Computer Vision and Image Understanding 000 (2016) 1–12

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Improved scene identification and object detection on egocentric vision of daily activities

Gonzalo Vaca-Castano[a,*], Samarjit Das[b], Joao P. Sousa[b], Niels D. Lobo[a], Mubarak Shah[a]

[a] Center for Research in Computer Vision, University of Central Florida, United States
[b] Robert Bosch LLC, Research and Technology Center, North America

ARTICLE INFO

Article history:
Received 16 December 2015
Revised 26 September 2016
Accepted 19 October 2016
Available online xxx

Keywords:
Scene classification
Object detection
Scene understanding
First camera person vision

ABSTRACT

This work investigates the relationship between scene and associated objects on daily activities under egocentric vision constraints. Daily activities are performed in prototypical scenes that share a lot of visual appearances independent of where or by whom the video was recorded. The intrinsic characteristics of egocentric vision suggest that the location where the activity is conducted remains consistent throughout frames. This paper shows that egocentric scene identification is improved by taking the temporal context into consideration. Moreover, since most of the objects are typically associated with particular types of scenes, we show that a generic object detection method can also be improved by re-scoring the results of the object detection method according to the scene content. We first show the case where the scene identity is explicitly predicted to improve object detection, and then we show a framework using Long Short-Term Memory (LSTM) where no labeling of the scene type is needed. We performed experiments in the Activities of Daily Living (ADL) public dataset (Pirsiavash and Ramanan,2012), which is a standard benchmark for egocentric vision.

1. Introduction

Egocentric vision has recently got significant interest from the vision community since the advent of wearable vision sensors and their potential applications. From the applications standpoint, egocentric videos are a key enabler for a number of technologies ranging from augmented reality to context–aware cognitive assistance, which could improve our daily lives dramatically. Current assistance systems like Siri, lack the ability to understand the visual context – where you are in your house, what objects you are working with. This shortcoming limits its capabilities to *help us* in many of our day-to-day activities. Egocentric vision, with its ubiquity, has the capacity to be the provider of such knowledge. Consequently, in this paper, we study some computer vision techniques that help to exploit inherent constraints of first-person camera video of individuals performing daily activities.

In the case of activities of daily living, the actions typically are performed in common places associated with human residences such as bathroom, corridor, patio, kitchen, among others, which will be referred as the scenes. Then, we are interested in the frame level scene identification problem, where the goal is to find the correct scene identity for all the frames of the egocentric video. We note that temporal constraints can be exploited to improve frame level scene identification performance. The location where an activity is performed remains consistent for several frames until the user changes his/her current location. Given a frame, several trained scene classifiers are evaluated and a decision about the identity is taken based on the classification scores. However, the scores obtained for individual frames can lead to wrong scene identification since these scores are agnostic with respect to the temporal constraints associated with egocentric vision. In this paper, we propose a formulation that uses the scene identification scores of temporally adjacent frames to improve the scene identity accuracy. The formulation is based on a Conditional Random Field (CRF).

We are also interested in the problem of improving the detection of objects. Object detection task attempts to find the location of objects in a frame. Traditional approaches use human labeled bounding boxes of objects as positive training data while visual features not included in the positive training bounding box are part of the negative data. However, in the real world, the objects are part of a scene. Consider, for example, Fig. 1(a) which shows a picture from a kitchen. Fig. 1(b) shows a list of possible objects that could be interesting to detect. It is obvious for humans that some types of objects are unlikely to be found in the observed scene,

* Corresponding author.
E-mail addresses: gonzalo@knights.ucf.edu (G. Vaca-Castano), Samarjit.Das@us.bosch.com (S. Das), JoaoP.Sousa@us.bosch.com (J.P. Sousa), niels@cs.ucf.edu (N.D. Lobo), shah@crcv.ucf.edu (M. Shah).

http://dx.doi.org/10.1016/j.cviu.2016.10.016
1077-3142/

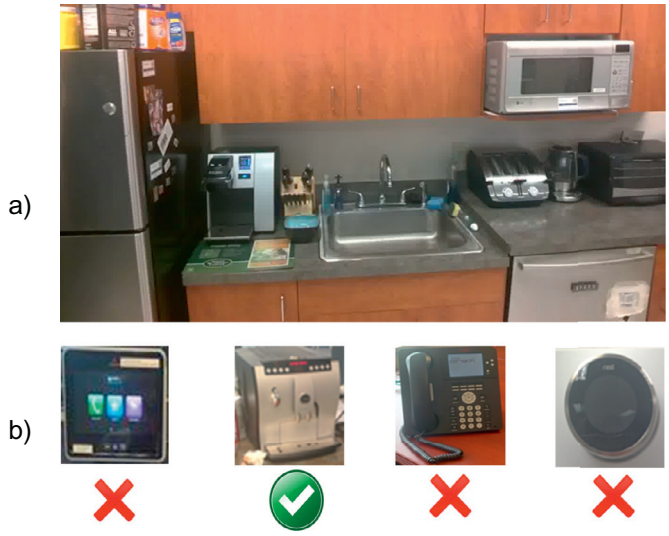

Fig. 1. Example of how object detection is influenced by the scene context. Figure a) contains an image taken in a kitchen. Figure b) shows a list of possible objects that could be detected. From the list, only the coffeemaker makes sense in the observed context.

while a coffeemaker is an object that most likely can be found in this type of scene.

