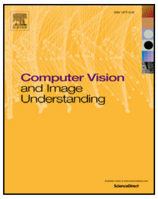
Computer Vision and Image Understanding 000 (2016) 1–12

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/© 2016 Elsevier Inc. All rights reserved.

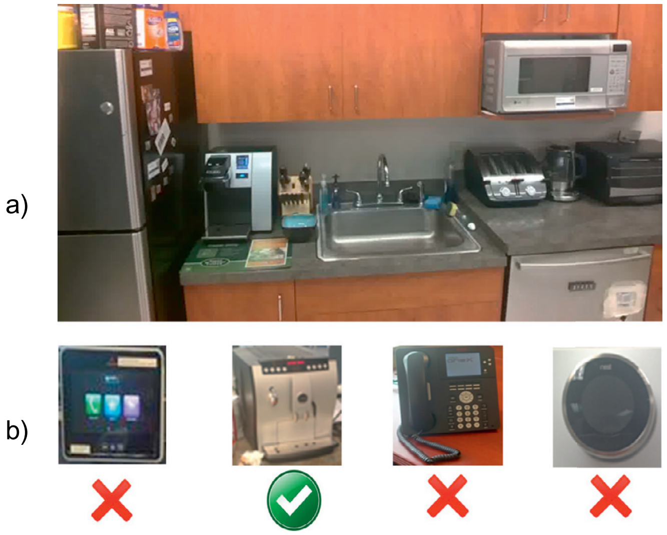


Fig. 1. Example of how object detection is influenced by the scene context. Figure a) contains an image taken in a kitchen. Figure b) shows a list of possible objects that could be detected. From the list, only the coffeemaker makes sense in the observed context.

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 input video frames and generate target tracks by connecting the detection outputs corresponding to identical objects across frames. The main difference among MOT trackers is the utilized detection-association mechanism. MOT does not overlap with the proposed mechanism in this paper to improve object detection, being in fact, complementary. We will not focus on the MOT problem in this paper.

Recently, Han et al. (2016) proposed a heuristic method for re-ranking bounding boxes in video sequences, linking bounding boxes temporally that have a high overlap from frame to frame. They achieved the third place in the video object detection (VID) task of the ImageNet Large Scale Visual Recognition Challenge 2015 (ILSVRC2015). Unfortunately, this approach assumes the availability of detections that overlap in every frame. This condition only can be achieved with high sampling rates which is prohibitive in long sequences as ours. In contrast, we process frames sampled approximately every second. Additionally, the mentioned approach does not consider the scene context associated with the objects in the frame.

The role of context in object recognition has been analyzed from a cognitive science perspective (Oliva and Torralba, 2007), but also from a computer vision perspective (Carbonetto et al., 2004; Divvala et al., 2009; Heitz and Koller, 2008; Park et al., 2010; Song et al., 2010; Soomro et al., 2015; Torralba et al., 2010, 2003). Heitz and Koller (2008) used a terminology coined by Forsyth et al. (1996) known as “thing” and “stuff” (TAS), linking discriminative detection of objects with unsupervised clustering of image regions. Other approaches like (Song et al., 2010) achieve a boost in object detection by iteratively switching between the classification task and detection using each other output as context. Divvala et al. (2009) studied several sources of context, and incorporate some of them to improve object detection. An approach more directly related to ours is the work of Torralba et al. (2003), where the global scene context and its influence over object recognition is considered by representing the scene as a low-dimensional global image representation (GIST), and this is used as contextual information to introduce strong priors that simplify object recognition.

The scene identification problem is essentially an image classification problem with a domain specific type of images. Over many years approaches based on Bag of Words paradigm (Csurka et al., 2004; Sivic and Zisserman, 2003) were the dominant state of the art. Further improvement was achieved by including spatial information using pyramids (Grauman and Darrell, 2005; Lazebnik et al., 2006) in association with new types of encoding (Jegou et al., 2010; Perronnin and Dance, 2007; Perronnin et al., 2010; Wang et al., 2010). Huge improvements have been obtained in classification and detection (almost double in less than 2 years according to the comprehensive evaluation of the ILSVRC challenge reported in Russakovsky et al. (2014)) after the generalized use of Convolutional Neural Networks (CNN). Most of these new approaches are based on the extension of the CNN architecture presented by Krizhevsky et al. (2012) for the ILSVRC classification challenge. A number of recent works (Girshick et al., 2014; Oquab et al., 2014; Razavian et al., 2014; Sermanet et al., 2014) had shown that CNN features trained on sufficiently large and diverse datasets, can be successfully transferred to other visual recognition tasks such as scene classification and object localization, with only a limited amount of task-specific training data. To the best of our knowledge the work of Gong et al. (2014) is the current state of the art for scene classification, where global CNN features are encoded together by concatenating multiple scale levels CNN features pooled by orderless Vector of Locally Aggregated Descriptors (VLAD). In our work, we show that scene identification methods can be improved by considering its egocentric video intrinsic temporal constraints.

3. Egocentric vision clues

In this work, we focus on two important building blocks towards the goal of using a first-person camera for context acquisition and scene understanding: a) improving scene identification by using temporal information, and b) improving the object-detection through the utilization of the visual appearance of the scene (either scene identity or global context).

We use the egocentric video temporal consistency constraint to improve scene identification accuracy by means of a Conditional Random Field (CRF) formulation, which penalizes short-term changes of the scene identity. This formulation is covered in detail in Section 3.1.

Assuming that we have a method for object detection that provides bounding boxes and their confidence scores, we show that it is possible to increase the performance of the detector by incorporating the information about the particular type of the scene for the frame that is being tested. We learn from the training data, to modify the confidence scores of the object detectors according to the type of scene identity. Detection scores for objects that are unlikely to appear in a particular kind of scene are re-scored with lower values, while the scores of categories commonly associated with the type of scene are increased. Section 3.2 covers the details in improving object detection by incorporating information about the scene to re-score the original object detection results. We propose two approaches. The first one is a greedy algorithm, and the second is an algorithm based on Support Vector Regression (SVR).

Finally, Section 3.3 presents a framework for improving object detection scores that simultaneously considers the temporal information and the global context, with the additional benefit of not requiring an explicit scene labeling of the video frames.

3.1. Improving scene identification

Given a set of training videos containing N_s type of scene identities, one scene classifier is trained for each type of scene. Under the assumption that other frames do not influence the scene identity of the current frame, each sampled frame is evaluated independently to determine the scene identity by comparing the scores of each one of the trained scene classifiers, and selecting the classifier with maximum score. However, we are dealing with first-person camera videos where the scene identity of a frame is influenced by the identities of previous frames. It is evident that a person requires some time to move from one scene to another, therefore, if a person is known to be in a particular scene, it is very likely that the individual will remain on the same stage during some additional frames.

We use a Conditional Random Field (CRF) formulation to model the temporal constraint of scene identities associated with first-person videos. The goal is to find the scene labels $\mathbf{y} = y_1, y_2, \cdots, y_N$ for a video sequence with N frames, that best fit the scores of the scene classifiers while enforcing the temporal constraint.

