Hand Segmentation for Gesture Recognition in EGO-Vision

Giuseppe Serra
giuseppe.serra@unimore.it

Marco Camurri
marco.camurri@yahoo.it

Lorenzo Baraldi
baraldi.lorenzo@gmail.com

Michela Benedetti
michela.benedetti89@gmail.com

Rita Cucchiara
rita.cucchiara@unimore.it

Dipartimento di Ingegneria "Enzo Ferrari"
Università degli Studi di Modena e Reggio Emilia
Via Vignolese 905, 41125 Modena - Italy

ABSTRACT

Portable devices for first-person camera views will play a central role in future interactive systems. One necessary step for feasible human-computer guided activities is gesture recognition, preceded by a reliable hand segmentation from egocentric vision. In this work we provide a novel hand segmentation algorithm based on Random Forest superpixel classification that integrates light, time and space consistency. We also propose a gesture recognition method based Exemplar SVMs since it requires a only small set of positive sampels, hence it is well suitable for the egocentric video applications. Furthermore, this method is enhanced by using segmented images instead of full frames during test phase. Experimental results show that our hand segmentation algorithm outperforms the state-of-the-art approaches and improves the gesture recognition accuracy on both the publicly available EDSH dataset and our dataset designed for cultural heritage applications.

Categories and Subject Descriptors

I.4 [**Image Processing and Computer Vision**]: [Segmentation, Scene Analysis, Applications]; I.4.6 [**Segmentation**]: [Pixel classification]

Keywords

Hand segmentation, Gesture Recognition, Exemplar SVM, Random Forest, Ego-vision

1. INTRODUCTION AND RELATED WORK

The recent progresses in sensor development and mobile computing, and the increasing availability of wearable computers (e.g. Google Glass and Vuzix SmartGlass) has raised the interest of the research community toward the new field of egocentric vision.

IMMPD'13, October 22 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2505483.2505490.

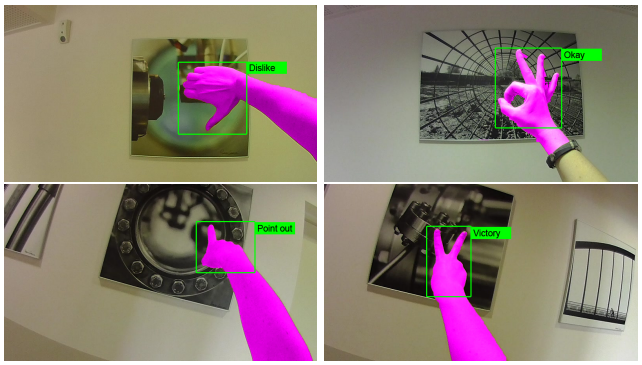


Figure 1: Sample results of the proposed hand segmentation and gesture recognition algorithm.

Egocentric vision, or ego-vision, is a paradigm that joints in the same loop human and wearable devices to augment the human vision capabilities by automatically process videos acquired with a first-person camera. Initial efforts have been made on the definition of methodologies to automatically understand human actions, objects and people interactions in egocentric vision. Systems that perceive what you perceive, see what you see and understand what you can understand or more will be used for many human augmentation applications. Some examples could be systems which recognize people around you, which understand dangerous situations, and provide assistance for activities such as surgery, sport, entertainment and more. The egocentric paradigm presents many new challenges, such as background clutter [14], large ego-motion and extreme transitions in lighting, but it also has some unique advantages. Egocentric videos are recorded from the same person over a continuous temporal space, there is no need to place multiple fixed cameras on the environment; furthermore objects and gestures are less likely to be occluded.

In this paper, we are interested in exploring the usage of ego-vison devices for cultural heritage domain: the museum experience, for example, could be enhanced by developing innovative human-machine interfaces such as new kinds of self-guided tour that can integrate information from the local

environment, Web and social medias. Furthermore these interfaces can help users to generate and share content in real time. In this scenario, hand detection and gesture recognition play a fundamental role, since this kind of applications should substitute other physical controller devices. Since gestures are strictly related to a specific scenario or application, it necessary to build a set of new classifiers using information gathered during a fast setup phase involving the user.

