Egocentric Object Recognition leveraging the 3D Shape of the Grasping Hand

Yizhou Lin, Gang Hua, Philippos Mordohai

Department of Computer Science
Stevens Institute of Technology
{ylin8, ghua, pmordohai}@stevens.edu

Abstract. We present a systematic study on the relationship between the 3D shape of a hand that is about to grasp an object and recognition of the object to be grasped. In this paper, we investigate the direction from the shape of the hand to object recognition for unimpaired users. Our work shows that the 3D shape of a grasping hand from an egocentric point of view can help improve recognition of the objects being grasped. Previous work has attempted to exploit hand interactions or gaze information in the egocentric setting to guide object segmentation. However, all such analyses are conducted in 2D. We hypothesize that the 3D shape of a grasping hand is highly correlated to the physical attributes of the object being grasped. Hence, it can provide very beneficial visual information for object recognition. We validated this hypothesis by firstly building a 3D, egocentric vision pipeline to segment and reconstruct dense 3D point clouds of the grasping hands. Then, visual descriptors are extracted from the point cloud and subsequently fed into an object recognition system to recognize the object being grasped. Our experiments demonstrate that the 3D hand shape can indeed greatly help improve the visual recognition accuracy, when compared with the baseline where only 2D image features are utilized.

Keywords: Mobile and Wearable Systems, Egocentric and First-Person Vision, Activity Monitoring Systems, Rehabilitation Aids

1 Introduction

The motivating healthcare application of this research is the advancement of hand rehabilitation following neuromuscular injury such as stroke [10]. Typically, rehabilitation is planned based on evidence gathered in constrained clinical settings. Consequently, the usage pattern of the paretic (partially paralyzed) hand in patients' daily activities remains largely unknown, despite the importance of quantitatively measuring hand impairment in daily activities to assist the therapist. In addition to stroke, hand impairment can arise from a variety of diseases or injuries, including spinal cord injury, sclerodema, Parkinson's disease, and radial nerve damage. For stroke alone, approximately 800,000 individuals incur a cerebrovascular accident each year in the U.S. [6]. According to a survey by Nowak [14], only 40% of people who survive a stroke experience full functional recovery [7]. For the remaining 60% with chronic hemiparesis, hand dysfunction is one of the most common sources of disability [11]. Occupational therapy is usually

undertaken in an effort to at least partially restore function. Unfortunately, the true effectiveness of different therapy regimens is difficult to gauge. At present, the translation of clinical treatment to improved use of the affected hand remains largely unknown. Clinical assessments can be performed and user questionnaires can be administered, but objective measurements of what is actually occurring in daily life are lacking. Increasingly, therapists are encouraged to follow evidence-based practice to foster better outcomes, but the data of interest, actual functioning within the community and home, is difficult to acquire.

This paper is a step in the direction of using wearable, egocentric computer vision technology to aid hand rehabilitation. Before beginning any research with patients and therapists, however, several technical obstacles have to be overcome. Here, we address two of the relevant challenges:

- the 3D reconstruction of the grasping hand from egocentric stereo images and
- establishing that there is a correlation between the shape of the grasping hand and the identity of the object that is about to be grasped.

Having established this relationship, we can now envision the development of an approach for monitoring paretic users in their daily activities, recording information relevant to the functionality of their hands and then presenting this information to physical therapists who will determine the next steps in the users' rehabilitation. Numerous challenges will have to be addressed to realize this long-term goal. The remainder of the paper describes the current state of our approach, which has only been tested on unimpaired users so far.

Recognizing objects from an egocentric perspective is a very important task in understanding the users' daily behavior and activities. It enables continuous monitoring of the users' hands, in our cases, in natural settings and can provide potentially much more useful information than what can be collected in brief consultations with a physical therapist. We strongly believe that an egocentric approach is essential for the success of our research. An advantage of egocentric vision is that there is a large amount of contextual information centered around the user, such as hand position and gaze [17, 5], to help with the recognition tasks. For example, the skin tone of the user's hand interacting with an object has been exploited by Pirsiavash and Ramanan [17] for localizing the object of attention to improve object detection. Fathi *et al.* [5] leveraged gaze information to help identify the object being handled by the user for daily action recognition.

Notwithstanding the demonstrated success of these previous works, two limitations are present in them. On one hand, the contextual information is more utilized as a prefilter to scope the location of the target object to be detected or recognized. In other words, it is not an integral part of the core recognition algorithms. On the other hand, all visual reasoning in these previous works is conducted in 2D. Some important visual information that could potentially be provided by 3D information has largely been neglected.

Research in kinesiology and neuroscience reveals that the physical attributes, such as the shape, size and weight, of the object to be grasped by the user determine the posture of the grasping hand, even before contact is made [21, 22, 29]. These findings largely motivated the research conducted in this paper. We conjecture that the hand posture in an object grasping action would provide additional cues to enhance the visual

recognition of the object being grasped. In this paper, we present a first systematic study to verify such a conjecture.

Fig. 1. A user is grasping an object, while the grasp is recorded by the egocentric stereo cameras.

