Pixel-level Hand Detection in Ego-Centric Videos

Cheng Li
Tsinghua University
Beijing, China
licheng09@mails.tsinghua.edu.cn

Kris M. Kitani
Carnegie Mellon University
Pittsburgh, PA, USA
kkitani@cs.cmu.edu

Abstract

We address the task of pixel-level hand detection in the context of ego-centric cameras. Extracting hand regions in ego-centric videos is a critical step for understanding hand-object manipulation and analyzing hand-eye coordination. However, in contrast to traditional applications of hand detection, such as gesture interfaces or sign-language recognition, ego-centric videos present new challenges such as rapid changes in illuminations, significant camera motion and complex hand-object manipulations. To quantify the challenges and performance in this new domain, we present a fully labeled indoor/outdoor ego-centric hand detection benchmark dataset containing over 200 million labeled pixels, which contains hand images taken under various illumination conditions. Using both our dataset and a publicly available ego-centric indoors dataset, we give extensive analysis of detection performance using a wide range of local appearance features. Our analysis highlights the effectiveness of sparse features and the importance of modeling global illumination. We propose a modeling strategy based on our findings and show that our model outperforms several baseline approaches.

Figure 1. Pixel-level hand detection under varying illumination and hand pose.

1. Introduction

In this work we focus on the task of pixel-wise hand detection from video recorded with a wearable head-mounted camera. In contrast to a third-person point-of-view camera, such as a mounted surveillance camera or a TV camera, a first-person point-of-view wearable camera has exclusive access to first-person activities and is an ideal viewing perspective for analyzing fine motor skills such as hand-object manipulation or hand-eye coordination. Recently, the use of ego-centric video is re-emerging as a popular topic in computer vision and has shown promising results in such areas as understanding hand-eye coordination [5] and recognizing activities of daily living [17]. In order to achieve more detailed models of human interaction and object manipulation, it is important to detect hand regions with pixel-level accuracy. Hand detection is an important element of such tasks as gesture recognition, hand tracking, grasp recognition, action recognition and understanding hand-object interactions.

In contrast to previous work on hand detection, the ego-centric paradigm presents a new set of constraints and characteristics that introduce new challenges as well as unique properties that can be exploited for the task of first-person hand detection. Unlike static third-person point-of-view cameras typically used for gesture recognition or sign language analysis, the video acquired by a first-person camera undergoes large ego-motion because it is worn by the user. The mobile nature of the camera also results in images recorded over extreme transitions in lighting, such as walking from indoors to outdoors. As a result, the large im-

age displacement caused by body motion makes it very difficult to apply traditional image stabilization or background subtraction techniques. Similarly, large changes in illumination conditions induce large fluctuations in the appearance of hands. Fortunately, ego-centric videos also have the property of being user-specific, where images of hands and the physical world are always acquired with the *same camera* for the *same user*. This implies that the intrinsic color of the hands does not change drastically over time.

The purpose of this work is to identify and address the challenges of hand detection for first-person vision. To this end, we present a dataset of over 600 hand images taken under various illumination and different backgrounds (Figure 1). Each image is segmented at pixel resolution and the entire dataset contains over 200 million labeled pixels. Using this dataset and a publicly available indoor ego-centric dataset, we perform extensive tests to highlight the pros and cons of various widely-used local appearance features. We evaluate the value of modeling global illumination to generate an ensemble of hand region detectors conditioned on the illumination conditions of the scene. Based on our finding, we propose a model using sparse feature selection and an illumination-dependent modeling strategy, and show that it out-performs several baseline approaches.

2. Related Work

We give a review of work that aims to generate pixel-level detections of hand regions from moving cameras. Approaches for detecting hand regions can be roughly divided into three approaches: (1) local appearance-based detection, (2) global appearance-based detection and (3) motion-based detection.

In many scenarios, local color is a simple yet strong feature for extracting hand regions and is the most classical approach for detecting skin-color regions [9]. Jones and Rehg [8] proposed a mixture of Gaussian model to model skin and non-skin regions. Their approach was shown to be effective for extracting skin regions in internet images. Color models have also been combined with trackers to take into account both the static and dynamic appearance of skin [16, 23, 1, 11].

Global appearance-based models detect hands using a global hand template, where dense or sparse hand templates are generated from a database 2D images [25] or 2D projections of a 3D hand model [18, 24, 15]. These methods can be especially efficient when only a small number of hand configurations need to be detected [10]. However, when hands must be detected in various configurations, this approach usually requires a search over a very large search space and it may be necessary to enforce a tracking framework to constrain the search.