This problem, in ego-vision scenario, has been addressed only recently by the research community. Khan and Stoettinger in [8] studied color classification for skin segmentation and pointed out how color-based skin detection has many advantages, like potentially high processing speed, invariance against rotation, partial occlusion and pose change. The authors tested Bayesian Networks, Multilayers Perceptrons, AdaBoost, Naive Bayes, RBF Networks and Random Forest. They demonstrated that Random Forest classification obtains the highest F-score among all the other techniques. Fathi et al. [4] proposed a different approach to hand detection, exploiting the basic assumption that background is static in the world coordinate frame. Thus foreground objects are detected as to be the moving region respect to the background. An initial panorama of the background is required to discriminate between background and foreground regions: this is achieved by fitting a fundamental matrix to dense optical flow vectors. This approach is shown to be a robust tool for skin detection for hand segmentation in a limited indoor environment but it performs poorly with more unconstrained scenes. Li and Kitani [9] provide an historical overview about approaches for detecting hands from moving cameras. They define three categories: local appearance-based detection, global appearance-based detection, where a global template of hand is needed, and motion-based detection, which is based on the hypothesis that hands and background have different motion statistics. Motion-based detection approach requires no supervision nor training. On the other hand, this approach eventually identifies as hand an object manipulated by the user, since it moves together his hands. In addition they proposed a model with sparse feature selection which was shown to be an illumination-dependent strategy. To solve this issue, they trained a set of random forests indexed by a global color histogram, each one reflecting a different illumination condition. Recently Bagdanov et al. [2] propose a method to predict the status of the user hand by jointly exploiting depth and RGB imagery.

All the presented previous works present good characteristics, but lack of generality, since they take into account only few aspects to model user hand appearance and they are not integrated with a gesture recognition system. In this paper we present a novel method for hand segmentation and gesture recognition that can be used as basis for ego-vision applications. Hand detection is based on Random Forest classifiers learned by color and gradient features which are computed on superpixels. In order to improve the detection accuracy we present two strategies that incorporate temporal and spatial coherence: temporal smoothing and spatial consistency. Hand detection masks is then used as input for the gesture recognition step in order to reduce misclassification. We propose to use Exemplar SVMs to recognize gestures since it requires a only small set of positive samples, hence it well suitable for the ego-vision application domain. Experimental results show that our hand segmentation algorithm outperforms the state-of-the-art approaches on publicly available EDSH dataset [9] and demonstrate that segmentation step improves the gesture recognition accuracy. Fig. 1 shows sample results of the proposed hand segmentation and gesture recognition algorithm.
Here, we mainly provide our contributions:

- we define a novel hand segmentation algorithm that differently from [9], uses a superpixel features, and integrate not only illumination invariance but also temporal and spatial consistency improves the state-of-the-art results in the publicly available EDSH dataset.

- we develop of a gesture recognition algorithm based on Exemplar SVM technique that, even with a few positive samples, permits to reach competitive results.

2. METHOD OVERVIEW

2.1 Hand segmentation

Ego-vision applications require a fast and reliable segmentation of the hands; thus we propose to use random forest classifiers, as they are known to efficiently work even with large inputs [3]. Since using a per-pixel basis in label assignment has show to be inefficient [7], we adopt segmentation method which assign labels to superpixels, as suggested in [16]. This allows a complexity reduction of the problem and also gives better spatial support for aggregating features that could belong to the same object.

To extract superpixels for every frames we use the Simple Linear Iterative Clustering (SLIC) algorithm, proposed in [1] as memory efficient and highly accurate segmentation method. The SLIC super-pixel segmentation algorithm is a k-means-based local clustering of pixels in a 5D space, where Lab color values and pixel coordinates are used. A parallel implementation of the SLIC super-pixel algorithm is available in [13].

We represent superpixels by features to encode color and gradient information. As pointed out by previous works, HSV and LAB color spaces have been proven to be robust for skin detection. In particular, we describe each superpixel with mean and covariance matrix of its pixel values, and a 32-bin color histogram both in HSV and Lab color spaces. To discriminate between objects with a similar color distribution of skin we include following gradient information: Gabor feature obtained with 27 filters (nine orientations and three different scales: 7×7, 13×13, 19×19) and a simple histogram of gradients with nine bins.

