Weakly Supervised Learning of Object Segmentations from Web-Scale Video

Glenn Hartmann[1], Matthias Grundmann[2], Judy Hoffman[3], David Tsai[2], Vivek Kwatra[1], Omid Madani[1], Sudheendra Vijayanarasimhan[1], Irfan Essa[2], James Rehg[2], and Rahul Sukthankar[1]

[1] Google Research
[2] Georgia Institute of Technology
[3] University of California, Berkeley

Abstract. We propose to learn pixel-level segmentations of objects from weakly labeled (tagged) internet videos. Specifically, given a large collection of raw YouTube content, along with potentially noisy tags, our goal is to automatically generate spatiotemporal masks for each object, such as "dog", without employing any pre-trained object detectors. We formulate this problem as learning weakly supervised classifiers for a set of independent spatio-temporal segments. The object seeds obtained using segment-level classifiers are further refined using graphcuts to generate high-precision object masks. Our results, obtained by training on a dataset of 20,000 YouTube videos weakly tagged into 15 classes, demonstrate automatic extraction of pixel-level object masks. Evaluated against a ground-truthed subset of 50,000 frames with pixel-level annotations, we confirm that our proposed methods can learn good object masks just by watching YouTube.

1 Introduction

We are motivated by the question: What could a computer learn about the real world solely from watching large quantities of internet video? We believe that internet videos, with their potentially noisy tags, can provide sufficient weak supervision to learn models of visual concepts. Specifically, our goal is to learn models that can perform pixel-level spatiotemporal segmentation of objects (*e.g.*, "dog") when trained only using video-level tags.

To force us to tackle the core challenges, in this paper we adopt an extreme stance characterized by several desiderata. Our models are *tabula rasa* and must learn concept models from large numbers of raw, potentially low-quality internet videos. The only training signals that can be provided to the system must be in the form of video-level tags, which indicate that the concept occurs somewhere within the video. Video tags can be corrupted by some degree of label noise (*e.g.*, some videos labeled "dog" may not contain dogs and there may be videos containing dogs that are missing the "dog" tag). Although the labels are video-level, the evaluation is on a spatiotemporal segmentation task with pixel-level error metrics, such as the precision/recall of pixel masks for a concept, measured

Fig. 1: Video object segmentation: (a) Stabilized frame; (b) Spatiotemporal oversegmentation. (c) Seeds from segment classifier. (d) Spatiotemporal object mask.

on a set of manually annotated ground truth videos. The proposed methods should be capable of scaling, both in the number of training videos and the number of object classes that we recognize.

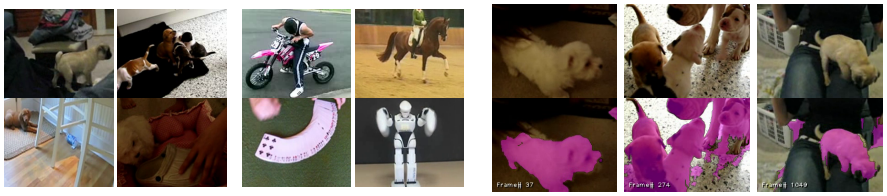
Figure 1 presents an overview of our object segmentation pipeline. Given a video tagged with a label, say "dog", it is first processed to extract spatiotemporal segments. Then segment-level classifiers (trained from raw video using weakly supervised learning) identify segments for given object categories in the video. These detected segments serve as seeds for extracting pixel-level object masks. The spatiotemporal segments ensure that the target concept is localized in both space and time. This mechanism of going from a tagged YouTube video to a pixel mask summarizes our goal of automatically distilling a large corpus of noisily-tagged video into a smaller collection of spatially- and temporally-segmented object instances with associated high-precision labels. Ours is the first work to tackle weakly supervised training of pixel-level object models solely from large quantities of internet video, where the only labels are potentially noisy video-level tags.

2 Related Work

The area of learning visual concepts from weakly supervised video is still in its infancy. Ramanan *et al.* [1] construct a single part-based animal model from video. Ommer *et al.* [2] learn from controlled, hand-recorded video and classify at the frame level. Ali *et al.* [3] build an appearance model from a single video. Leistner *et al.* [4] employ weak video primarily to regularize detectors trained using images. Our work is closest in spirit to recent work by Prest *et al.* [5], which trains on a combination of fully annotated images and manually curated labeled video; the task we address is more extreme as we learn exclusively under weak supervision from raw video with noisy labels.