Motion-based approaches explicitly take into account the ego-motion of the camera by assuming that hands (foreground) and the background have different motion or appearance statistics. The advantage of these motion-based approaches is that they require no supervision or training [22]. However, since there is no explicit modeling of the hand, objects being handled by hands are often detected as foreground. When there is no hand motion or camera motion, there is no way to disambiguate the foreground from the background. Methods that attempt to model moving backgrounds are effective when camera motion is limited and video stabilization methods can be used to apply classical background modeling techniques [7, 6]. However, when there is significant parallax caused by camera motion and nearby objects, it becomes very difficult to build robust background models.

In the greater context of activity analysis for ego-centric vision, the task of extracting hand regions with pixel-level accuracy will be a critical preprocessing step for many high-level tasks. As such, we have focused on a local appearance-based strategy due to the extremely dynamic nature of ego-centric videos. However, our work is certainly complementary to other approaches and can used as a strong cue for initialization.

3. Modeling Hand Appearance

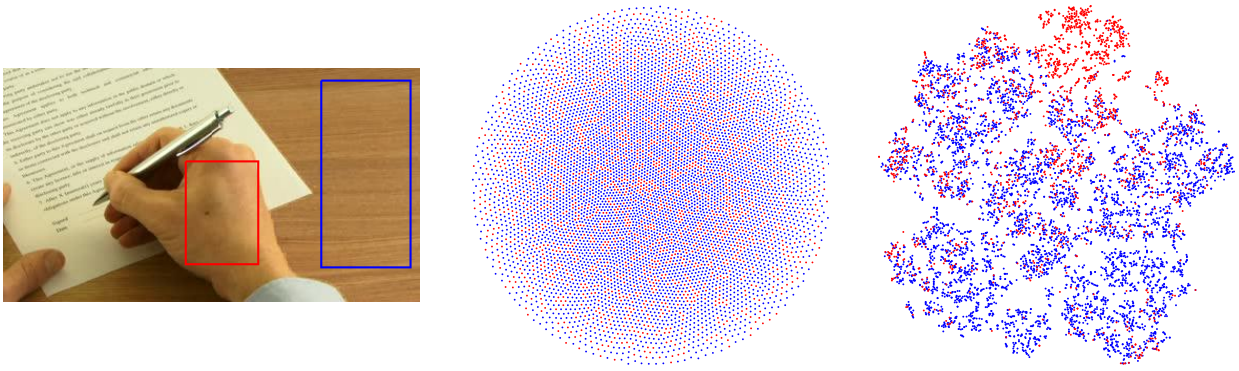
We are interested in understanding how local appearance and global illumination should be modeled to effectively detect hand regions over a diverse set of imaging conditions. To this end we evaluate a pool of widely used local appearance features, to understand how different features affect detection performance. We also examine the use of global appearance features as a means of representing changes in global illumination. We explain our local features and global appearance-based mixture model below.

3.1. Local Appearance Features

Color is a strong feature for detecting skin and has been the feature of choice for a majority of previous work [9]. Here we evaluate the RGB, HSV and LAB colorspaces which have been shown to be robust colorspaces for skin color detection. In contrast to previous work [8] using only a single pixel color features, we are interested in understanding how local color information (color of pixels surrounding the pixel of evaluation) contributes to detection performance.

We use the response of a bank of 48 Gabor filters (8 orientations, 3 scales, both real and imaginary components) to examine how local texture affects the discriminability of skin color regions. One of the typical limitations of color-based skin detection approaches is the difficulties encountered with attempting to discriminate against objects that share a similar color distribution to skin.

Figure 2 shows a visualization of the color feature space and the color+texture feature space for selected portions of



(a) Image regions (b) Color features (c) Color + texture features

Figure 2. Visualization of feature spaces with t-SNE [27]. Skin features in **red** and the desk features in **blue**. Texture features allows for better separation.

an image. Pixel features extracted from a portion of the hand (marked in red) and a portion of the desk (marked in blue) are visualized in 2D. Notice how the pixel features extracted from the hand and desk are completely overlapped in the color space (Figure 2b). However, by concatenating the response of 32 Gabor filters (4 orientations, 4 scales) to the color feature, we can see that the visualization of the feature space shows better separation between pixels of the hand and pixels of the desk (Figure 2c). This visualization suggests that low-level texture can help to disambiguate between hands and other similar colored objects.