We define a graph with scene label nodes y_i for each frame of the video, which are connected temporally through edges with their r neighbors frame labels. Each frame label has a number of possible observations (scene classifiers) associated $x^i_{j\in[1\cdots N_s]}$. Fig. 2 presents a particular case, where the two previous frames are connected.

Let $Pr(\mathbf{y}|G;\omega)$ be the conditional probability of the scene label assignments $\mathbf{y}$ given the graph $G(S_p, Edge)$ and a weight ω, we need to minimize the energy equation

$$log(Pr(\mathbf{y}|G;\omega)) = \sum_{s_i\in S_p} \psi(y_i|s_i) + \omega \sum_{s_i,s_j\in Edge} \phi(y_i,y_j|s_i,s_j), \qquad (1)$$

where ψ are the unary potentials, and ϕ are the pairwise edge potentials.

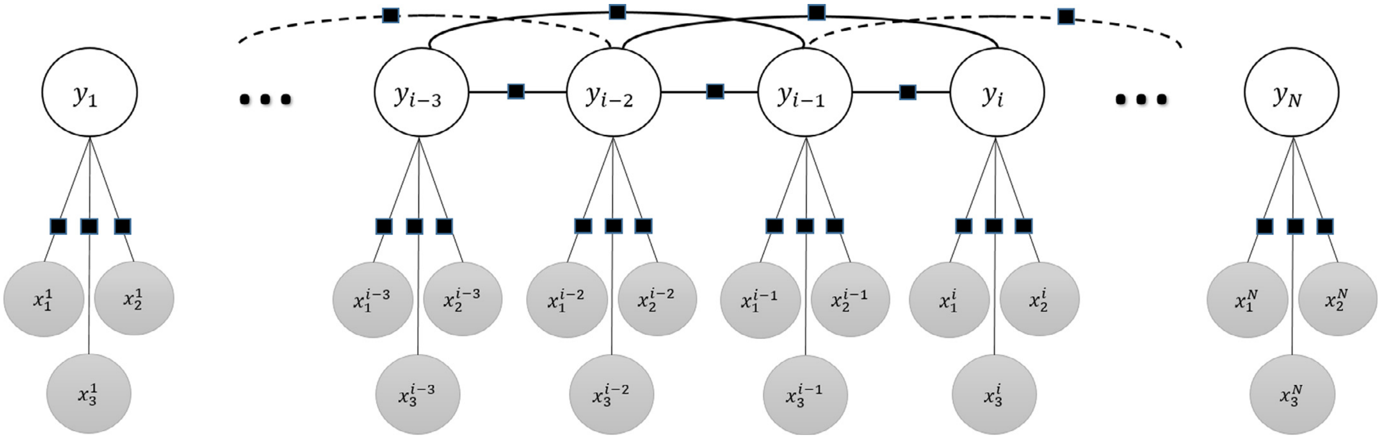


Fig. 2. Example of a graphical model representing temporal dependencies for scene labeling in a first-person camera video. A total of $r = 2$ previous observations and three possible scene identities are represented in the figure. The figure shows the observations (scene scoring) as shadowed nodes $x_{y_i}^i$ and label assignments as white nodes y_i. Experiments in Section 4.1 were performed with $r = 7$.

The energy function to minimize can be represented as

$$E(\mathbf{y}) = \sum_{p=1\cdots N} \psi(y_p) + \sum_{q=1\cdots N} \sum_{p=1\cdots N} w_{p,q} V(y_p, y_q), \tag{2}$$

where $w_{p,\,q}$ is an adjacency matrix, that indicates which nodes are connected by edges and how much influence any of the r neighbor frames has on the current frame.

In our problem the unary potential is determined by a normalized scene classification score $\mathbf{x}_{y_i}^i$ as

$$\psi(y_i) = 1 - \mathbf{x}_{y_i}^i, \tag{3}$$

which privileges scene labels with high scores.

The pairwise edge potential is given by a matrix $V(y_p,\, y_q)$. The matrix $V(y_p,\, y_q)$ is defined with zeros in its diagonal, implying that the energy is not affected if the scene identity remains the same, and with positive values in positions off diagonal of the matrix to penalize changes in the scene identity. This enforces the temporal continuity of scene identities for frames linked by edge potentials in the graph.

We will discuss choices for matrix $V(y_p,\, y_q)$ and adjacency matrix $w_{p,\,q}$ in the experimental section.

3.2. Improving object detection

Object detection is the process of finding a set of bounding boxes that delimits the regions which contain the objects of interest. When we are running object detectors, the detection scores signify the matching between the visual model and the testing bounding box content. Typically, object detectors consider at most only the local context, which corresponds to the surrounding regions of the bounding box where the object is localized, but rarely examine global information about the scene.

Consider a typical object used in ADL video, for instance, a microwave. A microwave is commonly found in the kitchen but is very unusual in other locations such as bedroom, bathroom or a laundry room. Consequently, in cases where it is possible to obtain information about the identity of the scene of the current frame, we could re-score the results of the object detector to penalize detections in scenes that typically do not contain the object that we are looking for. Overall, it is possible to increase the performance of the detector by incorporating the information about the particular type of the scene for the frame that is being tested.

The objective is to learn from the training data how much the detection score should be increased or decreased to account for the chances of having the object in a type of scene. The scene identity and the localization of the objects in every frame from the training videos of the Activities of Daily Living (ADL) dataset (Pirsiavash and Ramanan, 2012) are known in advance. Assuming that an object detector is available, we can obtain bounding boxes and their associated detection scores. We also could determine how much overlap exists between the candidate bounding box and the ground-truth bounding box of the searched object. The resulting measurement is called overlap score.

Fig. 3 clarifies the concept behind our method. We focus on the microwave object in this discussion, but it applies to any other object such as, refrigerator, tv, bed, computer, etc. In all the subfigures, the X–axis represents the detection scores produced for the different candidate bounding boxes, and the Y–axis represents the overlap score on the ground-truth bounding boxes measured using the same criteria as PASCAL VOC challenge (Area Overlap / Area Total). A detection is considered valid when the bounding box overlap score exceeds 0.5. Each dot in any of the figures represents a candidate bounding box. They are computed from object detectors trained using Fast R-CNN framework (Girshick, 2015). The color represents the scene identity. In this example, green color represents kitchen, while red color accounts for a bedroom.

From Fig. 3(a), it is clear that many valid detections (i.e., overlap score (Area Overlap / Area Total) is over 0.5) can be found in the kitchen scenes. The figure also shows that there is not a single valid microwave detection in bedroom scenes for the training dataset, which is consistent with our common sense understanding.

If we select a threshold for the object detection score that captures most of the valid detections in the kitchen, then such a threshold produces lots of false microwave detections in the bedroom scene; but if we set up a high threshold for microwave detection (in order to avoid adding invalid detection of the bedroom scenes), then a lot of correct detections from the kitchen will be ignored. Fig. 3(b) shows a possible re–scoring for the object detection scores based on the scene identity that deals with the fact that microwaves rarely appear in a bedroom. As can be appreciated from the figure, we have performed a simple shifting of the detection scores appearing in bedroom scenes. As a result, the detections from the bedroom scenes do not add any false positives which allows improving the results of object detection.