Our research bears superficial similarity to recent approaches to semi-supervised online learning of object detectors during tracking in video, such as [6]. However, rather than improving the model for a specific tracked object, our goal is to learn broader classes of concepts, without initialization, from raw internet video.

The video segments employed in our work are related to spatiotemporal representations such as Ke *et al.*'s oversegmented videos [7], Niebles *et al.*'s human motion volumes [8] and Brendel & Todorovic' 2D+t tubes [9]. We leverage recent work in video segmentation based on motion, such as Xiao & Shah [10], Brox & Malik [11] and Grundmann *et al.* [12, 13] to generate our representation.



(a) Positive Set (b) Negative Set (c) Inference on "dog" videos

Fig. 2: We learn to locate objects by training only on video-level labels. (a) samples from "dog" videos; (b) samples from background; (c) sample detections.

The weakly supervised learning task bears some similarity to multi-instance learning. In the vision community, related work in this vein includes: Zha *et al.*'s work on multi-label MIL for image classification [14]; Zhou & Zhang's MIML-BOOST and MIML-SVM applied to scene classification; Viola *et al.*'s MILBoost algorithm [15] and Chen *et al.*'s MILES system [16]. However, we focus on high-precision retrieval of instances rather than bag-level classification and are forced to contend with significantly greater label noise *at the bag level*.

Our work contrasts with that of Ren & Gu [17], who employ domain-specific cues (*e.g.*, hands) to segment objects in egocentric video; our methods are most suited for learning models from moving objects in scenes with relatively little background motion. Duchenne *et al.* [18] perform action annotation from weakly labeled data. However, their work is restricted to temporal localization and requires movie scripts that are closely aligned with the scene (and much less noisy than our video-level labels). We differ from existing weakly supervised approaches in video, such as Liu *et al.* [19], which require pixel-level labeling in a sparse set of frames; our work strives to learn object masks without *any* frame-, segment- or pixel-level supervision.

3 Problem Formulation

Our goal can be formalized as the following (see Figure 2). Given a set of object class labels $\mathcal{Y} = \{y_1, y_2, \ldots, y_n\}$ and a large set of weakly tagged videos $\{(v, \mathcal{Y}_v) : \mathcal{Y}_v \subseteq \mathcal{Y}\}$, we seek to learn a model for each concept y_j that can output pixel-level masks for test videos, localized in both space and time. We aim to learn concept models from raw, full-length internet videos containing multiple scenes and several topics. The video-level annotations simply indicate that the given concept y_j occurs somewhere in the video, possibly multiple times, at unspecified spatial and temporal locations. We recognize that each concept can exhibit a diversity of appearances due to intra-class variations (*e.g.*, dog breeds) and that most of the pixels in a video labeled y_j will be unrelated to y_j. While our notion of "concept" is general, our methods are applicable only to semantic labels that correspond to concepts with bounded spatiotemporal extent, such as objects and actions, as opposed to tags that demand scene-level understanding or higher-level domain-specific knowledge. Figure 2 gives a high-level idea of our framework.

To bound our exploration and to enable direct comparisons, in this paper we focus on a restricted set of approaches to our problem. Specifically, all of our proposed methods strictly adhere to the following general strategy:

1. We assume that while label noise can be significant (*e.g.*, 20%), it is independent of the given video v or concept y_j;
2. We learn each concept y_j separately, allowing any given video v to contain multiple concepts $\mathcal{Y}_v \subseteq \mathcal{Y}$;
3. We assume that each video v can be partitioned into a set of spatiotemporal segments $\mathcal{S}_v$, that each segment $s_i \in \mathcal{S}_v$ can be represented by aggregations of a variety of local features, and that each s_i can be independently classified;
4. Rather than directly incorporating the spatiotemporal dependencies between segments in our models (*e.g.*, using a CRF), we account for these in a more computationally scalable object mask refinement phase.

These principles guide us to computationally efficient algorithms that learn from large quantities of video ($> 10^8$ frames) using parallelized implementations. Specifically, our weakly supervised learning operates independently on instances that are spatiotemporal segments, represented using a set of features (bags of quantized features, with responses aggregated over the segment). In other words, the core problem can be formulated as a segment selection task, where the set of selected segments can be converted to pixel-level object masks.

In the following discussion, for a given concept y_j, the term *positive videos* refers to those videos in the labeled set that have the clip-level tag y_j and *negative videos* to those that do not.

