Understanding Egocentric Activities

Alireza Fathi[1], Ali Farhadi[2] , James M. Rehg[1]
[1] College of Computing, Georgia Institute of Technology
{afathi3,rehg}@cc.gatech.edu
[2] Computer Science Department, University of Illinois at Urbana-Champaign
afarhad2@illinois.edu

Abstract

We present a method to analyze daily activities, such as meal preparation, using video from an egocentric camera. Our method performs inference about activities, actions, hands, and objects. Daily activities are a challenging domain for activity recognition which are well-suited to an egocentric approach. In contrast to previous activity recognition methods, our approach does not require pre-trained detectors for objects and hands. Instead we demonstrate the ability to learn a hierarchical model of an activity by exploiting the consistent appearance of objects, hands, and actions that results from the egocentric context. We show that joint modeling of activities, actions, and objects leads to superior performance in comparison to the case where they are considered independently. We introduce a novel representation of actions based on object-hand interactions and experimentally demonstrate the superior performance of our representation in comparison to standard activity representations such as bag of words.

1. Introduction

Understanding human activities from video is a fundamental problem in computer vision which has spawned a rich literature [22, 32]. Much of the initial work in this area has been focused on analyzing movement patterns, and has resulted in near perfect performance on simple, standard datasets such as KTH [29]. In contrast to these early datasets, people in realistic scenarios manipulate objects as a natural part of performing an activity, and these object manipulations are important part of the visual evidence that should be considered. In addition, attempts to position fixed cameras in homes or offices to capture naturally-occurring activities is challenging due to the inherently-limited field of view of a fixed camera and the difficulty of keeping all relevant body parts, including fingers and hands, in focus and at sufficient resolution at all times.

An alternative to the conventional "third person" video

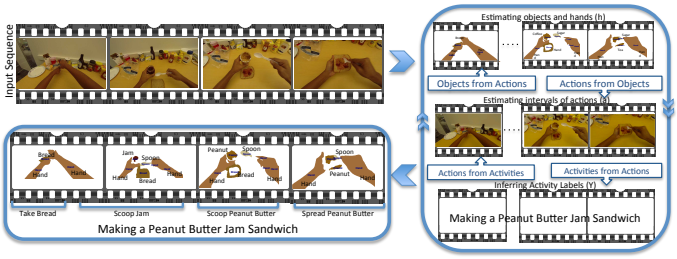

Figure 1: An overview of our approach.

capture paradigm is to mount a camera on the head of a subject and record activities from an egocentric perspective (i.e. from the subject's own point of view). There has been significant recent interest in the egocentric approach to vision [26, 30, 7, 31]. We believe that the egocentric paradigm is particularly beneficial for analyzing activities that involve object manipulation, for three reasons: First, occlusions of manipulated objects tend to be minimized as the workspace containing the objects is always visible to the camera. Second, objects tend to be presented at consistent viewing directions with respect to the egocentric camera, because the poses and displacements of manipulated objects are consistent in workspace coordinates. Third, actions and objects tend to appear in the center of the image and are usually in focus, resulting in high quality image measurements for the areas of interest.

In this paper we address the problem of understanding daily activities like meal preparation in an egocentric setting. Most day-to-day activities consist of actions that involve manipulating objects like pouring water into a cup, opening a peanut-butter jar, etc. Interactions between objects and hands contain important discriminative cues for action recognition. This suggests representing actions by objects and their interactions with hands. This approach is in contrast to traditional action recognition methods where body configurations and movements are the main features.

A key aspect of our approach is the use of the semantic relationships between activities, actions, and objects to prune the search space arising in video interpretation. We

show an overview of our framework in Fig 1. For example, the ability to detect objects being manipulated reduces the space of actions under consideration to those which are consistent with the object's affordances. Object identity also constrains the space of possible actions. For example, knowing that we are manipulating a cup rules out actions corresponding to making a sandwich and make actions corresponding to making coffee more probable. We are therefore interested in visual feature respresentations and classifiers which support the incorporation of such additional domain knowledge in the decision process.

