Viewpoint Integration for Hand-Based Recognition of Social Interactions from a First-Person View

Sven Bambach[1], David J. Crandall[1], Chen Yu[2]
[1]School of Informatics and Computing
[2]Department of Psychological and Brain Sciences
Indiana University, Bloomington, IN, USA
{sbambach, djcran, chenyu}@indiana.edu

ABSTRACT

Wearable devices are becoming part of everyday life, from first-person cameras (GoPro, Google Glass), to smart watches (Apple Watch), to activity trackers (FitBit). These devices are often equipped with advanced sensors that gather data about the wearer and the environment. These sensors enable new ways of recognizing and analyzing the wearer's everyday personal activities, which could be used for intelligent human-computer interfaces and other applications. We explore one possible application by investigating how egocentric video data collected from head-mounted cameras can be used to recognize social activities between two interacting partners (e.g. playing chess or cards). In particular, we demonstrate that just the positions and poses of hands within the first-person view are highly informative for activity recognition, and present a computer vision approach that detects hands to automatically estimate activities. While hand pose detection is imperfect, we show that combining evidence across first-person views from the two social partners significantly improves activity recognition accuracy. This result highlights how integrating weak but complimentary sources of evidence from social partners engaged in the same task can help to recognize the nature of their interaction.

Categories and Subject Descriptors

H.5 [**Information Interfaces and Presentation**]: Miscellaneous; I.4 [**Image Processing and Computer Vision**]: Segmentation, Applications

General Terms

Human Factors; Experimentation

Keywords

Google Glass; wearable devices; activity recognition; egocentric video; viewpoint integration; hand detection

ICMI 2015, November 9–13, 2015, Seattle, WA, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3912-4/15/11 ...$15.00.
DOI: http://dx.doi.org/10.1145/2818346.2820771

1. INTRODUCTION

Thanks to recent advances in technology, wearable devices are becoming increasingly common in our everyday lives. More and more people are wearing these devices, from activity trackers like FitBit that record workouts, to communication devices like the Apple Watch that serve as convenient interfaces to smartphones, to cameras like GoPro Hero and Autographer that let people record egocentric imagery and video, to heads-up display devices like Google Glass that augment the real visual world. Despite their novel form factors, wearable devices are just computers featuring CPUs, network connectivity, and sensors including cameras, microphones, GPS receivers, accelerometers, thermometers, light sensors, and so on. Wearable devices are thus able to sense data from the environment, allowing them to detect and monitor the wearer's activities throughout his or her everyday life [5]. Personal activity recognition in everyday contexts has a wide range of applications for intelligent human-computer interfaces [2, 5, 7, 11, 12].

The aim of the present study is to examine how video data collected from egocentric camera devices like Google Glass can be used to recognize social activities. As the first study to explore this new research direction, we focus on recognizing activities in dyadic interactions in which two social partners perform joint tasks. Using a dataset with four tasks and four actors, we show that egocentric video from a first-person perspective contains particularly informative data that can be used for activity recognition. We present two novel contributions. First, inspired by previous studies on egocentric video [6] that show that one's own hands are almost always present in the first-person view, we develop a new method of activity recognition based on hand detection and segmentation. This omnipresence occurs because the human cognitive system needs real-time visual information to produce visual-guided hand actions [8] and therefore people usually keep their hands in the egocentric view. The advantage of relying on just hands is that we avoid the need to solve much more complicated vision problems, like recognizing all possible objects or environments. Second, we show that integrating synchronous visual information captured from the two views of the social partners can dramatically improve recognition accuracy.

2. HAND INTERACTIONS

We begin by describing several building blocks of our study, including a dataset of first-person video captured from interacting people, and computer vision techniques to automatically detect and segment hands.