4 Training Segment Classifiers and Object Segmentation

We tackle this weakly supervised problem using the two establised approaches, described below. The learning techniques take as input segments that are either positive or negative, that is each segment inherits the binary label of the video it is in. The learned models then score and rank the segments of a given test video. Each segment is described by bags of local features. We also present a training variant based on one-vs-one class comparisons and a post-processing technique that takes as input the individually ranked segments and improves the final object masks by exploiting spatiotemporal consistency

4.1 Discriminative Segment Classifier

The most direct approach to learning under weak supervision is to train a discriminative one vs. rest model for a segment from each concept using all of the available data (labeled segments), which effectively treats the background segments present in each positive video as label noise. The intuition is that since similar background segments are present in both positive and negative data, a linear model (with its limited capacity) should largely ignore such segments and focus more on the desired concept (whose segments are unlikely to appear in the negative videos). The challenge is whether such an approach can work even

if the fraction of segments that relate to the concept is small (*e.g.*, 20% of the total). Thus, the input to the classifier is the set of features for a given segment and the ouput is a single real-valued output indicating the classifier's confidence that this segment is an instance of the concept.

We employ Fan *et al.*'s LIBLINEAR (linear SVM) classifier [20], trained independently (one vs. rest) on each concept using 200,000 positive and 400,000 negative segments, sampled uniformly from concept and background (negative) videos, respectively (sampling enables us to retain the training set in memory).

4.2 Multiple Instance Learning (MIL)

To explore MIL on our task, We adapt the MILBoost algorithm with ISR criterion [15]. We use sparse boosting with decision stumps [21] as the base classifier. All of the instance (segment) weights are updated by multiplying with the corresponding bag weights. Viola *et al.* noted [15] that the ISR criterion can lead to competition among instances in the same bag, but this is a reasonable choice for our problem because: 1) the target concept can occur in only a very small fraction of the pixels in a positive video, and 2) the tags for our videos are themselves noisy. We train using 500,000 positive and 50,000 negative segment instances.

4.3 One-vs-one Training Variant

Many of the segments within the positive videos (tagged by a specific desired concept, the target of learning) belong to concepts that *co-occur* with the desired concept. Such segments, may help detect the desired concept, but they are not *part* of the desired concept. When we take as the negative videos a subset of all videos, these frequently associated concepts tend to be learned, because they are not sufficiently represented in the negative videos. The problem of associated or co-occurring concepts is pervasive to weakly-supervised learning. Focusing the learner on what makes the concept what it is, by showing videos drawn randomly from different distributions corresponding to other concepts should help focus the learner on the desired concept.

We realize this idea by training one-versus-one linear classifiers for each class pair. Let $s_{i,j}(\mathbf{x})$ denote the score that the binary classifier, trained on segments from video tagged by concepts i and j, assigns to class i when applied to a segment with feature vector $\mathbf{x}$ of a segment. Then, when scoring segments from a video tagged by class i, the score of a segment x is defined as the minimum over all classifier scores:

$$s_i(x) = \min_{j \neq i} s_{i,j}(\mathbf{x}).$$

For each class pairing i and j, the segments are taken from videos tagged with i and j only (about 100,000 segments each in our experiments), and raw classifier scores are calibrated (to obtain probabilities) on 20% of such segments. Taking the minimum score is intuitive, as we seek those segments that are least like any other concept. Note that this is slightly different from traditional 1-vs-1 multi-class SVM training, which votes across many pairs of subsets of classes.

4.4 Object Segmentation from Ranked Segments

The segment-level classifiers described above output a set of segments for each video ranked by the likelihood of being instances of the concept.[4] Given such a list of segment "seeds" for a video, our goal is now to refine these into object masks using both appearance and spatial consistency. To construct such a dense labeling, we adopt a graph-cut based segmentation formulation, summarized briefly due to space considerations.

Our formulation employs a unary appearance (color and local texture) potential that is obtained using two Gaussian Mixture Models trained on foreground (pixels in selected seeds) and background (pixels sampled far from seeds). The pairwise term is standard and designed to enforce smoothness. The energy function is efficiently minimized using [22] for each frame in the test video.

5 Evaluation

We present both qualitative and quantitative evaluations of our method on a large corpus of partially groundtruthed internet video. Additional results examining the role of different features, type of video over-segmentation and comparisons with other weakly supervised classifiers are omitted here due to space limitations.

