Computer Vision and Image Understanding 000 (2016) 1–14

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Interactive multiple object learning with scanty human supervision

Michael Villamizar*, Anaís Garrell, Alberto Sanfeliu, Francesc Moreno-Noguer

Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Spain

ARTICLE INFO

Article history:
Received 15 April 2015
Revised 6 January 2016
Accepted 18 March 2016
Available online xxx

Keywords:
Object recognition
Interactive learning
Online classifier
Human-robot interaction

ABSTRACT

We present a fast and online human-robot interaction approach that progressively learns multiple object classifiers using scanty human supervision. Given an input video stream recorded during the human-robot interaction, the user just needs to annotate a small fraction of frames to compute object specific classifiers based on random ferns which share the same features. The resulting methodology is fast (in a few seconds, complex object appearances can be learned), versatile (it can be applied to unconstrained scenarios), scalable (real experiments show we can model up to 30 different object classes), and minimizes the amount of human intervention by leveraging the uncertainty measures associated to each classifier.

We thoroughly validate the approach on synthetic data and on real sequences acquired with a mobile platform in indoor and outdoor scenarios containing a multitude of different objects. We show that with little human assistance, we are able to build object classifiers robust to viewpoint changes, partial occlusions, varying lighting and cluttered backgrounds.

1. Introduction

Over the last decade, we have witnessed the enormous progress in the field of object recognition and classification in images and video sequences. At present, there are methods that produce impressive results in a wide variety of challenging scenarios corrupted by lighting changes, cluttered backgrounds, partial occlusions, viewpoint and scale changes, and large intra-class variations (Ali and Saenko, 2014; Felzenszwalb et al., 2010; Hinterstoisser et al., 2011; Malisiewicz et al., 2011; Schulter et al., 2014; Tang et al., 2012; Villamizar et al., 2012a).

This progress in object recognition has had a positive impact in many application fields such as robotics, where computer vision algorithms have been used for diverse robotics tasks such as object recognition and grasping (Alenyà et al., 2014; Amor-Martinez et al., 2014), detection and tracking of people in urban settings (Bellotto and Hu, 2009; Merino et al., 2012; Portmann et al., 2014), human-robot interaction (Ragaglia et al., 2014; Tamura et al., 2014), and robot localization and navigation (Corominas et al., 2008; Ferrer et al., 2013; Hornung et al., 2010).

The standard method for recognizing objects in images consists in computing object specific classifiers during an offline and time-consuming training step, where large amounts of annotated data are used to build discriminative and robust object detectors (Felzenszwalb et al., 2010; Malisiewicz et al., 2011). However, there are situations in which offline learning is not feasible, either because the training data is obtained continuously, or because the size of the training data is very cumbersome, and a batch processing becomes impractical. This is particularly critical in some robotics applications, specially those related to human-robot interaction, where the robots need to compute object detectors on the fly, in real time, and with very little training data.

In these cases, online learning methods which use their own predictions to compute and update a classifier have been proposed (Godec et al., 2010; Grabner and Bischof, 2006; Moreno-Noguer et al., 2008). Yet, although these approaches have shown great adaptation capabilities, they are prone to suffer from drifting when the classifier is updated with wrong predictions. This has been recently addressed by combining offline and online strategies (Gall et al., 2010; Kalal et al., 2010), semi-supervised boosting methods (Grabner et al., 2008), or by using human intervention during learning so as to assist the classifier in uncertain classification cases (Villamizar et al., 2012b; Yao et al., 2012).

In preliminary versions of this work, we already proposed online object detection approaches in which the human assistance is integrated within the learning loop in an active and efficient manner (Garrell et al., 2013; Villamizar et al., 2012b, 2015). In Garrell et al. (2013); Villamizar et al. (2012b) the proposed approach was focused for single object detection, and further extended in (Villamizar et al., 2015) to multiple instances using an adaptive

* Corresponding author. Tel.: +34934015806.
E-mail addresses: mvillami@iri.upc.edu (M. Villamizar), agarrell@iri.upc.edu (A. Garrell), sanfeliu@iri.upc.edu (A. Sanfeliu), fmoreno@iri.upc.edu (F. Moreno-Noguer).

http://dx.doi.org/10.1016/j.cviu.2016.03.010
1077-3142/

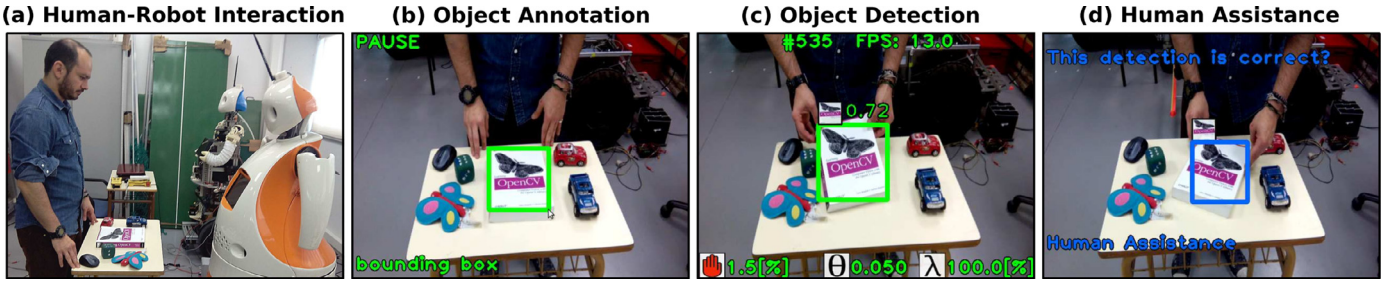

Fig. 1. Interactive and real-time approach for learning and detecting multiple objects through human-robot interaction. (a) Interplay between the robot and the human user. (b) Object annotation using a bounding box provided by the user (green box). (c) Output of the proposed detection method (green box). (d) Human intervention for unknown and difficult cases. The user provides the label for current object prediciton (blue box). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

