Lending A Hand: Detecting Hands and Recognizing Activities in Complex Egocentric Interactions

Sven Bambach Stefan Lee David J. Crandall
School of Informatics and Computing
Indiana University
{sbambach,steflee,djcran}@indiana.edu

Chen Yu
Psychological and Brain Sciences
Indiana University
chenyu@indiana.edu

Abstract

Hands appear very often in egocentric video, and their appearance and pose give important cues about what people are doing and what they are paying attention to. But existing work in hand detection has made strong assumptions that work well in only simple scenarios, such as with limited interaction with other people or in lab settings. We develop methods to locate and distinguish between hands in egocentric video using strong appearance models with Convolutional Neural Networks, and introduce a simple candidate region generation approach that outperforms existing techniques at a fraction of the computational cost. We show how these high-quality bounding boxes can be used to create accurate pixelwise hand regions, and as an application, we investigate the extent to which hand segmentation alone can distinguish between different activities. We evaluate these techniques on a new dataset of 48 first-person videos of people interacting in realistic environments, with pixel-level ground truth for over 15,000 hand instances.

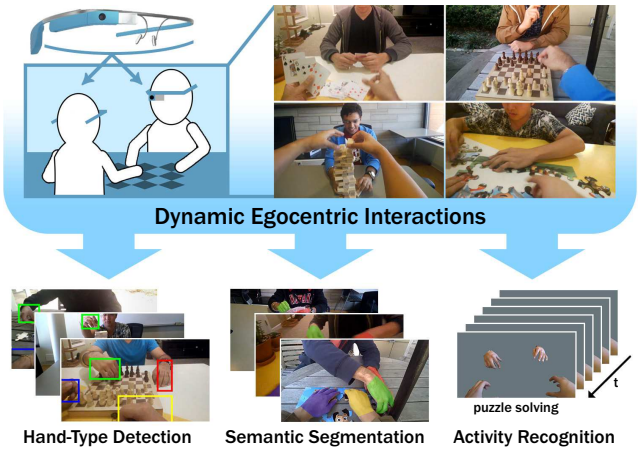

Figure 1: We present a CNN-based technique for detecting, identifying, and segmenting hands in egocentric videos of multiple people interacting with each other. To illustrate one specific application, we show that hand segments alone can be used for accurate activity recognition.

1. Introduction

Wearable cameras are starting to catch on, with devices like Google Glass, GoPro Hero, Narrative Clip, and others hitting the consumer market in the last few years. These products are being used to capture the adventures of sports enthusiasts [36], to help people suffering from memory loss by creating visual logs of their day [7], to enhance public safety when worn by police officers [33], to collect data on human behavior for scientific studies [4], or just for fun. These devices record huge volumes of images and video, so people will need automatic techniques to help browse, search, and visualize the egocentric imagery they collect. We need computer vision techniques that can handle the challenges of first-person video, including highly dynamic camera motion and poor imaging conditions.

While egocentric video captures a huge variety of objects, activities, and situations, one specific object is omnipresent in nearly every frame: the hands. We use our hands as our main channel of interaction with the physical world, for manipulating objects, sensing the environment, and expressing ourselves to other people. Our hands are almost always in our field of view, and their pose and configuration reflects what we are doing and what we intend to do next. In addition, understanding the activities and intentions of our social partners requires that we also detect their hands, disambiguating them from our own. This means that hand detection and tracking are fundamental problems of egocentric vision, both for computers and people; in fact, neuroscientists have discovered specific parts of the brain that respond to identifying our own hands (since "feeling of ownership of our limbs is a fundamental aspect of self-consciousness" [8]). We believe that almost any egocentric computer vision problem, from object detection to activity recognition, will thus require accurate hand detection.

Recent work in egocentric computer vision on recog-

nizing manipulated objects and activities has incorporated hand pose either explicitly or implicitly [11, 26, 28], while other work has specifically studied hand detection and segmentation [6, 20, 21]. However, these pioneering papers make assumptions that may limit their applicability in practice. Most work assumes that no other people appear in the videos, so that all hands belong to the camera owner. For instance, the hand segmentation of Li *et al.* [20, 21] is based on detecting skin pixels, not hands *per se*, so it would also find faces or any other bare skin regions. But real-world egocentric video is full of interactions with other people, so multiple hands must be found and disambiguated from one another. Lee *et al.* [18] allow for interacting people, but only in a highly-constrained lab setting where hands can be identified by their position within the egocentric frame.