In our approach, actions are represented as relations between objects and hands. Hand features, such as proximity and directionality of movement, are defined in an object-centric coordinate frame, thereby capturing the key properties of object manipulation. Activities are then modeled as a set of temporally-consistent actions. For example, an activity like making a peanut butter sandwich starts with taking a slice of bread and then opening a jar of peanut butter, followed by scooping peanut butter out of the jar and onto the bread. Our constraints on actions capture the fact that peanut butter cannot be scooped from the jar before it is opened. A key barrier to the use of semantic information in activity recognition is the need to hand-label object and actions across a large video corpus to support supervised learning. We address this issue by leveraging our previous work on unsupervised learning of object models from egocentric videos [7]. We augment automatically-learned object models with weak annotations of actions to complete the training data for an activity model. We present a fully-automatic method for learning activity models from such weakly-labeled data.

This paper makes three contributions: 1) We present a novel representation for egocentric actions based on hand-object interactions. 2) We develop a novel approach for automatically constructing a joint model of activities, actions and objects, in which the context provided by each element enhances the ability to recognize the others. 3) We provide experimental evaluations on an egocentric test bed and demonstrate benefits of joint modeling of actions, activities, and objects comparing to independent models.

We demonstrate the advantages of our semantic representations of actions and activities in comparison to state of the art feature based representation.

2. Previous Work

Action and activity recognition have been the subject of a vast amount of research in computer vision literature [22, 32]. We categorize the previous work on action recognition into two groups based on the kind of action classes they study. The first class of works consider body movements such as walking, running, etc, in which no other objects are involved other than the human body [4, 5]. The other class of works consists of actions such as drinking, smoking, opening, etc in which the object context plays an important role [23, 33]. In this paper, our focus is on recognition of actions in which objects are manipulated.

Action recognition methods can be further categorized into multiple groups based on the features they use to represent actions: features based on the entire human figure [4, 5], local space-time features [14, 24], features based on interaction of objects and hands [9, 23] and features based on point trajectories [21]. Even though space-time features and tracklets have obtained impressive results on challenging and realistic datasets [15, 25], they are not associated with semantic descriptions. In this paper, our goal is to develop semantically-meaningful features that model actions based on the interaction between hands and objects.

There have been various attempts in the past to model object context for action recognition. Mann et al. [18] use kinematic and dynamic properties of objects to understand their interactions. Moore et al. [23] use object context to classify hand actions. Li and Fei-Fei [16] use the object categories that appear in an image to identify an event. Wu et al. [33] perform activity recognition based on temporal patterns of object use, using RFID-tagged objects to bootstrap the appearance-based classifiers. Ryoo and Aggarwal [27] combine object recognition, motion estimation and semantic information for the recognition of human-object interactions. Gupta et al. [9] use a Bayesian approach to analyze human-object interactions with a likelihood model that is based on hand trajectories.

More recently there have been various attempts that use context to enhance action recognition. Marszalek et al. [19] demonstrate that the use of scene context improves action recognition performance. Yang et al. [34] treat the pose of the person in an image as a latent variable and use it to enhance action recognition. Yao and Fei-Fei [35] use the mutual context of object and human pose to recognize activities in images.

In contrast to these previous works, we recognize actions from the Egocentric view-point. Starner and Pentland [31] were one of the first to address action recognition from an egocentric viewpoint. Their system recognizes American sign language from a wearable camera. Spriggs et al. [30] segment and classify daily activities from a first-person view using accelerometers and visual information. A key difference between our work and these others is that we don't utilize any pretrained detectors for objects or hands. More recent examples of egocentric video analysis are Kitani et al. [12] who detect actions in outdoor environments and Aghazadeh et al. [1] who extract surprising events from life log videos.

Instead, our work introduces a framework in which activity, action and objects are recognized at the same timfe, and we show that the recognition results for each group en-

hances the others. Our learning and inference approaches consist of multiple stages, where information is first propagated from objects to actions and from actions to activities, and then after activity labels are fixed, the information is sent back to fix actions and finally to fix objects.

3. Model

Our task is to analyze an image sequence of a person performing an activity like making a tuna sandwhich. This entails inferring the activity label, segmenting the activity to a series of consecutive actions, and assigning object and hand labels to image regions in each frame. As a result, each input sequence contains a set of intervals $\mathbf{v} = \{u_1, ..., u_U\}$, where each interval u_i consists of F_i images $u_i = \{I_1, ..., I_{F_i}\}$, and each image I_j consists of m_j super-pixels. Throughout the paper, we refer to super-pixels as regions. Each super-pixel is represented with a multi-channel feature vector x_i, that includes color, texture and shape.

