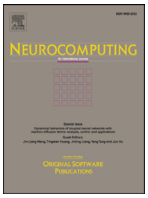
Neurocomputing 000 (2018) 1–13

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



MirBot: A collaborative object recognition system for smartphones using convolutional neural networks

Antonio Pertusa[*], Antonio-Javier Gallego, Marisa Bernabeu

Departamento Lenguajes y Sistemas Informáticos, Universidad de Alicante, San Vicente del Raspeig, Alicante E-03690, Spain

ARTICLE INFO

Article history:
Received 9 June 2017
Revised 15 February 2018
Accepted 1 March 2018
Available online xxx

Communicated by Deng Cai

Keywords:
Object recognition
Image datasets
Convolutional neural networks
Transfer learning
Multimodality
Human computer interaction

ABSTRACT

MirBot is a collaborative application for smartphones that allows users to perform object recognition. This app can be used to take a photograph of an object, select the region of interest and obtain the most likely class (dog, chair, etc.) by means of similarity search using features extracted from a convolutional neural network (CNN). The answers provided by the system can be validated by the user so as to improve the results for future queries. All the images are stored together with a series of metadata, thus enabling a multimodal incremental dataset labeled with synset identifiers from the WordNet ontology. This dataset grows continuously thanks to the users' feedback, and is publicly available for research. This work details the MirBot object recognition system, analyzes the statistics gathered after more than four years of usage, describes the image classification methodology, and performs an exhaustive evaluation using handcrafted features, neural codes, different transfer learning techniques, PCA compression and metadata, which can be used to improve the image classifier results. The app is freely available at the Apple and Google Play stores.

1. Introduction

Object recognition is a highly active topic in computer vision and can be particularly useful for mobile devices [1,2] as regards retrieving information about objects on the fly. Visually impaired persons can also benefit from these systems [3].

MirBot is a smartphone app that users can train to recognize any object. The objects are categorized according to lemmas (such as chair, dog, laptop, etc.) selected from the WordNet lexical database [4]. A user can employ MirBot to take a photograph and select a rectangular region of interest (ROI) in which the target object is located. The image, the ROI coordinates and a series of associated metadata are then sent to a server, which performs a similarity search using k-Nearest Neighbors (kNN) and returns the class of the most likely image, as shown in Fig. 1. The app users can validate the system response in order to improve the classification results for future queries, and this feedback allows the database to grow continuously with new labeled images.

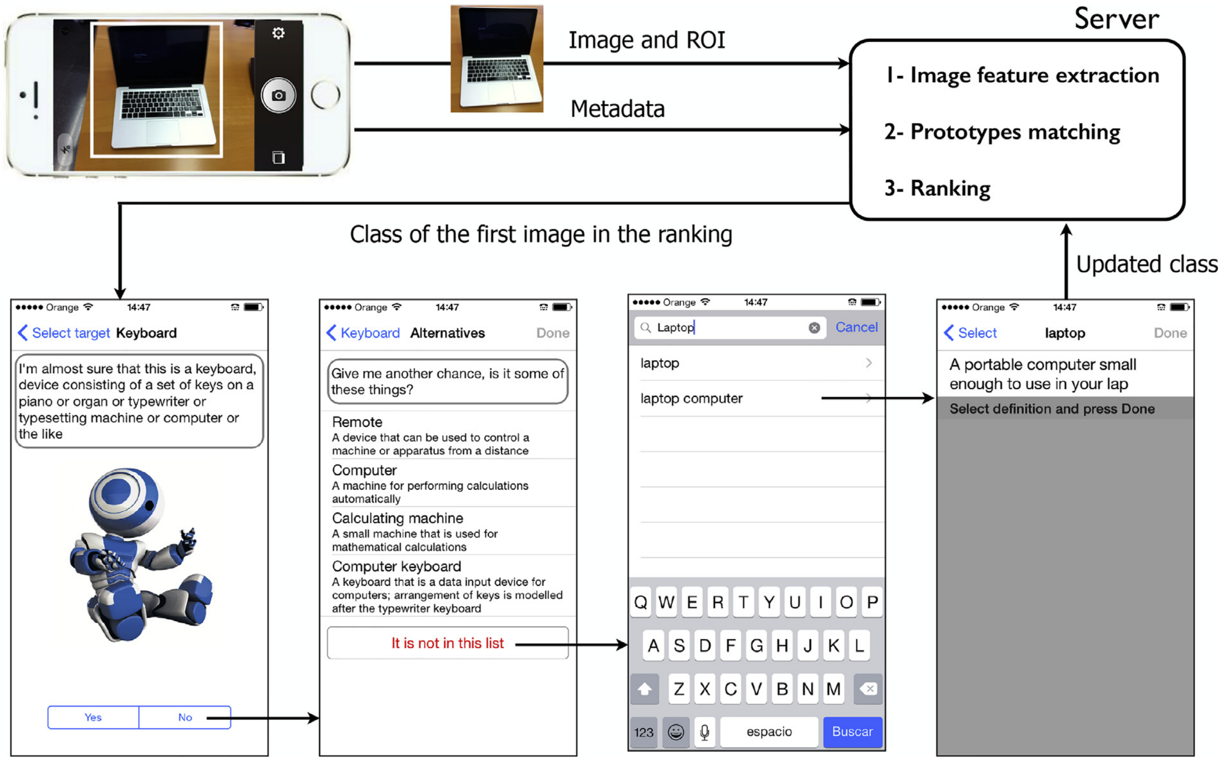
MirBot is designed as a pedagogical game in which a simulated robot can be trained in order to involve users in an automatic learning process. As pointed out in [5,6], developing machine learning tasks through games has proven to be a successful approach to make users enjoy labeling data. From the users' perspective, the main distinctive feature of MirBot with regard to other apps is that it allows them to train a personal image search system, thus making the dataset dynamic and user driven.

