IMPROVING EGOCENTRIC VISION OF DAILY ACTIVITIES

Gonzalo Vaca-Castano

Center for Research in Computer Vision
University of Central Florida

Samarjit Das, Joao P Sousa

Bosch Research and Technology Center
Pittsburgh PA

ABSTRACT

In this paper, we investigates the interplay between scene and objects on daily activities under egocentric vision constraints. The nature of egocentric vision implies that the identity of the current scene remains consistent for several frames. We showed that this constraint can be used to improve several scene identification baselines including the current state of the art scene identification method. We also show that the scene identity can be used to improve the object detection. In generic object detection, models for objects typically only considers local context, ignoring the global scene context; however in daily activities, objects are typically associated to particular types of scenes. We exploited this context clue to re-score the object detectors. Re-scoring function is learned from scene classifiers and object detectors in a validation set. In testing time, models of objects are weighted according to the scene identity score (context) of the tested frame, improving the object detection as measured by mAP, respect to object detectors without the scene identity clue. Our experiments were performed in the Activities of Daily Living (ADL) public dataset [1] which is a standard benchmark for egocentric vision.

Index Terms— Egocentric vision, Activities of Daily Living, Scene Identification, Object Detection

1. INTRODUCTION

Egocentric camera video processing has earned a lot of attention lately, due to the capability from modern wearable devices to capture videos, perform processing, and present results to the user. From an applications standpoint, egocentric video is a key enabler for a number of technologies ranging from augmented reality to context aware cognitive assistance. We particularly investigate intrinsic constraints of egocentric vision that can be exploited to improve scene identification and ego-centric object detection.

With regard to scene identification, we note that temporal constraints can be exploited to improve frame level scene identification performance. 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 many frames can lead to wrong scene identification since many of the frames capture noisy scene information, or can contain non-discriminative scene information. However, egocentric video enforces temporal smoothness constraints to some degree i.e. the scene identity is retained until user location state changes. In this paper, we will exploit the mentioned constrained to use the scene identification scores of surrounding frames to improve the scene identity accuracy.

We also look at the problem of improving the detection of objects . Object detection task attempts to find the location of objects in a frame. 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 such as kitchen, living rooms, laundry room, among others. Most of the objects are associated to some type of scenes. For instance, a fridge is an object that most likely can be found in a kitchen, compared to other objects like a tv or toothbrush. This observation is used as a constraint in our problem formulation to improve the quality of object detectors. The initial object detector is improved after identifying the scene where the frame was recorded and vary the model in favor of objects that are most probably present in the scene.

The main contributions of this paper are as follows. Firstly, we propose the use of temporal consistency constraint to improve scene identification accuracy in egocentric video by means of a Conditional Random Field (CRF) formulation. Secondly, we presents an algorithm to improve the object detection results by creating an adaptive score object modeling that is applied according to the type of scene identified. Additionally, we release scene identity frame annotations for the Activities of Daily Living (ADL) public dataset [1], where our experiments were performed.

2. RELATED WORK

Egocentric vision has recently got significant interests from the vision community since the advent of wearable vision sensors and their potential applications. Recent efforts [2, 3, 1] in egocentric vision have focused on object recognition, activity detection/recognition and video summarization, however none of these efforts have focused on scene identification and its relation with object detection. Fathi et al. [3] observed that object of interest tend to be centered and covers a large space

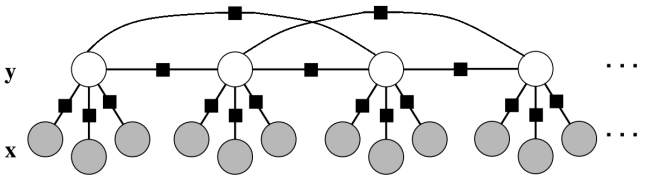

Fig. 1. Graphical Model representing temporal dependencies for scene labeling in a first-person camera video. The figure shows the observations (scene scoring) as shadowed nodes and label assignations as white nodes. A total of $r = 2$ previous observation are represented in the figure and three possible scene identities.

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 learn from the training dataset. Objects become part of the background after manipulation is competed. In [1], a new dataset of activities of daily living (ADL) in first person camera is presented.

The role of context in object recognition has been analyzed from a cognitive science perspective [4], but also from a computer vision perspective [5, 6, 7, 8, 9, 10, 11]. Heitz and Koller [6] used a terminology coined by Forsyth et al. [12] known as TAS "thing" and "stuff", linking discriminative detection of objects with unsupervised clustering of image regions. Other approach like [7] achieve boost in object detection by iteratively switching between classification task and detection using each other output as context. Divvala et al. [11] 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. [5] 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 use this as contextual information to introduce strong priors that simplify object recognition.