Inference involves assigning an activity label y to each sequence $\mathbf{v} = \{u_1, ..., u_F\}$, an action label a_i to each interval u_i, and an object, hand or background label h_j to each super-pixel x_j. Each y is a member of a set of possible activity labels, for example, $\mathcal{Y} = \{$making a peanut-butter sandwich, making a cheese sandwich, making coffee, etc$\}$. Each a_i is a member of a set of possible actions $\mathcal{A} = \{$pour water into cup, spread peanut-butter on the bread, etc$\}$. Each action consists of a verb (e.g. pick, pour, open, etc) and a set of object names (e.g. water, cup, bread, peanut-butter, hand, background, etc). Finally each super-pixel is assigned an object or hand label from the set $\mathcal{H} = \{$hand, cup, coffee, bread, water-bottle, etc$\}$.

We believe that objects, actions and activities should interact. We model this interaction by the graphical model depicted in Fig 2. Action labels interact with activity, object and hand labels. During training we observe action and activity labels and have access to weak labels of objects. During inference we only observe features from superpixels and infer the object labels, action intervals and activity labels.

4. Learning and Inference

To setup the notation, given a set of training sequences $\mathbf{x}^{(n)}$ and labels $\{y^{(n)}, \mathbf{h}^{(n)}, \mathbf{a}^{(n)}\}$, our task is to build a model, that given a new sequence $\mathbf{x}^{(n)}$ produces the true set of activity, action and object labels $\{y^*, \mathbf{h}^*, \mathbf{a}^*\} = \{y^{(n)}, \mathbf{h}^{(n)}, \mathbf{a}^{(n)}\}$. Joint learning of these variables require an unmanageably large training set and very expensive approximate inference. There are also conceptual difficulties with the joint learning of these variables, as it is not clear how to weigh the losses of each component against the others. As an alternative, we propose to exploit the independence structure of the domain and factor this model into four interacting modules:

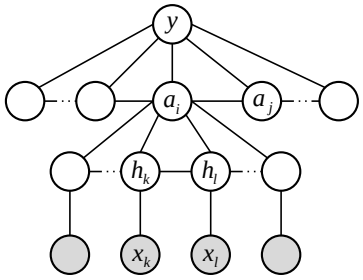

Figure 2: Our framework's model. During testing, the feature vectors of the regions x_k are observed, and our goal is to assign the object labels h_k to regions, action labels a_i to intervals and activity label y to sequences.

1. Learning to predict the intervals of actions based on hand-object interactions.

2. Learning to classify activities using an action based representations of activities.

3. Learning to modify intervals of actions given the activity labels, as well as hand and object interactions.

4. Learning to modify estimates of objects and hands given the action labels.

Our procedure is depicted in Fig 1. We start by initial estimates of appearance models for objects and hands in a weakly supervised setting. We use the multiple instance learning approach presented in our previous work [7] to provide initial object, hand and background models. We then use the module 1 to obtain action labels and use them to infer activity labels. Once we fixed the activity labels, we modify the actions accordingly. Having finalized action intervals we update our estimates of objects and hands. Our approach is similar to Expectation-Conditional Maximization [20] where the M step in the EM is replaced by conditional maximization steps. Below we describe each of these modules.

4.1. Learning Actions based on Object Interactions

This module estimates a discriminative score for assigning an action label a to an interval containing a vector of regions $\mathbf{x} = \{x_1, ..., x_m\}$, each of which are assigned an object or hand label h_i. This is a sub-problem of the original task which can be modeled by removing the top level of the graphical model (activities) as well as the connection between adjacent actions.

We want to learn a discriminative function $f_{h \to a}(a, \mathbf{h}, \mathbf{x})$, which returns a real number for any

action assignment a to an interval consisting of image regions $\mathbf{x}$ and object labels $\mathbf{h}$. We initialize $\mathbf{h}$, which is the initial object label assignments to regions using the classifiers learned from the weakly supervised object recognition. We extract object and hand interaction features and learn $f_{h\to a}(a, \mathbf{h}, \mathbf{x})$ by training a discriminative classifier on those features for each action class a. We use Adaboost [28] for classification. In contrast to the popular interest point features used for action recognition, our features are capable of capturing the semantics of interactions in the scene. Here we describe the set of object and hand interaction features used in our system.

