Egocentric Recognition of Handled Objects: Benchmark and Analysis

Xiaofeng Ren and Matthai Philipose
Intel Research Seattle
1100 NE 45th Street, 6th Floor, Seattle, WA 98052
{xiaofeng.ren,matthai.philipose}@intel.com

Abstract

Recognizing objects being manipulated in hands can provide essential information about a person's activities and have far-reaching impacts on the application of vision in everyday life. The egocentric viewpoint from a wearable camera has unique advantages in recognizing handled objects, such as having a close view and seeing objects in their natural positions. We collect a comprehensive dataset and analyze the feasibilities and challenges of the egocentric recognition of handled objects.

We use a lapel-worn camera and record uncompressed video streams as human subjects manipulate objects in daily activities. We use 42 day-to-day objects that vary in size, shape, color and textureness. 10 video sequences are shot for each object under different illuminations and backgrounds. We use this dataset and a SIFT-based recognition system to analyze and quantitatively characterize the main challenges in egocentric object recognition, such as motion blur and hand occlusion, along with its unique constraints, such as hand color, location prior and temporal consistency. SIFT-based recognition has an average recognition rate of 12%*, and reaches* 20% *through enforcing temporal consistency. We use simulations to estimate the upper bound for SIFT-based recognition at* 64%*, the loss of accuracy due to background clutter at* 20%*, and that of hand occlusion at* 13%*. Our quantitative evaluations show that the egocentric recognition of handled objects is a challenging but feasible problem with many unique characteristics and many opportunities for future research.*

1. Introduction

Monitoring human activity is at the heart of professions such as elder care, worker training and lifestyle coaching. Recent interests in automated support for these occupations have spurred research in developing perception-based systems that are capable of monitoring a variety of day-to-day activities in detail. The requirement for detail and variety means that traditional vision-based activity recognition, which focuses on body kinematics and location, is insufficient. On the other hand, recent work in the ubiquitous computing and AI communities [11, 30] based on wireless sensors suggests that the identity and manipulation of objects used may serve as robust, detailed indicators for a large variety of daily activities from brushing teeth to medication.

Figure 1. Recognition of handled objects in egocentric video. We use a wearable video camera to capture a continuous view from the user and to recognize the objects being manipulated in hands.

Although wireless sensors have yielded promising results in detecting object use, it is often infeasible to attach sensors to objects. Vision based object-use detection, which does not need per-object instrumentation, is therefore attractive. The recent work of [32] highlighted the potential of object-use tracking using a wall-mounted camera for kitchen tasks. Using fixed environmental cameras, however, is problematic in many respects. First, installing (multiple) environmental cameras is often unacceptable both for aesthetic and economic reasons. Second, people's bodies may often occlude the objects being manipulated. Third, important details of objects may be indistinct because objects are far from camera. A promising way to counter all these issues is to allow the user to wear the camera: an *egocentric* camera looking out from the chest towards the hands should find many of the above problems inherently easier than an environmental camera would.

Egocentric vision has been extensively studied [19, 28, 21, 5, 15, 6] in many contexts such as face recognition, gesture recognition, visual SLAM or virtual reality. Object recognition from a wearable camera typically detects am-

978-1-4244-3993-5/09/$25.00 ©2009 IEEE

bient or static objects in the environment[27, 20]. The pioneering work in handled-object recognition from egocentric video is that of Mayol and Murray [18]. They show the feasibility of detecting handled objects to summarize context using a 5-object, 600-frame dataset. The limited complexity of this early dataset is not sufficient to address the variety of challenges in using a wearable camera, such as limited optics and resolution, illumination variation, motion and blur, hand occlusion, pose change or background clutter.

We have collected a large dataset to investigate the feasibility of identifying handled objects from egocentric video in real-life settings. Using a high-quality wearable video camera, we have collected 10 sequences each of 42 day-today objects being handled in a manner typical for each object (Figure 1). The footage is high resolution (1024x768), 24-bit color at 15 frames per second, with $100,000$ frames total. The objects vary significantly in size, shape, textureness, specularity and rigidity. The video shots vary in illumination and background clutter. We believe that our dataset provides a comprehensive, realistic and challenging benchmark for identifying object use in daily activities.

We carry out an empirical analysis of this dataset and provide quantitative characterizations of egocentric handled object recognition. We discuss major challenges such as scale and textureness variation, motion blur and hand occlusion. We also identify unique constraints such as skin color, object and hand location priors and temporal consistency. We report benchmarking results of a SIFT-based recognition system and quantify the challenges of occlusion and background clutter through simulated experiments.

