCNN in Section V-C detects the objects in the image. Here, we use the *mean absolute error* to measure the accuracy of object cardinality prediction. The class-specific mean absolute error is displayed in Table VII.

Overall, the predicted object cardinality is very accurate, and the mean absolute error over the 17 object classes is 0.140. In detail, Table VII suggests that the performance of object cardinality prediction mainly relies on two factors: (1) the number of instances in the image, and (2) the number of nouns referring to the object class. To be specific, the three object classes of the lowest mean absolute error are MONITOR, HORSE and TABLE. Usually, there are less than 2 instances of these object classes in an image, and the nouns referring them are also limited. On the other hand, the three object classes of the largest mean absolute error are PERSON, CAR, and CHAIR, which often have multiple instances in an image and various aliases in nouns.

E. Localization of Objects

For object detection, the initial bounding boxes from the object detector are refined by the proposed scheme as in Section IV-E. We also test three object detection algorithms: DPM, ESVM and RCNN. Following the criterion of the PASCAL VOC Challenge, a detected bounding box is correct only if its intersection-over-union ratio with the ground truth is larger than 0.5. To evaluate the performance of the proposed scheme, we compare the average precision of the initial bounding boxes and refined bounding boxes of the three object detectors, and the class-specific average precision is displayed in Table VIII.

For the three object detectors, the average precision of the refined bounding boxes is higher than the initial bounding boxes from the object detectors for all object classes. By applying the proposed scheme, the mean average precision is improved by 38%, 56% and 9% for DPM, ESVM and RCNN, respectively, over the 17 object classes.

F. Impact of the Number of Nouns

In this experiment, we further evaluate the influence of the number of nouns to the performance of scene classification and object classification. Likewise, DPM is used as the object detector and SPM is used as the scene classifier. The nouns are sorted in descending order by their frequency of occurrence in the training set, and the top n nouns are used in the proposed model. We change n from 0 to 260, and derive the relations of the mean accuracy of scene classification and the mean AP of object classification with respect to the number of nouns, which is shown in Fig. 7.

Fig. 7 shows that the accuracy of scene classification increases as more nouns are used in the model, and almost converges when the number of nouns is larger than 200. For object classification, the mean AP also increases with the number of nouns, and converges when the number of nouns is larger than 50. As the UIUC dataset is object-oriented, the nouns used to describe the objects often have high frequency

TABLE VI
AVERAGE PRECISION OF OBJECT CLASSIFICATION USING DPM, ESVM AND RCNN AS OBJECT DETECTORS

	PLANE	BICYCLE	BOAT	BOTTLE	BUS	CAR	CHAIR	COW	TABLE
Text only	.581	.313	.552	.076	.720	.419	.330	.804	.638
DPM	.696	.523	.209	.567	.633	.903	.633	.343	.546
DPM + text	.917	.686	.656	.611	.845	.914	.584	.852	.786
DPM + scene	.950	.632	.784	.804	.783	.922	**.770**	.459	.854
DPM + scene + text	**.985**	**.720**	**.903**	**.813**	**.887**	**.922**	.619	**.876**	**.866**
ESVM	.552	.347	.188	.185	.589	.656	.422	.390	.441
ESVM + text	.686	.485	.538	.190	.763	.667	.451	.657	.759
ESVM + scene	.744	.413	.546	.270	.650	.673	.468	.440	.758
ESVM + scene + text	**.838**	**.522**	**.756**	**.273**	**.802**	**.682**	**.496**	**.678**	**.872**
RCNN	.872	.825	.834	.755	.894	.971	.920	.848	.908
RCNN + text	.932	.918	.887	.754	**.916**	.978	**.937**	.948	.909
RCNN + scene	.912	.849	.953	.755	.890	.958	.925	.838	**.926**
RCNN + scene + text	**.958**	**.925**	**.967**	**.756**	.915	**.980**	.935	**.950**	.924