2.1.1 *Illumination invariance, Temporal and Spatial Consistency*

In order to deal with different illumination conditions we train a collection of random forest classifiers indexed by a global HSV histogram with 32 bins, as described in [9]. Hence, training images are distributed among the classifiers by a k-means clustering on the feature space. By using a histogram over all three channels of the HSV color space, each scene cluster encodes both the appearance of the scene and its illumination. Intuitively, it models the fact that hands viewed under similar global appearance will share a similar distribution in the feature space. Given a feature vector $\mathbf{l}$ of a superpixel $\mathbf{s}$ and a global appearance feature $\mathbf{g}$, the posterior distribution of $\mathbf{s}$ is computed by marginalizing over different scenes c:

Figure 2: Comparison before (left image) and after (right image) Temporal smoothing.

Figure 3: Comparison before (left image) and after (right image) Spatial Consistency.

$$P(\mathbf{s}|\mathbf{l},\mathbf{g}) = \sum_{c} P(\mathbf{s}|\mathbf{l},c)P(c|\mathbf{g}), \quad (1)$$

where $P(\mathbf{s}|\mathbf{l},c)$ is the output of a global appearance-specific classifier and $P(c|\mathbf{g})$ is a conditional distribution of a scene c given a global appearance feature $\mathbf{g}$. In test phase, the conditional $P(c|\mathbf{g})$ is approximated using an uniform distribution over the five nearest models learned at training. It is important to underline that the optimal number of classifiers depends on the characteristics of the dataset: a training dataset with several different illumination conditions, taken both inside and outside, will need an higher number of classifiers than one taken indoor.

In addition to [9], we model the hand appearance not only considering illumination variations, but also including semantic coherence in time and space.

2.1.2 *Temporal smoothing*

We exploit temporal coherence to improve the foreground prediction of a pixel in a frame by a weighted combination of its previous frames, since past frames should affect the results prediction for the current frame.

The smoothing filter for a pixel $\mathbf{x}_t^i$ of a frame t (inspired by [10]) can thus be defined as follows:

$$P(\mathbf{x}_t^i = 1) = \sum_{k=0}^{d} w_k \big(P(\mathbf{x}_t^i = 1|\mathbf{x}_{t-k}^i = 1) \cdot \\ \cdot P(\mathbf{x}_{t-k}^i = 1|\mathbf{l_{t-k}}, \mathbf{g_{t-k}}) + P(\mathbf{x}_t^i = 1|\mathbf{x}_{t-k}^i = 0) \cdot \\ \cdot P(\mathbf{x}_{t-k}^i = 0|\mathbf{l_{t-k}}, \mathbf{g_{t-k}}) \big) \quad (2)$$

where $P(\mathbf{x}_{t-k}^i = 1|\mathbf{l_{t-k}}, \mathbf{g_{t-k}})$ is the posterior probability that a pixel in frame $t-k$ is marked as hand part and d is a number of past frames used. This likelihood can be defined as the probability $P(\mathbf{s}|\mathbf{l_{t-k}}, \mathbf{g_{t-k}})$, being $\mathbf{x}_t^i$ part of $\mathbf{s}$. In the same way, $P(\mathbf{x}_{t-k}^i = 0|\mathbf{l_{t-k}}, \mathbf{g_{t-k}})$ is defined as the probability $1 - P(\mathbf{s}|\mathbf{l}, \mathbf{g_{t-k}})$.

While $P(\mathbf{x}_t^i = 1|\mathbf{x}_{t-k}^i = 1)$ and $P(\mathbf{x}_t^i = 1|\mathbf{x}_{t-k}^i = 0)$ are prior probabilities estimated from the training set as follows:

$$P(\mathbf{x}_t^i = 1|\mathbf{x}_{t-k}^i = 1) = \frac{\#(\mathbf{x}_t^i = 1, \mathbf{x}_{t-k}^i = 1)}{\#(\mathbf{x}_{t-k}^i = 1)}$$

$$P(\mathbf{x}_t^i = 1|\mathbf{x}_{t-k}^i = 0) = \frac{\#(\mathbf{x}_t^i = 1, \mathbf{x}_{t-k}^i = 0)}{\#(\mathbf{x}_{t-k}^i = 0)}$$

where $\#(\mathbf{x}_{t-k}^i = 1)$ and $\#(\mathbf{x}_{t-k}^i = 0)$ are the number of times in which $\mathbf{x}_{t-k}$ belongs or not to a hand region, respectively; $\#(\mathbf{x}_t^i = 1, \mathbf{x}_{t-k}^i = 1)$ is the number of times that two pixels at the same location at frame t and $t-k$ belong to a hand part; similarly, $\#(\mathbf{x}_t^i = 1, \mathbf{x}_{t-k}^i = 0)$ is the number of times that a pixel in frame t belongs to a hand part and pixel in the same position in frame $t-k$ does not belong to a hand region. Figure 2 shows an example where temporal smoothing deletes blinking regions (i.e. the tea box brand and jar shadows on the right).