3. EGOCENTRIC VISION CLUES

In this work, we focus on two two important building blocks towards the goal of using 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 use of the scene identity.

3.1. Improving Scene Identification

Given a set of training videos containing C type of scene identities, one scene classifier is trained for each type of scene. Assuming that frame identity of a frame is independent of any other frame identity, each sampled frame is evaluated to determine the scene identity, by comparing the scores S_i of each one of the C trained scene classifiers, and selecting the class with maximum score for the particular frame. This assumption is distinctly erroneous since there is a dependence in the temporal domain when we are dealing with first camera videos. It is clear 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 person will remain in the same scene during some frames.

We use a Conditional Random Field (CRF) formulation to model temporal constraint and find a set of labels that best fit the scores of the C scene classifiers for a video sequence with N frames. We define a graph connecting frame labels temporally with their r previous frame labels, and each frame label depending on the current observations as is depicted in figure 1.

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

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

where ψ are the unary potentials, and ϕ are the pairwise edge potentials. In our problem the unary potential is determined by a normalized scene classification score S_i as

$$\psi(i) = 1 - S_i \tag{2}$$

which privileges scene labels with high scores.

The pairwise edge potential is simply given by a matrix $V(y_p, y_q)$ with ones in all their entries except in the diagonal where is zero. This matrix penalizes changes in the scene identity for frames linked by edge potentials in the graph, enforcing temporal continuity of scene identities.

The energy function to minimize can be represented as:

$$E(\mathbf{y}) = \sum_{p=1\cdots N} \psi(p, y_p) + \sum_{p=1\cdots N, q=1\cdots N} w_{p,q} V(y_p, y_q), \tag{3}$$

where $w_{p,q}$ is a weighted adjacency matrix, with weights equal to $1/r$ being r the number of previous frames that the current frame is connected to.

We use the graph-cuts based minimization method in [13, 14, 15] to obtain the optimal solution for equation 3, and improve the scene detection accuracy exploiting the inherent temporal constraint of egocentric vision.

3.2. Improving object detection

Assuming that we have a method for object detection that provides bounding boxes and their confidence scores, we show that is possible to increase the performance of the detector by incorporating the information about the particular type of the scene of the frame that is being tested. We learned from

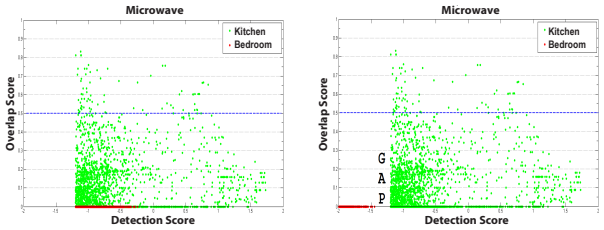

Fig. 2. Explanation of the main idea behind our method to improve object detection based on scene identity using training data of ADL dataset. Figures are generated from microwave detector, and show the detection score versus ground truth match score. Figure a) shows the detections for the kitchen in green and the results for a bedroom in red. Figure b) shows a re-scoring that improves the object detection.

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. Detection scores for objects that are unlikely to appear in a particular type of scene are re-scored with lower values, while scores of object detectors producing a good rate of true positives are increased relative to other object detectors of a different scene with lower detection rates.

The figure 2 explains the main idea of our method. In both figures, we use the object "microwave" and its associated ground–truth and DPM detection scores from the training videos of the Activities of Daily Living (ADL) dataset [1]. Here, the X axis in all the figures are the scores produced for each detection, and the Y axis represents how good is the match of the bounding box detection with respect to the groundtruth bounding boxes measured using same criteria as PASCAL VOC challenge (Area Overlap / Area Total). A correct detection is considered when the Bounding box PASCAL–overlap score exceeds 0.5. Each dot in any of the figures represents a candidate bounding box. The color represents the scene identity. In this example, green color represents kitchen , while red color represents a bedroom. From figure is clear that many valid detections(i.e. PASCAL–overlap score is over 0.5) can be found in the kitchen scenes. In the other side, the figure shows that there is not a single valid microwave detection in bedroom scenes for the training dataset, which is consistent with our common sense appreciation. If we select a threshold for the object detection score that capture most of the valid detections in the kitchen, then such 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 of correct detections from the kitchen will be ignored. The figure 2 b) shows a possible re–scoring for the object detection scores based on the scene identity that deal with the fact that microwaves are not located on the bedroom. As can be appreciated from the

	BoW CNN	MOP CNN	CNN L1	CNN L3
No Time	50.45	64.53	64.08	63.87
Proposed CRF	65.52	68.53	71.85	69.88
Improvement	+15.07	+4.00	+7.78	+6.01

Table 1. Scene identification overall accuracy for fourth baseline methods, and the improvement obtained after applying the proposed constraint

figure, a simple shifting in the score detection values based in the scene identity allows to improve the results of object detection. In this case the detections from the bedroom scenes do not add any false positives.