	HORSE	MOTORBIKE	PERSON	PLANT	SHEEP	SOFA	TRAIN	MONITOR
Text only	.775	.396	.749	.090	.758	.695	.590	.137
DPM	.428	.753	.909	.088	.522	.343	.765	.596
DPM + text	.891	.840	.944	.094	.830	.801	.900	.631
DPM + scene	.509	.838	.929	.160	.660	.733	.976	.723
DPM + scene + text	**.978**	**.892**	**.952**	**.163**	**.962**	**.804**	**.991**	**.856**
ESVM	.161	.676	.805	.190	.636	.202	.601	.487
ESVM + text	.365	.759	.833	.190	.833	.636	.754	.489
ESVM + scene	.185	.765	.809	.197	.706	.491	.822	**.498**
ESVM + scene + text	**.397**	**.829**	**.839**	**.202**	**.864**	**.798**	**.895**	.497
RCNN	.868	.912	.977	.684	.908	.771	.902	.901
RCNN + text	.911	.945	.985	.716	.944	.872	.937	.921
RCNN + scene	.872	.925	.979	.696	.919	.866	.957	.918
RCNN + scene + text	**.911**	**.953**	**.986**	**.724**	**.953**	**.910**	**.975**	**.934**

TABLE VII
MEAN ABSOLUTE ERROR OF OBJECT CARDINALITY PREDICTION

PLANE	BICYCLE	BOAT	BOTTLE	BUS	CAR	CHAIR	COW	TABLE
.098	.085	.098	.084	.067	.371	.217	.083	.060
HORSE	MOTORBIKE	PERSON	PLANT	SHEEP	SOFA	TRAIN	MONITOR	
.055	.129	.577	.073	.195	.064	.067	.054	

TABLE VIII
AVERAGE PRECISION OF OBJECT DETECTION USING DPM, ESVM, AND RCNN AS OBJECT DETECTORS

	PLANE	BICYCLE	BOAT	BOTTLE	BUS	CAR	CHAIR	COW	TABLE
DPM (all)	.392	.333	.018	.349	.439	.531	.320	.120	.351
DPM (refined)	**.481**	**.491**	**.067**	**.420**	**.530**	**.585**	**.384**	**.226**	**.692**
ESVM (all)	.163	.148	.001	.026	.147	.215	.143	.089	.029
ESVM (refined)	**.208**	**.293**	**.006**	**.122**	**.212**	**.316**	**.248**	**.164**	**.046**
RCNN (all)	.867	.696	.436	.490	.655	.625	.468	.613	.857
RCNN (refined)	**.889**	**.793**	**.482**	**.532**	**.676**	**.651**	**.504**	**.728**	**.882**

	HORSE	MOTORBIKE	PERSON	PLANT	SHEEP	SOFA	TRAIN	MONITOR
DPM (all)	.343	.492	.627	.034	.176	.207	.523	.677
DPM (refined)	**.567**	**.564**	**.674**	**.108**	**.241**	**.634**	**.668**	**.883**
ESVM (all)	.127	.336	.145	.013	.224	.004	.236	.341
ESVM (refined)	**.223**	**.398**	**.191**	**.064**	**.325**	**.028**	**.310**	**.564**
RCNN (all)	.750	.757	.629	.434	.583	.535	.729	.628
RCNN (refined)	**.851**	**.803**	**.637**	**.540**	**.630**	**.646**	**.753**	**.676**

of occurrence in the captions, thus, the curve of mean AP converges quickly as the number of nouns increases. However, in addition to the nouns that refer to the objects of interest, scenes can be characterized by a wider range of nouns. Hence, the curve of mean accuracy of scene classification converges until the number of nouns reaches 200.

G. Computational Complexity

To evaluate the computational complexity of the proposed method in the training phase and the testing phase, we conduct the experiments on a PC with 2.27GHz Intel Xeon E5520 CPU, 4GB RAM, and Ubuntu 14.04 LTS operating system. The algorithm is implemented with Matlab. The time for object detection and scene classification varies significantly with different off-the-shelf algorithms, and is not measured consequently. In the training phase, the extraction of nouns and the frequency of the objects and nouns in different scene classes can be fulfilled in few seconds. The computation of the matching probability and the optimization of the potential weights take about 3 minutes and 3.5 minutes, respectively. In the testing phase, the inference of the CRF model takes about