2. Related Work

Cameras have long been used in wearable computing for various applications. A series of projects by Pentland and his associates are early examples. Starner [29] investigated how to identify gestures made by social partners. Schiele [27] examined detecting ambient objects and retrieving media memories. Choudhury [22] showed how to combine face recognition and voice recognition to identify conversation partners. Clarkson [5] used spectral techniques on video and synchronized sensor data to identify routines in daily life. A lot of progress has been made on visual SLAM where location and environmental structure are simultaneously estimated from egocentric video, with recent successes including [6, 24]. Many of these efforts can be viewed as in line with the *active vision* paradigm [2].

In the ubiquitous and pervasive computing communities, inferring the activities of a user is an important and popular problem and has been extensively studied, often using alternative sensors such as accelerometers [14, 12] or GPS [16]. Environmental cameras for detecting the manipulation of special pointing devices in instrumented environments have been investigated in the human-computer interaction community [31, 15]. Object-use provides rich information about a user's daily activities wel beyond that from accelerometers or locations. Several works [23, 11, 30] investigated the use of tiny wireless sensors affixed to objects and had been successful in detecting handled objects and activities.

It is a challenging but attractive approach using vision to recognize objects being manipulated by a user. The recent work of [32] recognized handled objects and associated kitchen tasks from a fixed wall-mounted camera. The use of an environmental camera restricted its applicability to monitor relatively small areas. Mayol and Murray [18] first studied the detection of handled objects using a wearable camera toward event detection and summarization. They recognized objects using color histograms and tested on a small dataset of 5 objects and 600 frames. We clearly need a much larger dataset to investigate the feasibilities of egocentric recognition of handled objects and activities.

Object recognition is a central problem in vision and has seen a huge amount of progress in the recent years [25], with highly influential works including [17, 3, 4, 9, 33]. One key to this success is the availability of large comprehensive datasets and rigorous benchmarking evaluations. Caltech 101 [8] is very popular and one of the first large-scale benchmarks for category-level recognition using Internet photos. PASCAL [7] is another popular recognition challenge with a smaller number of categories but more challenging instances. LabelMe [26] brings annotation tools online and gathers together community contributions.

We believe we are the first to establish a comprehensive and realistic dataset for the egocentric recognition of handled everyday objects and benchmark it on state-of-the-art recognition techniques. There are several prior empirical studies that are particularly relevant. [27] presented an early example of wearable object recognition using receptive field histograms, tested on a dataset of 103 objects with clean, centered views. [32] used a wall-mounted camera and tested with 33 everyday objects and 3 videos. The GroZi dataset [20] consisted of 120 grocery products and 11000 images captured *in situ* in grocery stores.

3. A Benchmark for Egocentric Recognition of Handled Objects

The theme of this work is the collection and analysis of a dataset for egocentric recognition of handled object. We mount a PointGrey Grasshopper Firewire camera on a subject's left shoulder, pointing downward at the area in front of the body. The camera is fixed with a tripod and a lapel so that it moves (mostly) rigidly with the body. We capture and store uncompressed 24-bit RGB video at 1024x$768, 15$ frames per second. The data rate is 35.4 MB/sec, stretching the limit of the sustained write rate of a single hard-drive. A 4-disk RAID-0 configuration is used to ensure smooth data

(a)

(b)

Figure 2. We collect a large dataset of egocentric handled-object recognition. (a) 42 everyday objects with large variations in size, shape, color and textureness. (b) We take 10 video shots of each object, with large variations in illumination and scene background.

transfer and to avoid frame drops. To battle motion blur, we set the shutter speed between $1/50$ and $1/100$.

We use a list of 42 objects commonly found in everyday life, with examples including: a milk bottle, a cereal box, a plastic wrap, a lunch box, a water cup, a garden scissor, a stapler, and a digital camera. We collect data from two subjects in five environments, with distinctively different backgrounds and illuminations. The objects are shown in Figure 2(a), and examples of the environmental settings are shown in Figure 2(b).

To expedite the process the data is not collected "in-situ" or from complete daily activities. We simulate these activities and instruct the subjects to handle the objects as they would have in real life. From example, for a water cup, the subject will simulate the activity of holding and drinking from the cup; for a plastic wrap, the subject will pick up the wrap box and tear off a piece of the wrap from the box.

The total amount of video data we have collected is about 120 minutes or $100,000$ frames, of which about $70,000$ frames contain objects, about 1600 per object. We have also taken clean exemplar photos of the objects under varying poses (as in Figure 2(a)), averaging 13.1 per object. In addition, 420 frames are annotated with groundtruth segmentations of object vs background, and 40 frames are marked with segmentations of hand vs background.