3.2.1. Greedy algorithm

The goal of our algorithm is to find the optimal value to be added to the initial object detection score for each scene identity from the training data. If N_o is the number of different object

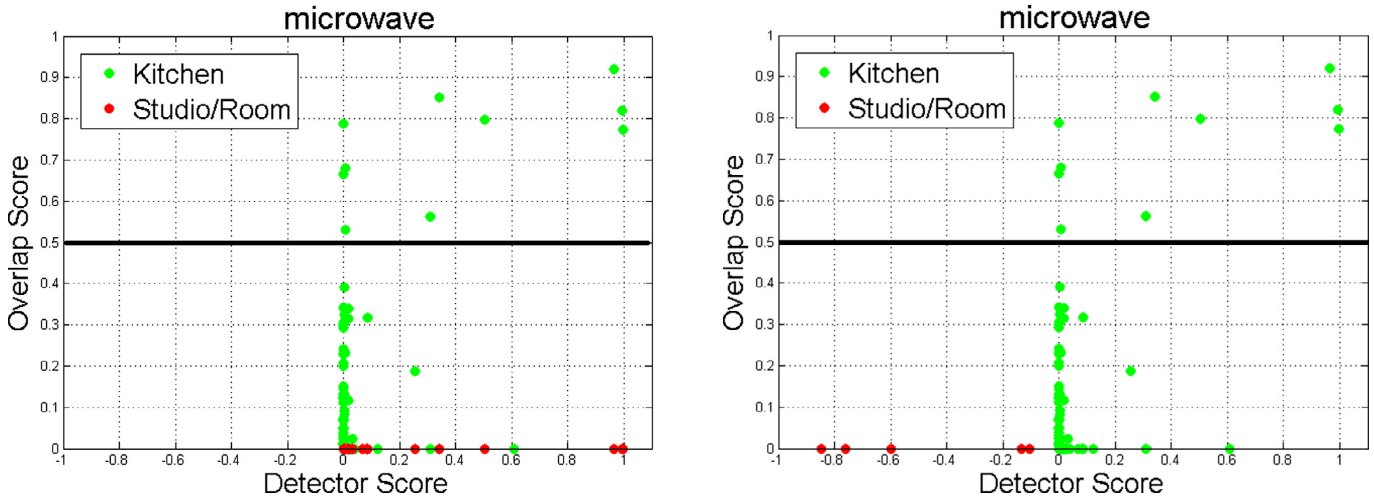


Fig. 3. Explanation of the main idea behind our method to improve object detection based on scene identity using training data of ADL dataset. Figures are generated from microwave detector, and show the detection score versus ground-truth match score. Figure a) shows the detections for the kitchen in green and the results for a bedroom in red. Figure b) shows a re-scoring that improves the object detection. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

detectors, there is a total of $N_s \times N_o$ values to be learned. The values are saved in a matrix $C_{N_s \times N_o}$, which contains the corrections that need to be added to the detection scores according to the type of scene and object detector. We fix the object detector and fill out the rows of the matrix $C_{N_s \times N_o}$ applying the procedure that is described below. Once the correction matrix is filled for the different scenes of a particular object detector, we repeat the same procedure with every object detector.

The procedure uses as input the detection scores and their corresponding overlap scores of the candidate object bounding boxes. The candidates are grouped according to the type of scene of the frame. The first step is to select a scene identity to be used as a reference by the other types of scenes to compute their corrections. We calculate the mean Average Precision (mAP) score of the object detector for the candidates in each type of scene and save them in a sorted list. The scene identity that has the highest mAP value is selected as the reference. Once the reference scene identity is selected, we process all the scenes that do not contain any valid detection according to the PASCAL–overlap criteria. This is the same case presented in Fig. 3(b). The magnitude of the correction is given by the difference between the lowest detection score value of a valid bounding box in the reference scene, and the value of the highest score of the new type of scene being processed. In practice, we also add a small fixed tolerance value ϵ, that ensure all the samples of the processed scene have scores lower than the lowest valid detection in the reference scene.

The remaining types of scenes are processed one by one starting from the scene with higher mAP in the sorted list of scenes computed in the first step that has not been processed yet. The intuition behind this choice is to assure that we adjust first the corrections of the type of scenes that need less adjustment in the correction value. At this point, we conduct a grid search of the correction value for the currently processed scene identity. The objective function is to maximize the mAP computed using the conjunction of candidates bounding boxes from previously processed scene identities and the currently processed scene identity. All the detection scores of the candidates involved in the computation of the mAP are adjusted according to their scene identities.

3.2.2. Support Vector Regression (SVR) algorithm

In this subsection, we present an algorithm to learn to re-rank the object detection scores depending of the scene identity of the tested frame. The algorithm is based on a Support Vector Regressor (SVR). The problem of regression is equivalent to finding a function which approximately maps from an input domain to the real numbers based on a training sample.

Our goal is to map the object detection score to a new score value considering the scene identity. Then, the input data must encode the current scene identity and also include the detection score. The scene identity is encoded as one-hot vector of scene identities i.e. a vector with dimension equal to the number of scenes, with an entry equal to one in the dimension representing the actual scene identity, and zeroes in all the others dimensions. Hence, the input data $x^i \in \Re^{N_s+1}$ is represented by the concatenation of the one-hot scene identity vector and the detection score of the candidate bounding box.

The output data $y^i \in \Re$ contains the target detection scores. With y^i having any one of these possible values:

$$y^i = \begin{cases} 1 & \text{if overlap score} \geq 0.5 \\ J^i & \text{otherwise} \end{cases} \qquad (4)$$

where J^i represents the overlap score of the candidate detection.

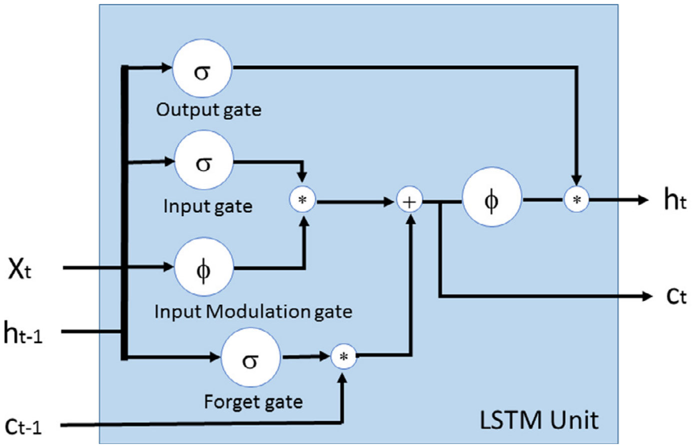
A different regressor is trained for every type of object in the dataset. During testing, the detection score and the output of the scene classifiers are used to encode the input vector. The regression output of the regressor associated to the type of object is used as the new score for the bounding box.

3.3. Improving object detection without scene identity labeling.

In this section, we present a framework to use the general visual information of the frame sequences, and impose temporal constraints with the purpose of estimating how likely certain types of objects are present in the frame (without using a specific object detection method). Such information is employed to improve the results of the existing object detectors.

