Egocentric Hand Detection Via Dynamic Region Growing

SHAO HUANG, University of Chinese Academy of Sciences, City University of Hong Kong
WEIQIANG WANG, University of Chinese Academy of Sciences
SHENGFENG HE, South China University of Technology
RYNSON W.H. LAU, City University of Hong Kong

Egocentric videos, which mainly record the activities carried out by the users of the wearable cameras, have drawn much research attentions in recent years. Due to its lengthy content, a large number of ego-related applications have been developed to abstract the captured videos. As the users are accustomed to interacting with the target objects using their own hands while their hands usually appear within their visual fields during the interaction, an egocentric hand detection step is involved in tasks like gesture recognition, action recognition and social interaction understanding. In this work, we propose a dynamic region growing approach for hand region detection in egocentric videos, by jointly considering hand-related motion and egocentric cues. We first determine seed regions that most likely belong to the hand, by analyzing the motion patterns across successive frames. The hand regions can then be located by extending from the seed regions, according to the scores computed for the adjacent superpixels. These scores are derived from four egocentric cues: contrast, location, position consistency and appearance continuity. We discuss how to apply the proposed method in real-life scenarios, where multiple hands irregularly appear and disappear from the videos. Experimental results on public datasets show that the proposed method achieves superior performance compared with the state-of-the-art methods, especially in complicated scenarios.

CCS Concepts: • **Computing methodologies** → **Scene understanding**; *Tracking*;

Additional Key Words and Phrases: Egocentric videos, egocentric hand detection, seed region generation, hand region growing

Fig. 1. Hand detection in hand-held and egocentric videos. Left: two frames from hand-held videos. Right: two frames from egocentric videos.

1 INTRODUCTION

In recent years, wearable cameras like the Google Glass are popular for capturing egocentric videos and have been used to record the daily life of the camera wearer in a first person point of view. The popularity of egocentric videos leads to novel applications such as gaze prediction [26], gaze analysis in multi-party conversations [21], summarization [16, 30], finger tracking [11], snap point prediction [40], facial attribute representation [39], and photographer identification [18]. The hands play a major role in a lot of applications, as humans are accustomed to interacting with objects using the hands [13]. This makes egocentric hand detection an essential task for egocentric video analysis.

To accurately locate the appeared hands, various methods have been proposed, either based on hand-crafted cues [13, 24] or deep architecture [3]. These works are designed for different scenarios, e.g., indoor [13, 24], outdoor [25, 34] scenes, or social interactions involving multiple people [3, 23]. In general, algorithms that can handle complex environments involving multiple persons impose high computational costs [3] in order to achieve satisfactory results. Algorithms that can only handle simple scenes likely fail in real-life scenarios [34]. In other words, existing methods cannot achieve a good balance between efficiency and accuracy. This problem motivates us to develop a new method that produces satisfactory performance at a low computational cost, especially in complex environments.

In this work, instead of assuming that only the wearer's hands appearing in the scene [25, 34], we observe that camera and hand motion patterns have distinct properties, which can be used to separate the hands from the video content. This is achieved by constructing an appropriate homography matrix among the correspondences of the successive frames. The seed regions, which are parts of the hands, are identified based on these motion patterns. The hand regions can then be progressively determined by dynamic region growing, starting from the seeds. The region growing algorithm is based on the proposed egocentric cues: contrast, location, position consistency and appearance continuity. We also discuss implementation issues on how to apply this work in complicated environments.

The main contributions of the proposed approach are summarized as follows:

- We propose an algorithm to automatically generate seed regions, by differentiating camera and hand motion patterns across successive frames. As these seeds are adaptively identified from the video itself, there is no need to make assumptions such as static background or limited number of hands.
- We propose a region growing algorithm by adding one adjacent superpixel with the highest score in each iteration. This score is computed based on the appearance and spatial constraints, in the current and previous frames.

- We propose to dynamically initialize, update and expire the appearance models to improve the hand detection in complex environments at low computational cost.

The rest of this paper is organized as follows. Section 2 introduces related works. Section 3 presents the proposed approach for egocentric hand detection and discusses implementation issues. Section 4 evaluates the proposed method on public datasets. Finally, Section 5 draws a conclusion of this work.

2 RELATED WORK

In this section, we first briefly summarize the latest progress made on egocentric video analysis. We then discuss related works on egocentric hand detection.

2.1 Egocentric Video Analysis

Egocentric videos are captured by a head-mounted camera. Thus, they largely differ from hand-held videos in many aspects. As the camera wearer moves his head, the content of the captured video changes, which indirectly reflects the wearer's actions. This distinct feature recently leads to many interesting research problems on egocentric videos. Some survey papers, e.g., [7] and Bambach [2], are published to summarize the latest progress. They explain how egocentric video analysis deals with the classical problems and its novel challenges, by dividing these works into multiple categories according to the various objectives. Recently, Molino *et al.* [9] compare the latest segmentation methods and selection algorithms. The popular egocentric datasets and evaluation metrics are also discussed to provide a comprehensive review. All the existing methods aim to analyze and understand egocentric videos by abstracting the content in different aspects (from the perspective of the wearer), thus saving human efforts in the time-wasting process of browsing the entire videos. Our work shares a similar objective via improving and accelerating egocentric hand detection, to benefit gesture recognition [4], pose recognition [35], action recognition [27] and understanding social interactions [3].

2.2 Egocentric Hand Detection

As shown in Fig. 1, the hands appear differently in hand-held and egocentric videos. Many researchers have recently made great efforts to investigate egocentric hand detection, by suggesting interesting assumptions and addressing the problem in confined environments, such as a single person performing daily activities indoors [13, 24] or outdoors [25, 34], or multiple persons involved in social interactions [3, 23]. Ren and Gu [34] are among the first to focus on hand detection in egocentric videos. They regard the problem as figure-ground segmentation, and locate hand regions by assuming that the hands have irregular optical flow patterns compared with the background. Fathi *et al.* [13] present the Georgia Tech Egocentric Activity (GTEA) dataset and conduct hand detection by incorporating multiple color features to segment the hands from objects. Li and Kitani [24] regard hand detection as a model recommendation task, where the n-best hand detectors are recommended based on the probe set, referring to a small amount of labeled data from the test distribution. To further address the limitation brought by the probe set, the virtual probes are proposed which can be automatically extracted from the test distribution. Another work presented by Li and Kitani [25] proposes a fully labeled indoor/outdoor egocentric hand detection benchmark dataset containing over 200 million labeled pixels, and gives extensive analysis of detection performance using a wide range of local appearance features. Lee *et al.* [23] model all hands in a social interaction instead of just the wearer's hands, and encode spatial

arrangements to disambiguate hand types using a probabilistic graphical model. Bambach *et al.* [3] further investigate the social interactions on their dataset named EgoHands, which contains 48 first-person videos of people interacting in realistic environments and also provides pixel-level ground truth for over 15000 hand instances. They use strong appearance models with Convolutional Neural Networks (CNNs), and introduce a simple candidate region generation approach to achieve state-of-the-art performance. Among the above works, [13, 34] assume that the background is static so that the optical flow patterns can be used for segmentation. Since the assumption is not true in real-world egocentric videos where the wearers likely walk around, it is difficult to achieve a satisfactory performance. Besides, [24, 25] assume no social interactions involved. Thus, their system generally performs poorly in scenes where the wearer is interacting with the partner, e.g., playing cards in the room. Although the deep-based method [3] achieves promising performance in complex environments, the disadvantage is the high computational cost and it requires a large amount of data for training. Hence, the performance is greatly degraded when insufficient training data is available in specific environments.

