In: Proceedings of IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Year: 2015; Pages: 1 - 6, DOI: 10.1109/ICMEW.2015.7169784

TOWARDS A UNIFIED FRAMEWORK FOR HAND-BASED METHODS IN FIRST PERSON VISION.

Alejandro Betancourt[1,2], *Pietro Morerio*[1], *Lucio Marcenaro*[1], *Emilia Barakova*[2], *Matthias Rauterberg*[2], *Carlo Regazzoni*[1]

[1]Information and Signal Processing for Cognitive Telecommunications Group.
Department of Naval, Electric, Electronic and Telecommunications Engineering.
University of Genoa, Italy

[2] Designed Intelligence Group.
Department of Industrial Design.
Eindhoven University of Technology.
Eindhoven, Netherlands.

ABSTRACT

First Person Vision (Egocentric) video analysis stands nowadays as one of the emerging fields in computer vision. The availability of wearable devices recording exactly what the user is looking at is ineluctable and the opportunities and challenges carried by this kind of devices are broad. Particularly, for the first time a device is so intimate with the user to be able to record the movements of his hands, making hand-based applications for First Person Vision one the most explored area in the field. This paper explores the more popular processing steps to develop hand-based applications, and proposes a hierarchical structure that optimally switches between each of the levels to reduce the computational cost of the system and improve its performance.

Index Terms— Hand-detection, Hand-segmentation, Hand-identification, Hands interactions, First Person Vision, Egovision, Werable cameras

1. INTRODUCTION

The emergence of new wearable devices such as action cameras and smart-glasses during the recent years has detonated an important chain effect between researchers, computer scientists, and high-tech companies [1]. The 90's futuristic dream of a wearable device that is always ready to be used in front of our eyes is nowadays technically possible [2]. In turn, this has increased the interest of the researchers and computer scientist to develop methods to process the recorded data. The more explored sensor is by far the video camera, which on one side, has enjoyed the benefits of a privileged location to record exactly what the user is seeing, but on the other side, has raised strong critics concerning privacy [3] and battery life issues. The videos recorded from this perspective are commonly called First-Person Vision (FPV) or Egocentric videos [4].

This work was partially supported by the Erasmus Mundus joint Doctorate in Interactive and Cognitive Environments, which is funded by the EACEA, Agency of the European Commission under EMJD ICE.

FPV videos offer important benefits and challenges to computer vision scientists. As the main benefit, it is the first time that a wearable device is recording exactly what the user have in front of him. However, being it mobile and wearable, implies highly variable environments with different illumination [5] and without any kind of static reference system [6]. In FPV, unlike in static cameras video processing, both the background and the foreground are in constant motion. An intuitive implication of the FPV camera location is the general belief that the user hands are being constantly recorded and thus the large number of studies based on their gestures and trajectories. The hand presence is particularly important in the field, because, for first time, hand gestures (conscious or unconscious) can be considered the more intuitive way of interaction with the device.

Hands have played an important role in a large group of methods, for example in activity recognition [7], user-machine interaction [8], or even to infer the gaze of the user [9]. In a recent work, the authors in [10] point out the effects of wrongly assume full time presence of the hands in front of the user.

Hand-based methods are traditionally divided in two large groups, namely model-based and data-based methods [11]. The former aims to find the best configuration of a computational hand model to match the image in the video, while the latter are lead by video features such as color histograms, shape, texture, orientation, among others. Figure 1 reviews some of the most relevant hand-based papers in FPV.

The classic taxonomy of hand-based methods is too broad and several authors have suggest further extensions according to the features used, the task addressed, and the required sensors. In [4, 10] the authors propose a hierarchical division of the processing steps that can be independently solved e.g. hand-detection, hand-segmentation, hand-tracking, etc. In practice, nowadays, it is common to find a well trained pixel-by-pixel hand-segmenter taking control of the whole system. The segmenter is thus responsible for understanding whether hands are present (after exhaustive negative classi-