2.1.3 *Spatial Consistency*

Given pixels elaborated by the previous steps, we want to exploit spatial consistency to prune away small and isolated pixel groups that are unlikely to be part of hand regions and also aggregate bigger connected pixel groups.

For every pixel $\mathbf{x}$, we extract its posterior probability $P(\mathbf{x}_t^i)$ and use it as input for the GrabCut algorithm [15]. Each pixel with $P(\mathbf{x}_t^i) \geq 0.5$ is marked as foreground, otherwise it's considered as part of background. After the segmentation step, we discard all the small isolated regions that have an area of less than 5% of the frame and we keep only the three largest connected components.

In Figure 3 an example with and without applying the Spatial Consistency method is depicted; notice this technique allows to better aggregate superpixels that are near the principal blob region.

2.2 Hand status recognition

Given a subregion of the image space whose pixel are likely to be a segmented hand, we want now to recognize the hand configuration, picking from a finite set of possible gestures. Thus, we use the Exemplar SVM (ESVM) approach proposed by Malisiewicz et al. [11]. This method involves two steps: first, for each class an independent training of a finite and small set of positive examples (the "exemplars") versus a huge set of negative examples is performed. After the training stage each independent SVM classifier is tuned to detect its specific positive exemplar, thus a further step to aggregate classifiers of each class is necessary. In the second stage, since the outputs of each classifier are not comparable, a calibration by fitting a probability distribution to a held-out set of negative and positive samples is performed, as described in [12]. The calibration can be interpreted as a simple rescaling and shifting of the decision boundary, and does not affect the ordering of the score, allowing to compare the outputs of multiple independently-trained Exemplar-SVMs. Thus, multiple Exemplar-SVMs can be combined to provide a model for each class.

Since Ego Vision applications are highly interactive, their setup step must be fast (i.e. few positive examples can be acquired) but they allow an *a priori* massive collection of negative examples (e.g. for a museum application, a large footage without hands can be early acquired to train the classifier). Thus, ESVMs is well suitable for our application and was preferred to Latent SVM (LSVM) proposed by Felzenszwalb et al. [5] that have shown good performance in image classi-

Features	EDSH_2	EDSH_kitchen
HSV	0.752	0.801
+ LAB	0.754	0.808
+ LAB hist.	0.755	0.823
+ HSV hist.	0.755	0.823
+ Grad hist.	0.758	0.828
+ Gabor	**0.761**	**0.829**

Table 1: Performance by incrementally adding new features.

fication competitions. Although LSVM could well model the hand deformability property, it's more complex, it requires a more balanced set of negative and positive examples during the training stage and exhibits similar performance w.r.t. ESVM method [11].

3. EXPERIMENTAL RESULTS

To evaluate the performance of proposed method we tested it on two datasets: EDSH and EGO-HSGR. The recent publicly available EDSH dataset [9] consists in egocentric videos acquired to analyze performance of several hand detection methods. It consists in three videos (EDSH_1, used as train video, and EDSH_2 and EDSH_kitchen used as test videos) that contain indoor and outdoor scenes with large variations of illumination, mild camera motion induced by walking and climbing stairs. All videos are recorded at a resolution of 720p and a speed of 30FPS. The dataset includes segmentation masks of hands, but it is not comprehensive of gesture annotations.

In order to analyze the performance of our method to recognize gestures, we generated a new dataset which contains 12 videos of indoor scenes (EGO-HSGR); it includes segmentation masks and gesture annotations. Videos have been recorded with a Panasonic HX-A10 Wearable Camcorder at a resolution of 800×450 with a 25FPS in two different locations: a library and department's exhibition area.

The aim of this dataset is to reproduce an environment similar to a museum for human and object interaction: paintings and posters are hung on the walls, true masterpieces or either its virtual images; the visitor walks and sometimes stops in front of an object of interest performing some gestures to interact with next generation wearable devices. We identify five different gestures that are used commonly: *point out*, *like*, *dislike*, *ok* and *victory*. These can be associated to different action or used for record social experience. Fig. 4 shows some frame examples.