In contrast, our work does not assume a confined environment or the persons involved. The proposed method learns this information directly from the video. Our observation is that the hands are frequently moving and interacting with objects. As their motions are different from head motions, our method can obtain the seed regions with high confidence. By jointly considering appearance and spatial constraints, it can progressively locate the hand regions.

3 METHODOLOGY

Our region growing approach for egocentric hand detection is based on gradually extending the seed regions into hand regions. To improve computation efficiency, it is built on superpixels, instead of pixels [25], as motivated by our observation that the over-segmentation method like [1] could accurately detect the boundaries between hand regions and the background. For the rest of this section, we first identify the seed regions by differentiating camera and hand motion patterns across successive frames in Section 3.1. We then present the iterative procedure for extending the seed regions into hand regions in Section 3.2. Finally, we discuss some implementation issues in Section 3.3.

3.1 Seed Region Generation

Egocentric videos differ significantly from hand-held videos. The movement and the change of viewpoint are more unpredictable, since the motion of the wearer's head, like swinging, shaking and nodding, may occur anytime. Besides, while moving objects in the background may frequently appear in hand-held videos [20] and some egocentric videos captured in outdoor scenes, the current egocentric datasets for detection/recognition focus on what the wearers are doing and are captured mostly indoors, where moving background objects are indeed rare [12, 13, 25, 32]. As suggested by [13, 25], the motions in egocentric videos are mainly produced by the mounted camera and the hands. Hence, we first establish the candidate correspondences across successive frames, and then classify the hand-related correspondences, which distribution leads to seed regions. Fig. 2 shows the steps for determining the hand-related correspondences.

In the preprocessing step, a given frame is first over-segmented using SLIC [1], which is known as simple linear iterative clustering, and adapts a K-means clustering approach to efficiently generate superpixels. The ORB [36] descriptors are then computed across

(a) All correspondences $\mathbf{M}$

(b) Candidate correspondences $\mathbf{M}_r$
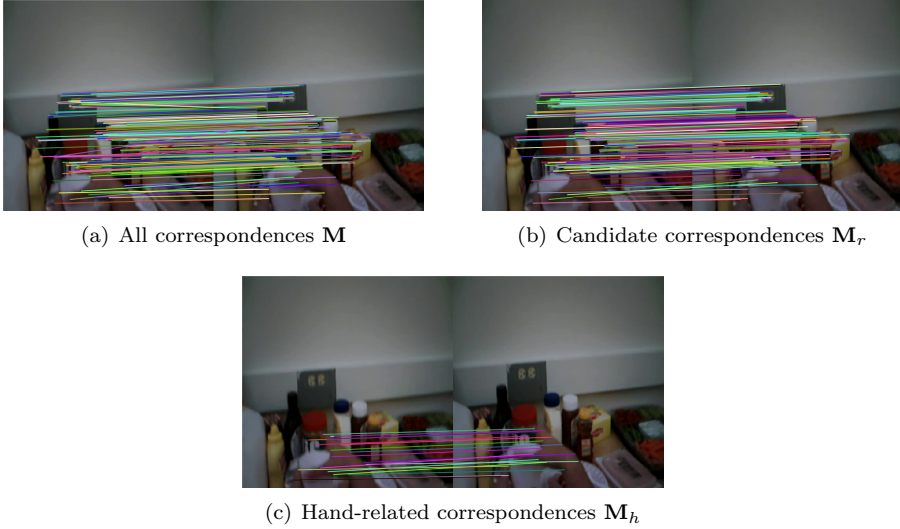

(c) Hand-related correspondences $\mathbf{M}_h$

Fig. 2. Steps for computing hand-related correspondences. The solid lines with different colors denote the correspondences across two successive frames.

successive frames. These descriptors are rotation invariant and resistant to noise, and can produce real-time performance. The extraction of these descriptors is much faster compared with SIFT [29] and SURF [5] descriptors. For a pair of successive frames p and q, the generated correspondences form the set $\mathbf{M} = \{\mathbf{m}_i | \mathbf{m}_i = (\mathbf{p}_i, \mathbf{q}_i), i = 1, 2, \cdots, N\}$, where $\mathbf{p}_i$ and $\mathbf{q}_i$ denote the ith 2D matching coordinates in p and q, respectively, as shown in Fig. 2(a).

Despite the efficiency of the ORB descriptors, some false correspondences may be introduced, which could be neglected because of the subtle motion caused by the small time interval, e.g., 1/30 of a second for the EgoHands dataset [3]. Thus, the set of obviously false correspondences $\mathbf{M}_f$ is considered as noise and directly excluded. Therefore,

$$\mathbf{m}_k \in \mathbf{M}_f, \ s.t. \ \|\mathbf{p}_k - \mathbf{q}_k\|_2 > \theta, \ k = 1, 2, \cdots, N, \tag{1}$$

where θ is the largest position difference allowed. It is adaptively learned for each video sequence, instead of for a single frame, as the behavior of the camera wearer, the environment and the configuration of the wearable camera tend to be consistent over the entire video. Suppose that the median value of all position differences in the video is ν. Then, $\theta = 10\nu$ is set in our system to differentiate false correspondences. Thus, the set of candidate correspondences $\mathbf{M}_r$ is generated. Therefore,

$$\mathbf{M}_r = \mathbf{M} \backslash \mathbf{M}_f. \tag{2}$$

It precisely reflects the motion between p and q, as shown in Fig. 2(b). For the camera motion pattern, all objects in the background are affected, e.g., bottles, chess, plates, as the visual field is determined by the head-mounted camera. For the hand motion pattern, only hand regions are affected. Hence, more correspondences agree with the camera motion pattern and distribute in the whole frame, while the hand-related correspondences distribute

in a few regions. We calculate the homography matrix from the candidate correspondences with RANSAC [14], and consider the camera-related correspondences $\mathbf{M}_c$ as "inliers".

