Unsupervised Categorization (Filtering) of Google Images Based on Visual Consistency

Pooyan Fazli
Department of Computer Science
University of British Columbia
Vancouver, BC, Canada V6T 1Z4
pooyanf@cs.ubc.ca

Ara Bedrosian
Department of Computer Science
University of British Columbia
Vancouver, BC, Canada V6T 1Z4
arabed@icics.ubc.ca

Abstract

The objective of this paper is to study the existing methods for unsupervised object recognition and image categorization and propose a model that can learn directly from the output of image search engines, e.g. Google Images, bypassing the need to manually collect large quantities of training data. This model can then be used to refine the quality of the image search, or to search through other sources of images. This integrated scheme has been implemented and optimized to be used in The Semantic Robot Vision Challenge as a new test-bed for research in the areas of image understanding and knowledge retrieval in large unstructured image databases.

Introduction

To furnish computers and their moving relatives, robots, with object recognition skill, a number of methods have been developed. These methods can be divided into two groups: recognition of individual objects and recognition of categories. Considerable progress has been made in the recognition of individual objects under different illumination and viewpoint conditions. Categories are more difficult to deal with and learning a model for them requires more sophisticated representations. In recent years extensive studies has been carried out on particular image categories like human faces, vehicles, pedestrians, and so on. The aim of our research, instead, is to develop techniques that work well with any known category. We have to be able to create object categories from a collection of highly noisy images with very few or even no training examples. Research done by (Sivic et al. 2005), (Fergus et al. 2005), and (Fei-fei, Fergus, and Perona 2006) are among the notable works in this domain.

In our project the initial collection of the images is extracted through Google Image Search Engine by entering a keyword. The resultant images are highly inconsistent and the good instances are mixed with unrelated images. This problem occurs because current Internet image search methods rely entirely on text cues such as the file name or surrounding HTML text, rather than image content. This is a fast way to find and collect images related to the given keyword but at the same time, practically, it is very poor to produce images with relevant visual content. We believe that the quality of returned images can be significantly improved by filtering the search results using a mutual consistency. According to the observations, the visual features of images relevant to the search topic are repeated frequently (i.e. popular visual features), while those features occur rarely in unrelated images. We call this phenomenon "visual consistency" and we will use this characteristic to re-rank the images. (Fergus, Perona, and Zisserman 2004)

The Semantic Robot Vision Challenge

"The Semantic Robot Vision Challenge is a new research competition that is designed to push the state of the art in image understanding and automatic acquisition of knowledge from large unstructured databases of images (such as those generally found on the web)"[1].

This competition was held for the first time at the Association for the Advancement of Artificial Intelligence (AAAI) conference in 2007. Basically the challenge is for a robot to enter a previously unvisited room, full of different objects and try to locate the query objects. Prior to entering the room, the robot has to use an image search engine such as Google to train classifiers for the query objects. Since the images returned by search engines are very noisy and inconsistent, the robot has to filter and re-rank them and try to find sets of images which are visually consistent. These sets of images and extracted feature descriptors will be used to find the actual objects in the room (real world).

Unsupervised Image Categorization

The important question in this research is how one can measure "visual consistency". By studying the existing methods it is observed that most of them are focused on categorizing the images with very high visual similarity. Research by Fergus et al. (Fergus, Perona, and Zisserman 2004) shows very good results if the images have very small viewpoint and context variation. These methods

[1] http://www.semantic-robot-vision-challenge.org/

discard all the images which are not similar or close to the mainstream images. The other issue with using these methods is the time required to learn the categories models. Most of the existing methods require a relatively long time to be able to build up an internal model before starting to do the recognition and classification part. Our research aims to develop a method which is fast, reliable and a perfect candidate to deploy in mobile robots.