To evaluate performance of our pixel-level hand detector a subset of six videos are used (three for training and two for testing). Segmentation masks are provided every 25 frames for a total of 700 annotations. For gesture analysis we extract all the keyframes and we manually annotated them distinguishing between gestures. The F-score (harmonic mean of the precision and recall rate) is used to quantity hand detection performance, while gesture recognition is evaluated in terms of mAP (mean Average Precision).

3.1 Features performance

First, we examine the effectiveness of our features to discriminate between hand and non-hand superpixels. Table 1 shows performance in terms of F-measure on EDSH dataset with different feature combinations: firstly we describe each superpixel with mean and covariance matrix of its pixel values in HSV color space, then we do the same using LAB color space and we add color histograms. Lastly, we include a histogram of gradients and Gabor feature. In order to analyze how visual features impact on the performance, in this experiment we do not include the temporal and spatial context information by using a single random forest classifier. Note that although color information plays a fundamental role for hand detection, some ambiguities between hands and other similar colored object still remain; these can be reduced by adding features based on gradient histograms. In fact, the usage of the full descriptor slightly improves the performance.

Features	EDSH_2	EDSH_kitchen
II	0.789	0.831
II + TS	0.791	0.834
II + TS + SC	**0.852**	**0.901**

Table 2: Performance comparison considering Illumination Invariance (II), Time Smoothing (TS) and Spatial Consistency (SC).

	EDSH_2	EDSH_kitchen
Hayman et al. [6]	0.211	0.213
Jones et al. [7]	0.708	0.787
Li et al. [9]	0.835	0.840
Our method	**0.852**	**0.901**

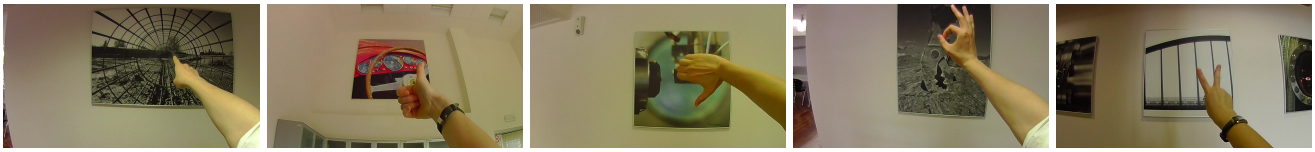
Table 3: Hand segmentation comparison with the state-of-the-art.

3.2 Temporal Smoothing and Spatial Consistency

In this experiment we validate the proposed techniques that take into account illumination variations, time dependence and spatial consistency. Table 2 shows the F-measure scores obtained on EDSH dataset incrementally adding Illumination Invariance (II), Time Smoothing (TS) and Spatial Consistency (SC). Note that there is a significant improvement in performance when all these three techniques are applied together. In particular, illumination invariance substantially increases the performance with respect to results obtained using only visual features and a single random forest classifier, while the improvement introduced by temporal smoothing is less pronounced. The main contribution is given by Spatial Consistency, that prunes away small and isolated pixel groups and merge spatially nearby regions, increasing the F-measure score of about six percentage points. The proposed technique is also tested in our EGO-HSGR dataset obtaining an F-measure score of 0.908 and 0.865 for the EGO-HSGR_4 and EGO-HSGR_5 videos.

3.3 Comparison to related methods

In Table 3 we compare our results to several approaches on EDSH dataset: a single-pixel color approach inspired by [7], a video stabilization approach based on background modeling using affine alignment of image frames inspired by [6] and an approach based on random forest, proposed by [9]. The single-pixel approach is a random regressor trained only using single-pixel LAB color values. The background modeling approach aligns sequences of 15 frames estimating



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

Figure 4: Our dataset consists of five gestures: a) point out; b) like; c) dislike; d) ok; e) victory.

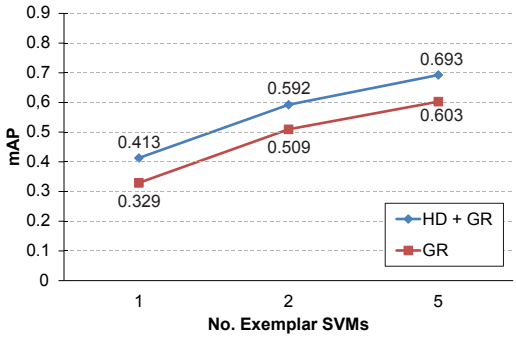


Figure 5: Mean Average Precision for different number of trained Exemplar SVMs.