Our framework is based on a feedback network called Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997). LSTM is a type of neural network that allows connections from units in the same layer, creating loops that enable the network to use information from previous passes, acting as memory. LSTM can actively maintain self-connecting loops without degrading associated information. Fig. 4 depicts the internal structure and the associated equations of the LSTM unit selected in our implemen-

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



$$i_t = \sigma(W_i X_t + U_i h_{t-1} + b_i)$$
$$f_t = \sigma(W_f X_t + U_f h_{t-1} + b_f)$$
$$g_t = \phi(W_g X_t + U_g h_{t-1} + b_g)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$
$$o_t = \sigma(W_o X_t + U_o h_{t-1} + b_o)$$
$$h_t = o_t \odot \phi(c_t)$$

Fig. 4. Internal representation of an LSTM unit.

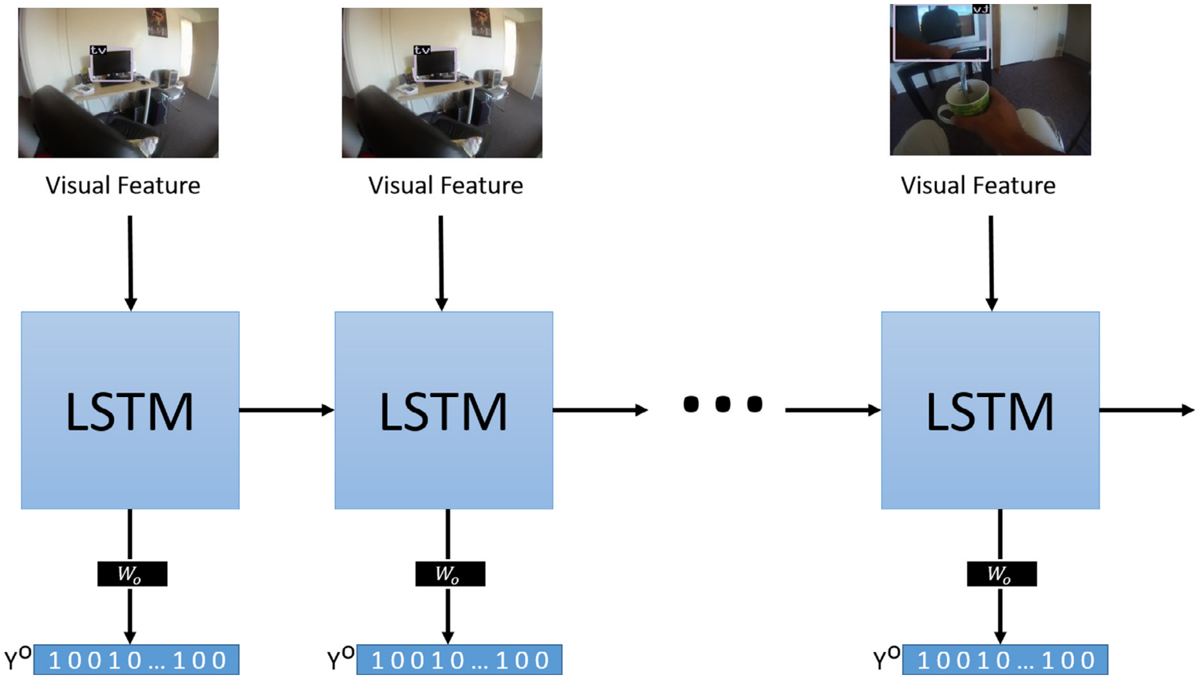

tv
tv
Visual Feature
Visual Feature
Visual Feature
LSTM
LSTM
LSTM
W_o
W_o
W_o
Y^O 1 0 0 1 0 ... 1 0 0
Y^O 1 0 0 1 0 ... 1 0 0
Y^O 1 0 0 1 0 ... 1 0 0

Fig. 5. Our framework to obtain the most likely objects from scene descriptor in a frame sequence. Visual features are used as inputs, while the target vector $Y^o = [y^o_1, y^o_2, \cdots, y^o_{N_o}]$ encodes the presence or absence of an object class in the frame.

tation. The LSTM unit takes an input vector X_t at each time step t and predicts an output h_t. In contrast to a simple Recurrent Neural Network (RNN) unit, the LSTM unit additionally maintains a memory cell c, which allows it to learn longer term dynamics. As a consequence, LSTM is a very effective technique to capture contextual information when mapping between input and output sequences.

Fig. 5 depicts the proposed framework. Every frame is preprocessed to obtain a visual image descriptor which feeds the Long Short-Term Memory (LSTM) network. The system is trained to produce the correct answer to the question: which objects are visible in the image?

The answer to the question is encoded also as a vector $Y^o = [y^o_1, y^o_2, \cdots, y^o_{N_o}]$, where N_o is the number of possible objects to be considered, and $y^o \in \{0, 1\}$. The vector Y_o has non-zero entries at the positions that indicate the indexes of existing objects in the frame. In training time, we use the information of every frame to fill out the vector Y^o, and the image descriptor X.

During testing, for each frame descriptor, we obtain a N_o dimensional output vector Y^o with values in the range [0, 1]. The N_o dimensions of the vector indicate how likely is finding a type of object given the visual information of the frame and its history. The output layer after the LSTM unit is shared across the time.

In practice, we use this likelihood as a way to re-score the results of object detectors, according to the general information of the scene by means of the simple re-scoring function

$$S^{new}_{p_j} = S_{p_j} + k * Y^o_p, \tag{5}$$

where $S^{new}_{p_j}$ is the new score for the instance j of object type p, S_{p_j} is the score result of the object detector j for object type p , Y^o_p is the output after the LSTM unit that indicates the likelihood of having the object p in the scene, and k is a constant that indicates the importance of the scene information in the final score. The value of k is determined from a small validation set containing ADL egocentric videos.

4. Experiments

We conducted our experiments in the Activities of Daily Living (ADL) dataset (Pirsiavash and Ramanan, 2012). ADL dataset captures High Definition (HD) quality video from 18 daily indoor activities such as washing dishes, brushing teeth, or watching television, performed by 20 different persons in their apartments. Each video has approximately 30 min length, and the frames are annotated every second with object bounding boxes of 42 different object classes. From the 42 annotated object classes, results of a trained Deformable Part-based Model (DPM) (Felzenszwalb et al., 2010) are provided for 17 of them. In addition to the provided DPM models, we trained object detectors using the Fast R-CNN framework (Girshick, 2015) and show that the proposed algorithms consistently achieve improvements independently of the type of object detector used.

The ADL dataset provides splits for separating training and testing data. From the twenty videos of the dataset, the first six of them were used as training data for object detection by the authors of the dataset. We followed the same splits on the data, then the first six videos were used to train scene classifiers, object detectors using deep networks, and the LSTM network for improving object detection without scene labels.

We performed scene identity annotations for all the video frames of the dataset. We identify eight types of scenes in the dataset. They are the kitchen, bedroom, bathroom, living room, laundry room, corridor, outdoor, and none of them (blurred frames, or non-identified place).

To evaluate the object detectors, we use the standard mean Average Precision (mAP) evaluation metric. We use the classical PASCAL VOC criteria, which establishes that at least a value of 0.5 on the overlap/union ratio among ground-truth and detection bounding box is needed to declare the bounding box as a valid detection.