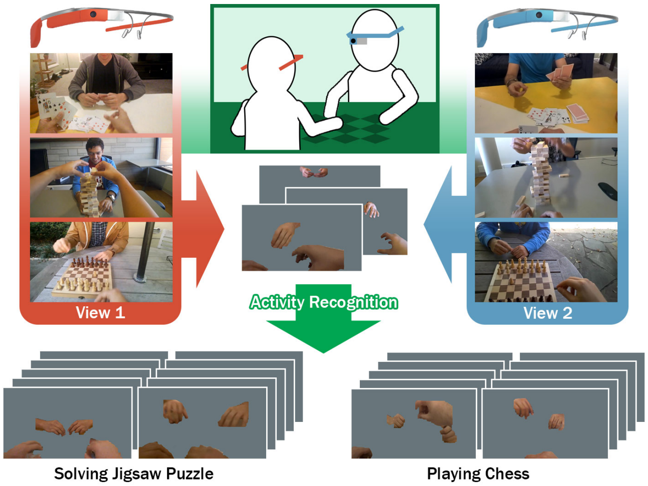


Figure 1: Two actors are engaged in different social interactions, while both wear Google Glass to capture video from each field of view. We present a vision-based framework that extracts hands from each view and jointly estimates the performed activity based on hand pose and position alone.

2.1 Dataset

We collected a dataset of first-person video from interacting subjects [1], using Google Glass to capture video (720p30) from each person's viewpoint, as illustrated in Figure 1. The subjects were asked to perform four different activities: (1) playing cards; (2) playing fast chess; (3) solving a jigsaw puzzle; and (4) playing Jenga (a 3d puzzle game). Sample frames for each activity are shown in Figure 2. To add visual diversity to the dataset, videos were collected in three different locations: a conference room table, an outdoor patio table, and a living room coffee table. We recorded four actors (all male, right-handed students) over several days, ensuring variety in actors' clothing and the surroundings. We systematically collected data from all four actors performing all four activities at all three locations while randomly assigning social pairs, resulting in $4 \times 4 \times 3 = 48$ unique combinations (24 for each viewpoint). Each video is 90 seconds long and synchronized across the two views.

To train and evaluate our computer vision techniques, we manually annotated hand position and shape in a subset of 100 random frames per video, resulting in pixel-level ground truth for 15,053 hands. Figure 2 shows some examples of annotated hands. For both the hand extraction and the activity detection tasks, the dataset was randomly split into a training set of 24 videos (12 pairs), a validation set of 8 videos and a test set of 16 videos such that the four activity classes were evenly distributed in each set.

2.2 Extracting Hands

We used a state-of-the-art computer vision algorithm to automatically extract hands from first-person videos. We briefly describe the approach here; more details, as well as an in-depth quantitative evaluation, are presented elsewhere [1]. The hand extraction process consists of two major steps: detection, which tries to coarsely locate hands in each frame, and segmentation, which estimates the fine-grained pixel-level shape of each hand.

Hand detection. Our hand detector applies convolutional neural networks (CNNs), which are the current state-of-the-art in general object detection [3]. CNNs are designed to solve image-level classification problems. To apply them to object detection, the typical approach is to use a lightweight classifier to generate a large set of image windows that may contain the object of interest (Figure 3a), and then to classify each of them using the more computationally demanding CNN (Figure 3b). In our case, we generate candidate hand regions using color features to identify likely skin regions, while also using the spatial biases inherent in hand positions in first-person videos (e.g. that the camera wearer's hands tend to be lower in the first-person view [6]).

To classify each candidate window, we trained a CNN with the architecture of Krizhevsky et al. [4], using ground-truth hand annotations from the set of training frames. The CNN was trained to perform a five-way classification task, categorizing image regions into background, left or right hand of the camera wearer, and left or right hand of the social partner. The network was trained using stochastic gradient descent until convergence on the validation set. During detection, regions with a sufficiently high classification score are marked as hand candidates, and then a non-maximum suppression step removes duplicate detections caused by overlapping candidates. The result of hand detection is a set of bounding boxes giving the location of up to four hands, corresponding to the two hands of the camera wearer and the two hands of the social partner (see Figure 3b).

Hand segmentation. Given the detected hand bounding box, we next extract finer-grained, pixel-level masks that capture the shape of the hands (Figure 3c). Our approach assumes that most pixels inside a detected hand window correspond with a hand, albeit with a significant number of background pixels caused either by detector error or because of the rectangular bounding boxes. This assumption lets us apply GrabCut, a well-known semi-supervised segmenta-

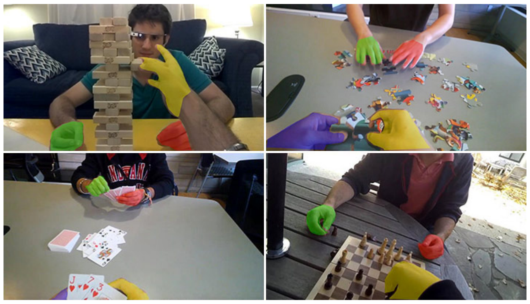


Figure 2: Sample frames from each activity in our experiments. Colored regions show ground truth hand masks.