Table 1: Summary of weakly supervised internet video dataset

Concept	Summary	Number
bike	motorbikes and bicycles, often with a rider	1,671
boat	a variety of watercraft including ships, boats and jetskis	1,283
card	playing cards, featured in magic tricks and card games	937
dog	dogs of various breeds, indoors & outdoors	1,336
helicopter	includes both full-size and toy helicopters in outdoors scenes	1,189
horse	typically horses being ridden in equestrian events	1,800
robot	a variety of robots, including toys, research & industrial machines	601
transformer	shape-shifting toys, often occluded by hands manipulating them	1,283
background	(from a variety of other tags detailed in text)	12,207

5.1 Dataset

Our dataset consists of full-length internet videos that are several minutes in length and contain multiple shots. To remain true to our goals, we perform no manual filtering or selection of the content. We have collected 20,000 public videos from YouTube, summarized in Table 1 along with additional background

[4] The common application case for object segmentation is that the given category occurs somewhere within the tagged test video; our method can be applied to untagged test videos by requiring a high-precision threshold on segment-level seeds and dropping videos without insufficient seed segments.

videos from several other tags, such as "stadium", "protest", "flower", "mountain", and "running". Additionally, a set of test videos from different classes has been manually annotated (at the pixel level) to generate a ground truth set of approximately 50,000 frames to generate precision/recall curves.

5.2 Experiments

We process each of the videos in the training set as follows to ensure uniformity. First, we scale each video to a consistent width of 240 pixels, maintaining its original aspect ratio. Next, we perform video stabilization [11, 13] to reduce camera motion that could corrupt motion features and shapes of spatiotemporal segments. We then perform hierarchical spatiotemporal segmentation[5] to identify segments (at multiple scales) that capture contiguous parts of objects and the background. To better understand the role of segmentation, we also repeat our experiments using a tesselation of *cuboids* (spatiotemporal generalization of patches), where each image is divided into 12×9 patches, 10 frames deep.

We represent each segment (and cuboid) using the following features: 1) RGB color histogram, quantized over 20 bins; 2) histogram of local binary patterns computed on 5×5 patches [23, 24]; 3) histogram of dense optical flow [25], with an additional fifth bin for near-zero flow; 4) heatmaps computed over a 8×6 grid to represent the (x, y) shape of each segment, summed over its temporal extent; 5) histogram of quantized SIFT-like local descriptors extracted densely within each segment.

Computational details: It is a challenging task to process videos at such a large scale. We distribute the job of video stabilization, spatiotemporal segmentation and feature extraction for each video to different machines using the MapReduce framework. Using our implementation, we are able to process our 20,000 videos using a cluster of 5000 nodes in less than 30 hours.

For the liblinear classifier, we present results based on a few regularization (C) values (and compare the classifiers using the same C). For MIL, we set the regularization term of sparse boost to 1.0 and used 1000 decision stumps.

Figure 3 presents pixel-level precision/recall curves[6] (overall and per class) for segment-level classification. Surprisingly, the choice of segment-level classifier is not critical. In particular, posing this problem in a multiple-instance learning (MIL) framework does not result in clear gains. Using the one-vs-one variant generates significant improvement over one-vs-rest, and in paired (per groundtruth video) tests, comparing precision at each of 5%, 10%, and 20% recall levels, we observe 17 or more wins vs. 9 or fewer losses.

The use of video segmentation also dominates cuboids in a similar fashion, both on average and in paired comparisons.

Weakly supervised learning of some visual concepts at the individual segment level is easier than others. For instance, "bike", "dog", "robot" and "transformers" seem to have sufficiently distinctive features that they can be separated from

[5] Using the web-based segmentation service at http://videosegmentation.com [12].

[6] Precision is the fraction of correctly classified pixels to classified pixels; recall is the fraction of correctly classified to groundtruth concept pixels.

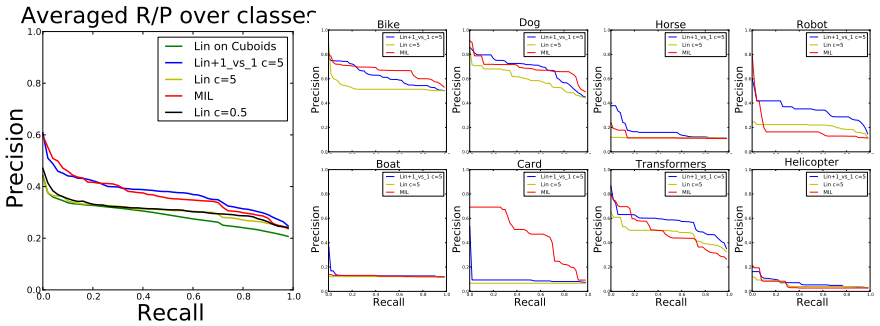


Fig. 3: Averaged & per-class results. Random pixel baseline precision: 16%.

the background class, at the segment level, under weak supervision. Among the difficult classes, spatiotemporal segmentation often undersegments helicopters. "Horse", "boat" and "card" seem to be difficult because of the problem of associated segments (e.g., water in the case of boat).