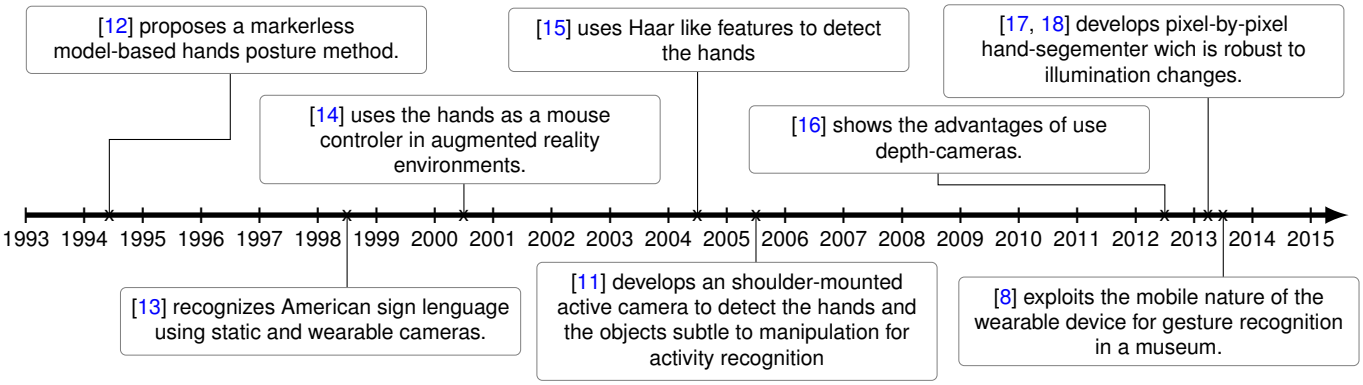
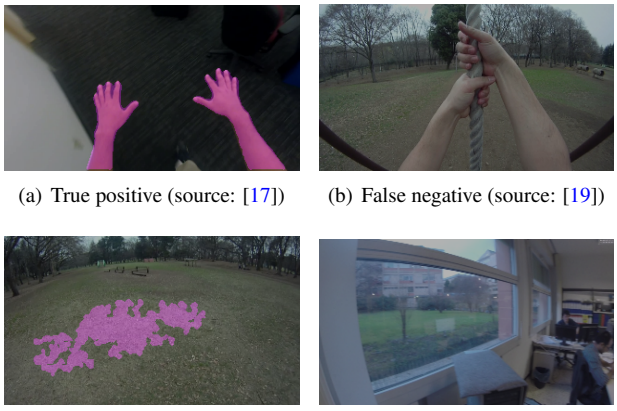


Fig. 1: Relevan papers of the hands

fication of every pixel), define the shape of the hands (task for which it is trained), and suggest the areas to be tracked keeping record across the frames of the segmented shapes. This approach achieves good results, particularly in finding hand-like pixels; however it rises several issues: i) it exhaustively uses computational resources even when the hands are not present, which is the common case in daily real videos; ii) it could disseminate noise in the system, produced by false-positives; iii) it usually does not exploit temporal correlation between frames. Figure 2 shows the possible results obtained by a pixel-by-pixel hand-segmenter.



(a) True positive (source: [17]) (b) False negative (source: [19])

(c) False positive (d) True negative

Fig. 2: Examples of the hand-segmentation.

Under this lines of thought, this paper attempts to highlight the importance of a proper fusion of the above-mentioned tasks in a unified hierarchical structure. In this structure, each level is developed to perform a specific task and provide its results to the rest of the levels. To optimize resources, the structure must switch between its components when it is required e.g. the hand-detector must work only when the hands are not present, while the hand-segmenter (along with a dynamic tracker) is active in the opposite case, but must give back the control of the system to the hand-detector when the hands leave the scene. Finally, the system design must give priority to shared features to optimize extra resources and make the real-time dream closer. The latter is not straightforward and explains why the current methods are usually evaluated in a post-processing framework, restricting the eventual applicability of the field.

This paper explores some of the ideas behind a unified framework for hand-based methods in FPV, and highlights some of the current challenges of the field for real-life applications. The remainder of the work is organized as follows: Section 2 conceptualizes a hierarchical structure to develop hand-based FPV methods. Later, section 3 shows some preliminary results for each of the levels. Finally, section 4, presents a discussion about the applicability in real scenarios of the FPV field, which, although devices are almost ready for final users, could require extra discussion from the computer vision community about its real scope.