We begin our study by building an egocentric 3D vision pipeline to segment and recover a dense 3D point cloud of the grasping hands from egocentric stereo images. The core hardware component of our egocentric 3D vision pipeline are the Vuzix 920AR augmented reality glasses equipped with a pair of egocentric cameras [1] [2]. The dense 3D point cloud provides a non-parametric representation of the shape of the hands. Since the 3D point cloud of the hand carries implicit information of the hand posture, in this paper we focus our analysis on the point cloud, to avoid the challenging task of estimating the hand posture, which we leave as our future work.

We have collected a dataset of stereo videos involving 36 users. The users were asked to wear the AR glasses and grasp 8 objects each from a different category three times. We then conduct a systematic evaluation on an egocentric object recognition task to recognize these 8 objects engaging or not engaging the 3D features extracted from the 3D point cloud of the grasping hand. Our analysis re-validated the correlations between the physical attributes of the object and the 3D shape of the grasping hand, and manifested that 3D hand shape can indeed enhance object recognition accuracy.

This paper presents the following contributions:

1. We are the first to introduce observations from kinesiology and neuroscience, regarding the correlation between the physical attributes of the object to be grasped and the grasping hand, to the egocentric vision community.
2. To the best of our knowledge, our work presents a first systematic study to reveal how the 3D shape of a grasping hand can help improve recognition of the objects being grasped from an egocentric point of view.
3. We collected a labeled dataset of stereo videos involving 36 users grasping 8 different categories of objects to carry out our study, which will be shared with the community to advance research on egocentric object recognition.

2 Related Work

The core assumption behind our work, that the shape of a grasping hand can reveal information about the object about to be manipulated, is based on findings in kinesiology

[1] http://www.vuzix.com/

[2] The Vuzix 920AR augmented reality glasses are equipped with a near-eye display which blocks the user's field of view. Ideally it should be a see-through display for use in daily activities.

and neuroscience. The shape, size and weight of the object to be grasped determined the posture of the hand even before contact is made [29]. A variety of grips that can be broadly distinguished into precision or power grips and further subdivided into finer categories have been identified in the relevant literature; see [21] and references therein. Schettino et al. [22] studied hand preshaping before an object is grasped and the effects on limited vision on the modulation of hand kinematics to adjust to object contours.

A prevalent and accurate way for obtaining hand pose and articulated motion is the use of data gloves [24]. Data gloves, however, are uncomfortable and prohibitively expensive making their use in everyday activities unappealing. Here, we will focus on vision-based solutions to these problems. Delamarre and Faugeras [1] used an articulated 3D model of the hand made of cones and spheres and matched it with stereo reconstructions of the hands using ICP to infer the pose of the fingers. Later, Dewaele et al. [2] presented a model-based approach for tracking the hands in stereoscopic videos that relied on motion cues in addition to shape estimates in each frame. In our current work, we have not taken advantage of the priors provided by articulated models or temporal information, but these are clearly fruitful directions for future research.

In recent years, we have observed a surge of interest in research on egocentric computer vision problems, which is the result of the combination of inspiration from early pioneering systems [23, 16] and the ever growing computing power in modern wearable platforms. Here, we review recent advancement of research in egocentric visual recognition. These works can largely be put into three categories, object/scene discovery and summarization, action/activity recognition, and object recognition.

One of the major applications of egocentric visual object/scene discovery is to summarize the egocentric video for daily life logging [8, 12]. Jojic *et al.* [8] proposed the structural epitome for summarize one's visual experiences. The resulting STEL Epitome image produced from the egocentric video is visually similar to a panoramic image. It automatically groups the same scene and objects into tightly connected regions in the epitome image. Lee *et al.* [12] proposed to combine a set of egocentric features with other visual cues to build a detector for category independent important object detection. The egocentric video is then summarized based on the important objects detected.

A larger body of recent research on egocentric computer vision has been devoted to recognizing the daily actions/activities from egocentric videos [27, 25, 17, 15, 3, 5]. The CMU-MMAC database [27] built at CMU is a multi-modal activity recognition database which incorporated the videos from a head-mounted forward-looking camera as one of the modality. Spriggs *et al.* [25] acknowledged the importance of objects in activity understanding, but did not pursue object recognition at the time. Ogaki *et al.* [15] explored the use of first-person eye movement and ego-motion for indoor activity recognition. The special set-up of their equipment is that there is an inward looking camera to capture the user's eye movement. Various egocentric cues have been explored to facilitate the recognition of activity, including the skin tone of the hands [17, 3] and the gaze [5].

As widely acknowledged in previous works, object recognition serves a very important role in activity understanding. The use of object recognition in egocentric vision systems dates back to the DyPERS system of Schiele *et al.* [23]. Mayol and Murray [13] recognized manipulation activities leveraging skin color and histogram-based objec-

t classification from a shoulder-mounted camera. Ren and Philipose [18] collected a large database of egocentric videos of objects to facilitate research in egocentric object recognition. The database was used by Ren and Gu [19] who segment objects that have been grasped by the user from the background. The approaches of [19, 4] rely primarily on optical flow motion cues.

