Embedding of Egocentric Action Videos in Semantic-Visual Graph

Davide Moltisanti, Michael Wray, Walterio Mayol-Cuevas and Dima Damen
Department of Computer Science, University of Bristol, United Kingdom
{Davide.Moltisanti, Michael.Wray, Walterio.Mayol-Cuevas, Dima.Damen}@bristol.ac.uk

Introduction

An egocentric camera captures rich and varied information of how the wearer interacts with their environment. The challenge for the visual understanding of this information is currently significant and not only limited by the enormous variety of such interactions but also by limitations in the current visual description (e.g. those rooted in motion or appearance) for distinguishing interactions. Supervised training, provided by labelled examples, alleviates some of these ambiguities by including information about the object being used (e.g. 'click-mouse' vs 'pick-up phone'), or perhaps less interestingly, by limiting the study to interactions that can be distinguished from one another (e.g. open vs drink).

Several datasets [6, 5, 3, 2] and methods [5, 8, 7] have attempted supervised object interaction recognition from egocentric videos. While these works differ in the features used and classification techniques adopted, they all assume a semantically distinct set of *pre-selected* verbs or verb-noun combinations in the ground-truth. Annotators are expected to assign a video segment to one of these *pre-selected* labels. When free annotations were used, from audio scripts [1] or textual annotations [2], a majority vote is used to select a single verb for each action class. Other annotations are treated as outliers, even when they represent a meaningful and correct annotation of a video segment.

In this *ongoing* work, we treat each free annotation as a valid description of an object interaction. For example, for one interaction such as lifting an object from a workspace, free annotations can result in multiple verb labels such as 'pick-up', 'lift', 'take', 'grab' or 'move'. In this context, recognition cannot be simplified as a one-vs-all classification task. Figure 1 shows a graphical abstract of our work. Given a dataset of egocentric object interactions with free annotations, we contribute three diversions from previous attempts: (i) We deal with all free annotations as valid correct labellings. (ii) A graph that combines labels with visual similarity is built, inspired by previous work on object class categories in images [4]. (iii) A test video can be embedded into the semantic-visual graph and the probability distribution over its possible annotations is estimated.

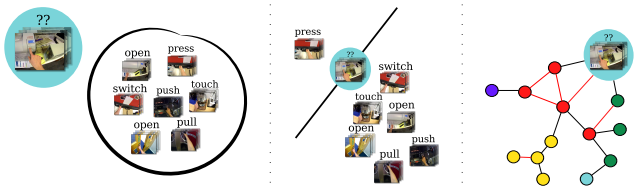

Figure 1. Given a dataset of free annotations, with potentially ambiguous verb labelling (left), we propose to deviate from the one-vs-all classical approach (middle) and build a graph that encapsulates label and visual similarities in the training set (right). Recognition amounts to embedding an unlabelled video into the graph and estimating the probability distribution over potential labels.

Embedding in Semantic-Visual Graph (SVG)

Building SVG: SVG is a representation of the training videos, with three sources of information encoded. First, annotations that are semantically linked, i.e. have a matching label, are also linked in SVG. Second, nodes that are visually similar, yet semantically distinct, should also be linked as these indicate visual ambiguities. Third, edge weights correspond to the normalised visual similarity, over neighbouring nodes, using a visual descriptor and a defined distance measure. First, an undirected graph, SVG_u is constructed, where one node $x_i \in SVG_u$ corresponds to one training video. Edges in SVG_u are created between nodes with a semantic relationship

$$x_i \frown x_j \in SVG_u \iff Label(x_i) = Label(x_j) \quad (1)$$

The undirected edge $x_i \frown x_j \in SVG_u$ is assigned a weight$w_{x_i \frown x_j}$ where $w_{x_i \frown x_j} = D_v(x_i, x_j)$ and D_v is a distance measure defined over the visual descriptor chosen. We then rank the distinct distance measures for all unconnected pairs of videos. Assume $rank(D_v(x_i, x_j))$ is a function that returns the relative position of the distance measure amongst all remaining pairings. Further links are added to SVG to encode visual ambiguity such that,

$$x_i \frown x_j \in SVG \iff rank(D_v(x_i, x_j)) \leq m \quad (2)$$

where m is the number of visual connections added to SVG_u that correspond to the top m visually similar and semantically dissimilar nodes in SVG_u. We differ from [4]

Name	Users	Sequences	OI Segments	Used Segments	Labelled Verbs	CNN						IDT					
						FV			BoW			FV			BoW		
						SVM	K-NN	SVG	SVM	K-NN	SVG	SVM	K-NN	SVG	SVM	K-NN	SVG
CMU-MMAC [3]	5	35	516	406	12	58.6	46.6	46.3	55.9	43.3	52.0	69.4	58.1	57.4	55.9	57.6	61.6
GTEA+ [5]	13	30	3371	1000*	25	15.6	30.0	31.0	25.1	33.5	33.6	43.6	43.4	42.1	27.8	34.5	40.3
BEOID* [2]	3-5	58	1488	1225	108	20.9	34.	37.5	15.2	19.1	19.6	38.7	36.0	37.4	34.8	39.6	45.0