In this paper we investigate hand detection, disambiguation, and segmentation in first-person videos of interacting people in realistic settings. Instead of making strong assumptions such as that all skin pixels correspond to hands or that the geometry of a scene is known ahead of time, we instead detect and disambiguate hands using strong appearance models using Convolutional Neural Networks. To make this efficient, we present a lightweight hand candidate generation approach based on sampling from a proposal distribution, and show that it produces better coverage than existing approaches like selective search [35] at a fraction of the computational cost. We then use these high-quality hand detections to perform pixel-level segmentation, outperforming existing first-person hand segmentation approaches. Finally, we test our hypothesis that hand configuration and pose give powerful evidence about camera wearer activity, showing that just these features alone yield impressive activity recognition performance. To make these experiments possible, we introduce a new dataset of 48 videos featuring different participants interacting in a variety of activities and environments. The dataset includes high-quality ground truth segmentation masks for over 15,000 hands.

In summary, our contributions in this paper include:

1. deep models for hand detection and classification in egocentric video, including fast domain-specific region proposals to dramatically cut computational cost;
2. a new technique for pixelwise hand segmentation;
3. a quantitative analysis of the power of hand location and pose in recognizing activities; and
4. a large dataset of egocentric interactions with fine-grained ground truth, which we have publicly released.

2. Related Work

Egocentric vision is becoming a popular research topic in computer vision, with recent papers dedicated to summarizing and describing events in life-logging photo data [19, 23, 32], recognizing handled objects [13, 27], and identifying activities [11, 12, 26, 30]. Most of this work is based on ideas from more traditional object and activity recognition, but with innovations to address the particular challenges of first-person imagery: highly dynamic and unpredictable camera motion, unusual composition and viewpoints, and noise from motion blur and poor illumination conditions.

Ren and Gu [27] were among the first to study hand detection in first-person video, specifically in the context of held object detection. They treat the problem as figure-ground segmentation, identifying regions with irregular optical flow patterns that may correspond to hands and held objects, versus regions with coherent flow in the background. Fathi *et al.* [13] extended this work to incorporate color features to segment hands from objects. The assumption in these papers is that the background is static so that optical flow can be used for segmentation. This assumption is often violated in real-world egocentric video, where interactions with other people create dynamic background environments and include hands other than the camera owner's.

Some very recent work has focused specifically on egocentric hand segmentation. Li and Kitani [20, 21] explicitly addressed illumination variation, proposing a model recommendation approach that picks the best local color feature model for each environment using scene-level feature probes. Their approach also assumes that there are no social interactions, so that the only hands in the video belong to the camera owner, and does not attempt to distinguish hands on a semantic level. They define a "hand" to include contiguous skin regions up to the sleeves, so they are really studying skin segmentation as opposed to trying to cleanly segment only hands. This is an important distinction from our work; we argue that in applications like hand pose classification or activity recognition, segmentations that are invariant to type of clothing are important.

Most relevant to our work is Lee *et al.* [18], which is the only paper to our knowledge that attempts to model hands in social interactions in first-person video. Their model encodes spatial arrangements to disambiguate hand types using a probablistic graphical model. However, they use a simplistic appearance model and evaluate their method primarily on highly constrained lab videos, and do not consider hand segmentation or activity recognition as we do here.

Although not directly related, other work has dealt specifically with hands in different domains. Mittal *et al.* [24] developed a system that uses deformable part models and skin heuristics to detect hands in the PASCAL VOC Person Layout challenge [10]. In contrast to egocentric imagery, these images are of people from a distance, and only those with visible heads are annotated in the ground truth. The system exploits these constraints by, for example, limiting the size of hand detections and using head detections to learn per-frame skin models. Another well-studied problem is hand pose estimation in 3D vision [9, 31], with some work starting to consider first-person data. For instance, Lin

Figure 2: *Visualizations of our dataset and ground truth annotations. Left:* Ground truth hand segmentation masks superimposed on sample frames from the dataset, where colors indicate the different hand types. *Right:* A random subset of cropped hands according to ground truth segmentations (resized to square aspect ratios for ease of visualization).

et al. [22] show that the 3D shape of a grasping hand can help improve recognition of the objects being grasped. We also study pose in the context of interacting with (and thus grasping) objects, but we do not require depth information.

3. EgoHands: A large egocentric hand dataset

We begin by presenting a new dataset of first-person video of pairs of interacting people, each with synchronized video from head-mounted cameras. Other first-person video datasets that have been proposed [13, 21, 26, 28] are designed to test recognition of activities or handled objects, with minimal interaction with other people. In contrast, our videos include more realistic and challenging social situations where multiple sets of hands appear in the view.

To create as realistic a dataset as possible while still giving some experimental control, we collected data from different pairs of four participants who sat facing each other while engaged in different activities. We chose four activities that encourage interaction and hand motion: (1) playing cards, specifically a simple version of Mau Mau [2]; (2) playing chess, where for efficiency we encouraged participants to focus on speed rather than strategy; (3) solving a 24- or 48-piece jigsaw puzzle; and (4) playing Jenga [1], which involves removing pieces from a 3d puzzle until it collapses. Sample frames for each activity are shown in Figure 1. We also varied context by collecting videos in three different locations: a table in a conference room, a patio table in an outdoor courtyard, and a coffee table in a home. We recorded over multiple days and did not restrict participant clothing so there is significant variety (e.g. both short- and long-sleeved shirts, etc.). We systematically collected data from four actors performing all four activities at all three locations while randomly assigning participants to one another for interaction, resulting in $4 \times 4 \times 3 = 48$ unique combinations of videos. Each participant wore a Google Glass, which recorded 720×1280 video at 30 Hz.