Our work is different than previous research in two aspects. First, most previous egocentric vision system, if not all of them, only leveraged a monocular forward-looking camera, while our system comprises an egocentric stereo camera. Second, our proposed egocentric object recognition method benefits from the 3D hand shape derived from egocentric stereo. To the best of our knowledge, this is an egocentric cue which has never been explored before for object recognition.

3 Egocentric Object Recognition

The functional goal of our egocentric computer vision system is to recognize the object being grasped or that is about to be grasped by the user's hand. Since the egocentric vision glasses we use have a pair of stereo cameras embedded in them, we are able to reconstruct dense 3D point clouds of the grasping hand (see Section 4 for details). Therefore, our visual feature extraction incorporates two processing channels, *i.e.*, the image channel and the hand (3D point cloud) channel, as shown in Figure 2. We introduce the detailed steps of feature extraction from these two channels, followed by different strategies that the object recognition can be conducted.

Feature extraction in the image channel. Given the pair of stereo images taken from the 3D egocentric camera, we densely extract SIFT descriptors from the pair of images. These SIFT descriptors are further quantized using a pre-trained visual vocabulary and subsequently aggregated to form a bag-of-features histogram. Then, recognition from the image channel can be carried out by training SVM classifiers from the bag-of-features histogram representation using linear, RBF, histogram intersection, and chi-square kernels.

Feature extraction in the hand channel. As we have discussed, from the egocentric stereo image pairs, we reconstruct dense 3D point clouds of the grasping hand using the pipeline detailed in Section 4. After we obtain the dense 3D point cloud, we extract the histogram of oriented normal vectors (HONV) feature [26] from the point cloud. The HONV feature, proposed by Tang *et al.* [26], aggregates the surface normals into azimuth and elevation bins. It has achieved outstanding performance in recognizing objects from images with depth information. Please refer to [26] for details on the HONV feature. We adopt it here as a descriptor to characterize the shape of the grasping hand to predict the object to be grasped. The fact that HONV features are viewpoint dependent is desirable in our settings since the users observe their hands from the same vantage point when performing a given action. Subsequent recognition of the object can be conducted, once again, by training SVM classifiers on HONV features with various kernel functions.

There are two different strategies that the visual features extracted from the image channel and the 3D hand channel can be combined to build a more robust system to recognize the egocentric object, *i.e.*, the pre-fusion strategy and the post-fusion strategy.

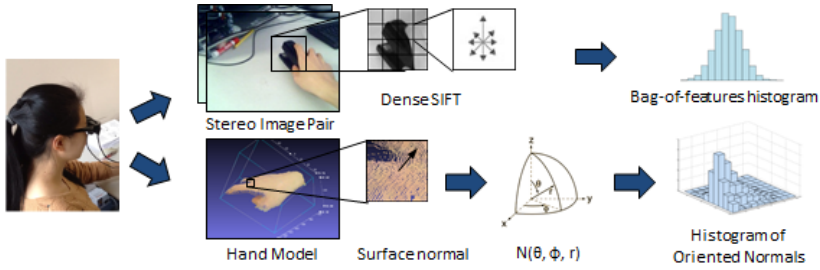


Fig. 2. The two feature processing pipelines in our egocentric object recognition system. The top row presents the feature representation from the image channel, where we extract a bag-of-words histogram as the feature representation for the object to be recognized based on densely extracted SIFT descriptor. The bottom row presents the feature representation from the 3D point clouds of the grasping hand. A histogram of oriented normals [26] is extracted to characterize the shape of the grasping hand.

The pre-fusion strategy. In pre-fusion, the features extracted from the image channel and the hand channel are concatenated together to form a larger feature vector to be used by the classification algorithm. The main challenge in the pre-fusion strategy is how to appropriately normalize features from the two different channels. We will examine the most appropriate normalization schemes in our experiments.

The post-fusion strategy. In post-fusion, we train two classifiers separately from features extracted from the two channels. Then the outputs from the two classifiers are linearly combined to form the final prediction by, for example, post training a linear SVM on the output scores from the two classifiers. Other more sophisticated combination algorithms, such as multiple kernel learning (MKL) SVM [28] can also be leveraged. Our experiments, however, do not show any improvement in performance compared with the simple combination scheme described above, despite the considerably more computationally expensive learning process.

We will conduct a systematic evaluation of these two fusion strategies in our experiments, and carefully examine if our conjecture, that the 3D shape of the grasping hand would provide beneficial information to recognize the object being grasped, is valid.