We focus on the identification of known objects, leaving the issue of category-level recognition for future work. Our main reasons are as follows: (1) identification is a good starting point for the egocentric recognition of handled objects and is challenging enough in real-life settings; (2) it is reasonable to assume that, in everyday life, most of the object instances we interact with are familiar; (3) intra-category variation is a topic for generic object recognition research and not unique to the egocentric viewpoint.

3.1. Benchmarking SIFT-based Recognition

We use a standard SIFT-based system, following the approach in [17], as the baseline to evaluate our dataset. Let $\{\mathbf{p}_i = (y_i, x_i), D_i\}$ be the set of SIFT keys and their descriptors in a frame, and let $\left\{\tilde{\mathbf{p}}_j = (\tilde{y}_j, \tilde{x}_j), \tilde{D}_j\right\}$ be the SIFT keys in an clean exemplar image. We find initial matchings between the features using SIFT distance and the ratio test, requiring the distance of the best match to be at most 0.6 of the second best match. In the set of initially matched features, we use RANSAC to search for the best perspective correspondence, i.e. the fundamental matrix $\mathbf{F}$ that allows most points to satisfy the epipolar constraint

$$[y_i \; x_i \; 1] \; \mathbf{F} \; [\tilde{y}_j \; \tilde{x}_j \; 1]^T = 0 \tag{1}$$

Having obtained the best $\mathbf{F}$, we go back and verify the SIFT matching, accepting those matches that fail the ratio test but are consistent with $\mathbf{F}$.

With the set of final matched features D_i and $\tilde{D}_j$, we convert distances between their SIFT descriptors to a scalar similarity score between the test image and the exemplar:

$$S = \sum \exp\left(-\alpha \|D_i - \tilde{D}_i\|\right) \tag{2}$$

For each exemplar z, we obtain a similarity S_z. The similarities $\{S_z\}$, after normalized to zero mean and unit variance (and capping at 4), are the input to a multi-class SVM classifier [1], which for m classes trains an all-pair set of $m(m-1)/2$ binary SVMs and use voting for prediction.

The benchmarking results of this system are shown in Figure 3, with an average recognition rate of $\mathbf{12.0}\%$. Comparing to a random chance of 2.4%, this result of 12.0% is reasonable but far from perfect, illustrating the challenges of our dataset and that of the egocentric recognition problem. As a comparison, we have also run the pyramid matching algorithm [9, 13] with the hierarchical clustering of SIFT features into words. The average accuracy there is 11.1%, comparable to that of our SIFT-based system.

4. Deconstructing Egocentric Recognition

Egocentric recognition of handled objects is a challenging problem that has many of its unique characteristics. The collection and analysis of our dataset has confirmed several well-known challenges and opportunities:

- Limited by form factor and cost, egocentric video is typically poor in quality comparing to photos. One would have to strike balances between shutter speed and sensor noise, or between resolution and frame rate.

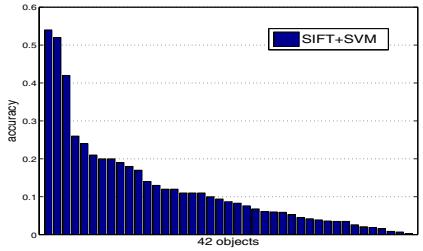


Figure 3. SIFT-based multi-class SVM classification on the 42-object dataset. Shown here are the recognition rate for each object. The graph shows huge variations in accuracy, ranging from over 50% for easy objects (e.g. cereal box) to mere chance 3% for hard objects (e.g. toothbrush or plastic spoon).

- Occlusion is a prevalent problem in egocentric video. Objects are being handled and are always occluded by the user's hands to varying degrees. Occlusion can be detrimental to recognition, especially for small objects.
- Everyday environments are often cluttered, filled with objects that may or may not be relevant to our task at hand. Figure-ground separation is non-trivial because the camera always moves and shakes with our body.

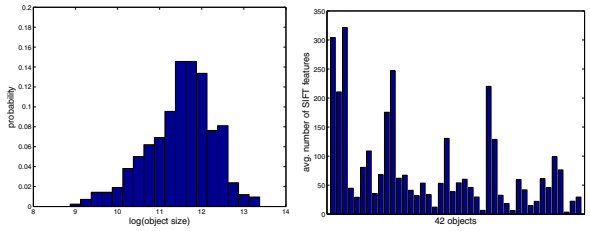
On the other hand, the egocentric setting has many interesting characteristics that help solve the recognition problem:

- The egocentric viewpoint is unique and provides strong constraints on location and scale. We manipulate objects in front of our body, making objects appear close to the center, at roughly a fixed distance.
- Hands are always present next to the objects. Skin color is known to be distinctive and can be very useful in locating the hands and in finding the objects.
- Egocentric data come in as continuous video streams. Video analysis can play an important role, helping both to separate scene background from moving objects and to improve recognition accuracy by enforcing the temporal consistency of labeling.