Hence, the set of hand-relevant correspondences $\mathbf{M}_h$ is formulated as

$$\mathbf{M}_h = \mathbf{M}_r \backslash \mathbf{M}_c. \tag{3}$$

These correspondences denote where the hands possibly lie, as shown in Fig. 2(c). Unfortunately, this method may fail to extract all hand-related correspondences, and may occasionally contain noise when the camera and hand motion patterns are indistinguishable in some frames. Therefore, it is not suitable to directly regard the superpixels containing $\mathbf{m} \in \mathbf{M}_h$ as hand regions. Instead, we locate the seed regions based on the distribution of $\mathbf{M}_h$.

Suppose a set of superpixels is obtained, i.e., $\mathbf{S} = \{\mathbf{s}_k | k = 1, 2, \cdots, L\}$ after the over-segmentation step using SLIC [1] for each frame, and superpixel $\mathbf{s}_k$ contains λ_k hand-related correspondences, $k = 1, 2, \cdots, L$. The neighbors of a superpixel are defined as adjacent superpixels sharing parts of the boundaries. $\mathbf{s}_k$ is regarded as a peak if λ_k is the largest among all its neighbors and the neighbors' neighbors, to avoid multiple peaks in a local region. For these peaks, $\lambda_k \geq 0.1\mu$ is set to remove noise, which contains only few hand-related correspondences, and μ is adaptively assigned the average number of hand-related correspondences of all frames in the input video sequence. The retained peaks form the set of seed regions $\hat{\mathbf{S}} = \{\hat{\mathbf{s}}_u | u = 1, 2, \cdots, T\}$, as shown in Fig. 3(a). Each seed $\hat{\mathbf{s}}_u \in \hat{\mathbf{S}}$ is assumed to have the priority to form a part of the hand.

(a) Seed region (green)   (b) Located hand regions (blue)
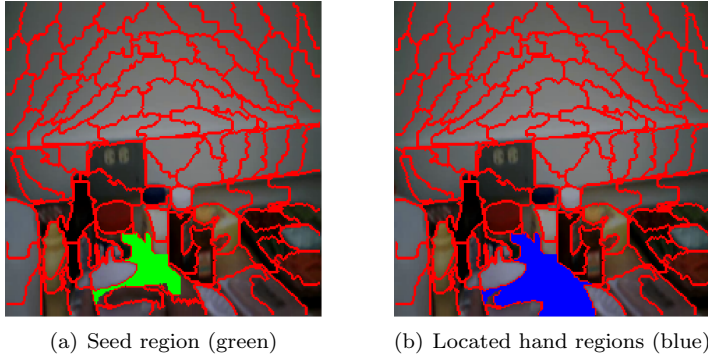
Fig. 3. The calculated seed region and the located hand regions.

3.2 Hand Region Growing

To locate the hand regions, we further utilize color and position information by jointly considering four egocentric cues: contrast, location, position consistency and appearance continuity. These cues are proposed based on appearance and spatial constraints, in the current frame as well as previous frames:

- **Contrast**: The contrast cue aims to build the appearance constraint in the current frame, between seeds $\hat{\mathbf{S}}$ and superpixels $\mathbf{S}$. A given frame is first converted into color space HSV. The normalized color histogram $\mathbf{h} = [h_1, h_2, \cdots, h_P]$ with P bins is then calculated for each superpixel, where $P = 16$ in our system for the sake of low computational cost. We use 8 bins for the H channel and 4 bins for S and V

channels. These bins are stacked to form the histogram. The score $S_1(\mathbf{s}_k, \hat{\mathbf{s}}_u)$ for superpixel $\mathbf{s}_k \in \mathbf{S}$ and seed $\hat{\mathbf{s}}_u \in \hat{\mathbf{S}}$ is calculated based on the relative entropy as:

$$S_1(\mathbf{s}_k, \hat{\mathbf{s}}_u) = \exp(-\sum_{j=1}^{P}(h_{k,j} \log \frac{h_{k,j}}{\hat{h}_{u,j}} + \hat{h}_{u,j} \log \frac{\hat{h}_{u,j}}{h_{k,j}})), \tag{4}$$

where $h_{k,j}$ and $\hat{h}_{u,j}$ denote the jth bin of the histogram on $\mathbf{s}_k$ and $\hat{\mathbf{s}}_u$, respectively.

- **Location**: The location cue aims to build the spatial constraint in the current frame, between seed $\hat{\mathbf{s}}_u \in \hat{\mathbf{S}}$ and superpixel $\mathbf{s}_k \in \mathbf{S}$. Suppose the centers of $\hat{\mathbf{s}}_u$ and $\mathbf{s}_k$ are denoted by $\hat{\mathbf{c}}_u$ and $\mathbf{c}_k$, respectively. This score $S_2(\mathbf{s}_k, \hat{\mathbf{s}}_u)$ is then defined as:

$$S_2(\mathbf{s}_k, \hat{\mathbf{s}}_u) = \exp(-\frac{\|\mathbf{c}_k - \hat{\mathbf{c}}_u\|_2}{l}), \tag{5}$$

where l is the length of frame diagonal, and $\|\cdot\|_2$ denotes the L2 norm;

- **Position Consistency**: As egocentric hand detection is conducted through the whole video instead of a single frame, temporal information helps build the spatial constraint between successive frames, by taking position consistency into account. For superpixel $\mathbf{s}_k \in \mathbf{S}$ and any $\bar{\mathbf{s}}'_w \in \bar{\mathbf{S}}'$ which denotes the calculated hand regions in the previous frame, the score $S_3(\mathbf{s}_k, \bar{\mathbf{s}}'_w)$ is defined as:

$$S_3(\mathbf{s}_k, \bar{\mathbf{s}}'_w) = \exp(\frac{\|\mathbf{c}_k - \bar{\mathbf{c}}'_w\|_2}{l}), \tag{6}$$

where $\bar{\mathbf{c}}'_w$ denotes the center of $\bar{\mathbf{s}}'_w$.

- **Appearance Continuity**: For egocentric videos, we observe that the appearance of a specific hand has only minor change over time, because the camera wearer would not be replaced and the illumination tends to be stable in a short period of time, as suggested by [3, 12, 13, 25, 32]. Multiple appearance models are built and updated as multiple hands may irregularly appear in the video. Similar to the contrast cue, the normalized color histogram is built for modeling appearance continuity. For superpixel $\mathbf{s}_k \in \mathbf{S}$ and any $\mathbf{a}_v^{t-1} \in \mathbf{A}^{t-1}$ which denotes the accumulated model at previous timestamp $t-1$, the score $S_4(\mathbf{s}_k, \mathbf{a}_v^{t-1})$ is defined as:

$$S_4(\mathbf{s}_k, \mathbf{a}_v^{t-1}) = \exp(-\sum_{j=1}^{P}(h_{k,j} \log \frac{h_{k,j}}{a_{v,j}^{t-1}} + a_{v,j}^{t-1} \log \frac{a_{v,j}^{t-1}}{h_{k,j}})), \tag{7}$$

where $a_{v,j}^{t-1}$ denotes the jth bin of the vth appearance model till previous timestamp $t-1$, and will be discussed in Section 3.3.