Object Frequency (f_1): contains the histogram of object labels (hand and background labels are included as well).

Object Optical Flow (f_2): we compute the average optical flow vector for each region. The vector of each region is discretized based on its orientation and magnitude.

Object Relative Location (f_3): we build an adjacency matrix for the regions. We quantize the relative location of the center of adjacent regions into bins. For every pair of object classes we compute the histogram of their relative location bins in the interval. We reduce the dimension using PCA.

Object Classification Score (f_4): sum of the classification scores for the regions assigned to each object type are concatenated to build this feature vector.

Object Pose (f_5): for each region we compute the pose based on its shape. We build a shape descriptor as a set of annular sections (similar to shape context), each of which can be thought of as a bin. We assign each bin's value to the total number of region pixels falling in that bin.

Hand Optical Flow (f_6): these features are similar to f_2. One is computed for left hand and one for right hand.

Hand Pose (f_7): similar to f_5, one for each hand.

Hand Location in Image (f_8): We split each image into multiple regions using horizontal and vertical cutting lines. We assign the number of left/right hand pixels falling in each region as its value.

Hand Size (f_9): The area of each hand in pixels.

Left/Right Hand Relative Location (f_{10}): for each image, if there are two hands in the image, we find their pair of closest points. We use their relative x and y distance as features. We concatenate these with the relative x and y location of the center of mass of the hands.

In Sec 5.1 we evaluate the performance of these features in the recognition of various action classes. Since there is a large number of actions (64) in our experiments, we break action recognition into two steps. We first estimate the action verbs (e.g. pour, dip, pick, etc) and then in the second step we estimate the object set. In the second step we use a probabilistic model on action verb label and object set classification scores. We learn classifiers for each object set using the same set of features mentioned above. We apply a probabilistic model to infer the object set given action verb label and object set scores.

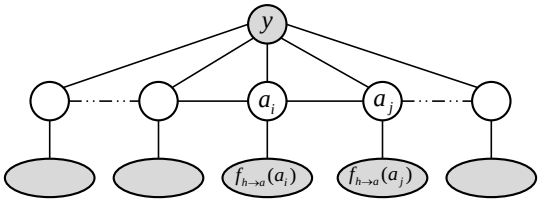

Figure 3: Model showing the decomposition of activities into actions. We refine the actions given the estimated activity label and action classification scores computed in the first stage of the algorithm.

4.2. Learning Activities from Actions

Given the set of action labels $\mathbf{a}$ we want to estimate the activity label y. We want to learn a discriminative function $f_{\mathbf{a}\to y}(\mathbf{a}, y)$ that receives the set of actions $\mathbf{a}$ in a sequence and an activity label y and returns a real number. We learn a classifier for each activity y given a histogram of action classes. We use Adaboost algorithm to learn the classifiers. During the test we build the feature vector from the action classes computed in previous stage. We assign the activity label with the highest classification score to the sequence.

4.3. Learning Actions from Activities

After fixing the activity label, we go back and enhance the action recognition results. Knowing the activity not only limits the set of possible actions, but also forces the appropriate ordering of actions. For example, pouring water is not an expected action in the making peanut-butter sandwich activity. Further, opening the peanut-butter is expected to happen before scooping peanut butter and spreading it on the bread. Given an activity label y assigned to a video and scores $f_{h\to a}$ computed in first stage, we want to assign action labels $\mathbf{a}$ to sub-intervals inside the sequence. A Conditional Random Field (CRF) chain [13] model (shown in Figure 3) is learned for every activity label y on action scores and the transition potentials between actions:

$$\arg\max_{\mathbf{a}} \sum_{(i,j)\in\mathcal{E}^{act}} \mathbf{w}_{y,a-a}^{\top} v(y, a_i, a_j) + \sum_{i\in\mathcal{V}^{act}} \mathbf{w}_{a,f}^{\top} f_{h\to a}(a_i, \mathbf{h}, \mathbf{x})$$

where $\mathbf{w}$ are weight vectors, $v(y, a_i, a_j)$ models the transition between adjacent actions a_i and a_j given activity label y, and $f_{h\to a}(a_i, \mathbf{h}, \mathbf{x})$ is the classification score of action a_i based on hand and object interactions computed in Sec 4.1. The parameters of the model are optimized using the quasi-Newton algorithm. During inference, the Viterbi algorithm is used to assign action labels.