4 3D Hand Shape from Egocentric Stereo

In this section, we present our algorithm for segmenting and reconstructing the hands given a stereo pair of images. An illustration of the steps of our current pipeline can be seen in Fig. 3. Processing begins by detecting potential skin regions using a generic color model, which is iteratively adapted to match the appearance of skin in the input frames. Stereo matching is then restricted to skin regions to improve accuracy and speed. The output of this stage is a point cloud of the hand, pre or post contact with the object, that does not contain the latter. Segmentation plays a significant role in the quality of our reconstructed models. Accurate segmentation allows us to pursue a model-free

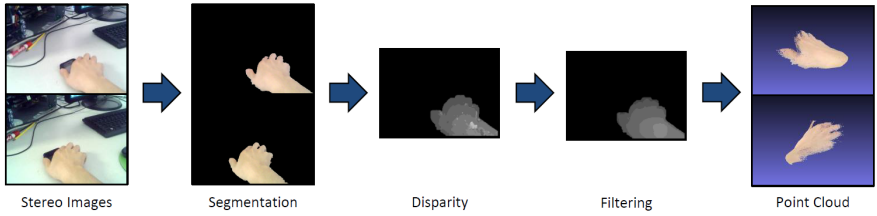
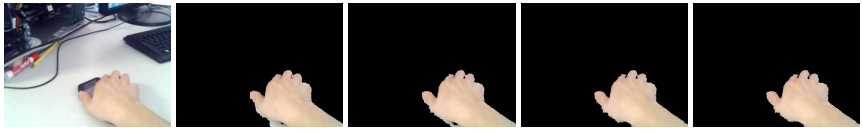
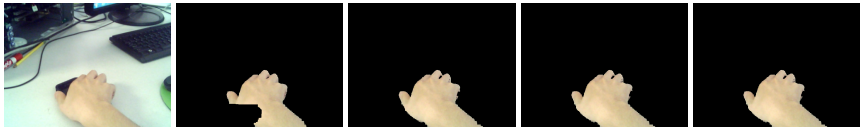


Fig. 3. Given a pair of images, the hand is iteratively segmented using graph cuts. Then, stereo matching, followed by bilateral filtering, is performed on the segmented images to generate point clouds which are used as inputs for recognition.



(a) Graph cut iterations after initial oversegmentation of the hand



(b) Graph cut iterations after initial undersegmentation of the hand

Fig. 4. The graph cut iterations from the initial segmentation using [9] to the final result.

approach, unlike previous work in stereo-based hand reconstruction [1, 2] that requires prior models which may have to be customized for each person.

Segmentation. First, we attempt to detect the hand by identifying potential skin regions using a generic skin-tone model [9]. We generate seeds for the hand region by selecting pixels with large probability for being skin. We also generate seeds for the background by selecting pixels that are extremely unlikely to be skin. The seeds are inserted as constraints in a binary MRF defined on a graph $G = (V \cup \{s, t\}, E)$. The weights of edges connecting nodes with the source and the sink are defined as follows:

$$e(v, s) = \begin{cases} w_{max} & P(\text{skin}) > \Theta_s \\ 0 & \text{otherwise} \end{cases}$$

$$e(v, t) = \begin{cases} w_{max} & P(\text{skin}) < \Theta_{\bar{s}} \\ 0 & \text{otherwise} \end{cases}$$

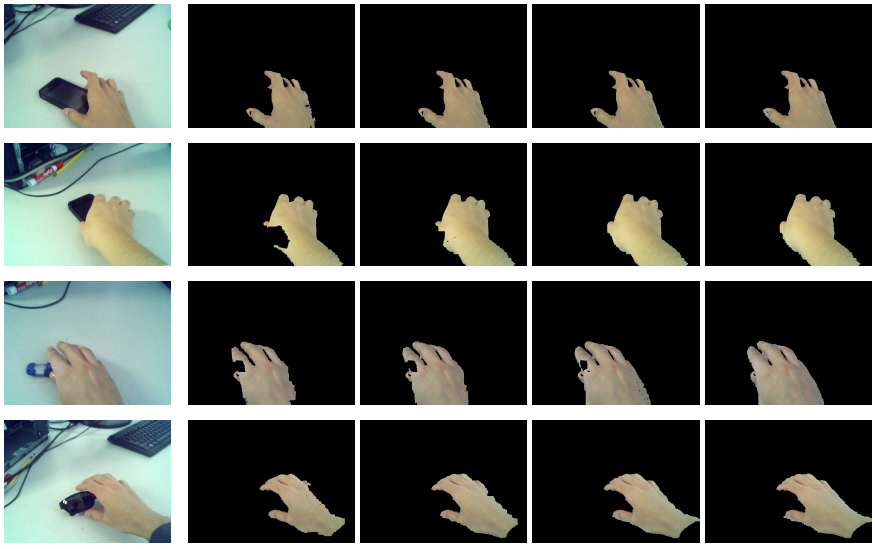


Fig. 5. Some more hand segmentation results from the iterative segmentation pipeline. From left to right, each row shows the original video frame and the segmentation results evolving with the iterations.

Edge weights between neighboring nodes v_p and v_q depend on the intensity difference between the corresponding pixels.

$$e(v_p, v_q) = \max(e^{-\frac{|I_{v_p} - I_{v_q}|}{\sigma}}, k)$$

We use the QPBO algorithm [20] to segment skin from non-skin regions. If multiple connected components labeled as skin are detected, we keep the one with the largest area. Due to the use of a generic skin-tone model, the results are not perfect, but because of the conservative construction of the graph weights, the detected skin regions tend to be pure. Using the initial segmentation, we estimate unimodal Gaussian color models for skin and non-skin pixels using maximum likelihood estimation. The new appearance models are used to recompute the probabilities of pixels being in the foreground or not and the graph cut is repeated with updated edge weights. We stop iterating when the result becomes stable, or the maximum number of iterations has been reached. It should be noted that due to the inherent difficulties in skin color detection, the model of Jones and Rehg [9] requires 16 Gaussian components for each class, but after our initial segmentation, one component is sufficient for a given video. The iterations are shown in Fig. 4, and more hand segmentation results from our proposed pipeline are shown in Fig. 5.