With the formulations for these proposed cues, score fusion is conducted to measure the likelihood of $\mathbf{s}_k$ being part of the hand. The contrast and location cues are considered jointly to avoid inconsistency. Hence, we jointly maximize $S_1(\mathbf{s}_k, \hat{\mathbf{s}}_u)$ and $S_2(\mathbf{s}_k, \hat{\mathbf{s}}_u)$ to select the optimal seed for $\mathbf{s}_k$. $S_3(\mathbf{s}_k, \bar{\mathbf{s}}'_w)$ and $S_4(\mathbf{s}_k, \mathbf{a}_v^{t-1})$ are maximized separately as $\bar{\mathbf{s}}'_w$ is independent of $\mathbf{a}_v^{t-1}$. Unlike the method proposed by Selective Search [38], where binary weights are used for score fusion, we adopt the weighted summation as:

$$S(\mathbf{s}_k) = \max_{\hat{\mathbf{s}}_u \in \hat{\mathbf{S}}}[\kappa_1 S_1(\mathbf{s}_k, \hat{\mathbf{s}}_u) + \kappa_2 S_2(\mathbf{s}_k, \hat{\mathbf{s}}_u)] + \max_{\bar{\mathbf{s}}'_w \in \bar{\mathbf{S}}'} \kappa_3 S_3(\mathbf{s}_k, \bar{\mathbf{s}}'_w) + \max_{\mathbf{a}_v^{t-1} \in \mathbf{A}^{t-1}} \kappa_4 S_4(\mathbf{s}_k, \mathbf{a}_v^{t-1}),$$
$$\tag{8}$$

where $\kappa_1 = \kappa_4 = 0.3$ and $\kappa_2 = \kappa_3 = 0.2$ are empirically set.

ALGORITHM 1: Pseudo-code for Hand Region Growing.

Input: each seed $\hat{\mathbf{s}}_u \in \hat{\mathbf{S}}$ with the corresponding score $S(\hat{\mathbf{s}}_u)$;
Output: the set of hand regions $\bar{\mathbf{S}}$;
initialize $\bar{\mathbf{S}} = \hat{\mathbf{S}}$, $\eta = \max\limits_{u=1,\cdots,T} S(\hat{\mathbf{s}}_u)$, $i = 0$;
while *true* **do**
    update the set of adjacent superpixels $\mathbf{N}(\bar{\mathbf{S}})$ of $\bar{\mathbf{S}}$;
    calculate the largest score $\epsilon = \max\limits_{\mathbf{s} \in \mathbf{N}(\bar{\mathbf{S}})} S(\mathbf{s})$;
    if $\epsilon < \alpha \times \eta$ **then**
       | break;
    end
    else
       update $i \leftarrow i + 1$;
       select the superpixel $\mathbf{s}_0 = \arg \max\limits_{\mathbf{s} \in \mathbf{N}(\bar{\mathbf{S}})} S(\mathbf{s})$;
       update $\bar{\mathbf{S}} \leftarrow \bar{\mathbf{S}} \cup \{\mathbf{s}_0\}$;
       update $\eta = \epsilon$;
    end
end
calculate connected components among $\bar{\mathbf{S}}$;
remove tiny ones of less than β pixels;

The hand regions are gradually located by extending from the seed regions $\hat{\mathbf{s}}_u \in \hat{\mathbf{S}}$. We add the one with the highest score among the remaining adjacent superpixels in each iteration. Algorithm 1 summarizes the proposed method. The impacts of parameters α and β are analyzed in Section 4.6. Fig. 3(b) shows an example of the hand regions obtained.

3.3 Implementation Issues

Here, we discuss four implementation issues of the proposed method:

- **Appearance Update**: We formulate the appearance update based on the assumption that the contribution of a frame is related to its distance from the current timestamp in the temporal domain. As suggested by [3, 23], if multiple hands belonging to different persons appear in a social interaction, they typically have different appearances. Thus, connected components obtained by Algorithm 1 contribute only to the relevant models. Concretely, each component in the current frame at timestamp t is represented by normalized histogram $\bar{\mathbf{h}}^t$ with P bins, and is assigned to the most similar model based on KL divergence. Suppose that the vth model $\mathbf{a}_v^{t-1} \in \mathbf{A}^{t-1}$ has n_v relevant components. It is then updated as:

$$\mathbf{a}_v^t = \delta \frac{1}{n_v} \sum_{i=1}^{n_v} \bar{\mathbf{h}}_{v,i}^t + (1-\delta)\mathbf{a}_v^{t-1}, \qquad (9)$$

  where $\delta = 0.4$ controls the decay over time, and $\bar{\mathbf{h}}_{v,i}^t$ is the ith relevant component at timestamp t assigned to $\mathbf{a}_v^{t-1}$. For the models with no assigned relevant components, $\mathbf{a}_v^t = \mathbf{a}_v^{t-1}$ is assigned. For the frame recognized as "no hand" due to having no seeds, $\mathbf{a}_v^t = \mathbf{a}_v^{t-1}$ is directly assigned for all models; no update nor region growing is conducted for computational efficiency.
- **Dynamic Initialization**: We have no access to the priori at the beginning of a video sequence, such as hand location bias (i.e., the hands have a higher chance

to appear in the central region than the boundary region of a frame), velocity distribution (i.e., the maximum/avarage hand and head motions), and restricted to only having the wearers' hands. Besides, it is unreasonable to assume that the hands would appear in the first frame or within the visual field all the time. In practice, multiple hands may appear and disappear irregularly, such as in [13]. Thus, the appearance models are dynamically updated and expired to keep only the active models as:

- The position consistency term in Eq. 8 is set to 0 if the previous frame is recognized as having "no hand". If seeds are found in the current frame, the hand position will then be initialized by Algorithm 1.
- The representation $\bar{\mathbf{h}}^t$ for a connected component at timestamp t may differ largely from any existing model $\mathbf{a}_v^{t-1} \in \mathbf{A}^{t-1}$, e.g., a hand appears in the visual field for the first time. In this case, a new model is initialized to $\bar{\mathbf{h}}^t$, under the condition that $\bar{\mathbf{h}}^t$ is continuously observed with only "subtle changes" at successive timestamps $t+1, \cdots, t+10$. Concretely, suppose that the average KL divergence of each component from its most similar model at timestamp t is σ^t, and the KL divergence of $\bar{\mathbf{h}}^t$ from its most similar model is ρ^t. The condition $\rho^t/\sigma^t \geq 3$ is then used to determine whether $\bar{\mathbf{h}}^t$ is distinctive enough to form a new model, and $\parallel \bar{\mathbf{h}}^t - \bar{\mathbf{h}}^{t+\gamma} \parallel_2 \leq 0.2$, $\gamma = 1, \cdots, 10$, is defined as the consistent constraint to depict "subtle changes". Note that occasionally a background region may be falsely recognized as a hand region in a one frame. This, however, would not result in a new appearance model $\bar{\mathbf{h}}^t$, as the error would not propagate to successive frames. Thus, the above conditions would not be satisfied.
- A model is considered to be expired if no related components are assigned to it for a period of time, e.g., 500 successive frames in our system. The expired model is removed from $\mathbf{A}^{1:t}$ (as a model expiry mechanism) to ensure that all current models are active.