Our focus in this paper is not on recognizing parallel actions and story telling [10, 3] or modeling temporal logical relations between intervals [2]. Instead, we focus on modeling the ordering between adjacent actions and the contextual relations between activities and actions.

4.4. Object Recognition using Action Context

We learn a probabilistic model for the objects given actions. The inputs to this model are object classification scores $\phi(x_i)$ of a region and the action label a. Output is the probability of each object label being assigned to the region x_i:

$$P(h_i|a, \phi(x_i)) \propto P(h_i|\phi(x_i))P(h_i|a)$$

We estimate $P(h_i|\phi(x_i))$ from learned classifiers on region appearance models and compute $P(h_i|a)$ from our training set.

Here we describe the method for computing object classification scores. The image annotations of the regions are unknown during both the training and testing phases. We are only provided with weak information on patterns of object-use in actions during training. For each action in training, we are given a verb and a set of nouns, corresponding to the objects used in that action. As a result, we know about the set of objects that are manipulated in action intervals. We use semi-supervised learning framework from [7] to build object classifiers given these weak informations.

In each interval, various objects might appear in the background. To only focus on objects being manipulated by hands, we segment the foreground from the background. Each foreground region contains hands, and might contain multiple regions corresponding to one or more objects. For example, in the action "scooping coffee into cup using spoon", the objects "spoon", "coffee" and "cup" might appear in the foreground simultaneously. We use a multi-class MIL framework to initialize a few regions belonging to each object class, by using the actions as positive bags for the set of their manupulated objects and negative bags for other objects. We expand these regions using a semi-supervised learning technique [6] and learn object classifiers using transductive SVM [11].

5. Experiments

In this section we present three sets of results to validate the performance of our method at its different stages: (1) object recognition, (2) action recognition and (3) activity recognition. We further analyze the performance of our semantic features for the task of action recognition.

We test our method on the GTEA (Georgia Tech Egocentric Activities) dataset [7]. We have augmented our dataset by adding the groundtruth action labels for activities. This dataset contains 7 kinds of daily activities recorded from a head-mounted camera as they were performed by 4 subjects. There are 16 kinds of objects used in these activities. The duration of each activity is about 1200 frames recorded at 15 fps. Each action consists of a verb and a few object names. There are 64 kinds of actions in the dataset, consisting of 11 different verbs (Fig 4). In this paper we use the activities performed by subjects 1, 3 and 4 for training, and use the activities performed by subject 2 for testing.

5.1. Action and Activity Recognition

We perform action recognition for every individual frame. We use features that capture the interaction of objects and hands as shown in Table 1. These features are described in detail in Sec 4.1. For each of the features, we learn multiple binary classifiers (one for each action verb class) using the Adaboost algorithm [28]. During the test we return the action class with the highest score as the action label. We have compared the accuracy of our features for different classes in Table 1. Since our features have semantic meaning, we can come up with interesting interpretations for how each feature should perform on each action class. In general, features based on the hand pose and hand location perform the best. While in traditional action recognition, the location of hands in the image is considered as a mis-leading feature, it performs the best in our domain because the Egocentric action is always recorded from the same vantage point.

For each frame, we concatenate the following features (f_2, f_6, f_7, f_8, f_9, f_{10}) to make our action representation feature vector (adding more features does not enhance the performance). We learn our action classifiers using 200 iterations of Adaboost algorithm. We compare the performance of our features with STIP and SIFT bag of words. In Fig 4 we show that our semantic representations of objects and hands provides a significant boost in recognition accuracy in comparison to widely used features like STIP and SIFT bag of words. Our features perform frame-based action recognition with 45% accuracy, while STIP performs with 14.4% and SIFT performs with 29.1% accuracy. It is interesting that SIFT bag of word features perform better than STIP. We believe this is because (1) in daily activities objects play a discriminative role in recognizing actions, (2) the same action can produce a variety of different movement patterns (imagine all the different ways one can close a water-bottle, e.g. hold with left hand and close with right hand, do it only with right hand, etc). To build the bag of word for SIFT and STIP features, we cluster them using Affinity Propagation [8]. We tune the number of words to achieve the best result.