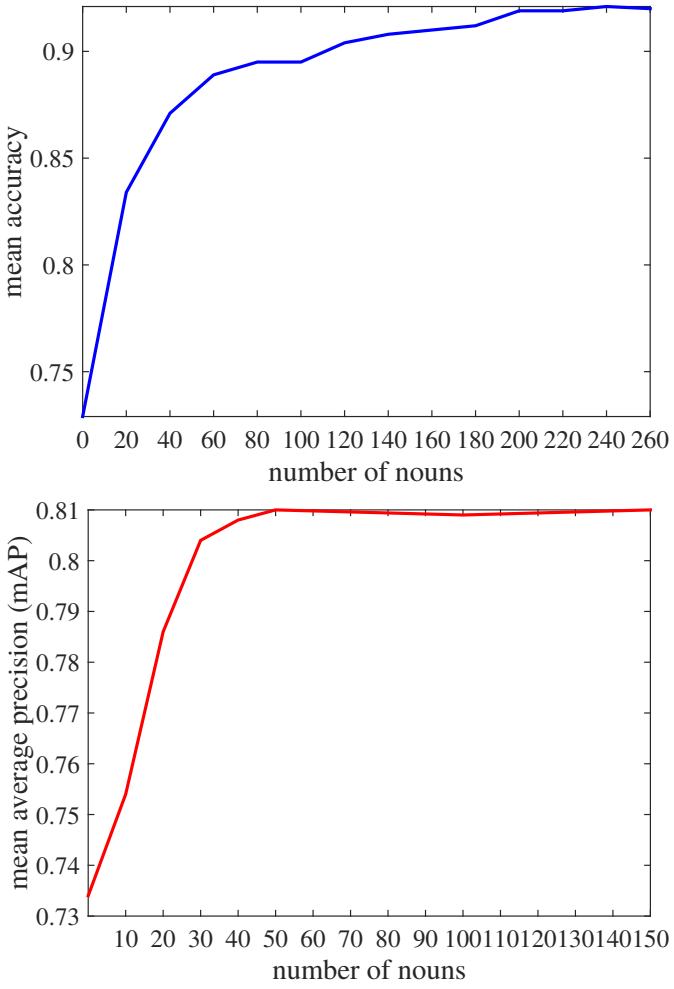


Fig. 7. The influence of the number of nouns to the mean accuracy of scene classification and the mean AP of object classification.

2 minutes over the 1260 testing samples. It can be seen that the proposed algorithm is highly efficient, although the off-the-shelf object detection and scene classification may take quite a long time.

VI. Conclusion

This paper proposes an integrated model to jointly recognize scenes and objects by leveraging the associated textual descriptions, and presents a learning algorithm to estimate the model efficiently. The learning process requires only coarsely labeled images without instance-level annotations. The key to the learning algorithm lies in that it can automatically infer the instance-level correspondence over the training set by solving a constrained bipartite matching problem. The proposed method can leverage the vast number of web images that come with textual descriptions, without requiring enormous amount of efforts to annotate them. By taking advantage of the cross-domain relations, comprehensive experiments on a real world dataset show that the proposed model can obtain remarkable performance improvement in comparison with the classifiers in isolation. However, the textual descriptions from the annotators are quite different from what people might actually use to describe images in real scenarios. For example, the image descriptions from social media only emphasize on one or two objects of interest in a more specific way, while the image annotators tend to describe everything in the image in plain language. Therefore, it would be a promising future work to develop text-aided image understanding algorithms on real-world basis.

References

[1] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, vol. 32, no. 9, pp. 1627–1645, Sept. 2010.

[2] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'06)*, New York, NY, USA, June 2006, pp. 2169–2178.

[3] E. Cheng, F. Jing, and L. Zhang, "A unified relevance feedback framework for web image retrieval," *IEEE Trans. Image Processing (TIP)*, vol. 18, no. 6, pp. 1350–1357, June 2009.

[4] M. Everingham, S. Eslami, L. Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge - a retrospective," *Int'l J. Computer Vision (IJCV)*, 2014.