2. A UNIFIED FRAMEWORK: MOTIVATION AND STRUCTURE

As already mentioned, one of the main challenges in FPV video analysis is to understand user's hand movements in uncontrolled activities. A proper interpretation of hands (e.g. trajectories, gestures, interactions) opens the door to advanced task such as activity recognition and user machine interaction, but more importantly it could be the cornerstone to move the wearable devices from experimental state to useful technology. Given the camera location and the user proximity, a system that is able to capture hand gestures could allow smart glasses to do things that other devices like smart-phones cannot. Incidentally, this technology could help to alleviate everyday difficulties of people with visual [20], speaking [13], or motor issues [21]

Current methods have reported remarkable results for tasks like detecting hands presence in front of the user [10], segmenting the silhouette [22, 18, 17, 23], recognizing basic

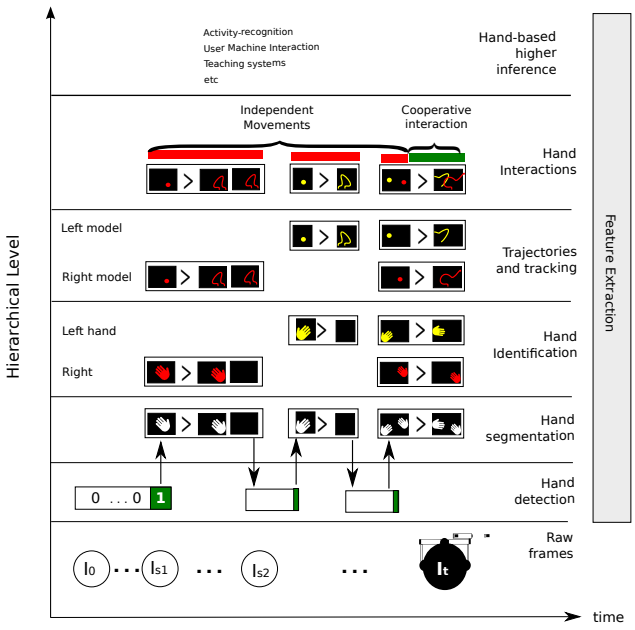


Fig. 3: Unified system for hand-based FPV methods.

gestures [8, 24], inferring hand posture [25, 26], and identifying whether a hand belongs to the user or to a third person [27]. In general, these subtasks could be considered partially solved, but for an ideal smart wearable camera they are supplied as independent pieces of software resting over different sets of assumptions. Two examples of this are: i) the already mentioned case of the pixel-by-pixel classifier, which despite of being developed to solve the hand-segmentation problem is used to detect, segment and track the hands on its own, at a high computational cost, ii) the hand-detector that, once it is sure about the hands presence, keeps working in parallel to detect if the hands leave, instead of using the detailed results of the hand-segmenter.

To design a unifying system for hand-based methods in FPV it is important to identify some of the more important components, standardize its inputs/outputs and define their role in the overall system. Our approach stands over the task division proposed in [10, 4] and is summarized in the hierarchical structure proposed in Figure 3. The figure shows the most important steps in hand-based methods. Some of them could be non necessary for some applications (e.g. not every application needs to identify the left and right hand) or extra levels could be included (e.g. pose recognition). In the bottom part of the diagram are the raw frames, while in the upper part lie the higher inference methods that search for patterns in the hand movements and trajectories. The diagram shows a feature extractor that can be re-used by all the levels: a system that is able to use the same features in multiple levels can save valuable computational resources and processing time.

The diagram makes evident the importance of a system that is able to optimally decide which is the minimum number of methods running in parallel for each time instance. This switching behaviour is crucial in the bottom levels (hand-detection and hand-segmentation), as well as in the identification and tracking levels to model each hand separately. In the first case, an optimized version of the sequential classifier proposed in [10] is used. The optimized version of this method switches from the hand-detection to the hand-segmentation level whenever the decision moves from "no hands" to "hands"; and from the hand-segmentation to the hand-detection level if there are no more positive pixels in the frame. In the second case, the switching models literature [28, 29] suggests useful strategies to decide which hand models need is to be used at that time. The hand-id (left-right) of the segmented hands emerges as a good switching variable. The hand-id can be estimated using the angles and the positions of the segmented shapes. The next section briefly summarizes each of each of the hierarchical levels, and states some of our findings and preliminary results.