- **Egocentric Hand Patterns**: As suggested by [26], four hand patterns may appear in a single egocentric frame. For "left hand only", "right hand only" and "intersecting hands", all hand regions are normally connected. Hence, fewer seeds would be generated, resulting in a single connected component using Algorithm 1. For "two separate hands", the generated seeds would be located in two different hands, resulting in two connected components. In this work, we consider an additional pattern called "no hand", as hands may disappear in the visual field, e.g., when the wearer is just sitting in front of a table. In this case, no seeds would be generated, as the condition $\lambda_k \geq 0.1\mu$ proposed in Section 3.1 is not satisfied due to the lack of hand-related correspondences. Hence, this frame can be quickly recognized as "no hand", without running the region growing algorithm. In addition, we also consider "multiple hands", which is introduced by [3] for social interactions, e.g., the wearer plays cards with friends. This hand pattern would produce more seeds distributed in multiple locations. Occasionally, false seeds that do not belong to any hands may be generated, if the camera and hand motion patterns are indistinguishable caused by the wearer's clothes or bystanders. Refinement is then applied at the end of Algorithm 1. Fig. 4 shows examples of our egocentric hand patterns.
- **Efficient Computation**: To improve the time efficiency, we benefit from the following techniques: ORB descriptors [36] for accelerating the correspondence computation

(a) Left hand only   (b) Right hand only   (c) Intersecting hands

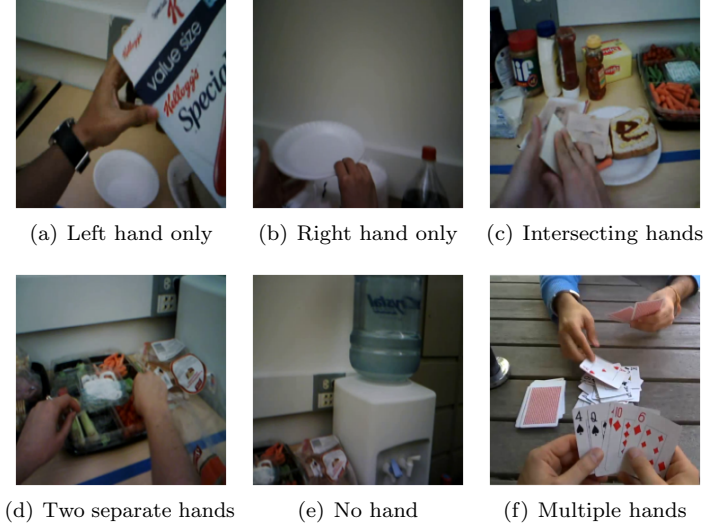(d) Two separate hands   (e) No hand   (f) Multiple hands

Fig. 4. Egocentric hand patterns.

across successive frames, computationally efficient histogram representation for modeling appearances, score computation of only the set of adjacent superpixels, model expiry mechanism for restricting the number of appearance models. Besides, we also employ an early rejection approach, following the idea of cascade of classifiers for detection [6, 8]: the condition, $\max_{\hat{\mathbf{s}}_u \in \hat{\mathbf{S}}} S_1(\mathbf{s}_k, \hat{\mathbf{s}}_u) \geq 0.5$, is set to efficiently remove the superpixels that obvious belong to the background superpixels as their appearances differ largely from the seeds. In indoor scenarios such as those in the GTEA [13] and ADL [32] datasets, the illumination does not change significantly. Thus, the contrast cue can be reliably used to exclude some obvious background superpixels which appearances differ largely from the seeds. For those regions that cannot be easily recognized as background or consist of both noise and hand regions, other cues (i.e., location, position consistency and appearance continuity) are further computed for classification.

4 EXPERIMENTS

In this section, we first introduce the datasets and the evaluation metrics in Section 4.1. We then evaluate the performance of the proposed method on different datasets in Sections 4.2, 4.3 and 4.4, followed by the ablation study in Section 4.5 and the parameter analysis in Section 4.6. Finally, we analyze the efficiency of the proposed method in Section 4.7.

4.1 Datasets and Evaluation Metrics

We evaluate the proposed method on three datasets: the GTEA dataset [13], the ADL dataset [32] and the EgoHands dataset [3]. These datasets and the evaluation metrics are summarized as follows:

- **GTEA dataset**: It is created primarily as an egocentric activity recognition dataset and contains 7 types of daily activities, each of which is performed by 4 different

camera wearers. The videos are taken in the same environment under static illumination. As only the wearer's hands are labeled with foreground hand masks in GTEA, social interactions with partners are not considered. We evaluate the performance on the GTEA dataset following the settings introduced in [25], by classifying all videos into coffee, tea and peanut categories. F-score is used to quantify the classification performance.

- **ADL dataset**: It contains 1 million frames of people performing everyday activities and is annotated with activities, object tracks, hand positions, and interaction events. It is challenging as it contains two types of actions: (1) long-duration actions, e.g., making tea that takes a few minutes, and (2) complex object interactions, e.g., opening the door of a fridge resulting in very different visual appearance. We evaluate the performance on the ADL dataset following the settings introduced in [24], and use F-score to quantify the performance.

- **EgoHands dataset**: it contains 48 Google Glass videos of complex, first-person interactions between two persons, i.e., the wearer and the partner, for 4,800 frames and more than 15,000 hands. The captured videos include realistic and challenging social situations where multiple hands appear. Four types of activities are involved: playing cards, playing chess, solving a jigsaw puzzle and playing Jenga. These activities are captured in three locations, and four camera wearers contribute to the dataset, resulting in a total of $4 \times 3 \times 4 = 48$ videos. In this dataset, the partner's hands appear in the vast majority of frames (95.2% and 94.0% for left and right), while the wearer's hands appear less often (53.3% and 71.1% for left and right). This shows that the wearer's hands are more frequently outside the visual field, and people tend to align attention with the dominant hand. We evaluate the performance on this dataset following the settings introduced in [3]: intersection over union (IoU) between the estimated hand regions and the ground truth annotations.