Spatially varying local gradient histograms are the feature of choice for many computer vision tasks such as object recognition, image stitching and visual mapping because they efficiently capture invariant properties of local appearance. We evaluate the 36 dimensional HOG [4] descriptor and the 128 dimensional SIFT [12] descriptor. We expect that these gradient histogram descriptors will capture local contours of hands and also encode typical background appearance to help improve classification performance.

Binary tests randomly selected from small local image patches indirectly encode texture and gradients, and have been proposed as a more efficient way of encoding local appearance similar to SIFT descriptors. We evaluate the 16 dimensional BRIEF [3] descriptor and a 32 dimensional ORB [20] descriptor to measure relative performance with respect to the task of hand region detection.

The use of small clusters of pixels, better known as superpixels, is a preprocessing step used for tasks such as image segmentation and appearance modeling for tracking [19, 28]. Since superpixels aggregate local appearance information through the use of various types of histogram features (*e.g.* keypoints, color, texture) they are robust to pixel-level noise. In our evaluation we encode color, space and boundary distance as features within a super pixel. The color descriptor is the mean and covariance of the HSV values within a super pixel (3+6 dimensions). We use the normalized second-order moment as the shape descriptor (3 dimensions). The boundary distance descriptor is the distance to the nearest super pixel boundary, where we expect there to be an image edge (1 dimension). For our work, we use 900 super pixels which we found to perform best after a grid search over various values.

3.2. Global Appearance Modeling

Using a single hand detector to take into account the wide range of illumination variation and its effect on hand appearance is very challenging. We show a 2D embedding of different scenes via the global color histograms using t-SNE [27]. The visualization shows the large variance in hand appearance across changes in illumination.

In order to account for different illumination conditions induced by different environments (*e.g.* indoor, outdoor, stairway, kitchen, direct sunlight) we train a collection of regressors indexed by a global color histogram. The posterior distribution of a pixel x given a local appearance feature $\boldsymbol{l}$ and a global appearance feature $\boldsymbol{g}$, is computed by marginalizing over different scenes c,

$$p(x|\boldsymbol{l},\boldsymbol{g}) = \sum_c p(x|\boldsymbol{l},c)\, p(c|\boldsymbol{g}), \tag{1}$$

where $p(x|\boldsymbol{l},c)$ is the output of a discriminative global appearance-specific regressor and $p(c|\boldsymbol{g})$ is a conditional distribution of a scene c given a global appearance feature $\boldsymbol{g}$.

Different global appearance models are learned using k-means clustering on the HSV histogram of each training image and a separate random tree regressor is learned for each cluster. By using a histogram over all three channels of the HSV colorspace, each scene cluster encodes both the appearance of the scene and the illumination of the scene. Intuitively, we are modeling the fact that hands viewed under similar global appearance will share a similar distribution in

Figure 3. Visualization of image space shows the diversity of hand appearance and scene illumination. Different scene categories are learned by clustering image histograms. Each scene category is used to learn scene-specific hand region detectors.

the feature space. At test time, the conditional $p(c|\boldsymbol{g})$ is approximated using an uniform distribution over the n nearest models (in our experiments we use n=5) learned at training.

4. Analysis

First we evaluate the performance of each local appearance feature by running three different experiments. We evaluate (1) different patch sizes for color features, (2) feature selection over feature modality and (3) feature selection over sparse descriptor elements. Second, we show how learning a collection of classifiers indexed by different scene models can increase robustness to changes in illumination and significantly improve performance. Third, we compare our approach to several baseline approaches and show how our approach is better equipped to handle a wide range of ego-motion and object manipulation.

4.1. Dataset

We compare our approach on both in-house and publicly available egocentric videos. We generated two datasets to evaluate the robustness of our system to extreme changes in illumination and mild camera motion induced by walking and climbing stairs. We denote these two videos as EDSH1 and EDSH2. They have been recorded in a similar environment but each video follows a different path including both indoor and outdoor scenes. Both hands are purposefully extended outwards for the entire duration of the video to capture the change in skin color under varying illumination (direct sun light, office lights, staircase, shadows, *etc*.). EDSH1 and EDSH2 are 6 and 3 minutes long, respectively.