In post-processing, we synchronized the video pairs to one another and cut them to be exactly 90 seconds (2,700 frames) each. For ground truth, we manually annotated a random subset of 100 frames from each video (about one frame per second) with pixel-level hand masks. Each hand pixel was given one of four labels: the camera wearer's left or right hand ("own left" or "own right"), or the social partner's left or right hand ("other left" or "other right"). The ground truth was created by six students who were told to label any hand pixels they could see, including very small hand regions caused by occlusion with objects or truncation at frame boundaries. Importantly, we defined the "hand" to stop at the wrist, in contrast to other work [20, 21] which has also included arms up to the participant's sleeves. We believe our definition is more useful and realistic in practice: if the goal is to detect hand pose and activities, for instance, the definition of what is a hand should not change dramatically depending on what a participant is wearing.

In total, our dataset contains around 130,000 frames of video, of which 4,800 frames have pixel-level ground truth consisting of 15,053 hands. The partner's hands appear in the vast majority of frames (95.2% and 94.0% for left and right, respectively), while the wearer's hands are seen less often (53.3% and 71.1% for left and right). This is likely because one's own hands are more frequently outside the camera's field of view, but right hands occur more often because people tend to align their attention with their dominant hand (and all our participants were right-handed). Figure 2 shows sample frames with ground truth.

To our knowledge, this is the largest dataset of hands in egocentric video, and we have released it on the web[1] with ground truth accessible through a Matlab API that we provide. We randomly partitioned the set of videos into training, validation, and test groups, such that actors, activities and locations are evenly distributed across partitions. This partitioning with 36 training, 4 validation, and 8 test videos is our "main split" that we use for most of our experiments.

[1] http://vision.soic.indiana.edu/egohands/

4. Hand Detection

In principle, finding hands in first-person video frames is simply an instantiation of one particular object detection task, for which we could apply any general object detection algorithm. But in practice, detecting hands requires some special considerations. Hands are highly flexible objects whose appearance and position can vary dramatically, but nonetheless we need models that are strong enough to discriminate between hand types (i.e., left vs. right hands, and the camera wearer's own hands vs. their social partner's).

Convolution Neural Networks (CNNs) offer very good performance for classification tasks [17]. For object detection, the now-standard approach is to divide an image into candidate windows, rescale each window to a fixed size, fine-tune a CNN for window classification [14,34], and then perform non-maximum suppression to combine the output of the region-level classifier into object detection results. Of course, the space of possible proposal windows is enormous, so it is important to propose regions that capture as many objects as possible in the fewest number of proposals.

In the context of detecting hands in egocentric views, there are strong spatial biases to hand location and size [5, 18], because of the way people coordinate head and hand movements: people are likely to center their active hand in or near their visual field as they perform a task, for example. We thus propose a simple approach to candidate window sampling that combines spatial biases and appearance models in a unified probabilistic framework.

4.1. Generating Proposals Efficiently

Our primary motivation is to model the probability that an object O appears in a region R of image I,

$$P(O|R,I) \propto P(I|R,O)P(R|O)P(O)$$

where $P(O)$ is the object occurrence probability of the object, $P(R|O)$ is the prior distribution over the size, shape, and position of regions containing O, and $P(I|R,O)$ is an appearance model evaluated at R for O. Given a parameterization that allows for sampling, high quality regions can then be drawn from this distribution directly.

Here we assume regions are rectangular, so they are parameterized by an image coordinate and width and height. For each of the four types of hands, we can then estimate $P(O)$ directly from the training data, and for $P(R|O)$ we fit a four-dimensional Gaussian kernel density estimator [15] again using the ground truth. For the appearance model $P(I|R,O)$ we define a simple model that estimates the probability that the central pixel of R is skin, based on a non-parametric modeling of skin color in YUV color space (disregarding the luminance channel). While simple, this model lets us sample very efficiently, by drawing a hand type O, and then sampling a bounding box from the KDE of $P(R|O)$, with the kernel weights adjusted by $P(I|R,O)$.

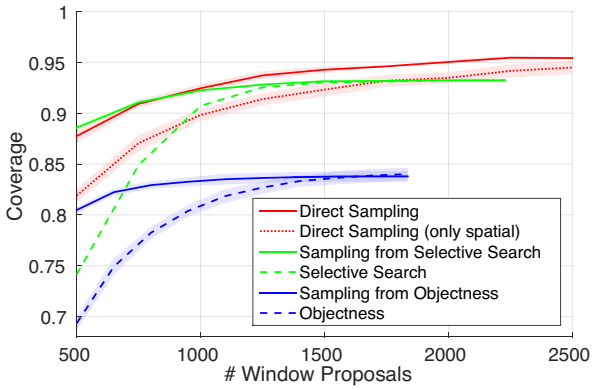

Figure 3: *Hand coverage versus number of proposals per frame,* for various proposal methods. The mean and standard deviation (shaded) across five trials are shown.