For each detected object, an optimal shifting of the detection score for each scene identity is learned using a simple algorithm. The algorithm uses as input the detection scores of the object detector and their bounding box proposal overlap scores with respect to the groundtruth bounding boxes (measured using Area Overlap / Area Total) for each type of scene. The detections are grouped according to type of scene of the frame. Firstly, The algorithm selects a type of scene to be used as reference to perform the detection score shifting. The scenes are sorted in descending order according to the mean Average Precision (mAP) score of the object detector, and the reference scene is selected from the top. Once the reference is selected, scenes that does not contain any valid detections according to the PASCAL–overlap criteria are processed first (same case as Figure 2 b) . The detection score for this kind of scene is negative and the magnitude of the shifting 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, plus a fixed GAP protection value. The remaining scenes are processed one by one starting from the scene with highest mAP from the sorted list of scenes. For each type of scene, the procedure is a exhaustive search of the shift value that produces the maximum mAP after adding the shifted detections of the current scene.

4. EXPERIMENTS

We extensively experimented our method in the Activities of Daily Living (ADL) dataset [1]. ADL dataset capture High Definition (HD) quality video from 18 daily indoor activities such as washing dishes, brushing teeth, or watching television, each performed by 20 different persons in their own apartments. Each video has approximately 30 minutes length, and all 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) [16] are provided for only 18 of them. We will use them and improve the object detection results provided.

Object	bed	book	bottle	cell	floss	deterg	dish	door	fridge	kettle	laptop	micro wave	pan	pitcher	soap	tap	remo te	tv
DPM	8.74	11.93	1.76	0.19	0.00	3.90	1.26	12.60	24.80	12.16	38.52	17.76	6.15	1.37	5.12	30.15	4.88	44.09
DPM Known scene	10.32	11.12	1.83	0.35	0.00	4.64	0.98	7.82	28.45	13.02	40.41	21.37	6.70	1.69	6.34	32.40	6.28	46.88
	+1.58	-0.8	**+0.07**	**+0.16**	**0**	**+0.74**	-0.3	-4.8	**+3.65**	**+0.86**	**+1.89**	**+3.61**	**+0.55**	**+0.32**	**+1.22**	**+2.25**	**+1.40**	**+2.79**
DPM + CNN-L1	9.01	12.11	1.73	0.18	0.00	4.02	1.53	12.83	25.95	11.43	38.99	18.88	6.23	0.68	5.43	30.19	5.14	45.70
	+0.27	**+0.18**	-0.03	-0.01	**0**	**+0.12**	**+0.27**	**+0.23**	**+1.15**	-0.7	**+0.47**	**+1.12**	**+0.08**	-0.7	**+0.31**	**+0.04**	**+0.26**	**+1.61**

Table 2. Results for the object detection of the ADL dataset using mAP metric (as percentage). Note that the use of scene information increases the mAP for 14 out of 18 object categories.

We performed scene identity annotations for all the video frames of the dataset. We identify 8 types of scenes in the dataset. They are: kitchen, bedroom, bathroom, living room, laundry room, corridor, outdoor, and none of them (blurred frames, or non identificated place). From the twenty videos of the dataset, the first six of them were used as training data following the original data splittings for object detection. In order to evaluate the object detectors, we use the standard mean Average Precision (mAP) evaluation metric.

4.1. Scene Identification

We perform frame by frame scene identification using four baseline methods, and apply them over the thirteen videos of the test dataset, then we show that the overall accuracy for scene identification methods is improved for all the baselines using the proposed CRF formulation.

Multi-Scale Orderless Pooling of Deep Convolutional Activation Features (MOPCNN) [17] is to the best of our knowledge, the current state of the art for scene classification. Therefore, we use this method as one of our baselines. MOPCNN operates in 3 scales, all of them using the sixth fully connected layer output of the Krizhevsky's CNN. In the full image scale, the descriptor is directly the output of the sixth layer, while for the other two scales the descriptor is created by VLAD encoding of periodically sampled CNN features in different scales and dimensional reduction.

The first baseline that we are going to use for scene identification is a Bag of Words (BoW) encoding of CNN features over object proposals (instead of periodically) selected by using the selective search window technique from [18], the second baseline is the the complete MOPCNN method, the third baseline is the full scale of the MOPCNN method (MOPCNN-L1) i.e. the global CNN descriptor, and finally the fourth baseline is the third scale of the MOPCNN (MOPCNN-L3) which uses VLAD encoding in the 64x64 pixels scale.