uncertainty classification threshold that reduces the human supervision. In this paper, we unify the formulation of these previous works and perform a more in-depth analysis of the method presented in Villamizar et al. (2015) through additional experiments in synthetic and real scenarios, while providing more comparisons against competing approaches.

More precisely, we propose a fast and online approach that interactively models several object appearances on the fly, using as few human interventions as possible, and still keeping the real-time efficiency (Villamizar et al., 2015). At the core of our approach, there is a randomized tree classifier (Criminisi et al., 2011; Ozuysal et al., 2010; P. Geurts and Wehenkel, 2006) that is progressively computed using its own detection predictions. Yet, to avoid feeding the classifier with false positive samples (i.e, drifting), we propose an uncertainty-based active learning strategy (Lewis and Gale, 1994; Settles, 2010) that gradually minimizes the amount of human supervision and keeps high classification rates. Note that this issue is critical in order to maintain long-term interactions with robots, as if the robot keeps asking for annotating images insistently, people tend to quickly give up the interaction (Garrell et al., 2013; Rani et al., 2006).

To make the proposed approach scalable for various object instances, multiple object specific classifiers are computed in parallel, but sharing the same features in order to maintain the efficiency of the method and to reduce the computational complexity in run time Villamizar et al. (2010).

As an illustrative example, Fig. 1 shows the operation of the proposed interactive method to learn and detect multiple object instances through human-robot interaction (Fig. 1a). Each time the human user seeks to model a new object of interest, he/she marks a bounding box around the object in the input image, via a mouse, keyboard or touchscreen (see Fig. 1b). The robot initializes a model for this new object and runs a detector on subsequent frames for this, and the rest of objects in the database (Fig. 1c). When the robot is not confident enough about the detections and class predictions, it requests the human assistance to provide the true class labels, which, in turn, are used to update the classifier, observe Fig. 1d. This procedure is performed continuously, and at each iteration, the performance and confidence of the classifier is increased whereas the degree of human intervention is reduced significantly.

The remainder of the paper is organized as follows: Section 2 describes the related work and our contributions, while Section 3 explains the proposed approach with all its main ingredients. Section 4 describes the experiments conducted to evaluate the proposed learning approach. We report results using synthetic an real data. The former are used to thoroughly assess the limits of the method in terms of number of classes it can handle or classification rate. Real experiments demonstrate the applicability of the method for diverse perception tasks in challenging scenarios.

2. Related work and contributions

In this section, we show and discuss our contributions along with the related work on the three main topics concerned with the proposed approach: human-robot interaction, the computation of online classifiers, and interactive learning techniques.

2.1. Human-robot interaction

Computer vision techniques for human-robot interaction have been mainly focused on recognizing people in urban scenarios (Bellotto and Hu, 2009; Merino et al., 2012; Portmann et al., 2014) as well as identifying human gestures and activities (den Bergh et al., 2011; Nickel and Stiefelhagen, 2007) to establish contact with people and perform particular robotics tasks such as guiding people in museums and urban areas (Garrell and Sanfeliu, 2010; Rashed et al., 2015; Thrun, 2000), providing information in shopping malls (Gross et al., 2002), or recognizing human emotions through classifying facial gestures (Bartlett et al., 2003). Although, these techniques have endowed the robot with remarkable interaction skills, they are commonly computed offline and using a potentially large training time. As a result, the robot is limited to perform tasks only for which it has been trained previously, missing the opportunity to learn and improve its perception skills through the interaction with humans.

Conversely, in this work we propose a very efficient interactive approach that combines human assistance and robot's detection predictions so as to build object detectors which can be applied for a wide range of robotics tasks. The approach exploits the interplay between robots and humans in order to compute and improve progressively the robot's perception capabilities.

2.2. Online classifiers

Despite showing impressive results, standard methods for object detection compute the classifiers using intensive and offline learning approaches applied to large annotated datasets (Felzenszwalb et al., 2010; Malisiewicz et al., 2011; Ozuysal et al., 2010; Villamizar et al., 2012a). Therefore, most of these offline approaches are not suitable for some particular applications requiring computing the classifier on the fly, either because the training data is obtained continuously, or the size of the training data is so large that it needs to be loaded progressively. To handle these situations, several online alternatives allowing to sequentially train the classifiers have been proposed (Avidan, 2007; Babenko et al., 2011; Grabner and Bischof, 2006; Hall and Perona, 2014; Santner et al., 2010).

In this work, the classifier we use is based on an online random ferns formulation (Kalal et al., 2010; Krupka et al., 2014; Ozuysal et al., 2010; Villamizar et al., 2012b, 2015), which has been show-

Please cite this article as: M. Villamizar et al., Interactive multiple object learning with scanty human supervision, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.03.010