To recognize the activities from the predicted actions, we learn a classifier on the histogram of action frequencies for each sequence. We learn multiple binary classifiers using Adaboost with 10 iterations. We can recognize 6 out of 7

Feature	Dimension	Total Accuracy	Pick	Open	Scoop	Close	Pour	Stir	Background	Spread	Put	Fold	Dip
Object Frequency: f_1	18	27.4	34	15	14	11	40	1	38	31	3	**9**	1
Object Optical Flow: f_2	18×8	31.7	25	15	18	**17**	**61**	5	41	**48**	3	3	3
Object Relative Location: f_3	$18 \times 18 \times 4$	25.2	30	11	11	9	34	3	41	23	2	4	3
Object Classification Score: f_4	18	25.2	38	12	**25**	9	37	8	35	9	4	2	3
Object Pose: f_5	18×4	26.9	31	11	10	11	39	1	45	21	6	1	0
Hand Optical Flow: f_6	2×8	30.8	20	25	0	10	54	0	**55**	25	0	0	0
Hand Pose: f_7	2×8	**34.1**	**71**	29	6	12	51	5	28	24	4	0	25
Hand Location in Image: f_8	$2 \times 3 \times 3$	**39.8**	45	38	23	**17**	8	2	44	22	1	0	**45**
Hand Size: f_9	2	25.8	33	**40**	0	7	17	23	43	3	0	0	14
Left/Right Hand Relative Location: f_{10}	4	26	4	39	7	3	32	**24**	50	19	**9**	0	2

Table 1: Classification accuracy of each feature on different action classes are shown (in percentages). Our interesting observation is that each action class is recognized with a particular feature the best. For example, during the action *pick*, the subjects always extend their hand to the end of table to take an object. As a result the hand shape discriminates this action class the best. The most interesting observation is that the feature vector extracted from the hand location in image performs the best in total. This is one of the benefits of egocentric footage. The second best performing feature is based on the hand pose. If the camera was not mounted on the head we weren't able to acquire high resolution images of hands to build these descriptors.

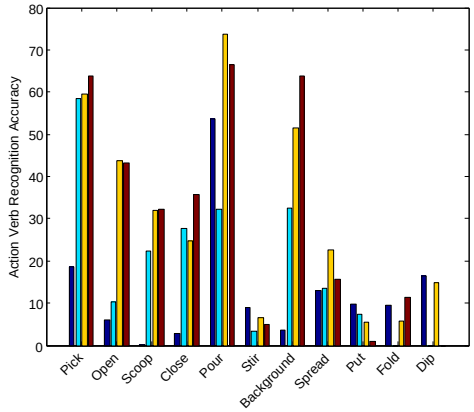

Figure 4: Action verb recognition results are compared between different methods: STIP [14] bag of words (blue), SIFT [17] bag of words (cyan), our features (yellow) and actions classification enhanced by our method after predicting the activity class (red). The total accuracy of different methods are as follows: STIP (14.4%), SIFT (29.1%), our features (45%) and our method (47.7%). There are 11 action verb classes which means the random classification accuracy is 9.1%. An interesting observation is that, since our method classifies the *making tea* activity as *making coffee*, it fails to recognize the *dipping* the tea-bag action in that sequence.

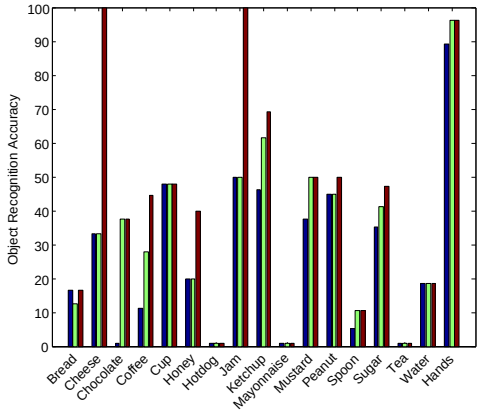

Figure 6: Object and hand recognition results are depicted. We compare the following three methods: object recognition results using the object classifiers only (blue), object recognition results using our method (green) and object recognition accuracy given the ground-truth action label (red). As is shown, knowing the action improves object recognition accuracy. Our system is capable of classifying 96.3% of the regions corresponding to hands correctly.

activities correctly. The only mistake is made by classifying *making tea* as *making coffee*. These two are very similar activities and contain very similar objects and actions.