To evaluate this candidate generation technique, we measured its *coverage* — the percentage of ground truth objects that have a high enough overlap (intersection over union) with the proposed windows to be counted as positive during detection. This is an important measure because it is an upper-bound on recall. Figure 3 shows coverage as a function of the number of proposed windows per frame for our method and two other popular window proposal methods: selective search [35] (which is the basis of the popular R-CNN detector [14]) and objectness [3]. The baselines were run using those authors' code, with parameters tuned for best results (for selective search, we used the "fast" settings given by the authors but with k set to 50; for objectness, we retrained the object-specific weights on our dataset). As shown in the figure, our direct sampling technique (red solid line) significantly outperforms either baseline (dashed green and blue lines) at the same number of candidates. Surprisingly, even our direct sampling without the appearance model (red dotted line) performed significantly better than objectness and about the same as selective search.

To further investigate the strength of the spatial consistencies of egocentric interaction, we also subsampled the baseline proposals biased by our learned model $P(O|R,I)$. For both baselines, incorporating our learned distribution improved results significantly (solid blue and green lines), to the extent that biased sampling from selective search performs as well as our direct sampling for lower numbers of proposals. However, our full technique offers a dramatic speedup, producing 1500 windows per frame in just 0.078 seconds versus 4.38 and 7.22 seconds for selective search and objectness. All coverage experiments were performed on a machine with a 2.50GHz Intel Xeon processor.

4.2. Window Classification using CNNs

Given our accurate, efficient window proposal technique, we can now use a standard CNN classification framework to classify each proposal (after resizing to the fixed-sized in-

put of the CNN). We used CaffeNet from the Caffe software package [16] which is a slightly modified form of AlexNet [17]. We also experimented with other network designs such as GoogLeNet [34], but found that when combined with our window proposal method, detection results were practically identical.

We found that certain adjustments to the default Caffe training procedure were important both to convergence and the performance of our networks. Only 3% of our proposed windows are positive so to avoid converging to the trivial majority classifier, we construct each training batch to contain an equal number of samples from each class. Also, we disabled Caffe's feature that augments the training data with horizontally and vertically flipped versions of exemplar images, since this reduced the classifier's ability to differentiate between left and right hands, for example.

The full detection pipeline consists of generating spatially sampled window proposals, classifying the window crops with the fine-tuned CNN, and performing per-class non-maximum suppression for each test frame. Each of these components has a number of free parameters that must be learned. For our window proposal method, we estimate the spatial and appearance distributions from ground truth annotations in the training set and sample 2,500 windows per frame to provide a high coverage. The CNN weights are initialized from CaffeNet excluding the final fully-connected layer which is set using a zero-mean Gaussian. We then fine-tune the network using stochastic gradient descent with a learning rate of 0.001 and momentum of 0.999. The network was trained until the validation set error converged. The overlap thresholds for non-max suppression were optimized for each class based on average precision on the validation set. To keep our technique as general as possible, we do not take advantage of the constraint that each hand type should appear at most once in a given frame, although this is an interesting direction for future work.

4.3. Detection Results

We evaluate the effectiveness of our detection pipeline in two contexts: detecting hands of any type, and then detecting hands of specific types ("own left", "own right", etc.). In both cases, we use the PASCAL VOC criteria for scoring detections (that the intersection over union between the ground truth bounding box and detected bounding box is greater than 0.5). Figure 4 shows precision-recall curves for both tasks, applied to the "main split" discussed in Section 3. For the general hand detection task (left), we obtain an average precision (AP) of 0.807 using our candidate window sampling approach, which is significantly higher than the 0.763 for selective search and 0.568 for objectness.

The right pane of Figure 4 shows Precision-Recall curves for distinguishing between the four hand types. For comparison, we also plot the performance of Lee *et al.* [18] on our

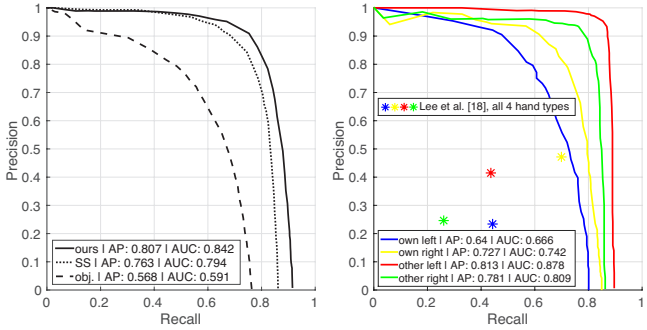

Figure 4: *Precision-Recall curves for detecting hands. Left:* General hand detection results with other window-proposal methods as baselines. *Right:* Results for detecting four different hand types compared with Lee *et al.* [18].

dataset. Direct comparison is not possible as their technique only estimates hand centroid positions, not bounding boxes; we make the comparison as favorable to them as possible by scoring a centroid as correct if it falls anywhere in the ground truth bounding box. They also generate only a single MAP estimate per frame, so performance is a P-R point instead of a curve. Our method outperforms significantly, likely due to our much stronger appearance models.