4.1. Scene identification

In this section, we show experiments on frame level scene identification and the improvements achieved by using the temporal information.

We performed frame scene identification on the video frames of the test dataset. The first baseline in our experiments is simply the results of the scene identification methods without considering the time constraint. A second more challenging benchmark considers the temporal constraint by using a moving average filter across the temporal domain. A third baseline examines a Hidden Markov Model (HMM). Finally, we show that the overall accuracy of scene identification methods is largely improved using the proposed CRF formulation.

We use four different frame level scene identification approaches in our experiments to show that the proposed formulation works well independent of the selected scene identification method. One approach is the traditional Bag of Words (BoW) representation, encoding CNN features computed over object proposals selected by using the selective search window technique by Cheng et al. (2014). We also performed experiments with the Multi-Scale Orderless Pooling of Deep Convolutional Activation Features (MOPCNN) (Gong et al., 2014) and the two additional variants described below.

Multi-Scale Orderless Pooling of Deep Convolutional Activation Features (MOPCNN) (Gong et al., 2014) is, to the best of our knowledge, the current state of the art for scene classification. MOPCNN operates in 3 scales, all of them using the sixth fully connected layer output of the Krizhevsky's convolutional network. In the full image scale, the descriptor is directly the output of the sixth layer, while the descriptor for the other two scales is created by VLAD encoding of periodically sampled CNN features at different scales followed by dimensional reduction.

Table 1
Comparison of the overall accuracy of four scene identification methods. The baseline 1 does not consider any temporal constrain, the baseline 2 uses a moving average filter in the time domain to decide the frame identity, and baseline 3 considers a HMM model. The proposed CRF is examined under two different choices of pairwise terms.

	BoW	MOP	CNN	CNN
	CNN	CNN	L1	L3
Baseline 1. No time	50.45	64.53	64.08	63.87
Baseline 2. Moving average	58.54	67.95	69.38	67.66
Baseline 3. HMM	61.21	68.97	70.92	69.79
Proposed CRF - 1. Uniform *V*, ω	**65.52**	68.53	71.85	69.88
Proposed CRF - 2. Non-uniform *V*, ω	62.27	**72.09**	**74.21**	**72.15**

The complete MOPCNN method is used as one of the tested scene identification methods, but also two variants of the method are also examined: a) the full scale of the MOPCNN method (MOPCNN-L1) i.e., the global CNN descriptor, and b) the third scale of the MOPCNN (MOPCNN-L3), which uses VLAD encoding in the 64×64 pixels scale. These two variants complete our four methods used for scene identification.

We used Caffe (Jia, 2013) to implement CNN feature extraction. For the Bag of Words implementation, a total of 200 object proposals were used, and the dictionary size was fixed in 5000 words. For all the scene identification methods, we use a linear SVM as the classifier. We use the graph-cuts based minimization procedure in Boykov and Kolmogorov (2004); Boykov et al. (2001); Kolmogorov and Zabih (2004) to obtain the optimal solution for the Eq. (2).

Table 1 shows the overall accuracies for the three baselines and the proposed CRF method. The baseline 1 in the table corresponds to the direct output of the scene classifiers. The baseline 2 corresponds to the moving average filtering in the time domain of the scene scores. The filter size is in some way a measure of how fast the person changes from the current scene to other scene. In our experiments, the sample rate is one frame per second (1 *fps*). We examined different filter sizes, finding that considering the $r = 7$ previous sampled frames on the currently tested frame produced best accuracies. These are the results reported in the second row of the table. The baseline 3 is a Hidden Markov Model (HMM) that predicts the sequence output of the scene identities.

The results of the proposed CRF method depends on the choice of the matrices $V(y_p,\ y_q)$ and $\omega_{p,\ q}$. Following the findings of the baseline 2, the presented results assume that information of the previous seven frames influences the current frame label.

We first consider the case where any of the seven previous frames have the same impact on the current frame label i.e., $\omega_{p,p-1} = \omega_{p,p-2} = \cdots = \omega_{p,p-7}$, and penalty is the same for any pair of scene identities, i.e., the $V(y_p,\ y_q)$ value is the same for any position off diagonal. The fourth row of Table 1 reports results for this uniform choice of *V* and ω.

We also considered the case where the influence of the most recent frames is stronger than the previous ones. Hence, we assumed that for each row of the matrix ω, its weights follow a Gaussian function with origin in the current frame. We also considered alternatives for matrix $V(y_p,\ y_q)$, where pairs of scene labels with more frequent transitions are penalized less severely than others pairs that rarely occurs. We use the ground-truth data to count for the possible transitions between scene identities which are normalized and represented as T_{y_p,y_q}. Values for the $V(y_p,\ y_q)$ entries are defined as $V(y_p, y_q) = 1 - T_{y_p,y_q}$. The last row of Table 1 shows the best results achieved with the selected non-uniform *V* and ω matrices.

In all the four scene classifiers, there is a visible improvement in the accuracy using the proposed CRF with respect to the baselines. The relative increase is more significant for the weakest scene classifier, the Bag of CNN features. As is expected, the state of the art method (MOPCNN) has the best accuracy between the scene classifiers before using any temporal constraint. However, after considering the temporal information, the improvement is superior in the scene detectors that only use one scale CNN as a classifier. As a result, the two variants of the MOPCNN method produce better accuracies than the complete MOPCNN method. This surprising result, indicates that in real life applications, a weaker but less computationally intense scene classifier can be used in place of expensive computational methods as long as the temporal constraint is exploited.

We also note that the CRF defined with a more complex pairwise relation (Non-uniform V and ω), that weights the importance of the closest frames to the tested frame, and considers the likelihood of scene transitions, produces significantly better results when the best scene classifiers (MOPCNN and their two variants) are used. The increase was a bit lower than the uniform V and ω CRF with the weakest scene classifier (BoW), but still considerably better than any of the baselines. We attribute this effect to the stochastic nature of the output generated by the noisiest BoW classifier that converts the output in a less predictable event.

4.2. Improving object detection

We perform experiments to demonstrate that the methods presented in this paper to improve object detection results, generalize to different kinds of object detectors. In this section, we use the DPM object detection results provided with the ADL dataset and also the object detection outputs of models trained using the Fast R-CNN framework.

The DPM models themselves are not provided, only the bounding boxes, and scores of the detections obtained by their models in the training and testing videos of the ADL dataset. A total of 17 types of objects is provided.

The Fast R-CNN models are trained using the VGG16 network (Simonyan and Zisserman, 2015) employing object proposals computed using EdgeBox (Zitnick and Dollr, 2014). We trained models for the 42 annotated objects. However, we only consider objects with an mAP of at least 5.00%. A total of 20 object detectors satisfies this condition.

We learned different matrices of corrections $C_{N_s \times N_o}$ and rescoring functions for the DPM and the Fast R-CNN detectors following the procedures described in Section 3.2. In the case of the greedy algorithm, the parameter ϵ was set to 0.05 for all the experiments. In the case of the SVR algorithm, we used a Radial Basis Function (RBF) as kernel. The parameters of the SVR were $C = 0.01$, and $\gamma = 0.1$, in order to have a smooth regression function.