We further compare our method which encodes interactions between activities, actions, and objects to the case of considering them independently. In our method, given the computed activity label, the action classification scores are given to a CRF model built for that activity. We use Viterbi algorithm to infer the action classes. Even though we had mis-classified the *making tea* activity as *making coffee* (1 mistake out of 7 activities), the action recognition results are improved (47.7%) in comparison to using the hand and object features alone (45%) as shown in Fig 4.

Our final recognition accuracy for the 64 classes is 32.4% compared to 4.8% for STIP and 11.6% for SIFT bag of words. Note that we are classifying every frame and chance in a 64-class classification problem is 1.6%.

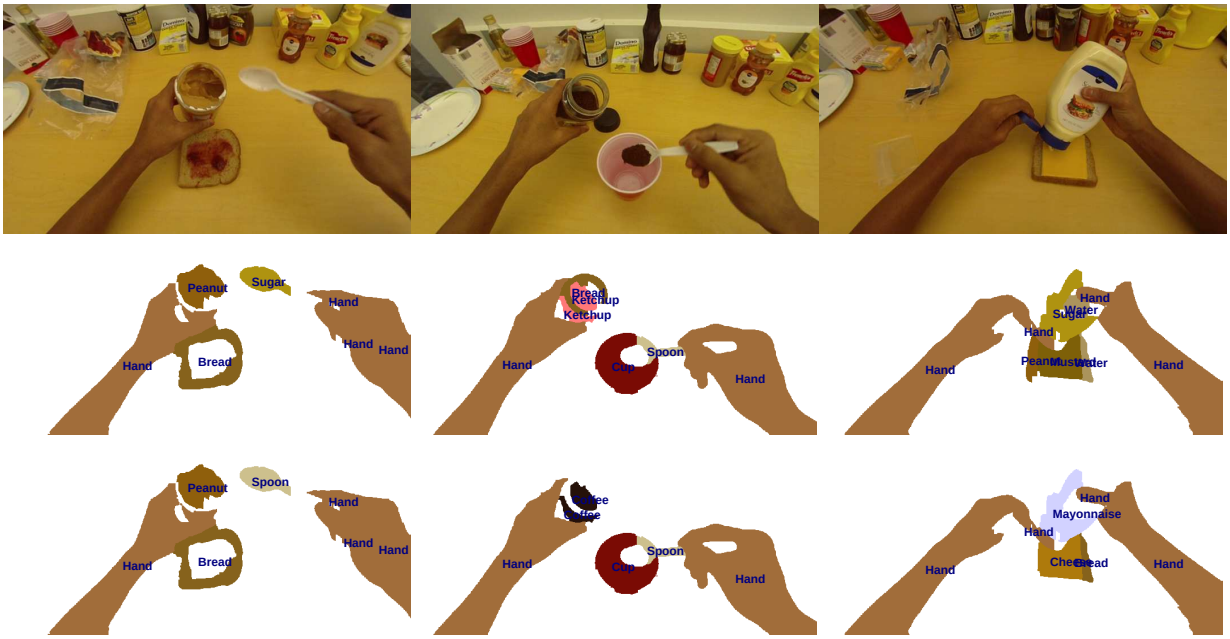

Figure 5: We show that object recognition accuracy is improved given the action as context. The images are shown in the first row. Object recognition results are shown in second row only based on object classification scores. In the third row, we show that knowing the action improves object recognition using our method in Sec 4.4.

5.2. Object Recognition

Object recognition is improved if the action is known. For example, if the action is "pouring", then "spoon" is not in the possible set of objects. We compare the object classification accuracies of classifiers learned from [7] with our method that uses action context. We present both qualitative and quantitative results demonstrating that object recognition is enhanced in our new framework.