We further compare the runtime with the baseline methods on all datasets, by considering the number of processed frames per second (fps). As most methods do not report this information except for the method proposed in [19], we either directly ran the available codes (for the method proposed in [3]) or asked the authors to help do it (for the methods proposed in [24, 25]). Failing these, we leave the runtime blank (for the methods proposed in [13, 17, 37]).

4.2 Performance Comparison on the GTEA Dataset

We compare the proposed method with six methods on the GTEA dataset [13]: (1) the single-pixel color approach [19] that utilizes a random regressor, (2) the video stabilization approach [17] based on background modeling by aligning a short sequence with affine transformation, (3) the foreground modeling [37] using feature trajectory-based projection which benefits from the KLT tracker [31], (4) the hybrid approach [13] with the combination of video stabilization and CRF, (5) the local and global appearance mixture model [25] that highlights the effectiveness of sparse features and the importance of modeling global illumination, and (6) the deep-based approach combined with strong appearance models [3]. Partial results for these methods have been reported in [25]. Table 1 summarizes the experimental results for comparison.

The single-pixel color approach [19] achieves surprisingly good performance even though it was proposed nearly 15 years ago. The video stabilization [17] and the foreground modeling [37] approaches perform poorly as they were designed for hand-held videos instead

Table 1. Performance comparison of different methods on the GTEA dataset based on F-score.

Method	F-score			Speed
	coffee	tea	peanut	
Single-pixel color [19]	0.837	0.804	0.730	∼7fps
Video stabilization [17]	0.376	0.305	0.310	-
Foreground modeling [37]	0.275	0.239	0.255	-
Stabilization + CRF [13]	0.713	0.812	0.727	-
Mixture of local and global [25]	0.933	0.943	0.883	∼10fps
Deep-based approach [3]	0.951	0.955	**0.918**	∼2fps
Ours	**0.952**	**0.959**	0.911	**∼32fps**

of egocentric videos. This demonstrates that there are indeed significant differences between the two types of videos, and simply applying the traditional methods on egocentric videos would result in unsatisfactory results. The fourth and the fifth methods [13, 25], which are proposed recently for egocentric hand detection, perform generally well. The deep-based approach [3] performs the best among all existing methods. Our method perform slightly better on coffee and tea but slightly worse on peanut. This shows that our method is comparable to the deep-based approach. However, in terms of efficient, our method performs much faster than the deep-based approach. In fact, ours is the fastest among all methods. This experiment demonstrates that the proposed method performs well in terms of accuracy and efficiency on the GTEA dataset [13], which contains videos of mostly static illumination with small camera motion and no social interactions. The detection errors from the proposed method are mainly due to over-segmentation, especially for the fingers as they are easily confused with other objects and the background. Fig. 5 shows some qualitative results on the GTEA dataset [13].
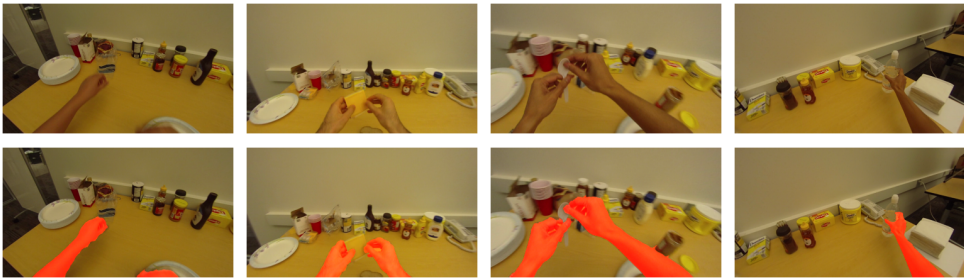

Fig. 5. Qualitative results on the GTEA dataset. Top: original frames. Bottom: detected hands (red).

4.3 Performance Comparison on the ADL Dataset

We compare the proposed method with two baseline methods on the ADL dataset [32], i.e., the local and global appearance mixture model [25] and the model recommendation approach with virtual probes [24] and using different recommendation strategies. Results for these baselines have been reported in [24], and Table 2 summarizes the experimental results for comparison.

Table 2. Performance comparison of different methods on the ADL dataset based on F-score.

Method	F-score	Speed
Mixture of local and global [25]	0.346	∼10fps
Model recommendation (NMF) [24]	0.322	∼1fps
Model recommendation (SC) [24]	0.252	∼1fps
Model recommendation (KNN) [24]	**0.384**	∼1fps
Model recommendation (RF) [24]	0.357	∼1fps
Ours	0.375	**∼32fps**

Table 3. Comparative results on EgoHands dataset. IoU score against baseline methods

Method	IoU score	Speed
Mixture of local and global [25]	0.478	∼6fps
Deep-based approach [3]	**0.556**	∼1fps
Ours	0.527	**∼18fps**

We can see that the proposed method outperforms the mixture model [25], which emphasizes the effectiveness of sparse features and the importance of modeling global illumination. The proposed method also outperforms the model recommendation approach [24] when NMF, SC or RF is used as the recommendation strategy. On the other hand, its performance is comparable to it when KNN is used, which applies a non-linear model to capture complex feature mapping. However, the proposed method is again the most efficient method. This experiment demonstrates that the proposed method achieves the favorable performance with higher efficiency compared with the baselines on the ADL dataset, which contains videos of real-life scenarios with complex object interactions in daily activities.

4.4 Performance Comparison on the EgoHands Dataset

We compare the proposed approach with two baseline methods on the EgoHands dataset [3], i.e., the local and global appearance mixture model [25], and the recently proposed deep-based approach [3] combining with strong appearance models. In the experiments, we neither conduct the activity recognition, nor carry out the classification of "the left hand" or "the right hand", and only compare the results of hand segmentation. Table 3 shows the experimental results with the same setting used in [3], and Fig. 6 shows some qualitative results.

This work achieves a better result compared with [25] which is based on hand-crafted cues like ours, while performs a little worse than the deep-based approach [3] with much higher efficiency for the following reasons:

- [3] starts from the pre-trained CaffeNet, indicating it benefits from the additional data thus unfair for comparison.

- Different from other works [24, 25], the EgoHands dataset defines the "hand" to stop at the wrist instead of the sleeve. Similar to the failures of [25], the proposed method cannot effectively distinguish hands and arms, as they share similar motions and appearances. However, [3] succeeds in these cases by learning from sufficient labeled samples.

