ObjectSense: A Scalable Multi-Objects Recognition System Based on Partial-Duplicate Image Retrieval

Shuang Wang[1], Yunfeng Xue[1], Lingyang Chu[1], Yuhao Jiang[2], Shuqiang Jiang[1]
Key Lab of Intelligent Information Processing, Institute of Computing Tech., CAS[1], Beijing, China
College of Information Science and Engineering, Shandong University of Science and Technology[2], Qingdao, China
{shuang.wang, yunfeng.xue, lingyang.chu, shuqiang.jiang}@vipl.ict.ac.cn[1]; yhjiang125@gmail.com[2]

ABSTRACT

In this demo, we present ObjectSense, a scalable object recognition system that recognizes multiple objects present in a static image or in the camera frames. Instead of applying learning based recognition framework, this system identifies objects through Partial-Duplicate Image Retrieval (PDIR) based method. First, objects are identified by measuring the similarity between an incoming image and reference image corpus that are labeled with the objects. To compute image similarities, we explore the Consistency Graph Model (CGM), which robustly rejects spatially inconsistent feature matches with the advantage of orientations and positions of local features. Then a kNN voting method is used to decide the object category based on the quantized image similarities. ObjectSense is scalable with promisingly high recall and accuracy, which fits well into recognition-guided shopping and human computer interaction. We built ObjectSense on two platforms, PC and Android.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models, Search Process; I.4.8 [**Scene Analysis**]: Object Recognition

Keywords

Multi-Objects Recognition, Partial-Duplicate Image Retrieval, kNN, Accuracy, Scalability.

1. INTRODUCTION

In recent years, computer-vision-based object recognition has seen its widespread application requirements across industries. However, state-of-the-art approaches are still challenged by problems of accuracy and speed.

The popular visual recognition works employ multi-class image classification-based framework, which usually proceed in two stages. First, the potentially emerged objects are represented by extracted visual features. Second, for each object label, a one-versus-all classifier is applied to reach a decision regarding the represented object. For this framework, the problems are: 1) Current classification methods still bring unsatisfactory accuracy rate. 2) Classification performance does not scale well with object classes. 3) Classification needs large training data to ensure its efficiency. 4) It does not work well when there are multiple objects in the view.

ICMR'13, April 16–20, 2013, Dallas, Texas, USA.
ACM 978-1-4503-2033-7/13/04.

Our PDIR-based object recognition system has the following four advantages: 1) Scalable and stable. High accuracy rate is achieved for both small and large object corpus size. 2) High recall rate even when multiple objects appear. 3) No complex training or learning procedures are required. 4) Data corpus is incremental for new coming objects with few operations.

2. DATA CORPUS

For this demo, we generate a set of reference images for each object and take right the object as the ground-truth label for them. The reference images could be either downloaded from the Internet or taken with a camera of the real object. For each specified object, better performance will be achieved if reference images are under different imaging conditions (e.g. viewpoint, illumination, object distance). We consider such set of reference images to be an appropriate representation of corpus objects, some of which are displayed in Figure 1.

Figure 1: Objects from the data corpus

3. ALGORITHM

ObjectSense identifies objects by measuring the similarity between an incoming (or query) image and reference images of each known object in the data corpus and telling the most possible emerged objects based on the similarities.

We start with PDIR-based framework. In order to make this system robust to different imaging conditions such as viewpoint, illumination, occlusion and object distance(object size), the standard SIFT features [1] are extracted to represent the incoming image, which is expected to perform better for objects with rich texture information. Each SIFT will be quantized through a hierarchical vocabulary tree [2]. In this demo, we set the vocabulary size as 500,000. Two candidate SIFTs from query image and one corpus reference image are initially matched if they are assigned to the same visual word, known as a candidate match. However, because valid feature matches from the small duplicate object region, as a minority, may be overwhelmed by the rest noisy matches from large background region and thus weakening the relation between relevant images, we establish a mutual verification scheme between candidate matches through Consistency Graph Model (CGM), where each node corresponds

to a candidate match and each edge weight corresponds to the mutual spatial consistency. We measure the mutual spatial consistency of two candidate matches in a coarse-to-fine manner through evenly sectored polar coordinate systems which softly quantize and combine the orientations and positions of the SIFTs. Consequently, spatially inconsistent matches will be removed by finding the most strongly connected subgraph within the CGM. After that, valid SIFT matches are held for the image similarity scoring pipeline together with the quantized mutual spatial consistency. The CGM greatly removes false recognition and brings high accuracy. Any feature match that passes the CGM could vote for its owner reference image, thus making the voting procedures for each object independent of each other. The independence of voting, together with the CGM, makes it possible to stably recognize every object in the incoming image, i.e. high recall. An inverted file [3] is used here to make the feature matching step very fast and to enable great scalability. New objects can be added to this system for recognition by simply collecting corresponding reference images and indexing them incrementally into the inverted file.