In the rest of the section, we will use our dataset, groundtruth labels and the SIFT-based recognition system to empirically quantify these challenges and opportunities.

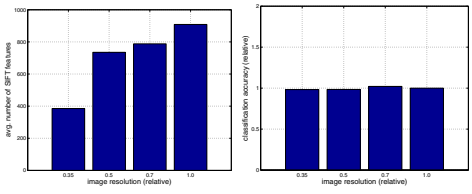
4.1. Challenges in egocentric recognition

Scale and texture variations. Our dataset covers a wide range of everyday objects. They vary significantly in scale and textureness. The object list includes very small objects (e.g. AA-battery, $<$ 3cm) and large ones (e.g. water filler, $>$ 30cm). This leads to a large variation in apparent object scale. Hand-marked object segmentations are used to compute object sizes. We show the empirical distribution of object scale in Figure 4(a). The variation is large, covering about two orders of magnitude, with the largest at 69% area of the entire frame, and the smallest at 0.9%.



(a) (b)

Figure 4. Object variations: (a) object size, with an average area of 10^5 pixels or 14% of the frame area; (b) textureness, measured by the average number of SIFT features detected in clean exemplars.



(a) (b)

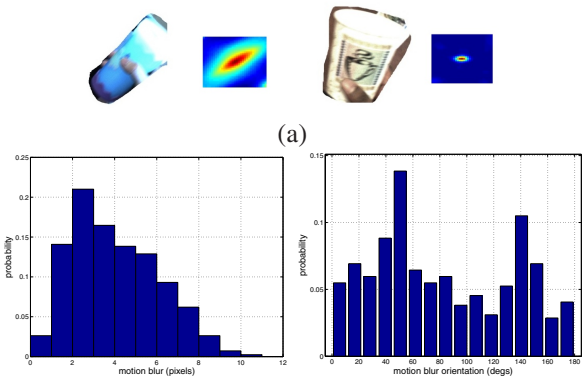
Figure 5. The impact of frame resolution: we downsize our video (from 1024x768) and evaluate the SIFT-SVM classification. (a) The average number of SIFT features detected, decreasing with image size. (b) Recognition rate relative to that of full-resolution, which stays about the same and does not decrease with resolution.

We also measure the textureness of the objects, by computing the average number of SIFT features detected in the clean exemplar photos, shown in Figure 4(b). There is also a large variation in how much texture we find in each object. Some objects, such as the water filler, lack texture and only host a small number of SIFT features. Half (21) of the objects in the dataset have an average number of SIFT features below 50. Lack of texture leads to poor SIFT-based classification on these objects.

Frame resolution is a major concern for egocentric video, as one cannot afford a resolution as high as that of modern photos. Is resolution a bottleneck for recognition performance? We empirically study the resolution issue by subsampling the original frames to a smaller size and evaluating our SIFT-based system on the subsampled frames. We experiment with four resolutions: full (1024x768), 0.7 (717x538), half (512x384), and 0.35 (359x269).

Figure 5(a) shows the average number of SIFT features for each resolution. The number drops with the resolution, as fine-scale features are now below the threshold and not detected. Somewhat surprisingly, the performance of the SIFT-based system does not drop with the resolution. Figure 5(b) shows the classification accuracy relative to the full-resolution case, based on $1/10$th of the training/testing data. We see that even at size 0.35, where the number of SIFT features is greatly reduced, the classification performance is close to that on full-resolution. These experiments indicate that resolution is not a bottleneck.

Motion blur is another major concern in egocentric



(a) (b) (c)

Figure 6. The analysis of motion blur: (a) examples of the non-parametric blur kernel estimated with maximum likelihood; (b) the amount of blur detected; (standard deviation of the blur kernel); (c) the orientation of otion blur, showing peaks at 45-degrees.

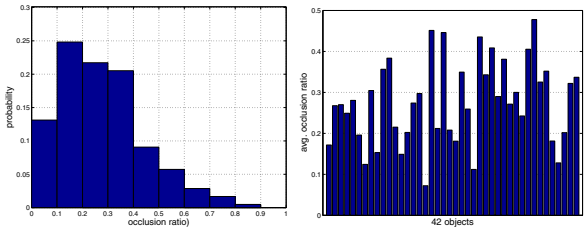
video. Lighting is typically limited, especially in-door, and hands can move rapidly when handling objects. Does manual shutter $(1/100)$ solve the motion blur problem? How much motion blur is there in the dataset?

We empirically analyze motion blur using blind image deconvolution. We use the standard maximum likelihood method [10] to estimate a non-parametric blur kernel for each image with groundtruth object mask, using the foreground object only (Figure 6(a)). We can then estimate the amount of blur (standard deviation of the blur kernel along the major axis) as well as the orientation of blur.