There is a curious asymmetry in our hand type detections, with our approach achieving significantly better results for the social partner's hands versus the camera owner's. Figure 5 gives insight on why this may be, presenting detection results from randomly-chosen frames of the test set. Hands of the camera wearer tend to have many more duplicate detections on subparts of the hands (e.g. in row 2, column 2 of the figure). We attribute this tendency to how frequently "own" hands are truncated by the frame boundaries and thus appear as single or only a few fingers in the dataset. Including these partial detections alongside fully visible hands during training encourages the network to model both appearances to minimize error. While this does result in a loss of precision, the system gains the ability to robustly detect hands that are occluded or only partially in the frame (e.g. row 3, column 3) which is often the case for egocentric video, due to the relatively narrow field of view of most cameras compared to that of humans.

Error analysis. A related question is whether the errors are primarily caused by failure to detect hands of different types or confusion between hand types once a hand is detected. An analysis of the per-window classifications showed that only 2% of hand windows are mislabelled as other hands. Similarly for detection, 99% of undetected hands at a recall of 70% are due to confusion with the background class. Generally, our predictions tend to be nearly uniform for windows with ambiguous hand types, which are then removed by reasonable decision thresholds and non-max suppression. The qualitative results in Figure 5 also suggest that there is little confusion between different hand types.

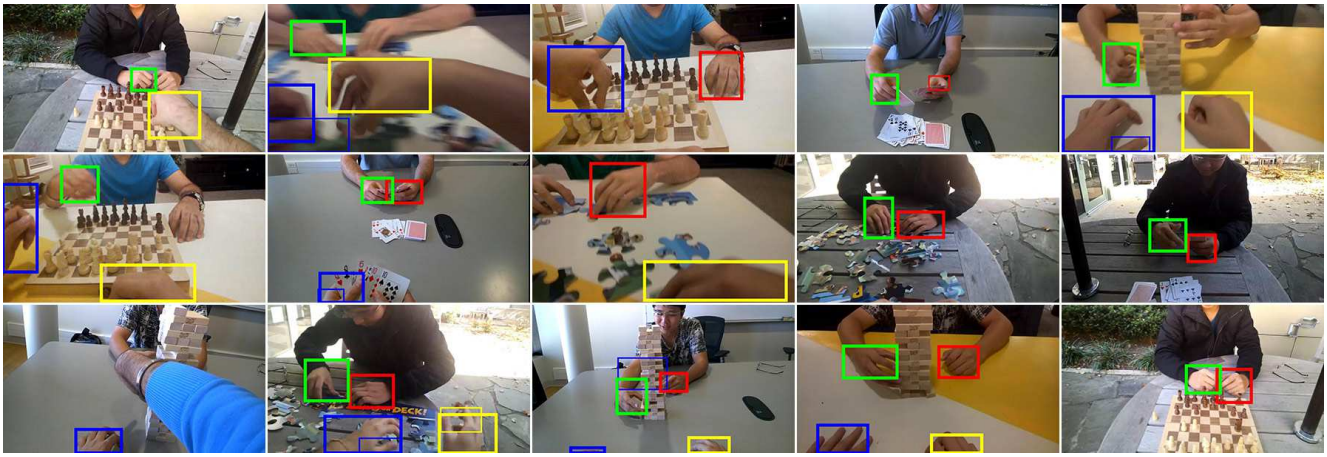

Figure 5: *Randomly-chosen frames with hand detection results,* for **own left**, **own right**, **other left**, and **other right** hands, at a detection threshold where recall was 0.7. Thick and thin rectangles denote true and false positives, respectively.

Generalizing across actors, activities, and locations. We next tested how well our classifiers generalize across different activities, different people, and different locations. To do this, we generated three different types of partitionings of our dataset across each dimension, where each split leaves out all videos containing a specific (first-person) actor, activity, or location during training, and tests only on the held-out videos. We also split on actor pairs and activities jointly, creating 18 divisions (as not all pairs did all activities). This stricter task requires the method to detect hands of people it has never seen doing activities it has never seen.