This work extends the contents of the MirBot system for object retrieval introduced in [7], which initially used handcrafted visual descriptors. The main contributions of this paper with regard to the initial work are a detailed description of the user interaction process, the statistics related to the database gathered after four years of usage, a new classification methodology based on CNN features (neural codes) obtained from pre-trained and fine-tuned models, the study of PCA compression on different neural networks, the inclusion of metadata to complement the neural codes and the evaluation results and conclusions.

When a user submits a photograph, visual descriptors are extracted and compared to the existing prototypes in the dataset in order to predict the class of the object. In the initial MirBot version [7], both local features and color histograms were extracted and combined to obtain the most likely class. In this work, several convolutional neural network (CNN) features have been added to these descriptors for use in evaluation. The gap between the results obtained using handcrafted descriptors and features extracted from convolutional networks led the traditional image descriptors in MirBot to be replaced with neural codes in June 2015.

[*] Corresponding author.
E-mail addresses: pertusa@dlsi.ua.es (A. Pertusa), jgallego@dlsi.ua.es (A.-J. Gallego), mbernabeu@dlsi.ua.es (M. Bernabeu).

https://doi.org/10.1016/j.neucom.2018.03.005
0925-2312/© 2018 Elsevier B.V. All rights reserved.



I'm almost sure that this is a keyboard, device consisting of a set of keys on a piano or organ or typewriter or typesetting machine or computer or the like

Give me another chance, is it some of these things?

Remote
A device that can be used to control a machine or apparatus from a distance

Computer
A machine for performing calculations automatically

Calculating machine
A small machine that is used for mathematical calculations

Computer keyboard
A keyboard that is a data input device for computers; arrangement of keys is modelled after the typewriter keyboard

laptop

laptop computer

A portable computer small enough to use in your lap

Select definition and press Done

Fig. 1. Architecture of the iOS app. This example corresponds to the longest user interaction sequence.

As pointed out in [8], finding images within large collections is an important topic for the computer vision community. Recent progress in object recognition has been built upon efforts to create large-scale, real-world image datasets [9,10] that are crucial for developing robust image retrieval algorithms, in addition to considering the large amount of data required in recent deep neural networks [11].

One of the main contributions of MirBot is a dataset with a similar structure to that of ImageNet [10], with the exception that all the images are gathered with smartphone cameras and stored with their associated metadata and with regions of interest. In October 2016 we had 25,292 validated images distributed in 1808 classes. Although the MirBot dataset still cannot be considered a very large collection, it is incremental and grows continuously thanks to its users' feedback. One important difference with regard to other datasets such as [8] and [10] is that, rather than employing images downloaded from the Internet, users take pictures specifically for object recognition. This signifies that MirBot images are focused on the target objects and gathered with minimum occlusions and plain backgrounds. Our team reviews the new images on a weekly basis in order to avoid inappropriate, unfocused or wrongly labeled samples, thus ensuring good quality data.

The MirBot dataset also includes a series of metadata that could be used to constrain the search space. These metadata, which are detailed in [7], are extracted from the smartphone sensors (angle with regard to the horizontal, gyroscope, flash, GPS, etc.), and have reverse geocoding information (type of place, country, closest points of interest, etc.) and EXIF camera data (aperture, brightness, ISO, etc.). All the images are stored with their associated metadata. We have evaluated the performance using these metadata in order to complement the image information.

This work begins by describing the user interaction interface in Section 2, and the methodology (Section 3) used for similarity search on the server side. The evaluation results are described and discussed in Section 4, followed by the conclusions in Section 5.

2. User interaction

As stated in [12], beyond the one-shot queries in the early similarity-based search systems, the next generation of systems attempts to integrate continuous feedback from the user in order to learn more about the query. MirBot is designed as an interactive game and the objective of its interface is to minimize the number of user interactions.

In order to avoid the dispersion that can be caused when using free object identifiers, users can only assign class names selected from the WordNet ontology. The main advantage of WordNet is its semantic structure, which prevents ambiguities in the labels. WordNet *synsets* (synonym sets) are unique identifiers for meaningful semantic concepts, and each of them is linked to a definition, although they can be related to different lemmas. Similarly to ImageNet, we use the WordNet synsets as class identifiers.

In MirBot, only portrayable objects can be chosen from the WordNet hierarchy, including the following root categories: animals (lexicographer identifier: (5)), food/drinks (13), plants (20), and objects, which include both WordNet natural objects (6) and artifacts (17). WordNet considers that an artifact is man-made whereas a natural object is not, but they have been merged in MirBot in order to simplify this concept for its users.

Fig. 2 shows the block diagram of the user interface. Before sending the query image to the server, users can find settings in the app that allow to choose whether they wish to perform the classification by considering only their own images, so as to constrain the search space, or by using the whole dataset.

Once the query has been submitted, the classes of the K most similar images are retrieved using the methodology described in Section 3, and the class of the first image in this ranking is given

Please cite this article as: A. Pertusa et al., MirBot: A collaborative object recognition system for smartphones using convolutional neural networks, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2018.03.005