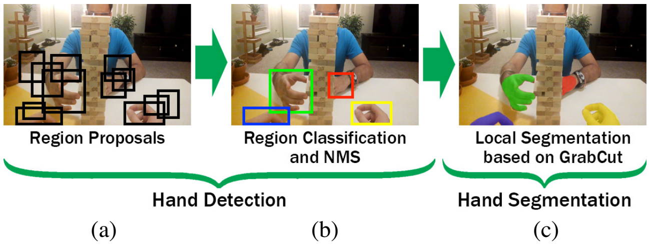

Region Proposals
Region Classification and NMS
Local Segmentation based on GrabCut
Hand Detection
Hand Segmentation
(a)
(b)
(c)

Figure 3: Overview of the hand extraction pipeline, consisting of a detection and a segmentation step.

tion algorithm [9]. Given an approximate foreground mask, GrabCut improves the segmentation by iteratively refining color-based appearance models of the foreground (hand) and the background, and relabeling foreground and background using a Markov Random Field.

While not perfect, this approach produces reasonable results, with a mean average precision (mAP) of 0.74 for detection and a pixel-level intersection-over-union of 0.56 for segmentation on our dataset [1].

3. ACTIVITY RECOGNITION AND VIEWPOINT INTEGRATION

Our main hypothesis is that hand poses by themselves reveal significant evidence about the objects people are interacting with and the activities they are doing. This would imply that automatic activity recognition systems could focus on accurately recognizing one type of object – the hands – instead of having to model and detect the thousands of possible objects and backgrounds that occur in real-world scenarios. While the hand poses in any given video frame may not necessarily be informative, we hypothesize that integrating hand pose evidence across frames and across viewpoints may significantly improve activity recognition results. We investigate (1) how well activities can be recognized in our dataset based on hand pose information alone, and (2) whether the two first-person viewpoints can be complementary with respect to this task.

3.1 Hand-based Activity Recognition

To explore if hand poses can uniquely identify activities, we created masked frames in which all content except hands were replaced by a gray color. Some examples of masked frames are shown in Figures 1 and 4. We trained a CNN with the architecture of [4] on a four-way classification task by feeding it the (rescaled) masked hand frames from the training set, i.e. the inputs to the first network layer were 224×224×3 buffers of normalized RGB pixel values. Each frame was labeled with one of the four activities. In this training phase, we used ground-truth hand segmentations to prevent the classifier from learning any visual bias not related to hands (e.g. portions of other objects that could be visible due to imperfect hand extraction). With 100 annotated frames per video and 24 videos in the training set, this led to a total of 2,400 training images (600 per activity). The network was trained with stochastic gradient descent using a batch size of 256 images until the accuracy on the validation videos converged. As is standard practice, the convolutional layers were initialized with weights from the ImageNet Visual Recognition Challenge [10], which led to convergence after 12 epochs.

To test the performance of the trained CNN, we first applied the hand extraction approach from Section 2 to each frame of all 16 videos in our test dataset, resulting in 16 × 2,700 = 43,200 test frames. Classifying each frame individually gave 53.6% accuracy on the four-way activity problem, nearly twice the random baseline (25.0%). This promising result suggests a strong relationship between hand poses and activities.

3.2 Viewpoint Integration

Of course, our automatic hand extraction is not perfect, and regularly suffers from false negatives (missing hands) and imperfect segmentations (inaccurate hand poses). Even when masks are perfect, they may not be informative: hands may be occluded or not in view at all (see Figure 4). However, even if the hands are occluded or inaccurately extracted in one frame, it is likely that another frame, either from the other person's view or from a nearby moment in time, yields an accurate estimate of the hand pose.