The first six videos of the ADL dataset were used to train the LSTM network. These videos contain information about which objects are shown in each one of the sampled frames. The Y^o vectors were generated by forming groups with duration of 20 s and an overlap of 8 s. We used the scene descriptor of the MOPCNN method to feed the network. The training was performed in batches of 16 groups executing a total of 30,000 epochs.

In testing phase, we feed each frame with the scene descriptor, and obtained a vector that indicates the likelihood of having the object (indexed in each dimension of the vector) given the general scene content. We used the Eq. (5) to re-score the object detection. The value of k in our validation set was 0.11 for both set of object detectors, the DPM and Fast R-CNN models.

Table 2
Results for the DPM object detection of the ADL dataset using mAP metric (as a percentage). The use of scene information increases the mAP for most of object categories. The best improvements are obtained when the scene identity is known. LSTM method performs better in comparison to the cases where the scene identity is estimated from scene classification.

		Scene known		*CNN-L1 scene*		
	DPM	Greedy	SVR	Greedy	SVR	LSTM
bed	8.74	10.32	9.28	9.01	9.34	9.37
book	11.93	11.12	10.98	12.11	11.21	12.54
bottle	1.76	1.83	2.05	1.73	2.01	1.69
cell	0.19	0.35	0.29	0.18	0.32	0.19
detergent	3.90	4.64	5.12	4.02	4.87	3.96
dish	1.26	0.98	1.35	1.53	1.04	1.38
door	12.60	7.82	8.64	12.83	9.79	14.24
fridge	24.80	28.45	29.18	25.95	26.05	26.36
kettle	12.16	13.02	12.67	11.43	12.56	13.01
laptop	38.52	40.41	37.81	38.99	32.93	39.81
microwave	17.76	21.37	22.13	18.88	21.86	19.57
pan	6.15	6.70	7.02	6.23	6.58	6.58
pitcher	1.37	1.69	1.65	0.68	1.79	1.27
soap	5.12	6.34	6.48	5.43	5.72	6.00
tap	30.15	32.40	33.38	30.19	31.84	29.59
remote	4.88	6.28	5.91	5.14	6.31	6.12
tv	44.09	46.88	48.21	45.70	47.19	45.12
Total	13.25	14.15	14.24	13.53	13.61	13.93

Table 3
Results for the Fast R-CNN object detectors on the ADL dataset using mAP metric (as a percentage). The LSTM method produces higher improvements compared to any of the other methods to re-score the object detection results.

		Scene known		*CNN-L1 scene*		
	Fast R-CNN	Greedy	SVR	Greedy	SVR	LSTM
book	12.83	13.62	13.88	13.12	14.14	13.33
bottle	11.28	12.32	9.96	8.70	9.81	11.71
cell	8.65	2.21	3.30	4.51	6.31	8.65
detergent	9.13	11.23	7.50	8.75	8.99	9.14
dish	11.19	13.03	13.85	12.01	12.96	11.95
door	5.59	5.69	5.85	5.61	5.24	5.74
fridge	24.95	27.54	26.25	25.07	25.41	26.75
kettle	23.83	31.11	26.79	27.12	27.20	27.28
laptop	37.46	41.17	33.16	43.91	37.37	48.84
microwave	32.35	36.85	36.78	33.62	34.53	32.37
mug/cup	13.24	14.67	14.21	12.51	12.90	14.29
oven/stove	43.02	47.73	54.58	49.54	52.66	52.54
pan	10.99	13.90	13.83	10.78	11.31	11.00
person	25.74	43.66	66.63	64.97	63.49	71.64
soap	18.77	19.09	20.53	17.05	16.94	18.62
tap	39.55	48.78	46.00	47.64	46.25	47.90
thermostat	9.01	9.63	6.27	6.00	7.83	8.99
remote	32.88	43.91	47.98	43.79	45.20	41.34
washer/dryer	38.86	47.17	45.09	39.09	40.42	40.52
tv	57.58	61.60	66.07	61.96	63.57	67.75
Total	23.35	27.24	27.91	26.79	27.15	28.49

The Figs. 6 and 7 show some qualitative results of five object detectors with a detection threshold fixed on −0.7 using the DPM object detector, for some random frames covering different scenes. The figures in column one show the detection results without using scene information, while the figures in the second column show the obtained detection after performing re–scoring considering the scene identity. The number of false microwave detections is reduced for the scenes in the bedroom, living room, and bathroom. In the same way, false positives objects such as tv are removed from the scenes in the kitchen, and bathroom.

Table 2 presents the results associated with the DPM object detectors and Table 3 displays the results related to the Fast R-CNN object detector. Tables 2 and 3 share the same structure. Each column of the tables presents the detection results in different scenarios. The columns of the tables present such scenarios.

The first column contains the results of the selected object detector applied on the sampled frames of the ADL dataset without considering any information about the scene.

The second and third columns present the results of improved object detection assuming the scene identities are known. Each