Now we have all corpus reference images scored based on their similarity to incoming image I_q. Next, we take kNN-like voting scheme to identify objects observed or labels for I_q. The j^{th} corpus reference image labeled with the i^{th} corpus object o_i is denoted by $I_j(o_i)$, $j \in [1, l_{o_i}]$. The final score of o_i will be calculated as follows:

$$s(o_i) = {}^{1}/_{n} \cdot \mathrm{Top}(\mathrm{SL}(o_i), n)$$

where

$$\mathrm{SL}(o_i) = \{ \mathrm{sim}(I_j(o_i), I_q) \mid j \in [1, l_{o_i}] \}$$

$\mathrm{sim}(\cdot,\cdot)$ is the similarity score between two images, $\mathrm{Top}(\cdot,n)$ takes the n largest elements. We consider o_i appears in I_q if $s(o_i) \geq t$, where t is a well-tuned threshold. Now let's assume that o_i is the ground truth label for I_q. If n is close to l_{o_i}, $s(o_i)$ will inevitably be small and false negative happens. If n is close to one, false positive rate rises. For this demo, we use $n = 3$.

Figure 2: Screenshot of ObjectSense PC interface

4. INTERFACE

ObjectSense takes the current frame from the video sequence captured by a connected camera as the input and perform recognition algorithm on it. Real-time recognition is achieved by repeating the steps above. To present the outputs clearly and user friendly, ObjectSense displays each recognized object with both its name and its thumbnail as shown in Figure 2. We have also built a similar Android version for ObjectSense.

5. EVALUATION

To evaluate the performance of ObjectSense, we asked 20 users to test our system with 100 real objects, of which 70 are within the corpus while the rest are not. Each user conducted 200 recognition activities by interacting with ObjectSense holding one or several random objects in hand at a time. The average recognition precision rate is above 95% and the recall rate is higher than 90%.

Now, ObjectSense supports several hundred object categories and is feasible to scale to thousands with few expenses. Figure 3 shows the scalability of our system. The average time cost of recognition per frame stays relatively stable as corpus size (total number of corpus objects) increases to even 400. Moreover, SIFT extraction takes up a huge part of recognition in terms of time cost and our recognition framework is proved quite efficient.

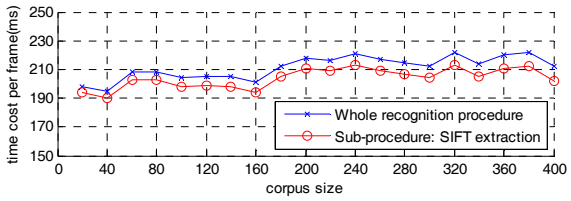

Figure 3: Speed test on different corpus sizes (total number of corpus objects)

6. CONCLUSION AND FUTURE WORK

In this demo, we have presented ObjectSense, a scalable multi-objects recognition system. We employ PDIR-based framework to score reference images in the data corpus and kNN-like scheme to vote for corpus objects based on the scores. ObjectSense works stably with high recall, accuracy and scalability.

In the future, we will experiment with some other image features to make ObjectSense more robust to objects with insufficient texture information. We will further bring ObjectSense closer to users' lives by making the best of the recognized objects such as e-shopping guide.

7. ACKNOWLEDGMENTS

This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400, in part by National Natural Science Foundation of China: 61070108, and 61035001, in part by the Key Technologies R&D Program of China under Grant no. 2012BAH18B02.

8. REFERENCE

[1] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, *60*(2), 91-110.

[2] Nister, D., & Stewenius, H. (2006). Scalable recognition with a vocabulary tree. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (Vol. 2, pp. 2161-2168). IEEE.

[3] Sivic, J., & Zisserman, A. (2003, October). Video Google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on* (pp. 1470-1477). IEEE.