We use Caffe [19] 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 baselines, we use a linear SVM classifier as classifier.

The overall accuracies for the baselines and the improvement obtained after applying the proposed method to exploit the egocentric temporal constraint is showed in the table 1. In all the cases, there is a clear improvement in the accuracy. The relative improvement is huge specially for the weakest scene classifier used as baseline, the Bag of CNN features. As is expected, the state of the art method (MOPCNN) has the best accuracy between the baselines before using the egocentric temporal constraint. After applying our method, the improvement is superior for the other methods that only use one scale CNN as classifier, producing a better accuracy than the complete MOPCNN method. This surprising result, indicates that in real life applications a weaker but less computational intense scene classifier can be used in replace of computational expensive methods as long as the temporal constraint is exploited.

4.2. Improving Object Detection

The table 2 presents a comparison of the mAP for each video of different object detectors for three different cases. 1) Object detection provided with the dataset. 2) Assuming perfect scene identification, and 3) Using the scene identification results from global descriptor (L1) and our CRF temporal constraint procedure. The value of the mAP increase in 14 out 18 available object detectors. The objects that had a decrease in the rate detection are actually very bad detectors like bottle (1.73mAP), cellphone (0.19mAP), or pitcher (1.37 mAP), where the detection scores are not from any help, while the improvement is consistent in good object detectors like fridge, laptop, microwave or tv.

5. CONCLUSIONS

In this paper, we have proposed a method for leveraging inherent constraints of egocentric vision towards improved scene and object recognition capabilities. We have demonstrated the use of temporal continuity of scene state for improved egocentric scene identification as compared to the state-of-the-art methods. We have also shown how egocentric object detection can be improved by utilizing scene contexts of the objects as priors and thus re-scoring the detectors accordingly. The presented algorithms were implemented and tested on the well known public ADL dataset, and new labeling of the type of scene for the dataset is released.

6. REFERENCES

[1] Hamed Pirsiavash and Deva Ramanan, "Detecting activities of daily living in first-person camera views," in *CVPR*, 2012.

[2] Xiaofeng Ren and Matthai Philipose, "Egocentric recognition of handled objects: Benchmark and analysis," in *CVPR Workshop*, 2009.

[3] Alireza Fathi, Xiaofeng Ren, and James M. Rehg, "Learning to recognize objects in egocentric activities," in *CVPR*, 2011.

[4] Aude Oliva and Antonio Torralba, "The role of context in object recognition," *TRENDS in Cognitive Sciences*, vol. 11, no. 12, pp. 520–527, 2007.

[5] Antonio Torralba, Kevin P. Murphy, William T. Freeman, and Mark A. Rubin, "Context-based vision system for place and object recognition," in *ICCV*, 2003.

[6] Geremy Heitz and Daphne Koller, "Learning spatial context: Using stuff to find things," in *ECCV*, 2008.

[7] Zheng Song, Qiang Chen, Zhongyang Huang, Yang Hua, and Shuicheng Yan, "Contextualizing object detection and classification," in *CVPR*, 2010.

[8] P. Carbonetto, N. de Freitas, and K. Barnard., "A statistical model for general contextual object recognition," in *ECCV*, 2004.

[9] A. Torralba, K. Murphy, and W. T. Freeman, "Using the forest to see the trees: object recognition in contex," in *Comm. of the ACM*, 2010.

[10] D. Park, D. Ramanan, and C. Fowlkes, "Multiresolution models for object detection," in *ECCV*, 2010.

[11] Santosh K. Divvala, Derek Hoiem, James H. Hays, Alexei A. Efros, and Martial Hebert, "An empirical study of context in object detection," in *CVPR*, 2009.

[12] D.A. Forsyth, J. Malik, M.M. Fleck, Greenspan, Leung T.K. H., S. Belongie, C. Carson, and C. Bregler, "Finding pictures of objects in large collections of images.," in *Object Representation in Computer Vision*, 1996.

[13] Yuri Boykov and Vladimir Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision.," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124–1137, September 2004.

[14] Yuri Boykov, Olga Veksler, and Ramin Zabih, "Efficient approximate energy minimization via graph cuts," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1222–1239, November 2001.

[15] Vladimir Kolmogorov and Ramin Zabih, "What energy functions can be minimized via graph cuts?," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 147–159, February 2004.

[16] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, 2010.

[17] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *ECCV*, 2014.

[18] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *CVPR*, 2014.

[19] Yangqing Jia, "Caffe: An open source convolutional architecture for fast feature embedding," 2013.