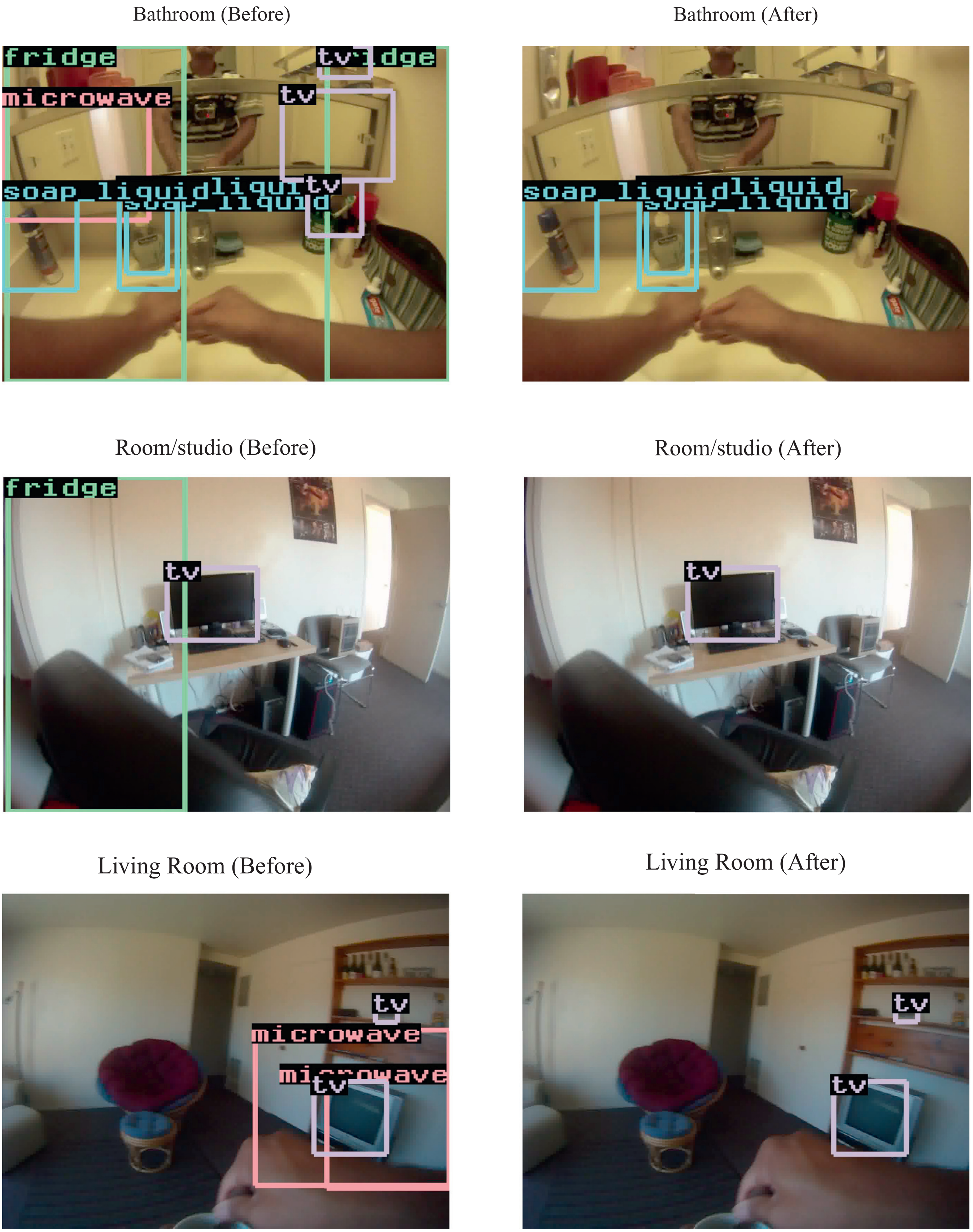


Fig. 6. Qualitative results of the object detection before and after re–scoring the detections based on the scene. Many false positives are removed after the proposed re-scoring.

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

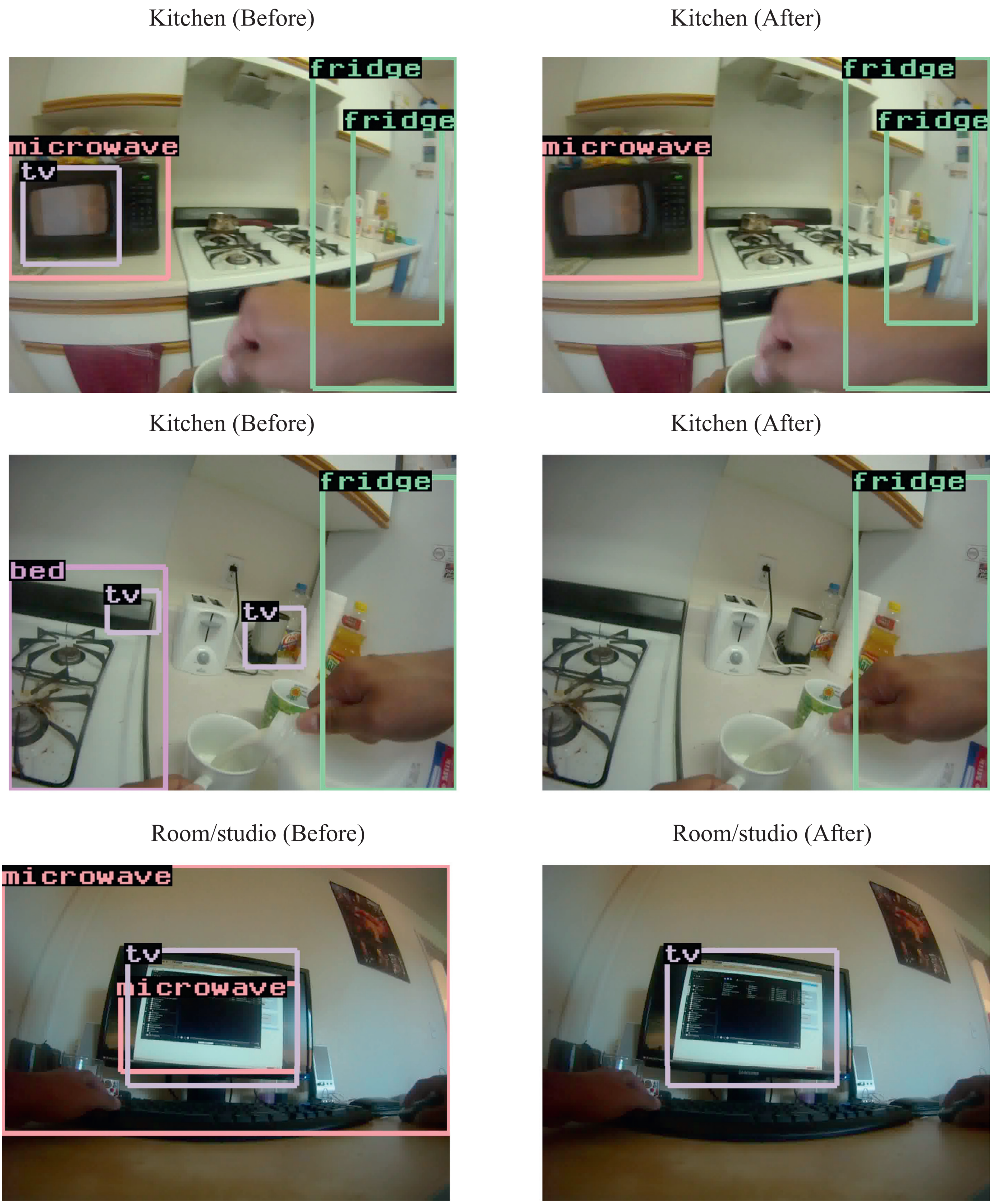


Fig. 7. More qualitative results of the object detection before and after re–scoring the detections based on the scene. Many false positives are removed after the proposed re-scoring.

column shows a different technique. They are the greedy algorithm and the SVR algorithm.

Instead of assuming the scene identities of the frames are known, the next two columns present the outcome of the greedy and the SVR algorithms, but this time using the best scene identification method that was obtained from the experiments of the previous section. This method is the model trained using the CNN features in full scales (L1) in conjunction with CRF. In the case of the greedy algorithm, the corrections values for a frame are computed as a weighted sum of corrections associated with the normalized scene identities scores for each type of object. The correction values are extracted from a column of the matrix $C_{N_s \times N_o}$, and the weights from the normalized scores of the scene identity classifiers. For the SVR algorithm, we follow a similar weighting strategy to estimate the new score values but using the scores obtained from the regression functions.

Finally, the last column shows the results of the proposed LSTM method to improve object detection without explicitly using scene labeling.

Trained Fast R-CNN object detectors produce better detector models than the provided DPM results. The performance is almost the double, as can be appreciated from comparing the first column of the tables. Besides, Fast R-CNN models can generate a longer number of good models, 20, compared to the 17 models provided using DPM.

When we include the information about the scene, we observe consistent improvements for all the presented scenarios independent of the detector model utilized. The gains are more notorious in the case of the Fast R-CNN detector models than for DPM. As is expected, using the exact information about the scene identity (columns 2 and 3) outperforms the results obtained when the scene identity is estimated (column 4 and 5) for both types of detector models. The increases are considerable in the case of Fast R-CNN models. SVR algorithm have slightly better overall performances compared to the greedy algorithm in all the tested scenarios.