The diagram shows a bottom-up design starting from the features and arriving to simplified trajectories to be used by different applications. However, it is worth to mention that a top-down analysis focused on the application field can remove some of the assumptions in different levels and lead to considerable improvements to the overall performance. As example, the authors in [27] take advantage of psychological experiments among kids and adults to relax the illumination assumption of their proposed method.

3. OUR APPROACH: PRELIMINARY ANALYSIS AND RESULTS

This section briefly describes each hierarchical level, suggests some approaches to face the problems that can be encountered, and reports some of our results. Some of the reported results are already published or are under review (references are provided), while others are still under experimental analysis and development. In the latter case we report our approach, summarizing the challenges faced and the preliminary conclusions.

Hand-detection: This level answers the yes-or-no question of the hands' presence in the frame. This problem is addressed in [10] as a frame by frame classification problem. In their experiments the authors report that the best result is achieved with the combination of Histogram of Oriented Gradients (HOG) features with a Support Vector Machine (SVM). One of the main problems of this frame-by-frame approach is its sensibility to small changes between frames, which makes unstable in time the decisions taken by the classifier. In recent experiments this issue is alleviated using a Dynamic Bayesian Network (DBN) that filters a real valued representation of the SVM classifier. Table 1 shows the performance of both approaches (HOG-SVM and the

DBN) for each of the 5 testing uncontrolled locations of the UNIGE-egocentric dataset. In the framework of the unified system, the hand-detector must be optimized to detect as fast as possible the frames on which the hands enter the scene.

Table 1: Comparsion of the performance of the HOG-SVM and the proposed DBN.

	True positives		True negatives	
	HOG-SVM	DBN	HOG-SVM	DBN
Office	0.893	0.965	0.929	0.952
Street	0.756	0.834	0.867	0.898
Bench	0.765	0.882	0.965	0.979
Kitchen	0.627	0.606	0.777	0.848
Coffee bar	0.817	0.874	0.653	0.660
Total	0.764	0.820	0.837	0.864

Hand-segmentation: It is probably the more explored problem in FPV. The main task is to delineate the silhouette of the hands at a pixel level. The more promising results are reported in [17, 18, 23] achieving F-scores around 83% under different illumination levels. The main challenge in the pixel-by-pixel approach is the computational complexity of the task, involving the decision for each pixel in each frame. For instance, the camera of the Google glasses has a resolution of 720p and records 30 frames per second, implying 928.800 pixel classifications per frame and a total of $27'864.000$ per second of video. A promising strategy to reduce this number is to simplify the frames as SLIC superpixels [30] and classify the simplified image as done in [8]. Within this approach, in [31] an optimized initialization of the SLIC algorithm is proposed. It allows to segment 13 frames per second, while the original SLIC is able to process only 1. Figure 4 shows an example of the optimized SLIC algorithm.

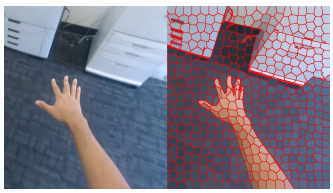
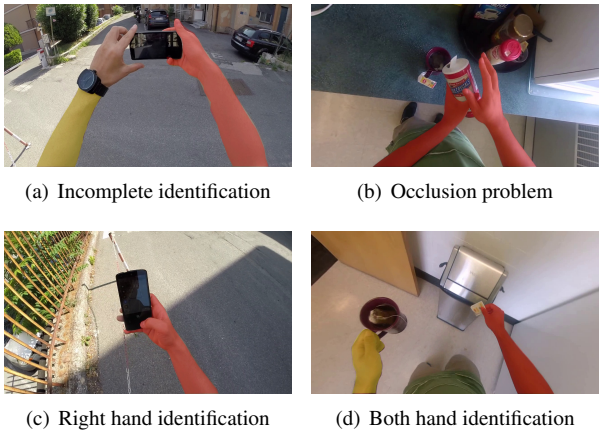


Fig. 4: Optimized superpixels of a frame with hands [31]