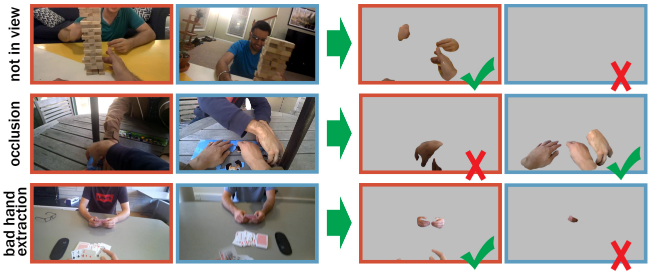


Figure 4: Sample moments where one viewpoint observes much more informative hand poses for activity recognition.

We integrate evidence across frames using a straightforward late fusion method at the decision level. Suppose we have a set $\mathcal{P}$ of actors, each of whom records a sequence of n frames, i.e. $\mathcal{F}_p = (F_p^1, F_p^2, ..., F_p^n)$ for each $p \in \mathcal{P}$. The frames are synchronized so that for any t and pair of actors $p, q \in \mathcal{P}$, F_p^t and F_q^t were captured at the same moment. Without loss of generality, we consider the specific case of two actors, $\mathcal{P} = \{A, B\}$. Suppose that our goal is to jointly estimate the unknown activity label H from a set of possible activities $\mathcal{H}$. By applying the CNN trained in the last section on any given frame F_p^t, we can estimate (using only the evidence in that single frame) the probability that it belongs to any activity $h \in \mathcal{H}$, $P(H = h|F_p^t)$.

Temporal integration. We integrate evidence across the temporal dimension, given the evidence in individual frames across a time window from t_i to t_j in a single view p,

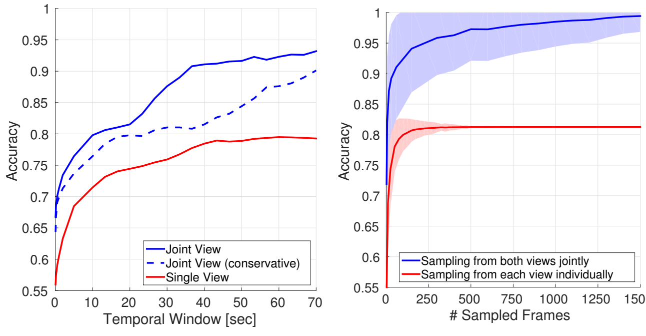
$$\begin{aligned}\hat{H}_p^{t_i,t_j} &= \arg\max_{H\in\mathcal{H}} P(H|F_p^{t_i}, F_p^{t_i+1}, ..., F_p^{t_j}) \\ &= \arg\max_{H\in\mathcal{H}} \prod_{k=t_i}^{t_j} P(H|F_p^k),\end{aligned}$$

where the latter equation follows from assumptions that frames are conditionally independent given activity, that activities are equally likely *a priori*, and from Bayes' Law. We evaluated this approach by repeatedly testing classification performance on our videos over many different time windows of different lengths (different values of $|t_j - t_i|$). The red line in Figure 5a shows that accuracy increases with the number of frames considered. For instance, when observing 20 seconds of interacting hands from a single viewpoint, the system predicts the interaction with 74% accuracy.

Viewpoint integration. Next we take advantage of the coupled interaction by integrating evidence across viewpoints,

$$\hat{H}^{t_i,t_j} = \arg\max_{H\in\mathcal{H}} \prod_{k=t_i}^{t_j} P(H|F_A^k)P(H|F_B^k),$$

which makes the additional assumption that the viewpoints are independent conditioned on activity. We again test over many different temporal windows of different sizes in our videos, but now using frames from both viewpoints. The re-



(a)

(b)

Figure 5: Comparison of activity recognition accuracy using one (red) and both (blue) viewpoints, using a sliding temporal window (a) and sampling nonadjacent frames (b).

sults are plotted in blue in Figure 5a and clearly outperform the single view approach, showing that the two views are indeed complementary. However, this fusion method has the potentially unfair advantage of seeing twice as many frames as the single view method, so we also show a more conservative line (dashed blue) that considers half the temporal window size (so that for any position on the x-axis, the red line and dashed blue line are seeing the same number of frames). This conservative comparison still outperforms the single view, demonstrating that observing $x/2$ seconds of hand interactions from both viewpoints is more informative than observing x seconds from only one.