In general, a valid observation of the experiment is that when the object detectors have good models (mAP over 20%), the improvements of the results by using the scene information are consistently higher than for weaker object detectors.

Finally, we highlight the results of the improved object detection without explicitly using the label of the scene. Besides of reducing the labeling effort, we note that the performance achieved using the proposed LSTM formulation outperform the results reached when we estimate scene labels from scene classifiers. In fact, for the case of the Fast R-CNN detectors, the results are superior to the ones obtained using the knowledge about the scene identity directly.

5. Conclusions

In this article, we presented algorithms for leveraging inherent constraints of egocentric vision towards improved scene identification and object detection capabilities. Firstly, we notice that the scene identity of a first-person video remains consistent for several frames. Subsequently, we presented a CRF formulation that improves the frame level scene identification results of different methods for scene identification. Secondly, we identified the association between some type objects with some scene locations and proposed two re-scoring algorithms to improve the object detection according to the scene content. For the case where an explicit scene labeling is not available, we proposed a LSTM formulation that directly estimates the likelihoods of having some objects given a sequence of scene descriptors. Such formulation was used to improve the object detection scores of the DPM and Fast R-CNN object detection outputs. The presented algorithms were implemented and tested on the well-known public ADL dataset.

References

Andriyenko, A., Schindler, K., 2011. Multi-target tracking by continuous energy minimization. CVPR.
Boykov, Y., Kolmogorov, V., 2004. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision.. IEEE Trans. Pattern. Anal. Mach. Intell. 26 (9), 1124–1137.
Boykov, Y., Veksler, O., Zabih, R., 2001. Efficient approximate energy minimization via graph cuts. IEEE Trans. Pattern. Anal. Mach. Intell. 20 (12), 1222–1239.
Carbonetto, P., de Freitas, N., Barnard., K., 2004. A statistical model for general contextual object recognition. ECCV.
Cheng, M.-M., Zhang, Z., Lin, W.-Y., Torr, P., 2014. Bing: Binarized normed gradients for objectness estimation at 300fps. CVPR.
Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C., 2004. Visual categorization with bags of keypoints.. ECCV Workshop on Statistical Learning in Computer Vision..
Divvala, S.K., Hoiem, D., Hays, J.H., Efros, A.A., Hebert, M., 2009. An empirical study of context in object detection. CVPR.
Everingham, M., Eslami, S.M.A., Gool, L.V., Williams, C.K.I., Winn, J., Zisserman, A., 2014. The pascal visual object classes challenge a retrospective. Int. J. Comput. Vis..
Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. 88, 303–338.
Fathi, A., Ren, X., Rehg, J.M., 2011. Learning to recognize objects in egocentric activities. CVPR.
Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D., 2010. Object detection with discriminatively trained part based models. IEEE Trans. Pattern Anal. Mach. Intell. 32 (9).
Forsyth, D., Malik, J., Fleck, M., Greenspan, H., L.T., Belongie, S., Carson, C., Bregler, C., 1996. Finding pictures of objects in large collections of images. Object Representation in Computer Vision.
Girshick, R., 2015. Fast r-cnn. In: International Conference on Computer Vision (ICCV).
Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR.
Gong, Y., Wang, L., Guo, R., Lazebnik, S., 2014. Multi-scale orderless pooling of deep convolutional activation features. ECCV.
Grauman, K., Darrell, T., 2005. The pyramid match kernel: Discriminative classificationcation with sets of image features. ICCV.
Han, W., Khorrami, P., Paine, T.L., Ramachandran, P., Babaeizadeh, M., Shi, H., Li, J., Yan, S., Huang, T.S., 2016. Seq-NMS for Video Object Detection. Technical Report. Technical Report for Imagenet VID Competition 2015.
Heitz, G., Koller, D., 2008. Learning spatial context: Using stuff to find things. ECCV.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Jegou, H., Douze, M., Schmid, C., Perez., P., 2010. Aggregating local descriptors into a compact image representation. CVPR.
Jia, Y., 2013. Caffe: an open source convolutional architecture for fast feature embedding.
Kolmogorov, V., Zabih, R., 2004. What energy functions can be minimized via graph cuts? IEEE Trans Pattern Anal Mach Intell 26 (2), 147–159.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. NIPS.
Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. CVPR.
Oliva, A., Torralba, A., 2007. The role of context in object recognition. TRENDS Cogn. Sci. 11 (12), 520–527.
Oquab, M., Bottou, L., Laptev, I., Sivic, J., 2014. Learning and transferring mid-level image representations using convolutional neural networks. CVPR.
Park, D., Ramanan, D., Fowlkes, C., 2010. Multiresolution models for object detection. ECCV.
Perronnin, F., Dance, C., 2007. Fisher kernels on visual vocabularies for image categorization. CVPR.
Perronnin, F., Snchez, J., Mensink, T., 2010. Improving the fisher kernel for large-scale image classification. ECCV.
Pirsiavash, H., Ramanan, D., 2012. Detecting activities of daily living in first-person camera views. CVPR.
Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S., 2014. Cnn features off-the-shelf: An astounding baseline for recognition. CVPR DeepVision Workshop.
Ren, X., Philipose, M., 2009. Egocentric recognition of handled objects: Benchmark and analysis. CVPR Workshop.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., Fei-Fei., L., 2014. Imagenet large scale visual recognition challenge. ArXiv:1409.0575.
Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y., 2014. Overfeat: Integrated recognition, localization and detection using convolutional networks. ICLR.
Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. ICLR.
Sivic, J., Zisserman, A., 2003. Video google: A text retrieval approach to object matching in videos. ICCV.

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

Song, Z., Chen, Q., Huang, Z., Hua, Y., Yan, S., 2010. Contextualizing object detection and classification. CVPR.

Soomro, K., Idrees, H., Shah, M., 2015. Action localization in videos through context walk. In: IEEE International Conference on Computer Vision.

Stauffer, C., 2003. Estimating tracking sources and sinks. CVPR Workshop, 4.

Torralba, A., Murphy, K., Freeman, W.T., 2010. Using the forest to see the trees: Object recognition in contex. Comm. of the ACM.

Torralba, A., Murphy, K.P., Freeman, W.T., Rubin, M.A., 2003. Context-based vision system for place and object recognition. ICCV.

Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T., Smeulders, A.W.M., 2013. Selective search for object recognition. Int. J. Comput. Vis. 104, 154–171.

Vaca-Castano, G., Das, S., Sousa, J.P., 2015. Improving egocentric vision of daily activities. In: IEEE International Conference on Image Processing (ICIP).

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y., 2010. Locality-constrained linear coding for image classification. CVPR.

Zamir, A.R., Dehghan, A., Shah, M., 2012. GMCP-Tracker: Global Multi-object Tracking Using Generalized Minimum Clique Graphs. ECCV.

Zhang, L., Li, Y., Nevatia, R., 2008. Global data association for multi-object tracking using network flows. CVPR.

Zitnick, C.L., Dollr, P., 2014. Edge boxes: Locating object proposals from edges. ECCV.

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