We do not have object labels during training and testing. As a result, to measure the object recognition accuracy quantitatively, we manually annotate the ground-truth object label corresponding to each super-pixel in the foreground region. We perform these annotations for a subsample of frames (every 50 frames) of the test sequences. For each object class, we measure the accuracy of assigning the true label to the super-pixels corresponding to that class. The results are shown in Fig 6. We demonstrate that object recognition accuracy improves when the groundtruth action labels are available. For example, in the case of hands, regions are classified with 96.3% accuracy. We show qualitative results of improved object recognition given action labels in Fig 5.

6. Conclusion

We describe a novel approach to the analysis of activities in egocentric video. Our method constructs a description of an activity in terms of the objects and actions with which it is performed. We leverage the inherent coherence of views and appearance that arises from the egocentric context. We show that object and action models can be learned with very little supervision, by exploiting the joint properties of objects, hands, and actions. We propose a hierarchical inference architecture in which bottom-up propagation of evidence for objects and actions is used to predict the activity category, followed by top-down refinement of object and action descriptions based on the activity model. We demonstrate that our approach can produce superior results in comparison to standard bag-of-words type representations for activity categorization.

7. Acknowledgment

Portions of this research were supported in part by NSF Award 0916687 and ARO MURI award 58144-NS-MUR.

References

[1] O. Aghazadeh, J. Sullivan, and S. Carlsson. Novelty detection from an ego-centric perspective. In *CVPR*, 2011.

[2] J. F. Allen. Towards a general theory of action and time. In *Artificial Intelligence*, 1984.

[3] A. Bobick and Y. Ivanov. Action recognition using probabilistic parsing. In *CVPR*, pages 196–202, 1998.

[4] A. F. Bobick and J. Davis. The recognition of human movement using temporal templates. *PAMI*, 23(3):257–267, 2001.

[5] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, 2003.

[6] A. Fathi, MF. Balcan, X. Ren, and J. M. Rehg. Combining self training and active learning for video segmentation. In *BMVC*, 2011.

[7] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *CVPR*, 2011.

[8] B. Frey and D. Dueck. Clustering by passing messages between data points. In *Science*, 2007.

[9] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: using spatial and functional compatibility for recognition. In *PAMI*, 2009.

[10] A. Gupta, P. Srinivasan, J. Shi, and L. S. Davis. Understanding videos, constructing plots: learning a visually grounded storyline model from annotated videos. In *CVPR*, 2009.

[11] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, 1999.

[12] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*, 2011.

[13] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[14] I. Laptev. On space-time interest points. In *IJCV*, 2005.

[15] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[16] L. J. Li and L. Fei-Fei. What, where and who? classifying event by scene and object recognition. In *CVPR*, 2007.

[17] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[18] R. Mann, A. Jepson, and J. M. Siskind. Computational perception of scene dynamics. In *ECCV*, 1996.

[19] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009.

[20] X. L. Meng and D. B. Rubin. Maximum likelihood estimation via the ecm algorithm: a general framework. In *Biometrika Trust*, 1993.

[21] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV*, 2009.

[22] T. B. Moeslund, A. Hilton, and V. Kruger. A survey of advances in vision-based human motion capture and analysis. In *CVIU*, 2006.

[23] D. Moore, I. Essa, and M. Hayes. Exploiting human actions and object context for recognition tasks. In *ICCV*, 1999.

[24] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC*, 2006.

[25] M. Raptis and S. Soatto. Tracklet descriptors for action modeling and video analysis. In *ECCV*, 2010.

[26] X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *CVPR*, 2010.

[27] M. Ryoo and J. Aggarwal. Hierarchical recognition of human activities interacting with objects. In *CVPR*, 2007.

[28] R. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. In *COLT*, 1998.

[29] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *17th International Conference on Pattern Recognition*, 2004.

[30] E. H. Spriggs, F. De La Torre, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In *Egovision Workshop*, 2009.

[31] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. In *PAMI*, pages 1371–1375, 1998.

[32] P. Turaga, R. Chellapa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: a survey. In *IEEE Transaction on Circuits and Systems for Video Technology*, 2008.

[33] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. Rehg. A scalable approach to activity recognition based on object use. In *CVPR*, 2007.

[34] W. Yang, Y. Wang, and G. Mori. Recognizing human actions from still images with latent poses. In *CVPR*, 2010.

[35] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010.