An additional video was taken in a kitchenette area which we denote as EDSH-kitchen, which features large amounts of ego-motion and hand deformations induced by the activity of making tea. This video was designed to evaluate the performance of our approach under large variations in camera motion. All in-house videos were recorded at a resolution of 720p and a speed of 30 FPS. 442 labeled frames of EDSH1 dataset were used exclusively for model training.

We also compare our approach on a publicly available dataset of egocentric activities from the Georgia Tech Egocentric Activity (GTEA) dataset [6]. We used the foreground hand masks available on the project site to compute scores for their proposed hand detection algorithm. The labeled portion of the GTEA dataset includes results for a single user performing three activities, which include making tea, making a peanut butter sandwich and making coffee. Since the GTEA dataset was created primarily as an activity recognition dataset, it contains very little camera motion and is taken in the same environment (*i.e*. sitting at a desk) under static illumination. The videos range from 2 to 4 minutes in length and the resolution was down-sampled to 720p to match our in-house data. The coffee sequence was used as training when testing on the tea and peanut butter sequence and the tea sequence was used for training when testing on the coffee sequence.

As a performance metric, the F-score (harmonic mean

Table 1. LAB color feature patch size evaluation using F_1-score.

Patch Size →	1×1	3×3	5×5	7×7	9×9
EDSH2	0.732	0.755	0.765	**0.769**	0.768
EDSH-Kitchen	0.769	0.794	0.800	0.796	**0.805**
GTEA-Coffee	0.858	0.873	0.880	0.884	**0.888**
GTEA-Tea	0.827	0.857	0.865	0.870	**0.880**
GTEA-Peanut	0.743	**0.767**	0.761	0.757	0.764

of the precision and recall rate) is used to quantify classification performance. We use a random forest regressor [2] with a maximum tree depth of 10, as our base model for all experiments.

4.2. Evaluating Local Color Features

In this experiment we examine the effects of increasing the spatial extent of color descriptors to detect hand regions. We extend the spatial extent of the color feature by vectorizing a $m \times m$ pixel patch of color values to encode local color information. We performed experiments using the RGB, HSV and LAB colorspaces but report only the LAB results since it performed the best. Table 1 shows the F-score across datasets for differing LAB features by patch sizes. Our results show that when color is the only feature type, modeling only a single pixel [8] does not always yield the best performance. Three out of the five datasets yielded the best results with a 9×9 image patch. On average, using a small patch can increase performance over a single pixel classifier by 5%. This result also confirms our intuition that observing the local context (*e.g.*, pixels surrounded by more skin-like pixels) should help to disambiguate hand regions.

4.3. Feature Performance over Modality

In this experiment we analyze the discriminative power of each feature modality using a forward feature selection evaluation criteria. In order to determine the features that have the greatest influence on our classification problem, we begin with an empty set of features and repeatedly add a new feature mode, such that it maximizes the performance on the training data using cross-validation.

Based on previous work we expect that color will play an influential role in detecting skin regions but we are also interested in how texture, gradient features and superpixel statistics can contribute to performance. Figure 4 shows the evolution of the F-measure on two test sets, by incrementally adding a new feature mode. We can see that initially color features (*i.e.* LAB, RGB, SP, HSV) are the most discriminative features, with the exception of the Gabor filter bank. The low-level Gabor filter bank is selected as the third most discriminative feature.

We see that high-order gradient features such as HOG and BRIEF are added after color features and enable an increase in overall performance latter in the pipeline. We learn here that while high-order texture features alone may not be very discriminative, they can improve performance when combined with the appropriate low-level features. We also observed a small initial dip in performance (Kitchen dataset) as more color features were added into the feature pool. The increase in the dimensionality of the color features initially causes over-fitting but the effect is counterbalanced (a mechanism of the RF learning algorithm) as more texture features are added to the pool. While we observed empirically that using multiple color spaces was useful for filtering out artificially colored objects like cardinal signs and red package markings, we also must be cautious of over-fitting when using many color-based features.

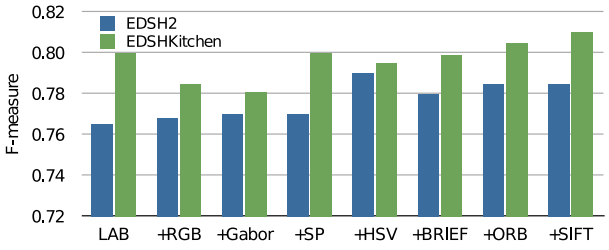


Figure 4. Performance of feature selection by feature mode.