An additional step is required in order to restrict the reconstructed model to the palm and fingers, excluding the forearm which is part of the same segment if the person is not wearing long sleeves. We begin by detecting the orientation of the arm in the image, using the eigenvector of the hand region associated with the largest eigenvalue.

Fig. 6. Visualizations of the reconstructed dense 3D point clouds of the grasping hand along with original video frames. From left to right, each row presents the original video frame and the dense 3D point cloud of the grasping hand viewed from different angles and at different scales.

Given this direction of maximum elongation, we estimate the width orthogonally to it. We then reject all pixels that are not contained in a region whose length is at most 20% larger than the width of the hand starting from the furthest point, typically a finger. This prevents inconsistent hand segments, which may or may not contain parts of the forearm, from being passed on to the 3D reconstruction, feature extraction and classification modules.

3D Reconstruction. Leveraging the skin masks to restrict the search range for stereo matches allows us to use a simple matching technique. Moreover, since we only allow skin pixels to match other skin pixels and do not attempt to match the background, matches are sought for a fraction of the pixels and only a fraction of the disparity range is valid for each pixel. This reduces the probability of error and the amount of computation.

The results shown throughout this paper are obtained using Normalized Cross Correlation (NCC) in square windows. More sophisticated methods could have been used, but stereo matching for a single surface is much easier than the general problem of stereo, even in the presence of self-occlusions as in our case.

A bilateral filter is applied on the disparity map to correct disparity errors due to differences in illumination and camera response function between the two cameras, poor lighting, motion blur and lack of texture on the hand. The bilateral filter we use is defined according to:

$$d(\mathbf{x}) = \sum_{\mathbf{y}} w(\mathbf{x},\mathbf{y}) d(\mathbf{y}) e^{-\frac{(D(\mathbf{x},\mathbf{y}))^2}{2\sigma_s^2}} e^{-\frac{(C(\mathbf{x},\mathbf{y}))^2}{2\sigma_c^2}}, \quad (1)$$

where $w(\mathbf{x},\mathbf{y})$ is a rectangle window function, $D(\mathbf{x},\mathbf{y})$ is the distance function between $\mathbf{x}$ and $\mathbf{y}$,

$$D(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}_x - \mathbf{y}_x)^2 + (\mathbf{x}_y - \mathbf{y}_y)^2},$$

$C(\mathbf{x},\mathbf{y})$ is the color difference function between $\mathbf{x}$ and $\mathbf{y}$,

$$and C(\mathbf{x},\mathbf{y}) = |\mathbf{x}_R - \mathbf{y}_R| + |\mathbf{x}_G - \mathbf{y}_G| + |\mathbf{x}_B - \mathbf{y}_B|.$$

Finally, 3D point clouds are reconstructed based on the filtered disparity map and calibration information, obtained using images of a checkerboard and standard techniques [30]. Screenshots of the dense 3D point clouds of the grasping hand extracted from our proposed pipeline are shown in Fig. 6. As can be observed, the results are quite robust to unsatisfactory illumination conditions.

5 Data Collection

Fig. 7. The objects in our dataset: cup, statue, stapler, tea can, pen, iPhone, flash disk and sunglasses.

We have 8 objects from different categories in our dataset as shown in Fig. 7 and we asked 36 subjects to grasp them. For each user and each object, at least 3 grasps were performed. We did not instruct the users to grasp the objects using certain hand gestures, other than to request that they use their right hand. Between two consecutive grasps, the subject was asked to remove his or her hand completely from the object, so that every grasp could be considered independent. After all the data were collected, we manually labeled those frames in which the user's hand just touched the object, and the user's hand was about to leave from the object. Frames with heavy motion blur were not labeled.

6 Experiments

We performed a number of object recognition experiments on a dataset comprising 36 people grasping the 8 objects in a total of 767 frames. The data was split so that 25 people and 546 frames were in the training set and the remainder in the test set.

We note that in our experiments, all the frames of a single user are either in the training set or in the testing set, e.g. we do not perform testing on a user's frames while

different frames of the same user in the training set. Moreover, the number of video frames for each object category in the training and testing sets are shown in Table. 6.

	Cup	Pen	iPhone	Flash disk	Stapler	Sunglasses	Statue	Tea can
# Training frames	27	46	86	81	71	72	89	74
# Testing frames	16	21	35	27	36	25	37	24

Table 1. Number of video frames in the training set and testing set.

All experiments were performed with constant values for the parameters. For the MRF, the values used were $\Theta_s = 0.9$, $\Theta_{\bar{s}} = 0.01$, $\sigma = 25$ and $k = 0.05$. During stereo matching, we have a strong prior about the location of the hand with respect to the egocentric cameras which allows us to restrict the disparity range between 80 and 200. This interval has been effective for all subjects in our experiments. The NCC window size was 19×19 and the size of the bilateral filter was 50×50 with $\sigma_s = 30$ and $\sigma_c = 50$. To speed up the disparity calculation, we downsample the images to $\frac{1}{2}$ of their original size (305×225) and upsample them to their original size afterwards.