- The proposed method would generate inaccurate seeds when the moving bystanders or objects appear in the background, as we assume that the motions are mainly produced by the mounted camera and the hands, which is not satisfied in the above scenario.
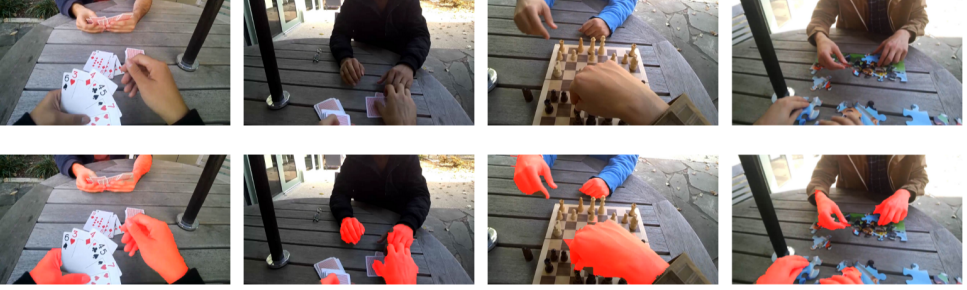

Fig. 6. Qualitative results on the EgoHands dataset. Top: original frames; bottom: hand detection results (red).

4.5   Ablation Study

To evaluate the impacts of the proposed egocentric cues (Section 3.2), we have conducted an experiment to compare different versions of the proposed method, with each version having one cue (i.e., one term in Eq. 8) removed. Table 4 shows the results on the three datasets. From these results, we can draw the following conclusions:

- The appearance constraints, i.e., contrast and appearance continuity, are observed to have larger impacts compared with the spatial constraints, i.e., location and position consistency, on the three datasets. This indicates that the skin color plays an important role on egocentric hand detection.
- For the GTEA dataset, the appearance constraints are more critical as the illumination in the videos tends to be static, meaning that the contrast and appearance continuity cues are more reliable here.
- For the ADL dataset, the impacts of the appearance constraints decrease, due to the wider range of indoor imaging conditions, while the impacts of the spatial constraints increase, due to the long-duration actions and complex object interactions (which require more accurate hand positioning and tracking).
- For the EgoHands dataset, "appearance continuity" stands out from all cues with more models constructed compared with the GTEA dataset, as multiple hands are involved in social interactions. "Location" and "position consistency" also affect the performance because of the strict definition of hands.

4.6   Parameter Analysis

In Algorithm 1, the hand regions are gradually located from the seed regions with two parameters. While α defines the termination condition of region growing, β defines the threshold for noise component removal. In general:

- A large α value increases the rejection rate. Thus, fewer superpixels are retained, resulting in the incompleteness of the hand regions, e.g., the missing fingers. On the other hand, a small α may produce some redundant regions that are hand-irrelevant, e.g., the grasped bread.

Table 4. Ablation study on the GTEA, ADL and EgoHands datasets.

Dataset \\ Method	GTEA			ADL	EgoHands
	coffee	tea	peanut		
No "contrast"	0.762	0.768	0.728	0.305	0.467
No "location"	0.914	0.921	0.872	0.334	0.486
No "position consistency"	0.876	0.883	0.836	0.347	0.492
No "appearance continuity"	0.791	0.796	0.755	0.316	0.448
Ours	**0.952**	**0.959**	**0.911**	**0.375**	**0.527**

- A large β value increases the risk of falsely removing components that belong to the hands, e.g., the partially occluded palm, while a small β may not be effectively enough to exclude noisy components, e.g., the chess.

To study how the results are affected by α and β, we compare the performances of having different combinations of α and β on the ADL dataset. Table 5 shows the comparison results, which agrees with the above analysis. In addition, these results show that within an appropriate range of α and β, there is only a small change in performance, indicating that the proposed method is robust to parameter variations.

Table 5. Parameter analysis by adopting various combinations of α and β on the ADL dataset.

β \\ α	300	350	400	450	500
0.5	0.351	0.359	0.363	0.361	0.358
0.55	0.361	0.367	0.370	0.368	0.364
0.6	0.367	0.372	**0.375**	0.373	0.368
0.65	0.363	0.368	0.371	0.368	0.362
0.7	0.357	0.361	0.364	0.360	0.352

4.7 Runtime Analysis

We have implemented the proposed method on a PC with an i7 2.6GHz CPU and 16GB RAM. The average rate is ∼32fps, indicating that it can run in real-time (640 × 360). The region growing step takes more time compared with the seed generation step (∼21ms vs. ∼10ms). Without the early rejection step discussed in Section 3.3 to avoid unnecessarily computing all cues, an additional ∼23ms would be needed. The early rejection step only causes a 1.3% drop in performance on the ADL dataset. Compared with the existing methods, a significant speedup is observed, e.g., ∼1s using the model recommendation approach [24] (640 × 360, on a 2.6GHz CPU), ∼100ms using the local and global appearance mixture model [25] (640 × 360, on a 2.6GHz CPU), ∼2s using DP-DPM, and 9s using R-CNN as reported in [10] (500 × 400, on a 2.9GHz CPU and Titan X GPU), ∼40ms using region growing as reported in [22] (320 × 180, on a 2.80GHz CPU), ∼29ms using YOLO growing as reported in [33] (640 × 480, on a 3.5GHz 6-core CPU and Titan X GPU).

We have further conducted an experiment to study the efficiency of the proposed method on two older PCs with limited computational capabilities and memory space, to simulate the computation environment of wearable devices. The following experimental setups are tested:

(1) A PC with Pentium IV 2.53 GHZ CPU and 512M RAM, video resolution at 320×240: We achieve a framerate of ∼13fps.
(2) A PC with i3 3.3 GHZ CPU and 2G RAM, video resolution at 320×240: We achieve a framerate of ∼38fps.
(3) A PC with i3 3.3 GHZ CPU and 2G RAM, video resolution at 640×360: We achieve a framerate of ∼15fps.

These results indicate the high efficiency of the proposed method, even when compared with relevant works in virtual reality [15] and augmented reality [28].

5 CONCLUSION

In this paper, we have proposed a seed region generation algorithm and novel egocentric cues for egocentric hand detection, by growing the detected seed regions into hand regions iteratively. We have discussed some implementation issues in applying the proposed method in complex environments. Experimental evaluations on public datasets demonstrate that the proposed method performs comparably against the state-of-the-art methods, while being significantly more efficient. Ablation study and parameter analysis have been conducted to study the impacts of the proposed cues and the important parameters. As a future work, we are currently applying the proposed method on some related tasks like gesture recognition and hand-based action recognition.