Figure 6(b) and (c) show the empirical distributions of the blur. We find that most motion blurs are small, with the peak at sigma $= 2$ pixels; the distribution also has a heavy tail with large blurs occurring up to 10 pixels. Interestingly, the orientation distribution also has two peaks, roughly at 45^o and 135^o. These two directions may arise from how the subjects move certain objects (e.g. cups and coke cans).

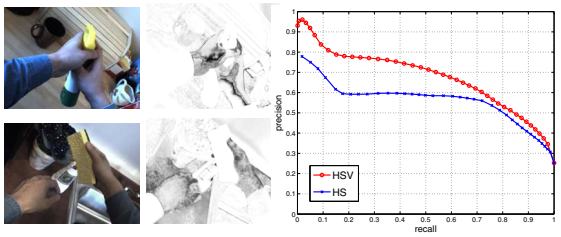
Occlusion-by-hand occurs in every instant as we manipulate objects. We measure the degree of occlusion as follows: our groundtruth object segmentation marks object boundaries even when they are occluded, and gives us the total 2D area of the object in the scene. To estimate how much of the object is occluded by the hands, we build a mixture-of-Gaussian hand color classifier and classify the pixels into hand vs non-hand (object). (More details of the hand color classifier are discussed in the next section.)

We find out that the hand color detector works reasonably well and suffices for our purpose, as we only need a crude estimate of hand occlusion. Figure 7(a) shows the empirical distribution of the occlusion ratio, hand area over total object area. It is approximately log normal, with the expected occlusion ratio around 20%, and the standard deviation about a factor of 2. Figure 7(b) shows the average occlusion ratio for individual objects, which varies signif-



(a) (b)

Figure 7. We combine groundtruth segmentation and hand-color model (Gaussian mixture) to calculate the occlusion ratio. (a) The distribution of the occlusion ratio. (b) The average occlusion ratio per object, with large variations across objects.



(a) (b)

Figure 8. We use mixture-of-Gaussian models, trained from groundtruth, to detect hand colors. (a) Hand color detection is reasonably good but not perfect. (b) Precision-recall curves evaluating the performance of hand color detection.

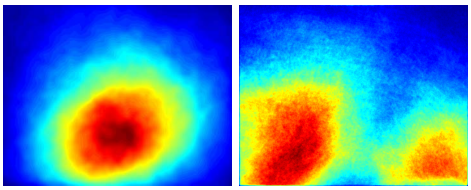
icantly depending on both the object size and the manners we interact with these objects.

4.2. Opportunities in egocentric recognition

Hand color is an important cue for handled objects as we perform most activities with hands, and skins color is known to be robust. We use the standard mixture-of-Gaussian model for hand color, in HSV space, with 10 components. Groundtruth segmentations of the hands (from 40 frames) are used to both train and evaluate the models.

Figure 8(a) shows a few examples of the Gaussian mixture hand color detection. Figure 8(b) shows the precision-recall curve for the mixture model in the HSV space, along with that in the HS space (discarding brightness). The precision-recall curves confirm our observation, that color works reasonably well, but far from perfect, with average precision around 67%. A simple color model cannot address many complications in real-life video, such as illumination, saturation, or the presence of near-skin-color objects. Detecting and tracking hands in these videos may prove plausible yet non-trivial.

On the other hand, we manipulate objects in the front of our body, so the canonical viewpoint of the egocentric camera provides strong information about where hands and objects are without having to track hands. Figure 9(a) shows the **location prior** for objects, where we take the groundtruth object segmentations and average their spatial



(a) (b)

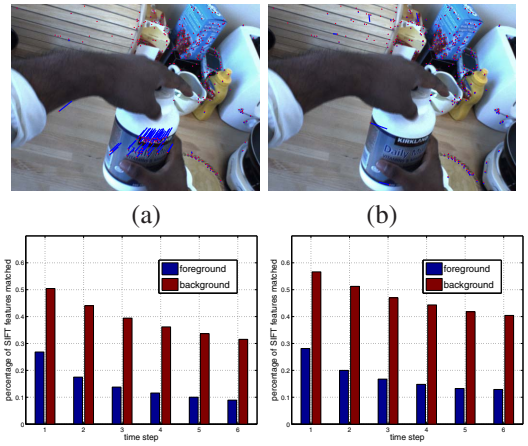
Figure 9. Location prior: (a) empirical distribution of object location, averaged from groundtruth segmentations; (b) empirical distribution of hand location, averaged from hand color detection.

support. Objects indeed tend to appear at the lower center region of the frame, and only rarely toward the periphery. Similarly, Figure 9(b) shows the prior distribution of hand location. Without enough hand segmentations, we compute the hand color probabilities using the HSV mixture model (Figure 8) and average them. We clearly see two hands in the frame, with a bigger left hand because it is closer to the camera on the left shoulder. Such prior information can be of great help for detecting and tracking objects and hands.