Hand-identification: It is an intuitive but challenging task. The objective is to identify the left and the right hand. The hand-identification problem is extended in [27], proposing a Bayesian method to identify, using the relative positions, the hands of the user as well as the hands of a third person in the video. At this point it is worth to mention the robustness of the proposed hand-detector to the presence of third person hands. However, in the segmentation level, extra effort must be done to segment only the user hands. Assuming a reliable hand-segmentation it is possible to build a simple identification system based on the angle and the side of the frame from which the hand appears. We found that in realistic scenarios this approach properly differentiate the left and the right hand in almost all the frames at low computational cost. Two difficult scenarios of this approach are: i) The hands are close enough to create a single shape; ii) the appearance of hands is divided by an external object as a bracelet or a watch, creating several hand-like shapes. Figure 5 shows an example of our identification algorithm based on manually segmented shapes.



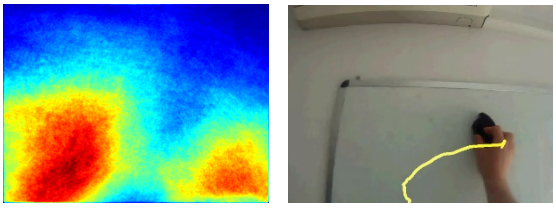
(a) Incomplete identification (b) Occlusion problem

(c) Right hand identification (d) Both hand identification

Fig. 5: Hand-Identification.

Tracking and trajectories: For a wearable camera it is important to record, track and denoise hands trajectories. An intuitive and straightforward approach is to keep history of the hands centroid as done in [23]. However, the use of dynamic filters could help to increase the accuracy of the trajectories, reduce the sampling rate (the lower the sampling rate the closer to real time performance), and manage the switching process between the hand-detection and the hand-segmentation level. Regarding the initial conditions (e.g. initial coordinates of the tracked hand) the best choice is to use dynamic filters like the h-filter, which only requires the empirical distribution of the initial coordinates [32]. The initial distribution can be found using empirical data as shown in [33] (Figure 6(a)). Regarding the dynamic model, our preliminary experiments suggest that a simple linear model can achieve promising results for high frequency sampling. However additional tests are required before a conclusion can be made. Figure 6(b) shows frame by frame tracking of a hand center.

Hands Interactions: Once hands are located and their trajectories are inferred, a possible next step is to understand the



(a) Empirical distribution of hand locations [33]

(b) Tracking the hand using the centroid

Fig. 6: Hand-Tracking.

(a) Google glass device (RGB device)

(b) Meta glass (Stereoscopic device)

Fig. 7: RGB and RGB-D wearable devices.

interactions between them. For instance, if each hand is performing an independent task (e.g. moving an object, making a gesture) or if both hands are cooperating to accomplish particular objective (e.g. making tea, spreading butter, driving, etc.). At this level important features can be found in the center of mass, the location of the handled objects, the distance between the hands and the relationship between the left and right trajectory. One of the most important works about hand-interaction is [7], where the spatial relation of the hands as well as the handled objects is used to infer some cooking task like (e.g. Pout, stir, spread, etc.).

Hand-based higher inference: in the upper level of the structure are the methods on which the results are built using the information of the hands, some examples are activity-recognition [7] and user-machine interaction [7]. At this point we highlight the relevance of a deep discussion about the real applicability of hand-based methods in FPV and which are the benefits of using single wearable RGB cameras over other systems, like stereoscopic cameras, the Kinect or the Leap-Motion.

On the one side, the miniaturization of RGB cameras makes them the most promising candidate to be worn. On the other side, in exchange of extra battery consumption and an increase in the size of the device, the use of other sensors can bring important improvements to hand-based methods. For example, the depth component can reduce the complexity of the hand-segmentation level and extra information like the pose of the hands can be straightforwardly inferred. Figure 7 shows an example of a RGB and a stereoscopic wearable device. Regarding external devices, like the Kinect or the Leap-Motion, they can, under certain conditions, acquire a wider perspective of the body and, as a result, provide a better understanding of hand movements. As a counterpart, the wearability of these external devices is highly restricted. In summary, external devices can represent a default choice for applications based on static locations without battery restrictions. However, if the application field includes a user moving around with restricted battery availability then a wearable RGB cameras is the most promising option.