4.4. Feature Performance using Sparse Features

Using the same feature selection process, we now evaluate each feature element (dimension) independently. This experiment gives us more insight into the interplay between individual feature elements. For efficiency reasons, we sample 100 random features at every iteration to generate a pool of candidate features for evaluation. Each regressor uses 10-fold cross validation over the training data, where each fold consists of a subset of at least one million data points from different image frames. The folds were generated over temporal windows of the training video to encourage more independence between folds.

Figure 5 shows the results of feature selection by selecting a single element from the pool of all 498 local appearance feature dimensions. In the top graph, we observe that performance plateaus after 20 to 30 feature elements. This suggests that only a few sparse features are needed to achieve near-optimal performance.

The bottom graph visualizes the number of dimensions used from each feature modality. Dark red denotes the highest count of features (20 features) and blue denotes a small count of features. The first four dimensions selected are color features, first from the LAB and then HSV. The Gabor filter response is the fifth dimension to be included. Notice that the number of LAB and HSV features continue to increase after other texture and gradient features are added. This indicates that local color information is discriminative when used together with texture and gradient features.

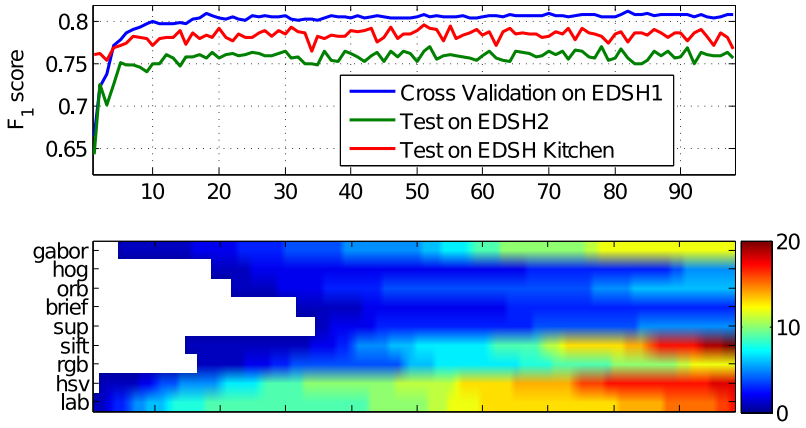


Figure 5. Performance of feature selection by adding single feature element per step (**top**) and distribution of used features over modes (**bottom**).

Other higher order gradient features are added between the 16th and 32nd iterations of the feature selection process. It is interesting that the elements of the BRIEF descriptor are infrequently utilized (mostly blue). This may be a reflection of its redundancy with the ORB descriptor. When the dimensionality of the feature is extended to 100, we observe that SIFT feature and HSV features are aggressively selected because they help to disambiguate the more difficult cases. This result reconfirms our previous result that higher-order gradient features are more discriminative when coupled with color features. When the application calls for only a small number of features, a sparse combination of HSV, LAB, Gabor and perhaps SIFT features would yield the best performance.

4.5. Number of Global Appearance Models

The appearance of the hands changes dramatically depending on the illumination of the scene, as can be seen in Figure 3. To address this dynamic nature of hand appearance, we have proposed a mixture model approach that adaptively selects the nearest set of detectors that were trained in a similar environment. In this experiment, we analyze the effect of the number of pre-trained detectors on the performance of the hand detector. The F-measure generated by different numbers of global appearance models is shown in Figure 6. We observe a big jump in performance after about 10 models and performance approximately plateaus thereafter. On our datasets, we observe that at least 10 different detectors (trained in different scenarios) are need to cover the variance in the appearance of the hands. It is also interesting to note that the performance is relatively stable for a wide range of k. Although we expect that the optimal number of scene clusters k will vary depending on the statistics of the dataset, we have gained an important insight that it is better to have multiple models of skin conditioned on the imaging conditions to achieve more robust performance.

4.6. Baseline Comparisons

We give the results of comparative analysis against several baseline approach in Table 2. We compare against four baseline approaches: (1) single-pixel color approach inspired by [8], (2) video stabilization approach inspired by [7] based on background modeling using affine alignment of image frames, (3) foreground modeling using feature trajectory-based projection of Sheikh *et al*. [22] and (4) hybrid approach [6] of Fathi *et al*. which uses a combination of video stabilization, gPb [14], super-pixel segmentation and graph cuts to extract hands.