The previous observation is used as a constraint in our problem formulation to improve the quality of object detectors. We concentrate on Activities of Daily Living (ADL), where most of the first person activities are performed in few prototypical scenes that are common to all the actors. ADLs are an extremely challenging scenario for object detection, since the objects suffer from notable changes on appearance due to radial distortion, pose change and actor influence over the object. We do not focus on direct improvements in the object detection. Instead, the results of object detection are improved after re-scoring the outcome of the object detection method. Objects that are most probably present in a type of scene get higher scores, while objects that are unusual in a type of scene get lower scores. In this paper, we present two type of formulations. Firstly, a formulation to manage the case, where the labels of the test videos are explicitly predicted from scene models learned in training data. Two algorithms are proposed for this case: a greedy algorithm, and a Support Vector Regression (SVR) based algorithm. Secondly, a formulation based on Long Short-Term Memory (LSTM), that directly infers the probability of having a type of object in a sequence, without an explicit knowledge of the label of the scenes. As we show in our experiments, the improvements are consistent for different types of scene detectors and two types of object detectors in both formulations.

To summarize, the main contributions of this paper are the following. Firstly, we propose the use of temporal consistency constraint to improve scene identification accuracy in egocentric videos, with the aid of a Conditional Random Field (CRF) formulation analyzed under two types of pairwise relations. Secondly, we present two algorithms to improve the object detection results, by modifying the object detection scores of the bounding box proposals according to the scene identity of the frame currently tested. Finally, in the case that scene labeling of the training data is not available, we present an LSTM formulation that predicts how likely a type of object will be present in the current frame of a video sequence. This prediction allows to re-score the object detection according to the scene context producing excellent results. We performed our experiments in the Activities of Daily Living (ADL) public dataset (Pirsiavash and Ramanan, 2012).

2. Related work

A relatively recent trend in computer vision community is the egocentric vision. Most efforts (Fathi et al., 2011; Pirsiavash and Ramanan, 2012; Ren and Philipose, 2009) in egocentric vision have focused on object recognition, activity detection/recognition and video summarization, however, with the exception of our previous work (Vaca-Castano et al., 2015), none of these efforts have focused on scene identification and its relation with object detection. Ren and Philipose (2009) collected a video dataset of 42 objects commonly found in every day life with large variations in size, shape, color, etc. They quantify the accuracy drop of object detectors after simulating background clutter and occlusion on clean exemplars. Fathi et al. (2011) observed that the object of interest tends to be centered and covers a large space of the image frame. Based on that observation they perform unsupervised bottom–up segmentation and divide each frame into hand, object, and background categories. A list of objects that are part of the video is provided, and an appearance model for them is learned from the training dataset. Objects become part of the background after the manipulation of the object is completed. In Pirsiavash and Ramanan (2012), a new dataset of videos of Activities of Daily Living (ADL) in first-person camera is presented. The dataset contains bounding boxes annotations for 42 different objects of frames sampled every second from the videos. The dataset also provides the results of Deformable Part Model (DPM) object detectors for some of those objects. The object detection models were trained from a subset of egocentric videos of the dataset, since models trained on standard object detection datasets like Imagenet (Russakovsky et al., 2014) or PASCAL VOC contain only iconic view of the objects, compared to the most challenging appearance of objects from egocentric videos. Many of the classes with available ground-truth were not reported in the object detection due to their insignificant performance.

Improvement in object detection has been fueled mainly by PASCAL VOC competition (Everingham et al., 2010), and more recently by ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2014). An extensive analysis of the results of the different competitions on PASCAL VOC challenge during years 2008 to 2012 was published (Everingham et al., 2014) by their organizers. Their analysis shows clearly that the reference method for object detection in VOC 2008–2012 was the Deformable Part-based Model (DPM) (Felzenszwalb et al., 2010), which won the detection contest on 2008 and 2009. DPM model uses a histogram of oriented gradients representation (HOG) to describe a coarse scale root filter and a set of finer-scale part templates that can move relative to the root. During testing, the model is applied everywhere in the image (sampled in different scales) using sliding window technique. A huge gain in performance was achieved later by Girshick (2015); Girshick et al. (2014) using a combination of selective search (Uijlings et al., 2013) and Convolutional Neural Networks (CNN). In that work, the Convolutional Neural Network trained by Krizhevsky et al. (2012) for the ImageNet (ILSVRC) classification challenge was used, but a fine tuning in the fully connected layers of the network was performed in order to adapt the domain to the PASCAL VOC dataset.

In spite of the significant performance gains of these methods for single image object detection, these methods under-perform on video object detection due to multiple factors such as motion blur, temporary occlusions, objects out of focus, among others. One focus of our paper is improving the results of object detectors on sampled frames using scene context. Once better object detectors are available, the tracking by detection framework of the Multiple Object Tracking (MOT) problem, could be incorporated to obtain better tracks and handle long-term temporal relations. Different MOT algorithms (Andriyenko and Schindler, 2011; Stauffer, 2003; Zamir et al., 2012; Zhang et al., 2008) use object detections on the

Please cite this article as: G. Vaca-Castano et al., Improved scene identification and object detection on egocentric vision of daily activities, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.10.016