6.1 Baseline Results

We use SIFT and the bag-of-features (BoF) histogram as our baseline. For SIFT, we scaled the images into three different sizes and generated patches of size 16×16 every 16 pixels. For BoF, we compute a $1024 - D$ feature histogram for each frame. We use SVMs with different kernels for the classification. The results are shown in Table 3.

6.2 3D Features only

We also did experiments on classifying the objects using only the 3D hand features. In other words, in these experiments we attempted to classify objects without extracting any features from them. Results are shown in Table 3. The confusion matrices from the two top performing kernels, namely the histogram intersection kernel and the χ^2 are shown if Figure 8(b) and Figure 9(b). We achieved nearly 35% accuracy when classifying 8 objects. This demonstrates that features from the 3D hand shape carry information about the grasped object. Although the recognition accuracy is not as high as the results obtained from the image channel, it is much better than the random guess baseline, which is 12.5%.

6.3 Pre-fusion Results

We combined features from the two processing channels using a pre-fusion strategy according to Section 3. We tried three normalization schemes: normalization 1 where we normalize each feature in the dataset to be in $[0, 1]$, normalization 2 where we normalize each feature so that its sum of squares is equal to 1, and normalization 3 where we normalize each feature to have 0 mean and sum of squares equal to 1.

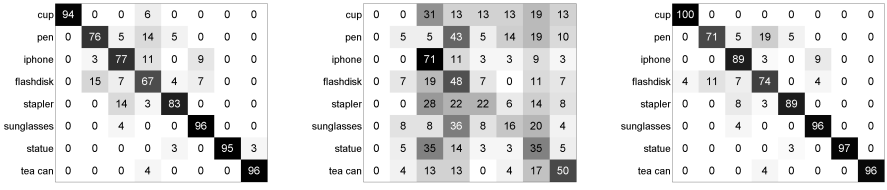
After normalizing the SIFT and HONV features, we just simply combine them into a longer vector. We tested the normalization scheme in histogram intersection kernel SVM. The result is shown in Table 2. Normalization 2 is used for all other experiments. Results are shown in Table 3 and Figure 8(c) and Figure 9(c).

Norm	SIFT	Pre-fusion
Norm1	84.61	87.78
Norm2	85.07	89.14
Norm3	85.97	89.14
No Norm	78.33	80.54

Table 2. Recognition accuracy using different normalization schemes

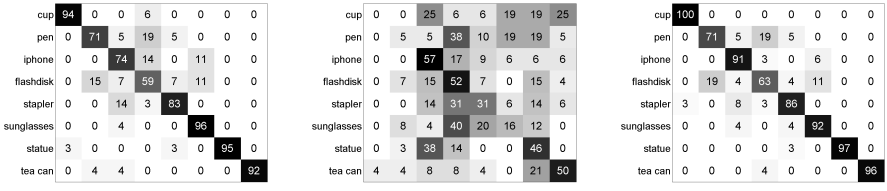
Kernel	SIFT	HONV-only	Pre-fusion
LN	78.73	30.77	83.26
HI	85.07	34.39	89.14
RBF	81.00	34.84	82.81
χ^2	82.81	35.75	87.33

Table 3. Recognition accuracy on classifying the objects



(a) SIFT 85.07% (b) HONV 34.39% (c) SIFT+HONV 89.14%

Fig. 8. The confusion matrix for histogram intersection kernel



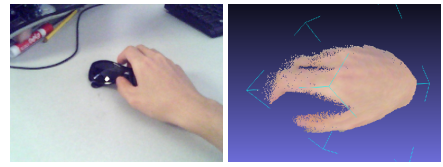
(a) SIFT 82.81% (b) HONV 35.75% (c) SIFT+HONV 87.33%

Fig. 9. The confusion matrix for Chi-Square kernel

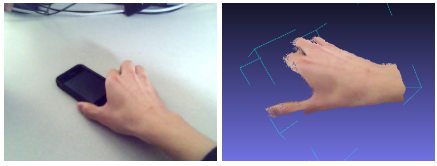
6.4 Discussion

From the results we can see that the additional 3D features of the hand have greatly improved the accuracy in classifying the objects. Using the histogram intersection kernel and the χ^2 kernel, improved the accuracy of 7 out of the 8 categories. Our experiment results also show that post-fusion does not improve the classification results. Please refer to Figure 8 and Figure 9 for details. This is a strong indication that the 3D

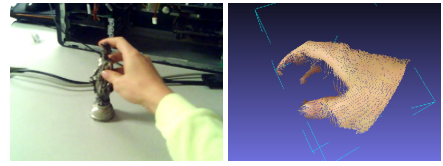
(a) Incorrectly classified as flash disk without 3D information



(b) Incorrectly Classified as Stapler without 3D information

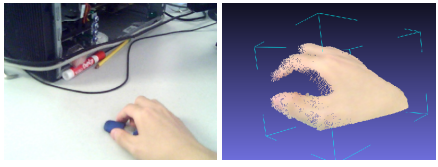


(c) Incorrectly Classified as flash disk without 3D information



(d) Incorrectly Classified as Cup without 3D information

Fig. 10. Some example frames that was incorrectly classified when using SIFT features, but was correct when we add additional 3D hand features.