Motion is another important characteristic of our problem as we take as input a continuous video stream. The camera constantly moves, and the hands undergo complex motions. On the other hand, we assume the background is static, based on the observation that egocentric camera points downward from the user's body and are subject to only minor interference from moving objects (e.g. people).

We focus on the motion or temporal correspondence of the SIFT features detected in each frame. We use the same SIFT matching algorithm as in Section 3.1, between frames that are close in time, based on SIFT distances and the ratio test. We find two sets of matchings: one per-feature matching, where we only apply the ratio test and do not use any model model; and one consistent-motion matching, where we enforce the epipolar constraint to find feature matchings from a single perspective motion (as in Eq. 1). Figure 10(a) and (b) show examples of these two matchings. Note that not all matched features in (b) are in (a): once we obtain a RANSAC estimate for the fundamental matrix, we accept all matches that are consistent with this motion, even those that fail the ratio test. To accommodate both foreground and background motions, We extract two motion layers by sequentially estimating two fundamental matrices.

In Figure 10 (c,d) we show the percentage of features that are matched between two nearby frames, varying with the distance in time (1 to 6 frames apart). We use groundtruth object segmentations to compute the percentage for both foreground features and background features. As expected, the foreground motions are fast and complex, and only a small percentage of the features can be reliably matched across time, with 27% for adjacent frames, and 9% for 6 frames apart ($1/3$ second at 15 fps). In comparison, the background is static and moves slowly, and a much higher



(a) (b) (c) (d)

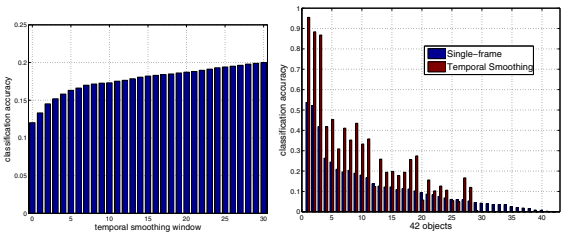
Figure 10. Matching SIFT features across frames: (a) SIFT features matched individually using the ratio test, red dots indicating the feature location and blue lines the displacement. (b) Features matched in a single motion layer by enforcing the epipolar constraint. (c) The percentage of features individually matched, for both foreground and background regions. (d) The percentage of features matched with the epipolar constraint.

percentage of features can be matched across time, with 51% for adjacent and 32% for 6 frames apart.

Comparing the results in (c) (no motion model) and (d) (perspective motion), we observe that enforcing a consistent motion improves background feature matching, allowing many features to match without passing the ratio test. The percentage of matched features increases, especially for longer-range matching (from 32% to 41% at distance 6). The effect is much smaller in the foreground case.

Finally, we exploit the continuous video setting and evaluate the benefit of enforcing the **temporal consistency of labeling**. Objects appear continuously in the videos and do not switch identity from frame to frame. We quantify this continuity using a simple voting approach: taking the classification results from the SIFT system, we look at a local temporal window of $[-K, K]$ frames, and choose the majority label in this window. Figure 11(a) shows the average classification accuracy vs the size of the voting window. We see that such a simple temporal smoothing greatly improves the performance, from average accuracy 12% with no smoothing to 20% at a window size of $K = 30$ (4 seconds). This illustrates the potential of temporal consistency.

In Figure 11(b) we show the per-class accuracy comparison of single-frame detection (12% average) to temporal smoothing (20% average). It is interesting to see that how aggressive temporal smoothing ($K = 30$) biases toward the easy and dominating objects, improving accuracy of many objects by a two-fold, while ignoring the hard cases. The hard objects, with per-frame labeling close to random, are smoothed out because their labels are rare and far apart.



(a) (b)

Figure 11. Exploiting temporal consistency of object labels. (a) Recognition improves when we average labels over a temporal window (using voting). It consistently improves with window size, from 12% to 20%. (b) Per-object comparison of single-frame recognition vs temporal smoothing. Accuracy greatly improves for most objects, while the hardest ones are "smoothed out".