[5] A. Farhadi, M. Hejrati, M. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *Proc. European Conf. Computer Vision (ECCV'10)*, Heraklion, Crete, Greece, Sept. 2010, pp. 15–29.

[6] Y. Wang and G. Mori, "A discriminative latent model of image region and object tag correspondence," in *Advances in Neural Information Processing Systems (NIPS'10)*, Vancouver, BC, Canada, Dec. 2010.

[7] R. Socher and L. Fei-Fei, "Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'10)*, San Francisco, CA, USA, June 2010, pp. 966–973.

[8] W. Lu, J. Li, T. Li, W. Guo, H. Zhang, and J. Guo, "Web multimedia object classification using cross-domain correlation knowledge," *IEEE Trans. Multimedia (TMM)*, vol. 15, no. 8, pp. 1920–1929, Dec. 2013.

[9] X. Benavent, A. Serrano, R. Granados, J. Benavent, and E. Ves, "Multimedia information retrieval based on late semantic fusion approaches: Experiments on a wikipedia image collection," *IEEE Trans. Multimedia (TMM)*, vol. 15, no. 8, pp. 2009–2021, Dec. 2013.

[10] V. Ordonez, J. Deng, Y. Choi, A. Berg, and T. Berg, "From large scale image categorization to entry-level categories," in *Proc. IEEE Int'l Conf. Computer Vision (ICCV'13)*, Sydney, NSW, Australia, Dec. 2013, pp. 2768–2775.

[11] A. Karpathy, A. Joulin, and L. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Advances in Neural Information Processing Systems (NIPS'14)*, Montreal, Quebec, Canada, Dec. 2014.

[12] M. Katsurai, T. Ogawa, and M. Haseyama, "A cross-modal approach for extracting semantic relationships between concepts using tagged images," *IEEE Trans. Multimedia (TMM)*, vol. 16, no. 4, pp. 1059–1074, June 2014.

[13] L. Li, R. Socher, and L. Fei-Fei, "Towards total scene understanding: Classification, annotation and segmentation in an automatic framework," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'09)*, Miami, FL, USA, June 2009, pp. 2036–2043.

[14] S. Fidler, A. Sharma, and R. Urtasun, "A sentence is worth a thousand pixels," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'13)*, Portland, OR, USA, June 2013, pp. 1995–2002.

[15] A. Gupta and L. Davis, "Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers," in *Proc. European Conf. Computer Vision (ECCV'08)*, Marseille, France, Oct. 2008, pp. 16–29.

[16] Y. Li, D. Crandall, and D. Huttenlocher, "Landmark classification in large-scale image collections," in *Proc. IEEE Int'l Conf. Computer Vision (ICCV'09)*, Kyoto, Japan, June 2009, pp. 1957–1964.

[17] G. Wang, D. Hoiem, and D. Forsyth, "Building text features for object image classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'09)*, Miami, FL, USA, June 2009, pp. 1367–1374.

[18] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'10)*, San Francisco, CA, USA, June 2010, pp. 902–909.

[19] G. Iyengar, P. Ircing, M. Krause, D. Petkova, P. Duygulu, S. Khudanpur, R. Manmatha, B. Pytlik, S. Feng, D. Klakow, H. Nock, and P. Virga, "Joint visual-text modeling for automatic retrieval of multimedia documents," in *Proc. ACM Int'l Conf. Multimedia (ACMM'05)*, Singapore, Nov. 2005, pp. 21–30.

[20] Y. Yeh, C. Huang, and Y. Wang, "Heterogeneous domain adaptation and classification by exploiting the correlation subspace," *IEEE Trans. Image Processing (TIP)*, vol. 23, no. 5, pp. 2009–2018, March 2014.

[21] K. Barnard, P. Duygulu, D. Forsyth, N. Freitas, D. Blei, and M. Jordan, "Matching words and pictures," *J. Machine Learning Research (JMLR)*, vol. 3, pp. 1107–1135, March 2003.

[22] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *J. Machine Learning Research (JMLR)*, vol. 3, pp. 993–1022, March 2003.

[23] D. Blei and M. Jordan, "Modeling annotated data," in *Proc. 26th Annual Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'03)*, Toronto, Canada, July 2003, pp. 127–134.