Based on our specific challenge we need to consider all the images representing our query object with different shapes, styles, and pose variations. By doing this we will be able to provide enough visual code words for our robot to assist him in locating the requested object in the room during the competition. We also need to be able to discard the unrelated images based on the fact that they might not be repeated in the search results and there is not any specific relationship between these irrelevant images. Using this assumption we collect the (visual) features from all images, where we keep the popular ones and discard the rest. This will serve as the basis for ranking the images. To make sure that these features are representing the main characteristic of our queried object, we will use a set of "background" (or negative) images to refine our collected visual feature set.

Proposed Approach

In our approach images are collected for a keyword (Positive Data) using Google image search engine or from known databases. The same program is used to collect clutter images (Negative Data). Some pre-processing work is done to fix the size of the images and convert them to gray scale. The DoG-SIFT keypoint detector (Lowe 1999) is then used for feature extraction to collect data from the two sets of our sample images (Positive and Negative data). During the approach three measures are computed for every feature:

- *Score:* Popularity of the feature among the images
- *Positive Match Number:* Number of similar features to this feature in sample images
- *Negative Match Number:* Number of similar features to this feature in Clutter Images

To organize the features, we have implemented a flat histogram and features are added to the histogram one by one. The KD-Tree method is used to sort the features based on their distance (SIFT difference) from the new feature being added to the histogram. After identifying a close match all the features within 5% distance of the match are selected and their scores are increased based on the measure of their similarity to the new feature (Positive Match). The score and Positive Match Number of this new feature are also initialized based on the scores and Positive Match Number of the nearest features. If the distance is more than a predefined threshold, none of the scores are changed and just this feature is added as a new one to the histogram. For negative matches, the nearest features scores are decreased based on the measure of their similarity to the new feature and also negative features are not added to the feature histogram. At the end, the list of top-ranked images is returned based on the sum of the scores of the features that have been appeared in those images. A sample of the re-ranked images is shown below:

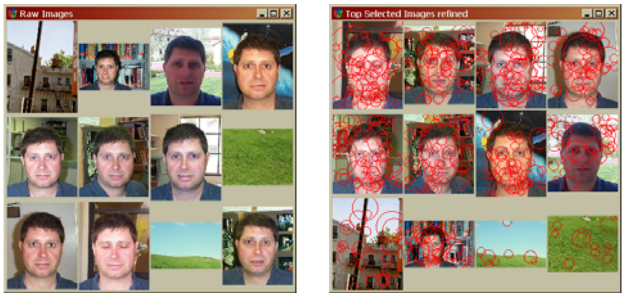

Figure 1. Original input images (Left) and re-ranked output images, circles show the popular features (Right).

Conclusion and Future Work

We have run a comprehensive set of tests with different values and image sets to tune the parameters for maximum performance. The preliminary results show that our method can manage to learn a model on manually constructed image collections with added noise and also on raw Google images with an acceptable accuracy. To extend this work, we would like to create a set of synthetic images where we can run our algorithm against them to measure the quality of the filtered image sets. This will give us a benchmark to compare different variations of the algorithm in a controlled environment.

For future work we intend to use other methods of interest point extraction and also consider spatial configuration (Shape and Geometry) of the features relative to each other. More sophisticated clustering algorithms can be used to find the "densest area" in the features space. Linguistic techniques based upon related words in the text of the webpage containing the image could also be employed to identify and remove mislabeled images.

References

Fei-Fei, L., Fergus, R., and Perona, P. 2006. One-shot Learning of Object Categories. In *IEEE Trans. PAMI*, Vol 28(4), pp. 591-611.

Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A. 2005. Learning Object Categories from Google's Image Search. In *Proc. ICCV*, pp. 1816-1823 Vol. 2.

Fergus, R., Perona, P., and Zisserman, A. 2004. A Visual Category Filter for Google Images. In *Proc. ECCV, Springer-Verlag*, pp. 242-256.

Lowe, D. 1999. Object Recognition from Local Scale-Invariant Features. In *Proc. ICCV*, pp. 1150–1157.

Sivic, J., Russell, B. C., Efros, A. A., Zisserman, A., Freeman, W. T. 2005. Discovering Object Categories in Image Collections. In *Proc. ICCV*.