Table 1. Method is tested on three public datasets with increasing number of free annotations. Number of users, sequences, Object-Interaction (OI) segments and used segments in the results are detailed. Number of verb annotations shows increased level of ambiguity with BEOID allowing free annotations. Results show improved performance for embedding as the number of labelled verbs increases when compared to classification using SVM or KNN. Improved results is particularly noted when using IDT for motion representation and BoW for encoding. *: We sampled 1000 videos randomly from GTEA+.

in that we ensure each node is connected to its top visually similar but semantically distinct node. The undirected graph SVG_u is then converted to a directed graph by replacing each edge by two directed edges.

$$x_i \frown x_j \in SVG_u \Rightarrow \{x_i \rightarrow x_j, x_j \rightarrow x_i\} \in SVG \quad (3)$$

The weights are normalised to define the probability of traversing from video x_i to x_j, $P(x_i \rightarrow x_j)$.

Embedding a test video in SVG: Given a test video, x, we begin by finding the set $\mathcal{R}$ which contains the z closest neighbours in SVG to x based on visual distance. We embed x into SVG by adding directed edges connecting x to nodes in $x_i \in \mathcal{R}$, with normalised weights $P(x \rightarrow x_i)$. We then use the Markov Random Walk (MRW) method from [4] to determine $Class\,(x)$. MRW attempts to traverse the nodes in the directed graph to estimate the probability of $Class(x)$. Given the Markovian assumption and a predefined number of steps t, we calculate the probability distribution of reaching a node x_i as follows

$$P(x_{i+t} \mid x) = \prod_{x_i \in R} \left(P(x \rightarrow x_i) \prod_{j=1}^{t} P(x_{i+j-1} \rightarrow x_{i+j}) \right) \quad (4)$$

We then select $\arg\max_{Class(x)} P(Class(x))$ as the semantic label of x.

Results

The method is tested on three public datasets with increasing ambiguity (or variability) in annotations. Table 1 details the three datasets including a new free annotations for BEOID (Fig 2). We compare our method against linear SVM and K-NN, and evaluate both Improved Dense Trajectories (IDT) and Convolutional Neural Network (CNN) features pre-trained on ImageNet, each encoded with both BagOfWords (BoW) and Fisher-Vectors (FV).

Preliminary results show that as the number of verb annotations increases, the performance of standard classifiers decreases for the various descriptors. Table 1 also shows that motion-based descriptors (IDT) outperform appearance based descriptors (CNN) for the task of action recognition when using verbs as classes (i.e. open vs close) without considering the object being used. SVG outperforms SVM and K-NN when using IDT features and BoW encoding. The percentage of improved performance increases as the number of verb annotations increases. Analysis of the sensitivity of the SVG approach to the number of visual links m, the number of closest neighbors z and the number of steps in MRW t are available.

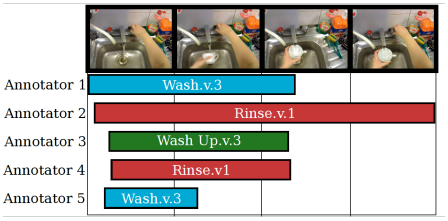

Figure 2. For the same action, five annotators gave different valid annotations using a free choice of verbs.

Conclusion To deal with semantically ambiguous free annotations, we propose to embed egocentric videos within a semantic-visual graph to estimate the probability distribution over possible labellings. Preliminary results on three public datasets, including new free annotations on BEOID will be discussed.

References

[1] J. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, and S. Lacoste-Julien. Unsupervised learning from narrated instruction videos. *CVPR*, 2016.

[2] D. Damen, T. Leelasawassuk, O. Haines, A. Calway, and W. Mayol-Cuevas. You-do, I-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *BMVC*, 2014.

[3] F. De La Torre, J. Hodgins, A. Bargteil, X. Martin, J. Macey, A. Collado, and P. Beltran. Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) database. *Robotics Institute*, 2008.

[4] C. Fang and L. Torresani. Measuring image distances via embedding in a semantic manifold. In *ECCV*. 2012.

[5] A. Fathi, Y. Li, and J. Rehg. Learning to recognize daily actions using gaze. In *ECCV*, 2012.

[6] J. Ghosh, Y. J. Lee, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012.

[7] T. McCandless and K. Grauman. Object-centric spatio-temporal pyramids for egocentric activity recognition. In *BMVC*, 2013.

[8] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012.