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

Sven Bambach, David Crandall, School of Informatics and Computing, Indiana University
Chen Yu, Department of Psychological and Brain Sciences, Indiana University

INDIANA UNIVERSITY BLOOMINGTON

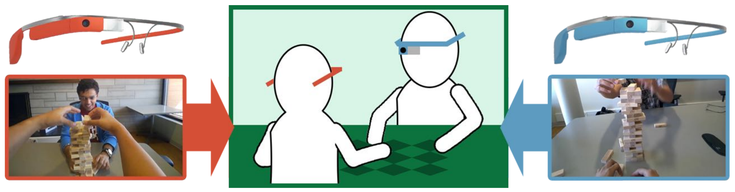
1. Introduction

- **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 equipped with **sensors** that enable new ways of recognizing/analyzing the wearer's everyday personal activities.
- Here, we aim to examine how video data collected from **Google Glass** can be used to recognize **social activities**.

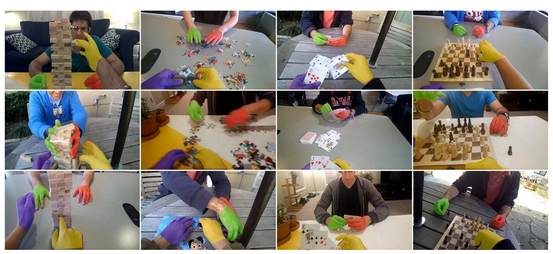
Two Google Glass devices as used in our experiments.

1. We show that **hands** are important cues in the first person view, and develop a new method of **activity recognition** based on **hand detection** and **segmentation**.
2. We show **that integrating synchronous visual information** captured **across views** of different socials partners can dramatically improve recognition accuracy.

2. Social Interactions Dataset



- We recorded synchronized first-person video from two interacting subjects, using **two Google Glasses**:
  - 4 actors x 4 activities x 3 locations = **48 unique videos**
  - Pixel-level ground truth annotations for **15,053 hands** in **4,800 frames** to train computer vision algorithms



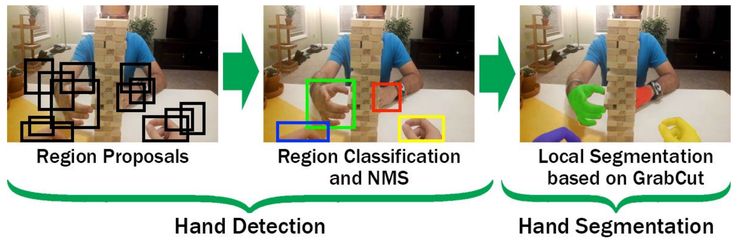
Some sample frames from the data with ground truth hand masks superimposed in different colors (indicating different hand types). Each column shows one activity: *Jenga*, jigsaw puzzle, cards, and chess.

The *EgoHands* dataset is publicly available:

http://vision.indiana.edu/egohands

3. Extracting Hands

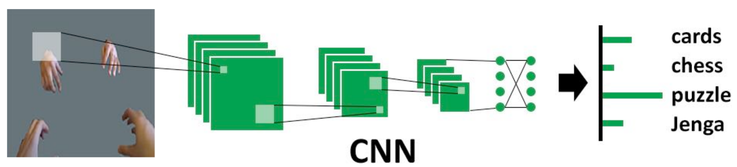
- We use a state-of-the-art **computer vision** algorithm to automatically extract hands from first-person videos.
- **Detection** step coarsely locates hands:
  - Lightweight classifier proposes a large set of image windows that may contain the object of interest **(a)**.
  - Convolutional neural network (CNN) classifies each window for the presence of specific hand type **(b)**.
- **Segmentation** step estimates fine-grained shape/pose:
  - Based on a coarse detection, the GrabCut algorithm iteratively refines the segmentation with a MRF **(c)**.



For more details about this hand extraction pipeline, as well as an in-depth quantitative evaluation, please check out our **ICCV 2015** paper on first-person hand detection: **http://vision.indiana.edu/lending-a-hand**

4. Activity Recognition

- We hypothesize that first-person hand poses can reveal significant information about the interaction.
- To test this, we created **masked frames** with everything but hands replaced by a gray color, and used them as input to a CNN:

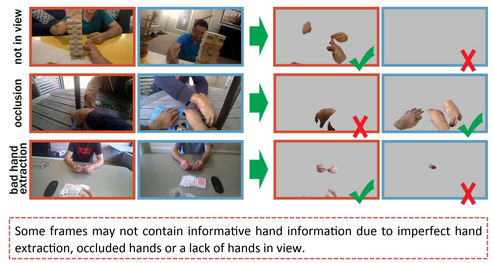


Each masked frame is labeled as belonging to one of the four activities in our dataset, training the network on a four-way classification task. During training only frames with (perfect) ground-truth segmentations are used. During testing, hands are extracted automatically in each frame.

- The trained network was tested on 16 videos
- Classifying each frame individually yielded **53.6% accuracy** (chance baseline = 25%)

5. Integrating Both Camera Viewpoints

- If one frame is not informative, it is likely that another frame, either from the other person's view or from a nearby moment in time, yields more information.



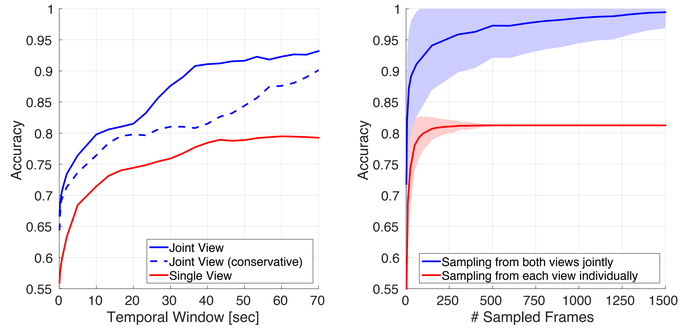
Some frames may not contain informative hand information due to imperfect hand extraction, occluded hands or a lack of hands in view.

- We **integrate evidence across frames** using a straightforward **late fusion** method at the decision level.
- Temporal integration: $\hat{H}_p^{t_i,t_j} = \arg\max_{H\in\mathcal{H}} P(H|F_p^{t_i}, F_p^{t_i+1}, ..., F_p^{t_j})$

H = Activity
$F_p^{t_i}$ = Frame from person p at time t_i
$P(H=h|F_p^{t_i})$ is given by applying CNN to frame $F_p^{t_i}$

$$= \arg\max_{H\in\mathcal{H}} \prod_{k=t_i}^{t_j} P(H|F_p^k)$$

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



Comparison of activity recognition accuracy using one (red) and both (blue) viewpoints, using a sliding temporal window (left) and sampling non-adjacent frames (right).

6. Conclusion

- We presented a system that can **recognize social interactions** between two partners based on extracted **hand poses** from **head-mounted cameras**.
- We show that **the two viewpoints are complementary** with respect to that task.
- For future work, we would like to extend this idea to more activities as well as more **fine-grained actions**.

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.*