Sampling Frames. Adjacent frames are highly correlated, so we hypothesize that it may be better to integrate evidence across wider time periods. A generalization of the above is to use a set $\mathcal{T} \subseteq [1, n]$ of times at which to observe frames,

$$\hat{H}^{\mathcal{T}} \quad = \quad \arg\max_{H \in \mathcal{H}} \prod_{k \in \mathcal{T}} P(H|F_A^k)P(H|F_B^k).$$

We tested this by repeatedly sampling different numbers of frames from each video. The red line in Figure 5b shows accuracy as a function of number of frames ($|\mathcal{T}|$) for single viewpoints (with the shaded area indicating standard deviation over 2,700 sampling iterations). After a high initial variance, accuracy converges to 81.25%, or 13 of 16 videos, at about 500 frames (about 20% of a video). Of the three incorrect videos, two are of chess (where we predict puzzle and cards) and the other is of cards (where we predict Jenga).

Finally, we combine both viewpoints together by sampling sets of corresponding frames from both views. The blue line in Figure 5b shows the results (plotted so that at any position on the x-axis, the red line sees x frames while the blue line sees $x/2$ frames from each viewpoint). Even for a small number of samples, this method dramatically outperforms single view, albeit with a large standard deviation, indicating that some paired samples are much more informative than others. More importantly, as the number of samples increases, the joint view method approaches 100% accuracy. This means that using the complementary information from both viewpoints helps correctly predict the three videos that were not correctly predicted with the single view.

4. CONCLUSION AND FUTURE WORK

We presented a system that can recognize social interactions between two partners who both wear head-mounted cameras (Google Glass) by automatically analyzing each actor's egocentric video stream using state-of-the art computer vision techniques. In particular, we show that the knowledge of hand pose and position alone can provide enough information to distinguish between the four activities present in our data. Further, we demonstrate that the two viewpoints are complementary and that predicting the interaction based on integrating evidence across viewpoints leads to better results than analyzing them individually.

We plan to extend our work in several directions, including testing on a larger set of activities as well as finer-grained actions (such as picking up a card or placing a piece on the chess board). We also plan to investigate more challenging social interactions involving more than two people, as well as integrating other sensor information from Google Glass, such as the accelerometer.

Acknowledgments. This work was supported by the National Science Foundation (CAREER IIS-1253549, CNS-0521433), the National Institutes of Health (R01 HD074601, R21 EY017843), Google, and the IU OVPR through the IUCRG and FRSP programs. It used compute facilities provided by NVidia, the Lilly Endowment through its support of the IU Pervasive Technology Institute, and the Indiana METACyt Initiative. SB was supported by a Paul Purdom Fellowship.

5. REFERENCES

[1] S. Bambach, S. Lee, D. Crandall, and C. Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *IEEE International Conference on Computer Vision*, 2015.

[2] T. M. T. Do, K. Kalimeri, B. Lepri, F. Pianesi, and D. Gatica-Perez. Inferring social activities with mobile sensor networks. In *ACM International Conference on Multimodal Interaction*, 2013.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[4] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep Convolutional Neural Networks. In *Neural Information Processing Systems*, 2012.

[5] J. R. Kwapisz, G. M. Weiss, and S. A. Moore. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, 12(2):74–82, 2011.

[6] S. Lee, S. Bambach, D. Crandall, J. Franchak, and C. Yu. This hand is my hand: A probabilistic approach to hand disambiguation in egocentric video. In *CVPR Workshop on Egocentric Vision*, 2014.

[7] N. Oliver, E. Horvitz, and A. Garg. Layered representations for human activity recognition. In *ACM International Conference on Multimodal Interaction*, 2002.

[8] J. Pelz, M. Hayhoe, and R. Loeber. The coordination of eye, head, and hand movements in a natural task. *Experimental Brain Research*, 139(3):266–277, 2001.

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

[10] O. Russakovsky, J. Deng, H. Su, et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 2015.

[11] Y. Song, L. Morency, and R. Davis. Multi-view latent variable discriminative models for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[12] Y. C. Song, H. Kautz, J. Allen, M. Swift, Y. Li, J. Luo, and C. Zhang. A markov logic framework for recognizing complex events from multimodal data. In *ACM International Conference on Multimodal Interaction*, 2013.