Table 1 summarizes our results, again in terms of average precision (AP), with averages across splits weighted by the number of hand instances. The table shows that the detector generalizes robustly across actors, with APs in a tight range from 0.790 to 0.826 no matter which actor was held out. This suggests that our classifier may have learned general characteristics of human hands instead of specific properties of our particular participants, although our sample size of four people is small and includes limited diversity (representing three different ethnicities but all were male). For locations, the courtyard and office environments were robust, but AP dropped to 0.648 when testing on the home data. A possible explanation is that the viewpoint of participants in this location is significantly different, because they were seated on the floor around a low table instead of sitting in chairs. For activities, three of the four (cards, puzzle, and chess) show about the same precision when held out, but Jenga had significantly lower AP (0.665). The Jenga videos contain frequent partial occlusions, and the tower itself is prone to be mistaken for hands that it occludes (e.g. row 3, column 3 of Figure 5). Finally, splitting across actor pairs and activities results in a sharper decrease in AP, although they are still quite reasonable given the much smaller (about 6x) training sets caused by this strict partitioning of the data.

	All hands	Own hands		Other hands	
		Left	Right	Left	Right
Main split	0.807	0.640	0.727	0.813	0.781
All activities but:					
cards	0.768	0.606	0.776	0.708	0.732
chess	0.851	0.712	0.788	0.821	0.808
Jenga	0.665	0.644	0.693	0.583	0.502
puzzle	0.803	0.747	0.813	0.675	0.681
weighted average	0.772	0.675	0.768	0.699	0.686
All actors but:					
B	0.799	0.669	0.773	0.779	0.796
H	0.816	0.718	0.772	0.756	0.740
S	0.790	0.709	0.798	0.799	0.696
T	0.826	0.689	0.783	0.770	0.789
weighted average	0.807	0.700	0.782	0.776	0.756
All locations but:					
courtyard	0.790	0.702	0.785	0.755	0.755
office	0.772	0.659	0.757	0.794	0.687
home	0.648	0.558	0.703	0.538	0.591
weighted average	0.737	0.639	0.748	0.698	0.678
Split across actor pairs and activities					
weighted average	0.627	0.492	0.598	0.513	0.542

Table 1: *Hand detection accuracy when holding out individual activities, participants, and locations,* in terms of average precision. For example, the training set for *all activities but cards* included all videos *not* containing card playing, while the test set consisted *only* of card playing videos.

5. Segmenting Hands

While simply detecting hands may be sufficient for some applications, pixelwise segmentation is often more useful, especially for applications like hand pose recognition and in-hand object detection [22]. Once we have accurately localized hands using the above approach, segmentation is relatively straightforward, as we show in this section. We use our detector both to focus segmentation on local image

regions, and to provide semantic labels for the segments.

Our goal in this section is to label each pixel as belonging either to the background or to a specific hand class. We assume most pixels inside a box produced by our detector correspond with a hand, albeit with a significant number of background pixels caused both by detector error and because hands rarely fill a bounding rectangle. This assumption allows us to apply a well-known semi-supervised segmentation algorithm, GrabCut [29], to our problem. Given an approximate foreground mask, GrabCut improves the segmentation by iteratively refining appearance models of the foreground and background pixels, and relabeling foreground and background using a Markov Random Field.

In more detail, for each detected hand bounding box, we use the simple color skin model described in Section 4.1 to estimate an initial foreground mask. We use an aggressive threshold so that all pixels within the box are marked foreground except those having very low probability of being skin. Note that we avoid running GrabCut on the entire image because arms, faces, and other hands would confuse the background color model. Instead, we use a padded region around the bounding box, ensuring that only local background content is modeled. We take the union of the output masks for all detected boxes as the final segmentation.

Segmentation results. Using the skin color model learned for the training set, we detected hands and produced segmentations for each frame in our test set. To put our results in context, we ran the publicly-available pixelwise hand detector of Li et al. [21], which was designed for first person data. We trained their technique with 900 randomly-sampled frames from our training set. As we mentioned before, that paper defines "hand" to include any skin regions connected to a hand, including the entire arm if it is exposed. To enable a direct comparison to our more literal definition of hand detection, we took the intersection between its output and our padded bounding boxes.

Table 2 presents segmentation accuracy, in terms of pixelwise intersection over union between the estimated segmentation mask and the ground truth annotations. Our technique achieves significantly better accuracy than the baseline of [21] (0.556 versus 0.478). A similar trend is present across the stricter actor pair and activity data splits. Figure 6 shows our segmentations on some randomly-sampled test frames. Examining the differences between our approach and the baseline lends some insight. Our GrabCut-based approach looks only at local image color distributions and leans heavily on the quality of our detections. The baseline method, however, learns classifiers that must perform well across an entire frame which is complicated by the close visual similarity between hands and other visible skin.