The single-pixel color classifier is a random regressor trained only on single-pixel LAB color values. The background modeling approach uses a sequence of 15 frames and aligns them using an affine transformation. After alignment, pixels with high variance are considered to be foreground hand regions. The motion-based foreground modeling of Sheikh *et al*. [22] using the KLT tracker [26, 13] in place of the particle video tracker [21]. We found that [22] is greatly dependent on the accuracy of the feature tracking. Although the performance was significantly degraded

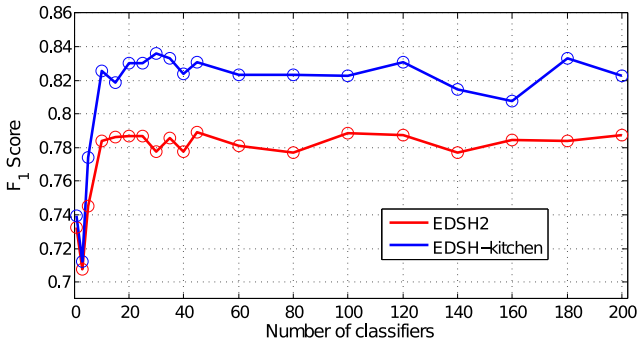


Figure 6. Performance for different number of scene categories.

by this change we include the performance as reference. The results of the hybrid approach [6] were computed using the hand masks available on the project website for the GTEA dataset.

Table 2 shows the results of our comparative experiment. We observed that the single pixel classifier works quite well compared to the other baselines. The video stabilization approach for background modeling is sensitive to ego-motion and therefore performs better on the GTEA dataset and worse on the EDSH dataset, which contains significant ego-motion. When we use our approach using all features, we see that our approach over-fits to the training data. When we use the results of feature selection and use only the top 50 features, we outperform all other baseline models. By incorporating 100 scene illumination models we get an additional increase in performance. We could not run this test on the GTEA dataset since there were not enough labeled images but we expect only modest gains since there is virtually no change in illumination.

Our approach generates stable detection around hand regions but also occasionally classifies small regions such as red cups and wooden table-tops. We applied a post-processing step, by keeping only the top three largest connect components (and removing all small contours) and we observed an additional increase in the F-measure (especially the precision rate). This version of the experiment suggests that running more sophisticated inference over the output of our pixel-wise model may also improve overall performance. On the most challenging EDSH2 dataset, our approach (*i.e*. post-processing, sparse features and scene-specific modeling) improves over a single color pixel approach score of 0.708 to 0.835, an 18% improvement. On average, compared to a single color feature approach, our our approach yields a 15% increase in performance over all datasets.

5. Conclusion

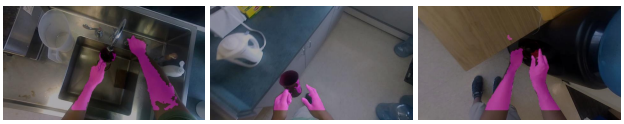
We have presented a thorough analysis of local appearance features for detecting hand regions. Our results have shown that using a sparse set of features improves the robustness of our approach and we have also shown that global appearance models can be used to adapt our detectors to changes in illumination (a prevalent phenomenon in wearable cameras). Our experiments have shown that a sparse 50 dimensional combination of color, texture and gradient histogram features can be used to accurately detect hands over varying illumination and hand poses. We have also shown that modeling scene-specific illumination models is necessary to deal with large changes in illumination. On average we observed a 15% increase in performance by applying our proposed approach on challenging indoor and outdoor datasets.

Dealing with extreme conditions such as complete saturation (*i.e*. parts of the scene and hands become pure white), very dark scenes, and high contrast cast shadows, is very challenging for local appearance based approaches (Figure 8). We believe these challenging situations can be address by the combined use of local appearance, global shape pri-

Baseline: Video stabilization

Proposed

Baseline: Feature trajectory projection [22] using KLT



Proposed

Baseline: Video stabilization+gPb+superpixel+graphcut [6]

Proposed

Figure 7. Qualitative comparison of hand region detection.

Figure 8. Failure cases caused by saturation, insufficient lighting and high contrast shadows.

Table 2. Comparative Results. F_1-score against baseline methods.