Figure 4 shows examples of object masks (magenta overlay) for each of the eight classes, as well as precision-recall curves for the corresponding video. These results include object mask refinement (dashed lines in P-R curves) as well as the raw P-R curves using individual segment classifiers. We see that, the object masks localize objects from different classes, even under challenging conditions: dog at the beach, complicated close-up of motorbike, etc. Object mask refinement works best for high-precision individual segment results. We note that rare objects, such as the beach ball in the horse video, are occasionally highlighted (false positives). Additional failure cases are shown in Figure 4 (last row).

6 Conclusion

This paper proposes the idea of learning spatiotemporal object models, with minimal supervision, from large quantities of weakly and noisily tagged video. Since we are the first to tackle this problem, particularly at large scale, we conduct an evaluation of several computationally scalable approaches to weakly supervised learning. We believe that weakly supervised learning from internet video has the potential to radically transform object and action recognition. This paper is just the first step towards that goal.

In future work, we plan to explore several directions. First, our current framework implicitly uses segment-level loss whereas the evaluation is at the pixel level; directly optimizing the latter is worth exploring. Second, we plan to investigate how our approach scales to thousands of concepts. Finally, we plan to use our object segmentation masks as strongly supervised training data for training traditional object detectors in both image and video domains.

Acknowledgments We thank C. Cortes, S. Kumar, K. Murphy, M. Ruzon, E. Sargin, G. Toderici, J. Weston, and J. Yagnik for many helpful discussions.