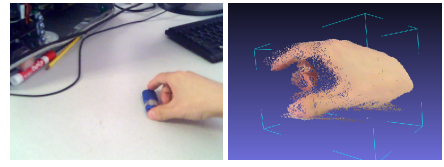


Fig. 11. These two frames of flash disks were incorrectly classified as stapler without using our 3D features. Due to occlusion from the hand, using SIFT feature only was not sufficient to classify them. The grasping shape of the hands are consistent for flash disks for almost all users, hence combining the 3D hand shape features helped to recognize them.

shape of the grasping hand can indeed provide beneficial complementary information to recognize the object being or about to be grasped.

We show some frames that were correctly classified after we used the additional 3D features, in Figure 10. In these cases, the object is severely occluded by the hand, but features from the hand itself can be used to correct the errors. Some more examples that were incorrectly recognized when we only use SIFT features extracted from the video frames are shown in Fig. 11 and Fig. 12. After we utilized the HONV feature extracted from the dense 3D point cloud of the hand, our recognition system made the correct prediction on these examples.

From our experiments we also observed that a few frames were correctly classified but when we added the 3D hand features, they were assigned the wrong labels. This is mainly because not all users use a common gesture to grasp the object. Figure 13 shows an example.

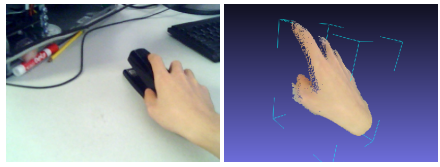
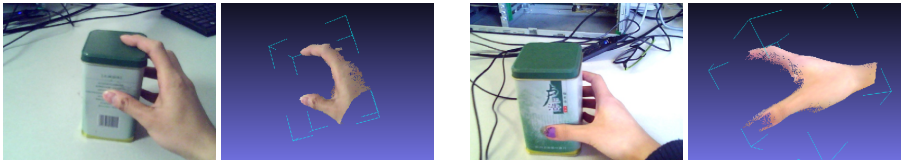


Fig. 12. This frame of a stapler was incorrectly classified as iPhone. In appearance, the stapler in this frame looks very like an iPhone. Because the grasping gesture of the hand is totally different when people are grasping an iPone, factoring the 3D features of the grasping hand again helped to recognize it correctly.



(a) Classified as Statue (b) Common gesture

Fig. 13. The left frame was incorrectly classified as a statue, because the user did not use the common gesture to grasp the object.

7 Conclusion and Future Work

We have shown the first results that demonstrate that objects can be recognized indirectly by observing the shape of the hand that is about to manipulate them form an egocentric perspective. Adding features from the hand to a standard appearance based recognition engine resulted in a boost in performance, even though the hand channel does not observe the object in the frame.

This study opens up many possibilities for future work. In the short term, one direction for improvement is the use of articulated hand models which can provide strong priors for pose estimation and robustness against occlusion. We also plan to consider and exploit the temporal aspect of our data collection. The temporal continuity and redundancy of the video are virtually guaranteed to lead to performance improvements compared with our current single-frame implementation. Moreover, we plan to tie segmentation and 3D reconstruction more closely together in a feedback loop.

In the longer term, we plan to investigate the inverse direction of the relationship between hand shape and object recognition. In this scenario, the objective would be to estimate the hand shape and measure how much it deviates from the ideal grasp that corresponds to the object the user intends to manipulate. Such a system would learn the user's grasps and observe their evolution in time as a means of assessing the functionality of the user's hand. As mentioned in the introduction, this is an ambitious goal, but successful research in this direction can have tremendous impact on a large fraction of the population.