	EDSH		GTEA		
	2	kitchen	coffee	tea	peanut
Single pixel color [8]	0.708	0.787	0.837	0.804	0.730
Video stabilization [7]	0.211	0.213	0.376	0.305	0.310
Trajectory projection [22]	0.202	0.217	0.275	0.239	0.255
Stabilization+gPb+superpixel+CRF [6]	—	—-	0.713	0.812	0.727
Ours (d=498, k=1)	0.706	0.707	0.728	0.815	0.738
Ours (d=50, k=1)	0.781	0.808	0.884	0.873	0.815
Ours (d=50, k=100)	0.826	0.810	—	—	—
Ours (post-process)	**0.835**	**0.840**	**0.933**	**0.943**	**0.883**

ors and more expressive global illumination models.

This work has shown that hand region pixels can be detected with reasonable confidence for a wide range of illumination changes and hand deformations. Based on the findings of this work, we believe that our proposed pixel-level detection approach can be used to enable a variety of higher level tasks such hand tracking, gesture recognition, action recognition and manipulation analysis for first-person vision.

Acknowledgement

This research was supported in part by NSF QoLT ERC EEEC-0540865. The work was done while Li was a summer scholar at the Robotics Institute at CMU. Li was also supported by the Sparks Program and the Department of Physics at Tsinghua University.

References

[1] A. Argyros and M. Lourakis. Real-time tracking of multiple skin-colored objects with a possibly moving camera. *ECCV*, 2004. 2

[2] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 5

[3] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. BRIEF: Binary robust independent elementary features. In *ECCV*, 2010. 3

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005. 3

[5] A. Fathi, Y. Li, and J. Rehg. Learning to recognize daily actions using gaze. In *ECCV*, 2012. 1

[6] A. Fathi, X. Ren, and J. Rehg. Learning to recognize objects in egocentric activities. In *CVPR*, 2011. 2, 4, 6, 7, 8

[7] E. Hayman and J.-O. Eklundh. Statistical background subtraction for a mobile observer. In *ICCV*, 2003. 2, 6, 8

[8] M. Jones and J. Rehg. Statistical color models with application to skin detection. In *CVPR*, 1999. 2, 5, 6, 8

[9] P. Kakumanu, S. Makrogiannis, and N. Bourbakis. A survey of skin-color modeling and detection methods. *Pattern recognition*, 40(3):1106–1122, 2007. 2

[10] M. Kölsch and M. Turk. Robust hand detection. In *FG*, 2004. 2

[11] M. Kolsch and M. Turk. Hand tracking with flocks of features. In *CVPR*, 2005. 2

[12] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 3

[13] B. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. In *International joint conference on Artificial intelligence*, 1981. 6

[14] M. Maire, P. Arbeláez, C. Fowlkes, and J. Malik. Using contours to detect and localize junctions in natural images. In *CVPR*, 2008. 6

[15] I. Oikonomidis, N. Kyriazis, and A. Argyros. Markerless and efficient 26-DOF hand pose recovery. *ACCV*, 2011. 2

[16] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. *ECCV*, 2002. 2

[17] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012. 1

[18] J. Rehg and T. Kanade. Visual tracking of high DOF articulated structures: an application to human hand tracking. *ECCV*, 1994. 2

[19] X. Ren and J. Malik. Tracking as repeated figure/ground segmentation. In *CVPR*, 2007. 3

[20] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: an efficient alternative to SIFT or SURF. In *ICCV*, 2011. 3

[21] P. Sand and S. Teller. Particle video: Long-range motion estimation using point trajectories. *IJCV*, 80(1):72–91, 2008. 6

[22] Y. Sheikh, O. Javed, and T. Kanade. Background subtraction for freely moving cameras. In *ICCV*, 2009. 2, 6, 7, 8

[23] L. Sigal, S. Sclaroff, and V. Athitsos. Skin color-based video segmentation under time-varying illumination. *PAMI*, 26(7):862–877, 2004. 2

[24] B. Stenger, P. Mendonça, and R. Cipolla. Model-based 3D tracking of an articulated hand. In *CVPR*, 2001. 2

[25] E. Sudderth, M. Mandel, W. Freeman, and A. Willsky. Visual hand tracking using nonparametric belief propagation. In *Workshop on Generative Model Based Vision*, 2004. 2

[26] C. Tomasi and T. Kanade. *Detection and tracking of point features*. School of Computer Science, Carnegie Mellon Univ., 1991. 6

[27] L. Van der Maaten and G. Hinton. Visualizing data using t-SNE. *JMLR*, 9(2579-2605):85, 2008. 3

[28] S. Wang, H. Lu, F. Yang, and M. Yang. Superpixel tracking. In *ICCV*, 2011. 3