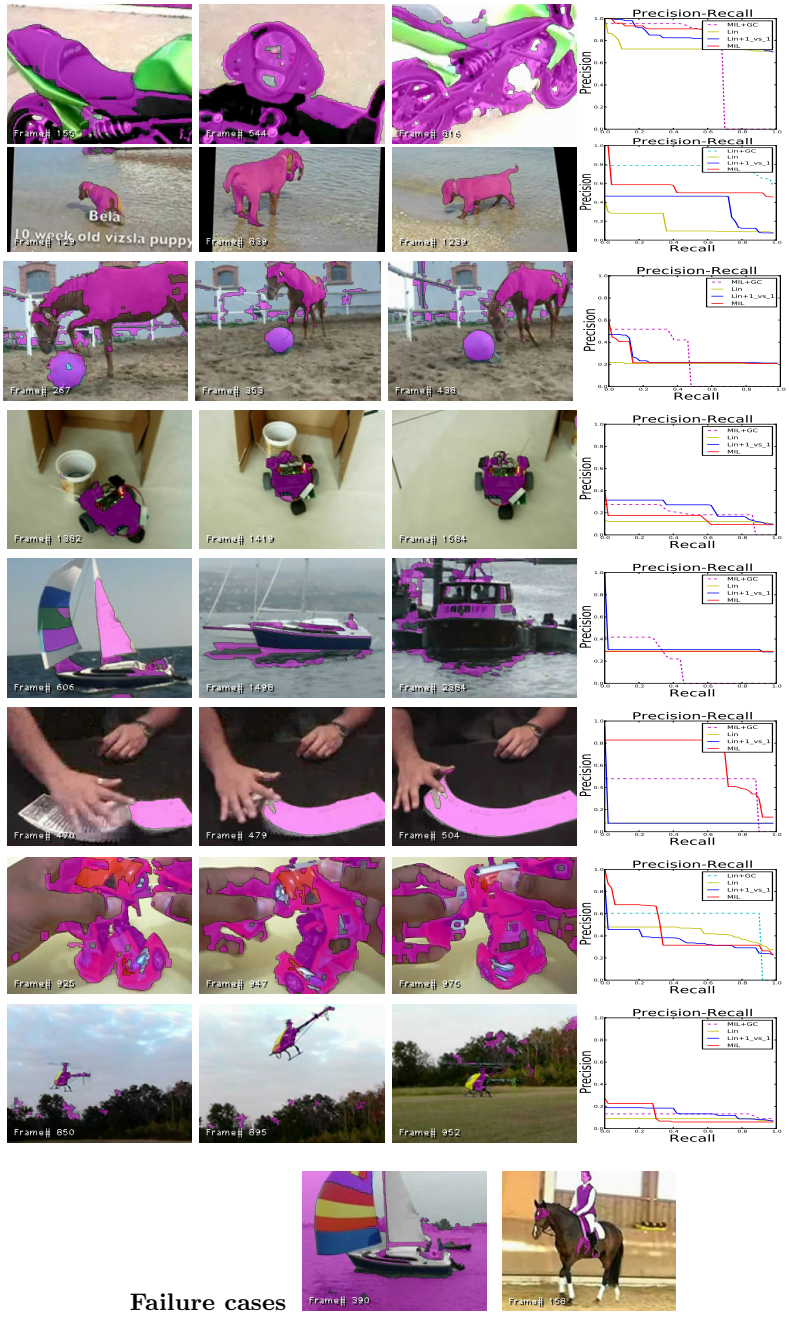


Fig. 4: Sample object segmentation results for each class and some failure cases.

References

1. Ramanan, D., Forsyth, D., Barnard, K.: Building models of animals from video. PAMI **28** (2006)
2. Ommer, B., Mader, T., Buhmann, J.: Seeing the objects behind the dots: Recognition in videos from a moving camera. IJCV **83** (2009)
3. Ali, K., Hasler, D., Fleuret, F.: FlowBoost—Appearance learning from sparsely annotated video. In: CVPR. (2011)
4. Leistner, C., Godec, M., Schulter, S., Saffari, A., Werlberger, M., Bischof, H.: Improving classifiers with unlabeled weakly-related videos. In: CVPR. (2011)
5. Prest, A., Leistner, C., Civera, J., Schmid, C., Ferrari, V.: Learning object class detectors from weakly annotated video. In: CVPR. (2012)
6. Kalal, Z., Matas, J., Mikolajczyk, K.: P-N Learning: Bootstrapping binary classifiers by structural constraints. In: CVPR. (2010)
7. Ke, Y., Sukthankar, R., Hebert, M.: Event detection in crowded videos. In: ICCV. (2007)
8. Niebles, J.C., Han, B., Ferencz, A., Fei-Fei, L.: Extracting moving people from internet videos. In: ECCV. (2008)
9. Brendel, W., Todorovic, S.: Learning spatiotemporal graphs of human activities. In: ICCV. (2011)
10. Xiao, J., Shah, M.: Motion layer extraction in the presence of occlusion using graph cuts. PAMI **27** (2005) 1644–1659
11. Brox, T., Malik, J.: Object segmentation by long term analysis of point trajectories. In: ECCV. (2010)
12. Grundmann, M., Kwatra, V., Han, M., Essa, I.: Efficient hierarchical graph-based video segmentation. In: CVPR. (2011)
13. Grundmann, M., Kwatra, V., Essa, I.: Auto-directed video stabilization with robust L1 optimal camera paths. In: CVPR. (2011)
14. Zha, Z.J., Hua, X.S., Mei, T., Wang, J., Qi, G.J., Wang, Z.: Joint multi-label multi-instance learning for image classification. In: CVPR. (2008)
15. Viola, P., Platt, J., Zhang, C.: Multiple instance boosting for object detection. In: NIPS. (2005)
16. Chen, Y., Bi, J., Wang, J.: MILES: Multiple-instance learning via embedded instance selection. PAMI **28** (2006) 1931–1947
17. Ren, X., Gu, C.: Figure-ground segmentation improves handled object recognition in egocentric video. In: CVPR. (2010)
18. Duchenne, O., Laptev, I., Sivic, J., Bach, F., Ponce, J.: Automatic annotation of human actions in video. In: ICCV. (2009)
19. Liu, D., Hua, G., Chen, T.: A hierarchical visual model for video object summarization. PAMI **32** (2010) 2178–2190
20. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. JMLR **9** (2008) 1871–1874
21. Duchi, J., Singer, Y.: Boosting with structural sparsity. In: ICML. (2009)
22. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. PAMI **23** (2001) 1222–1239
23. Ojala, T., et al.: Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In: ICPR. (1994)
24. Wang, X., Han, T.: An HOG-LBP human detector with partial occlusion handling. In: ICCV. (2009)
25. Chaudhry, R., et al.: Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems. In: CVPR. (2009)