References

1. Delamarre, Q., Faugeras, O.: Finding pose of hand in video images: a stereo-based approach. In: IEEE International Conference on Automatic Face and Gesture Recognition. pp. 585–590 (1998)
2. Dewaele, G., Devernay, F., Horaud, R.: Hand motion from 3d point trajectories and a smooth surface model. In: Pajdla, T., Matas, J. (eds.) Computer Vision - ECCV 2004, 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part I. pp. 495–507 (2004)
3. Fathi, A., Farhadi, A., Rehg, J.: Understanding egocentric activities. In: ICCV. pp. 407–414 (2011)
4. Fathi, A., Ren, X., Rehg, J.: Learning to recognize objects in egocentric activities. In: CVPR. pp. 3281–3288 (2011)
5. Fathi, A., Li, Y., Rehg, J.M.: Learning to recognize daily actions using gaze. In: Proc. European Conf. on Computer Vision (2012)
6. Go, A.S., Mozaffarian, D., Roger, V.L., Benjamin, E.J., Berry, J.D., Borden, W.B., Bravata, D.M., Dai, S., Ford, E.S., Fox, C.S., Franco, S., Fullerton, H.J., Gillespie, C., Hailpern, S.M., Heit, J.A., Howard, V.J., Huffman, M.D., Kissela, B.M., Kittner, S.J., Lackland, D.T., Lichtman, J.H., Lisabeth, L.D., Magid, D., Marcus, G.M., Marelli, A., Matchar, D.B., McGuire, D.K., Mohler, E.R., Moy, C.S., Mussolino, M.E., Nichol, G., Paynter, N.P., Schreiner, P.J., Sorlie, P.D., Stein, J., Turan, T.N., Virani, S.S., Wong, N.D., Woo, D., Turner, M.B.: Heart disease and stroke statistics–2013 update: A report from the american heart association. Circulation 127, 6–245 (January 2013)
7. Hankey, G.J., Jamrozik, K., Broadhurst, R.J., Forbes, S., Anderson, C.S.: Long-term disability after first-ever stroke and related prognostic factors in the perth community stroke study, 19891990. Stroke 33, 1034–1040 (2002)
8. Jojic, N., Perina, A., Murino, V.: Structural epitome: a way to summarize one's visual experience. In: NIPS (2010)
9. Jones, M., Rehg, J.: Statistical color models with application to skin detection. International Journal of Computer Vision 46(1), 81–96 (2002)
10. Kelly-Hayes, M., Robertson, J.T., Broderick, J.P., Duncan, P.W., Hershey, L.A., Roth, E.J., Thies, W.H., Trombly, C.A.: The American heart association stroke outcome classification: executive summary. Circulation 97, 2474–2478 (1998)
11. Kwakkel, G., Kollen, B.J., Wagenaar, R.C.: Long term effects of intensity of upper and lower limb training after stroke: a randomised trial. Journal of Neurology, Neurosurgery and Psychiatry 72, 473–479 (2002)
12. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: CVPR (2012)
13. Mayol, W., Murray, D.: Wearable hand activity recognition for event summarization. In: Ninth IEEE International Symposium on Wearable Computers. pp. 122–129 (2005)
14. Nowak, D.A.: The impact of stroke on the performance of grasping: Usefulness of kinetic and kinematic motion analysis. Neuroscience and Biobehavioral Reviews 32, 1439–1450 (2008)
15. Ogaki, K., Kitani, K., Sugano, Y., Sato, Y.: Coupling eye-motion and ego-motion features for first-person activity recognition. In: ECCV (2012)
16. Pentland, A.: Looking at people: sensing for ubiquitous and wearable computing. Pattern Analysis and Machine Intelligence, IEEE Transactions on 22(1), 107 –119 (jan 2000)
17. Pirsiavash, H., Remanan, D.: Detecting activities of daily living in first-person camera views. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (2012)

18. Ren, X., Philipose, M.: Egocentric recognition of handled objects: benchmark and analysis. In: First Workshop on Egocentric Vision (2009)
19. Ren, X., Gu, C.: Figure-ground segmentation improves handled object recognition in egocentric video. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. pp. 3137 –3144 (june 2010)
20. Rother, C., Kolmogorov, V., Lempitsky, V., Szummer, M.: Optimizing binary mrfs via extended roof duality. In: CVPR (2007)
21. Santello, M., Soechtino, J.F.: Gradual molding of the hand to object contours. Journal of Physiology 79(3), 1307–1320 (July 1998)
22. Schettino, L.F., Adamovich, S.V., Poizner, H.: Effects of object shape and visual feedback on hand configuration during grasping. Experimental Brain Research 151, 158–166 (2003)
23. Schiele, B., Oliver, N., Jebara, T., Pentland, A.: An interactive computer vision system DyPERS: Dynamic personal enhanced reality system. In: ICVS. pp. 51–65 (1999)
24. Simone, L.K., Sundarrajan, N., Luo, X., Jia, Y., Kamper, D.G.: A low cost instrumented glove for extended monitoring and functional hand assessment. Journal of Neuroscience Methods 160, 335–348 (2007)
25. Spriggs, E., De La Torre, F., Hebert, M.: Temporal segmentation and activity classification from first-person sensing. In: First Workshop on Egocentric Vision (2009)
26. Tang, S., Wang, X., Lv, X., Han, T.X., Keller, J., He, Z., Skubic, M., Lao, S.: Histogram of oriented normal vectors for object recognition with a depth sensor. In: ACCV (2012)
27. la Torre, F.D., Hodgins, J., Bargteil, A., Martin, X., Macey, J.: Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. Tech. rep., Carnegie Mellon University (2008)
28. Vishwanathan, S.V.N., Sun, Z., Theera-Ampornpunt, N., Varma, M.: Multiple kernel learning and the SMO algorithm. In: Advances in Neural Information Processing Systems (December 2010)
29. Winges, S.A., Weber, D.J., Santello, M.: The role of vision on hand preshaping during reach to grasp. Experimental Brain Research 152, 489–498 (2003)
30. Zhang, Z.: A Flexible New Technique for Camera Calibration. IEEE Trans. on Pattern Analysis and Machine Intelligence 22(11), 1330–1334 (2000)