Failure modes. Our method has two main possible failure modes: failure to properly detect hand bounding boxes, and inaccuracy in distinguishing hand pixels from background within the boxes. To analyze the influence of each, we perform an ablation study based on the ground truth annotations. Applying our segmentation approach to the ground truth detection boxes instead of the output of the hand classifier, our results rose from 0.556 to 0.73. On the other hand, taking the output of our hand detector but using the ground truth segmentation masks (by taking the intersection with the detected boxes) achieved 0.76. Each of the studies improve over our fully automatic approach by roughly 30-35%, indicating that neither detection nor segmentation is individually to blame for the decrease in accuracy, and that there is room for future work to improve upon both.

	Own hands		Other hands		
	Left	Right	Left	Right	Average
Main split					
Ours	0.515	0.579	0.560	0.569	0.556
Li et al. [21]	0.395	0.478	0.534	0.505	0.478
Split across actor pairs and activities					
Ours	0.357	0.477	0.367	0.398	0.400
Li et al.	0.243	0.420	0.361	0.387	0.353

Table 2: *Hand segmentation results,* in terms of intersection over union with ground truth.

Figure 6: *Hand segmentation results on random frames*, zoomed into areas containing hands.

6. Hand-based Activity Recognition

We now investigate one particular application of hand detection and segmentation in first-person video: activity recognition. Interacting with different objects affords different types of hand grasps, the taxonomies of which have been throughly studied [25]. Moreover, when multiple actors are interacting, it seems likely that the absolute and relative position of hands within in the field of view also reveals evidence about the activity that the actors are performing. An interesting question is whether activities can be detected based on hand pose information alone, without using any information about the appearance or identity of handled objects or the rest of the scene. Aside from aca-

demic interest, focusing on hands independently of scene could be valuable in recognition systems: while it may be impossible to model or anticipate every single handled object or visual environment, we have shown that it is very possible to accurately detect and segment hands. To what extent could hand pose alone solve activity recognition in first-person views?

To address this question, we fine-tuned another CNN to classify whole frames as one of our four different activities. To prevent the classifier from seeing any information other than hands, we used the ground truth segmentation to mask out all non-hand background (see bottom right of Figure 1). The network saw 900 frames per activity across 36 videos during training and 100 per activity across four videos for validation. The classifier achieved 66.4% per-frame classification accuracy, or roughly 2.7 times random chance, on our test dataset with non-hand regions blacked out. While these results are not perfect, they do confirm a strong connection between activities and hand location and pose.

To evaluate how well the technique would work in an automated system, we reran the above experiment using the output of our segmentation instead of the ground truth for the test set. The per-frame activity classification accuracy falls from 66.4% to 50.9%, but this is still roughly twice random chance. This decline is caused by two types of errors, of course: incorrect information about the spatial configuration of the hands due to imperfect detection, and incorrect hand pose information due to imperfect segmentation. We once again investigated the relative effect of these errors, similar to the ablation study described in Section 5, and found that replacing either detection or segmentation with ground truth increased the fully automatic performance by about nine percentage points. This suggests that capturing the spatial arrangement of hands and correctly predicting their pose are equally important to per-frame activity recognition using only hand information.

Incorporating temporal constraints. So far we have considered each frame independently, but of course much information about activity lies in the temporal dynamics of the hands over time. We tried a simple voting-based approach to incorporate some of this temporal structure: we classify each individual frame in the context of a fixed-size temporal window centered on the frame. Scores across the window are summed, and the frame is labeled as the highest scoring class. To again compare with the ground truth informed upper bound, we only consider labeled frames, so a window of k frames spans approximately k seconds.

Table 3 presents the results. Temporal information increases activity recognition accuracy significantly, with even a window of 5 frames improving results from 0.664 to 0.764 when using ground truth segmentations, and from 0.509 to 0.618 using the fully automatic system. Accuracy continues to improve with increasing window size, with 50 frames achieving 0.929 with the ground truth and 0.734 for the automatic segmentations. This improvement is likely due to two factors: certain hand poses may be more distinctive than others, and segmentation errors in any given frame can be disregarded as outliers. We also show results averaged over stricter splits, such that any actor seen in testing is not seen in training. This partitioning reduces the number of splits with enough test data to two, since not all pairs performed all activities. Though limited in scope, the results of this strict task are similar to the "main split."

	Window size (k)				
	1	5	15	30	50
Main split					
Segmentation mask	0.509	0.618	0.680	0.724	0.734
Ground truth mask	0.664	0.764	0.851	0.900	0.929
Split across actor pairs (average)					
Segmentation mask	0.570	0.639	0.679	0.687	0.671
Ground truth mask	0.661	0.742	0.790	0.814	0.847

Table 3: *Activity recognition accuracy from hand masks*, using a temporal window of k frames. See text for details.

Our results suggest that hand segmentation could deliver high activity recognition accuracy without the need to recognize objects or backgrounds; however, our ground truth experiments show that automated approaches would benefit from increased segmentation accuracy.