4.3. Simulations for performance analysis

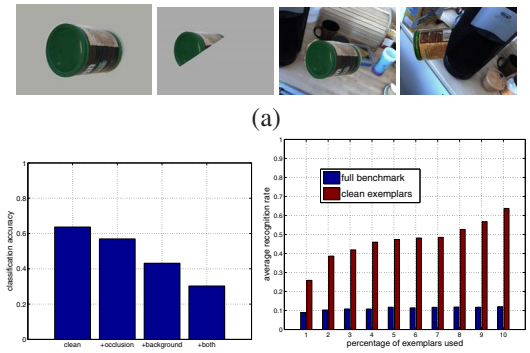
In this section we use simulated experiments to analyze **potential** recognition accuracies on our dataset. First, we want to bound the performance of SIFT-based recognition. We know it cannot be perfect as many objects in our dataset lack texture. To obtain an "upper bound", we use the SIFT recognition system and calculate the leave-one-out recognition rates for the **clean exemplar** images. The average accuracy in this case is 63.7%. Considering the lack of texture and the large variation in viewpoints, 63.7% provides a reasonable upper bound of how a SIFT-based matching system may expect to perform.

Next we analyze the impact of **occlusion**. As shown in Figure 7, we have a reasonable estimate of how much occlusion occurs in our dataset. We use the estimated log normal distribution of the occlusion ratio (average 20%, standard deviation a factor of 1.8) to simulate occlusion on clean exemplars. We randomly choose an occlusion ratio, randomly choose a projection direction, and remove part of the object from the clean exemplars (5500 total). The average recognition rate in this case is 57.0%. This is lower than the non-occlusion case 63.7% but not too far a drop.

Similarly we analyze the impact of **background clutter** using simulations. We pick random background scenes from the actual dataset (those with no object occurrence) and overlay clean exemplars on them. We then run the SIFT matching system, and obtain an average accuracy of 43.2%. Background clutter results in a 20% drop in accuracy.

We also combine both occlusion and background clutter in the simulation, occluding parts of objects and then adding a cluttered background. The average accuracy now drops to 30.3%. Interestingly, occlusion incurs a much larger decrease in accuracy in the cluttered case than the non-cluttered case. This confirms that SIFT matching is capable of handling partial occlusion to a certain extent, and this capability is reduced where clutter features abound.

Finally, we investigate whether the exemplar set (550 total) has enough coverage for object **pose variation**. We apply the k-medoids algorithm to select subsets of the exemplars, using pair-wise similarity scores (Eq. 2). As expected, recognition rates drop as we reduce the size of the exemplar set, for both the full benchmark and that of clean exemplars. (Figure 12(c)). The effect on the clean exemplar classification is much larger, from 63.7% of the full set to 25.9% using a subset of 55 exemplars (10%). In comparison the actual benchmark performance drops from 12.0% to 8.9%. These numbers show that pose coverage is indeed important, and we don't really have redundancy with 13.1 exemplars per object. On the other hand, mixed with many other difficulties, pose variation alone has a minor impact on benchmark performance.



(a) (b) (c)

Figure 12. (a) Examples of four simulation settings: clean exemplars, with occlusion, with background, and with occlusion+background. (b) Recognition results for the four cases, showing how much occlusion and background clutter affect accuracy. (c) We use k-medoids to select a subset of exemplar photos. Recognition rate decreases when we reduce the number of exemplars, much more for the case of clean photos.

5. Discussions

We have collected a dataset for the study of egocentric recognition of handled objects. Th dataset covers 42 everyday objects, 2 subjects and 5 environmental settings, with a total of 100,000 frames at 1024x768 and 24-bit RGB. This dataset provides a realistic and comprehensive benchmark for egocentric recognition and confirms many interesting challenges and opportunities for future research.

Egocentric recognition of handled objects in daily life is a challenging problem. A standard SIFT-based recognition system achieves an average accuracy of 12%. Main difficulties include scale and textureness variations, motion blur, background clutter, and hand occlusion. We have used the dataset to quantitatively explore the extent of these issues.

The egocentric viewpoint also enjoys unique advantages and presents many opportunities for progress. We have analyzed the dataset to show that hand color can be reliably detected, hand and objects appear at near the center of the view, background motion is consistent, and temporal con-

sistency can readily improve accuracy.

We have used clean exemplars and simulations to estimate the upper bound of SIFT-based recognition, at 64%, along with the impacts of background clutter (20%) and occlusion 13%. These results clearly suggest three directions for future research: (1) handling non-textured objects using edges and contours, such as shapecontext [3]; (2) using motion and location priors as well as hand color to remove background clutter; (3) and explicitly modeling occlusion with hand detection and removal. We believe that egocentric recognition of handled objects is challenging but feasible, and progress on this problem may soon lead to applications with far-reaching impacts on our everyday lives.

References

[1] OpenCV, http://sourceforge.net/projects/opencvlibrary/.

[2] J. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. *Int'l. J. Comp. Vision*, 1(4), 1988.