REFERENCES

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *TPAMI* 34, 11 (2012), 2274–2282.
[2] Sven Bambach. 2015. A survey on recent advances of computer vision algorithms for egocentric video. *arXiv:1501.02825* (2015).
[3] Sven Bambach, Stefan Lee, David J Crandall, and Chen Yu. 2015. Lending a hand: detecting hands and recognizing activities in complex egocentric interactions. In *ICCV*. 1949–1957.
[4] Lorenzo Baraldi, Francesco Paci, Giuseppe Serra, Luca Benini, and Rita Cucchiara. 2014. Gesture recognition in ego-centric videos using dense trajectories and hand segmentation. In *CVPR Workshops*. 688–693.
[5] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. SURF: speeded up robust features. In *ECCV*. 404–417.
[6] Alejandro Betancourt, Miriam Lopez, Carlo Regazzoni, and Matthias Rauterberg. 2014. A sequential classifier for hand detection in the framework of egocentric vision. In *CVPR Workshops*. 586–591.
[7] Alejandro Betancourt, Pietro Morerio, Carlo S Regazzoni, and Matthias Rauterberg. 2015. The evolution of first person vision methods: A survey. *TCSVT* 25, 5 (2015), 744–760.
[8] Hakan Cevikalp, Bill Triggs, and Vojtech Franc. 2013. Face and landmark detection by using cascade of classifiers. In *Automatic Face and Gesture Recognition*. 1–7.
[9] Ana Garcia del Molino, Cheston Tan, Joo-Hwee Lim, and Ah-Hwee Tan. 2017. Summarization of egocentric videos: A comprehensive survey. *IEEE Trans. on Human-Machine Systems* 47, 1 (2017), 65–76.
[10] Xiaoming Deng, Ye Yuan, Yinda Zhang, Ping Tan, Liang Chang, Shuo Yang, and Hongan Wang. 2016. Joint hand detection and rotation estimation by using CNN. *arXiv:1612.02742* (2016).
[11] Sylvia M Dominguez, Trish Keaton, and Ali H Sayed. 2006. A robust finger tracking method for multimodal wearable computer interfacing. *IEEE Trans. on Multimedia* 8, 5 (2006), 956–972.
[12] Alireza Fathi, Yin Li, and James M Rehg. 2012. Learning to recognize daily actions using gaze. In *ECCV*. 314–327.
[13] Alireza Fathi, Xiaofeng Ren, and James M Rehg. 2011. Learning to recognize objects in egocentric activities. In *CVPR*. 3281–3288.
[14] Martin A Fischler and Robert C Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24, 6 (1981), 381–395.

[15] Serkan Genç, Muhammet Baştan, Uğur Güdükbay, Volkan Atalay, and Özgür Ulusoy. 2015. HandVR: a hand-gesture-based interface to a video retrieval system. *Signal, Image and Video Processing* 9, 7 (2015), 1717–1726.

[16] Joydeep Ghosh, Yong Jae Lee, and Kristen Grauman. 2012. Discovering important people and objects for egocentric video summarization. In *CVPR*. 1346–1353.

[17] Eric Hayman and Jan-Olof Eklundh. 2003. Statistical background subtraction for a mobile observer. In *ICCV*. 67–74.

[18] Yedid Hoshen and Shmuel Peleg. 2016. An egocentric look at video photographer identity. In *CVPR*. 4284–4292.

[19] Michael J Jones and James M Rehg. 2002. Statistical color models with application to skin detection. *IJCV* 46, 1 (2002), 81–96.

[20] Mathias Kolsch and Matthew Turk. 2004. Fast 2D hand tracking with flocks of features and multi-cue integration. In *CVPR Workshops*. 158–158.

[21] Shiro Kumano, Kazuhiro Otsuka, Ryo Ishii, and Junji Yamato. 2017. Collective first-person vision for automatic gaze analysis in multiparty cnversations. *IEEE Trans. on Multimedia* 19, 1 (2017), 107–122.

[22] Jayant Kumar, Qun Li, Survi Kyal, Edgar A Bernal, and Raja Bala. 2015. On-the-fly hand detection training with application in egocentric action recognition. In *CVPR Workshops*. 18–27.

[23] Stefan Lee, Sven Bambach, David Crandall, John Franchak, and Chen Yu. 2014. This hand is my hand: a probabilistic approach to hand disambiguation in egocentric video. In *CVPR Workshops*. 543–550.

[24] Cheng Li and Kris Kitani. 2013. Model recommendation with virtual probes for egocentric hand detection. In *ICCV*. 2624–2631.

[25] Cheng Li and Kris Kitani. 2013. Pixel-level hand detection in ego-centric videos. In *CVPR*. 3570–3577.

[26] Yin Li, Alireza Fathi, and James Rehg. 2013. Learning to predict gaze in egocentric video. In *ICCV*. 3216–3223.

[27] Yin Li, Zhefan Ye, and James M Rehg. 2015. Delving into egocentric actions. In *CVPR*. 287–295.

[28] Hui Liang, Junsong Yuan, Daniel Thalmann, and Nadia Magnenat Thalmann. 2015. AR in hand: Egocentric palm pose tracking and gesture recognition for augmented reality applications. In *ACM Multimedia*. 743–744.

[29] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *IJCV* 60, 2 (2004), 91–110.

[30] Zheng Lu and Kristen Grauman. 2013. Story-driven summarization for egocentric video. In *CVPR*. 2714–2721.

[31] Bruce D Lucas, Takeo Kanade, and others. 1981. An iterative image registration technique with an application to stereo vision.. In *IJCAI*, Vol. 81. 674–679.

[32] Hamed Pirsiavash and Deva Ramanan. 2012. Detecting activities of daily living in first-person camera views. In *CVPR*. 2847–2854.

[33] Akshay Rangesh, Eshed Ohn-Bar, Mohan M Trivedi, and others. 2016. Driver hand localization and grasp analysis: a vision-based real-time approach. In *ITSC*. 2545–2550.

[34] Xiaofeng Ren and Chunhui Gu. 2010. Figure-ground segmentation improves handled object recognition in egocentric video. In *CVPR*. 3137–3144.

[35] Grégory Rogez, James S Supancic, and Deva Ramanan. 2015. First-person pose recognition using egocentric workspaces. In *CVPR*. 4325–4333.

[36] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. 2011. ORB: an efficient alternative to SIFT or SURF. In *ICCV*. 2564–2571.

[37] Yaser Sheikh, Omar Javed, and Takeo Kanade. 2009. Background subtraction for freely moving cameras. In *ICCV*. 1219–1225.

[38] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. 2013. Selective search for object recognition. *IJCV* 104, 2 (2013), 154–171.

[39] Jing Wang, Yu Cheng, and Rogerio Schmidt Feris. 2016. Walk and learn: facial attribute representation learning from egocentric video and contextual data. In *CVPR*. 2295–2304.

[40] Bo Xiong and Kristen Grauman. 2014. Detecting snap points in egocentric video with a web photo prior. In *ECCV*. 282–298.