7. Conclusion and Future Work

We showed how to detect and distinguish hands in first person video, by combining CNN-based classification with fast candidate generation based on sampling from a joint model of hand appearance and geometry. We then showed that these detections could also be used to yield state-of-the-art hand pose segmentation. We explored the potential of these segmentations by showing that activities can be successfully recognied in our first-person dataset based on the configuration and pose of hands alone. Finally, we introduced a novel first-person dataset with dynamic interactions between people, along with fine-grained ground truth.

In future work, we plan to generalize our hand-based activity recognition to larger sets of activities, including fine-grained actions (e.g. picking up vs. laying down a card). We will also consider more challenging social situations, e.g. with multiple interacting people moving around the room.

Acknowledgments. This work was supported by NSF (CAREER IIS-1253549, CNS-0521433), NIH (R01 HD074601, R21 EY017843), Google, and IU OVPR through IUCRG and FRSP grants. It used compute facilities provided by NVidia, the Lilly Endowment through support of the IU Pervasive Technology Institute, and the Indiana METACyt Initiative. SB was supported by a Paul Purdom Fellowship. We thank Benjamin Newman, Brenda Peters, Harsh Seth, and Tingyi Wanyan for helping with dataset collection, and Rob Henderson, Robert Henschel, Bruce Shei, and Abhinav Thota for systems support.

References

[1] Jenga. http://wikipedia.org/wiki/Jenga. 3

[2] Mau mau (card game). http://wikipedia.org/wiki/Mau_Mau_(card_game). 3

[3] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *PAMI*, 34(11):2189–2202, 2012. 4

[4] S. Bambach, D. Crandall, and C. Yu. Understanding embodied visual attention in child-parent interaction. In *ICDL-EPIROB*, 2013. 1

[5] S. Bambach, J. M. Franchak, D. J. Crandall, and C. Yu. Detecting hands in children's egocentric views to understand embodied attention during social interaction. In *Annual Conference of the Cognitive Science Society (CogSci)*, 2014. 4

[6] A. Betancourt, M. M. Lopez, C. S. Regazzoni, and M. Rauterberg. A sequential classifier for hand detection in the framework of egocentric vision. In *CVPR Workshops*, 2014. 2

[7] Y. Bhattacharjee. A little black box to jog failing memory. *The New York Times*, March 9 2010. 1

[8] H. H. Ehrsson, C. Spence, and R. E. Passingham. That's my hand! Activity in premotor cortex reflects feeling of ownership of a limb. *Science*, 305(5685):875–877, 2004. 1

[9] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding*, 108(1):52–73, 2007. 2

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 2

[11] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *CVPR*, 2011. 2

[12] A. Fathi, J. Hodgins, and J. Rehg. Social interactions: A first-person perspective. In *CVPR*, 2012. 2

[13] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *CVPR*, 2011. 2, 3

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*. 2014. 4

[15] A. Ihler. Kernel Density Estimation (KDE) Toolbox for Matlab. http://www.ics.uci.edu/~ihler/code/kde.html. 4

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 5

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012. 4, 5

[18] S. Lee, S. Bambach, D. J. Crandall, J. M. Franchak, and C. Yu. This hand is my hand: A probabilistic approach to hand disambiguation in egocentric video. In *CVPR Workshops*, June 2014. 2, 4, 5

[19] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012. 2

[20] C. Li and K. M. Kitani. Model recommendation with virtual probes for egocentric hand detection. In *ICCV*, 2013. 2, 3

[21] C. Li and K. M. Kitani. Pixel-level hand detection in egocentric videos. In *CVPR*, 2013. 2, 3, 7

[22] Y. Lin, G. Hua, and P. Mordohai. Egocentric object recognition leveraging the 3d shape of the grasping hand. In *ECCV Workshops*. 2014. 3, 6

[23] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *CVPR*, 2013. 2

[24] A. Mittal, A. Zisserman, and P. Torr. Hand detection using multiple proposals. In *BMVC*, 2011. 2

[25] J. R. Napier. The prehensile movements of the human hand. *Journal of bone and joint surgery*, 38(4):902–913, 1956. 7

[26] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012. 2, 3

[27] X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *CVPR*, 2010. 2

[28] X. Ren and M. Philipose. Egocentric recognition of handled objects: Benchmark and analysis. In *CVPR Workshops*. 2009. 2, 3

[29] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 23(3):309–314, 2004. 7

[30] M. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *CVPR*, 2013. 2

[31] T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, et al. Accurate, Robust, and Flexible Real-time Hand Tracking. In *Proc. CHI*, 2015. 2

[32] E. Spriggs, F. De La Torre, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In *ECCV Workshops*, 2009. 2

[33] R. Stross. Wearing a badge, and a video camera. *The New York Times*, April 6 2013. 1

[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014. 4, 5

[35] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *IJCV*, 104(2):154–171, 2013. 2, 4

[36] N. Wingfield. Gopro sees opportunity in its amateur daredevils. *The New York Times*, January 30 2014. 1