[24] D. Putthividhy, H. Attias, and S. Nagarajan, "Topic regression multimodal latent dirichlet allocation for image annotation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'10)*, San Francisco, CA, USA, June 2010, pp. 3408–3415.

[25] Y. Jia, M. Salzmann, and T. Darrell, "Learning cross-modality similarity for multinomial data," in *IEEE Int'l Conf. Computer Vision (ICCV'11)*, Barcelona, Spain, Nov. 2011, pp. 2407–2414.

[26] H. Larochelle and S. Lauly, "A neural autoregressive topic model," in *Advances in Neural Information Processing Systems (NIPS'12)*, Lake Tahoe, NV, USA, Dec. 2012, pp. 2717–2725.

[27] Z. Niu, G. Hua, X. Gao, and Q. Tian, "Semi-supervised relational topic model for weakly annotated image recognition in social media," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'14)*, Columbus, OH, USA, June 2014, pp. 4233–4240.

[28] Y. Gao, M. Wang, Z. Zha, J. Shen, X. Li, and X. Wu, "Visual-textual joint relevance learning for tag-based social image search," *IEEE Trans. Image Processing (TIP)*, vol. 22, no. 1, pp. 363–376, Jan. 2013.

[29] R. Cinbis, J. Verbeek, and C. Schmid, "Multi-fold mil training for weakly supervised object localization," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'14)*, Columbus, OH, USA, June 2014, pp. 2409–2416.

[30] H. Bilen, M. Pedersoli, and T. Tuytelaars, "Weakly supervised object detection with convex clustering," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'15)*, Boston, MA, USA, June 2015, pp. 1081–1089.

[31] Z. Shi, T. Hospedales, and T. Xiang, "Bayesian joint topic modelling for weakly supervised object localisation," in *IEEE Int'l Conf. Computer Vision (ICCV'13)*, Sydney, Australia, Dec. 2013, pp. 2984–2991.

[32] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'14)*, Columbus, OH, USA, June 2014, pp. 580–587.

[33] D. Klein and C. Manning, "Fast exact inference with a factored model for natural language parsing," in *Advances in Neural Information Processing Systems (NIPS'03)*, Whistler, BC, Canada, December 2003.

[34] L. Li, H. Su, E. Xing, and L. Fei-Fei, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Advances in Neural Information Processing Systems (NIPS'10)*, Vancouver, BC, Canada, Dec. 2010, pp. 1378–1386.

[35] T. Malisiewicz, A. Gupta, and A. Efros, "Ensemble of exemplar-SVMs for object detection and beyond," in *IEEE Int'l Conf. Computer Vision (ICCV'11)*, Washington, DC, USA, Nov. 2011, pp. 89–96.

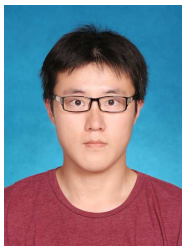
[36] O. Barinova, V. Lempitsky, and P. Kholi, "On detection of multiple object instances using hough transforms," *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, vol. 34, no. 9, pp. 1773–1784, Sept. 2012.

[37] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'09)*, Miami, FL, USA, June 2009, pp. 1794–1801.

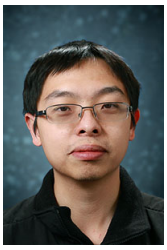
[38] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'10)*, San Francisco, CA, USA, June 2010, pp. 3360–3367.

[39] A. Abdel-Hakim and A. Farag, "CSIFT: A SIFT descriptor with color invariant characteristics," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'06)*, New York, NY, USA, June 2006, pp. 1978–1983.

[40] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *Proc. European Conf. Computer Vision (ECCV'10)*, Heraklion, Crete, Greece, Sep. 2010, pp. 776–789.



Botao Wang received the B.S. degree in electronic engineering, in 2010, from Shanghai Jiao Tong University, Shanghai, China, where he is currently working toward the Ph.D. degree. His main research interests include object detection, scene classification, and image understanding.