[3] S. Belongie, J. Malik, and J. Punicha. Shape matching and object recognition using shape contexts. *IEEE Trans. PAMI*, 24(4):509–522, 2002.

[4] A. Berg, T. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondence. In *CVPR*, volume 1, pages 26–33, 2005.

[5] B. Clarkson, K. Mase, and A. Pentland. Recognizing user context via wearable sensors. In *ISWC*, pages 69–76, 2000.

[6] A. Davison, I. Reid, N. Molton, and O. Stasse. Monoslam: Real-time single camera slam. *IEEE Trans. PAMI*, 29(6):1052–1067, 2007.

[7] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008). http://www.pascal-network.org/challenges/VOC/voc2008/workshop/.

[8] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In *Workshop on Generative-Model Based Vision*, 2004.

[9] K. Grauman and T. Darrel. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, pages 1458–65, 2005.

[10] T. Holmes. Blind deconvolution of quantum-limited incoherent imagery: maximum-likelihood approach. *J. Opt. Soc. Am.*, 9(7):1052–61, 1992.

[11] S. Intille, K. Larson, E. Tapia, J. Beaudin, P. Kaushik, J. Nawyn, and R. Rockinson. Using a live-in laboratory for ubiquitous computing research. In *Pervasive*, pages 349–365, 2006.

[12] N. Kern, B. Schiele, and A. Schmidt. Multi-sensor activity context detection for wearable computing. In *Ambient Intelligence*, pages 202–232, 2003.

[13] J. Lee. LIBPMK: A pyramid match toolkit. Technical Report MIT-CSAIL-TR-2008-17, MIT CSAIL, 2008.

[14] S. Lee and K. Mase. Activity and location recognition using wearable sensors. In *Pervasive Computing*, 2002.

[15] B. Leibe, T. Starner, W. Ribarsky, Z. Wartell, D. Krum, B. Singletary, and L. Hodges. The perceptive workbench: Toward spontaneous and natural interaction in semi-immersive virtual environments. In *VR*, pages 13–20, 2000.

[16] L. Liao, D. Fox, and H. Kautz. Location-based activity recognition using relational Markov networks. In *IJCAI*, 2005.

[17] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–57, 1999.

[18] W. Mayol and D. Murray. Wearable hand activity recognition for event summarization. In *ISWC*, pages 122–9, 2005.

[19] W. Mayol, B. Tordoff, and D. Murray. Wearable visual robots. In *ISWC*, pages 95–102, 2000.

[20] M. Merler, C. Galleguillos, and S. Belongie. Recognizing groceries in situ using in vitro training data. In *CVPR*, pages 1–8, 2007.

[21] A. Pentland and T. Choudhury. Face recognition for smart environments. *Computer*, 33(2):50–55, 2000.

[22] A. Pentland and T. Choudhury. Face recognition for smart environments. *IEEE Computer*, 33(2):50–55, 2000.

[23] M. Philipose, K. Fishkin, M. Perkowitz, D. Patterson, D. Fox, H. Kautz, and D. Hahnel. Inferring activities from interactions with objects. *IEEE Pervasive Computing*, 3(4):50–57, 2004.

[24] M. Pollefeys, D. Nistér, J.-M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S. J. Kim, P. Merrell, C. Salmi, S. N. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewénius, R. Yang, G. Welch, and H. Towles. Detailed real-time urban 3d reconstruction from video. *Int'l. J. Comp. Vision*, 78(2-3):143–167, 2008.

[25] J. Ponce, M. Hebert, C. Schmid, and A. Zisserman. *Toward Category-Level Object Recognition*. Springer-Verlag, 2007.

[26] B. Russell, A. Torralba, K. Murphy, and W. Freeman. LabelMe: a database and web-based tool for image annotation. Technical Report AIM-2005-025, MIT, 2005.

[27] B. Schiele, N. Oliver, T. Jebara, and A. Pentland. An interactive computer vision system - DyPERS: Dynamic personal enhanced reality system. In *ICVS*, pages 51–65, 1999.

[28] T. Starner, B. Schiele, and A. Pentland. Visual context awareness via wearable computing. In *ISWC*, pages 50–57, 1998.

[29] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(12):1371–1375, 1998.

[30] S. Wang, W. Pentney, A.-M. Popescu, T. Choudhury, and M. Philipose. Common sense based joint training of human activity recognizers. In *IJCAI*, pages 2237–42, 2007.

[31] A. Wilson and S. Shafer. Xwand: UI for intelligent spaces. In *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 545–52. ACM, 2003.

[32] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg. A scalable approach to activity recognition based on object use. In *ICCV*, pages 1–8. IEEE, 2007.

[33] J. Zhang, M. Marszaek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *Int'l. J. Comp. Vision*, 73(2):213–238, 2007.