Gestures	GR	HD + GR
Dislike	0.626	0.761
Point out	0.382	0.458
Like	0.817	0.754
Ok	0.698	0.951
Victory	0.490	0.541

Table 4: Average precision with five trained Exemplar SVMs per Gesture.

their mutual affine transformations; pixels with high variance are considered to be foreground hand regions. As can be seen, although the single-pixel approach is conceptually simple, is still quite effective. In addition, we observe that the low performance of the video stabilization approach is due to large ego-motion because the camera is worn by the user. The method recently proposed by [9] is more similar to our approach, but the use of superpixels, the selection of a new set of local features and the introduction of temporal and spatial consistency allow us to outperforms that results.

3.4 Hand Recognition

In order to evaluate the use of Exemplar SVM for gesture recognition we test our approach using a different number of trained classifiers on EGO-HSGR dataset. Since Ego-vision scenario requires a fast initial setup for the user, we analyze our approach with a very few positive samples. Figure 5 shows the mean Average Precision in two different settings: we apply our gesture recognition algorithm based on Exemplar SVM on the dataset frames directly (GR); we use the Exemplar SVM on the same frames processed by our hand segmentation algorithm (HD + GR). As expected, the mean Average Precision increases proportionally with the number of trained Exemplars in both settings. The performance obtained using our hand segmentation approach outperforms the gesture recognition without hand segmentation. In Table 4 we present the Average Precision per gesture obtained with five trained Exemplar SVM. Notice that using our hand detection technique provides a gain in performance for all gestures, except *Like*. This is due to fact *Like* gesture is more sensitive to erroneous hand segmentation that negatively effects the recognition step.

4. CONCLUSION

The work in this paper gives some initial but very promising results for the feasibility in adopting ego-vision to recognize human actions by first-person camera view. The proposed approach shows interesting results in term of accuracy for hand segmentation and gesture recognition which are novel in the panorama of multimedia-oriented computer vision dealing with the new paradigm of ego-vision.

Although the problem could become very challenging for a computer vision approach (e.g. cameras are not fixed and are freely moving, quality of images are poor and noisy due to the wearable camera), it can open new interesting scenarios for multimedia applications. For example user actions can be self-interpreted and integrated, since sensors are directly attached to the user.

5. REFERENCES

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2274–2282, 2012.

[2] A. D. Bagdanov, A. Del Bimbo, L. Seidenari, and L. Usai. Real-time hand status recognition from rgb-d imagery. In *Proc. of International Conference on Pattern Recognition (ICPR)*, 2012.

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

[4] A. Fathi, X. Ren, and J. Rehg. Learning to recognize objects in egocentric activities. In *Proc. of CVPR*, 2011.

[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[6] E. Hayman and J. Eklundh. Statistical background subtraction for a mobile observer. In *Proc. of ICCV*, 2003.

[7] M. Jones and J. Rehg. Statistical color models with application to skin detection. In *Proc. of CVPR*, 1999.

[8] R. Khan, A. Hanbury, and J. Stoettinger. Skin detection: A random forest approach. In *Proc. of ICIP*, 2010.

[9] C. Li and K. M. Kitani. Pixel-level hand detection in ego-centric videos. In *Proc. of CVPR*, 2013.

[10] K.-H. Liu, M.-F. Weng, C.-Y. Tseng, Y.-Y. Chuang, and M.-S. Chen. Association and temporal rule mining for post-filtering of semantic concept detection in video. *IEEE Transactions on Multimedia*, 10(2):240–251, 2008.

[11] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.

[12] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

[13] C. Y. Ren and I. Reid. gslic: a real-time implementation of slic superpixel segmentation. Technical report, University of Oxford, Department of Engineering Science, 2011.

[14] X. Ren and C. Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *Proc. of CVPR*, 2010.

[15] C. Rother, V. Kolmogorov, and A. Blake. "grabcut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, 2004.

[16] J. Tighe and S. Lazebnik. Superparsing - scalable nonparametric image parsing with superpixels. *International Journal of Computer Vision*, pages 329–349, 2013.