Dahua Lin received his Ph.D. from the department of EECS at Massachusetts Institute of Technology in 2012. He received his M.Phil. from the department of Information Engineering at the Chinese University of Hong Kong in 2007, and B.Eng. from the department of Electrical Engineering and Information Science at the University of Science and Technology of China in 2004. He was a research intern at Microsoft Research Silicon Valley, Microsoft Research Redmond, and Microsoft Research Asia, respectively in 2010, 2009, and 2004. He received the Best Student Paper Award at NIPS 2010, and the Outstanding Reviewer Awards at ICCV 2009 and ICCV 2011.

His research spans multiple areas in machine learning, data science, and computer vision. In particular, he is interested in developing new probabilistic models and machine learning techniques for large-scale data analysis, as well as their applications in image and text understanding. He has also worked on a variety of topics in computer vision and pattern recognition before joining CUHK.

Hongkai Xiong (M'01-SM'10) received the Ph.D. degree in communication and information system from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2003. Since then, he has been with the Department of Electronic Engineering, SJTU, where he is currently a distinguished Professor. From December 2007 to December 2008, he was with the Department of Electrical and Computer Engineering, Carnegie Mellon University (CMU), Pittsburgh, PA, USA, as a Research Scholar. From 2011 to 2012, he was a Scientist with the Division of Biomedical Informatics at the University of California (UCSD), San Diego, CA, USA.

His research interests include source coding/network information theory, signal processing, computer vision and machine learning. He has published over 140 refereed journal/conference papers. He is the recipient of the Best Student Paper Award at the 2014 IEEE Visual Communication and Image Processing (IEEE VCIP'14), the Best Paper Award at the 2013 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (IEEE BMSB'13), and the Top 10% Paper Award at the 2011 IEEE International Workshop on Multimedia Signal Processing (IEEE MMSP'11).

In 2014, he was granted National Science Fund for Distinguished Young Scholar and Shanghai Youth Science and Technology Talent as well. In 2013, he was awarded a recipient of Shanghai Shu Guang Scholar. From 2012, he is a member of Innovative Research Groups of the National Natural Science. In 2011, he obtained the First Prize of the Shanghai Technological Innovation Award for Network-oriented Video Processing and Dissemination: Theory and Technology. In 2010 and 2013, he obtained the SMC-A Excellent Young Faculty Award of Shanghai Jiao Tong University. In 2009, he was awarded a recipient of New Century Excellent Talents in University, Ministry of Education of China. He served as TPC members for prestigious conferences such as ACM Multimedia, ICIP, ICME, and ISCAS. He is a senior member of the IEEE (2010).

Yuan F. Zheng (F'97) received the MS and Ph.D. degrees in Electrical Engineering from The Ohio State University, in Columbus, Ohio in 1980 and 1984, respectively. His undergraduate education was received at Tsinghua University, Beijing, China in 1970. From 1984 to 1989, he was with the Department of Electrical and Computer Engineering at Clemson University, Clemson, South Carolina. Since August 1989, he has been with The Ohio State University, where he is currently Professor and was the Chairman of the Department of Electrical and Computer Engineering from 1993 to 2004. From 2004 to 2005, Professor Zheng spent sabbatical year at the Shanghai Jiao Tong University in Shanghai, China and continued to be involved as Dean of School of Electronic, Information and Electrical Engineering until 2008. Professor Zheng is an IEEE Fellow.

Professor Zheng's research interests include two aspects. One is in wavelet transform for image and video, and object classification and tracking, and the other is in robotics which includes robotics for life science applications, multiple robots coordination, legged walking robots, and service robots. Professor Zheng was and is on the editorial board of five international journals. Professor Zheng received the Presidential Young Investigator Award from Ronald Reagan in 1986, and the Research Awards from the College of Engineering of The Ohio State University in 1993, 1997, and 2007, respectively. Professor Zheng along with his students received the best conference and best student paper award a few times in 2000, 2002, and 2006, and received the Fred Diamond for Best Technical Paper Award from the Air Force Research Laboratory in Rome, New York in 2006. In 2004, Professor Zheng was appointed to the International Robotics Assessment Panel by the NSF, NASA, and NIH to assess the robotics technologies worldwide in 2004 and 2005.