4. CONCLUSION AND DISCUSSION

This paper highlights the importance of a systemic hierarchical approach to develop hand-based methods. The proposed hierarchical switching structure for hand-based methods in FPV can reduce the computational requirements and under further analysis of the sampling rates could help to reach real time performances. The latter would expand the application areas of FPV video analysis, which by now has been mainly focused on offline processing applications. Each level of the proposed structure address a well defined and scoped tasks, allowing us to strategically design each of our methods for a unique purpose e.g. hand-detection, hand-segmentation, hand-tracking, etc.

Based on the development process and the feedback of our previous work we point out the convenience of an application-based analysis of the field in order to better understand its real scope and the advantages of a particular sensor choice. We highlight the mobility and the battery cost as the main advantage of RGB cameras. However, if battery restrictions are removed, stereoscopic cameras could lead to more reliable results. From our point of view this discussion must lead the coming developments to be focused on tasks only achievable by wearable cameras and not by other devices like smart-phones, smart watches or static systems (e.g. Kinect, Leap-Motion).

We consider this as a good moment to analyze the lessons learned by the Glass Project and the current approaches of other companies like Microsoft with the Holo-Lens device. A brief analysis of the media reports about the glass project ending reveals two valuable lessons: i) People would be willing to use these device only if they are able to do things that existing technologies cannot do ii) There are big opportunities for the task oriented approaches, such as medical and industrial applications, on which privacy issues are minimum and the scenarios faced by the user can be partially restricted. On the other hand, the available information of the Holo-Lens project, sketches a device with an exhaustive use of hand-gestures as way of interaction. From this perspective hand-based methods would clearly play an important role in the future of this device.

5. REFERENCES

[1] T Starner, "Project Glass: An Extension of the Self," *Pervasive Computing*, vol. 12, no. 2, pp. 125, 2013.

[2] S Mann, "Wearable Computing: a First Step Toward Personal Imaging," *Computer*, vol. 30, no. 2, pp. 25–32, 1997.

[3] Robert Templeman, Mohammed Korayem, D.J. Crandall, and Kadapia Apu, "PlaceAvoider: Steering first-person cameras away from sensitive spaces," in *Network and Distributed System Security Symposium*, 2014, number February, pp. 23–26.

[4] A Betancourt, P Morerio, C.S. Regazzoni, and M Rauterberg, "The Evolution of First Person Vision Methods: A Survey," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2015.

[5] Z Lu and K Grauman, "Story-Driven Summarization for Egocentric Video," in *Computer Vision and Pattern Recognition*, Portland, OR, USA, June 2013, pp. 2714–2721, IEEE.

[6] J Ghosh and K Grauman, "Discovering Important People and Objects for Egocentric Video Summarization," in *Computer Vision and Pattern Recognition*. June 2012, pp. 1346–1353, IEEE.

[7] A Fathi, A Farhadi, and J Rehg, "Understanding Egocentric Activities," in *International Conference on Computer Vision*. Nov. 2011, pp. 407–414, IEEE.

[8] G Serra, M Camurri, and L Baraldi, "Hand Segmentation for Gesture Recognition in Ego-vision," in *Workshop on Interactive Multimedia on Mobile & Portable Devices*, New York, NY, USA, 2013, pp. 31–36, ACM Press.

[9] Y Li, A Fathi, and J Rehg, "Learning to Predict Gaze in Egocentric Video," in *International Conference on Computer Vision*. 2013, pp. 1–8, Ieee.

[10] A Betancourt, Lopez M, M Rauterberg, and C.S. Regazzoni, "A Sequential Classifier for Hand Detection in the Framework of Egocentric Vision," in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Columbus, Ohio, June 2014, vol. 1, pp. 600–605, IEEE.

[11] W Mayol and D Murray, "Wearable Hand Activity Recognition for Event Summarization," in *International Symposium on Wearable Computers*. 2005, pp. 1–8, IEEE.

[12] J Rehg and T Kanade, "DigitEyes: Vision-Based Hand Tracking for Human-Computer Interaction," in *Workshop on Motion of Non-Rigid and Articulated Bodies*. 1994, pp. 16–22, IEEE Comput. Soc.

[13] T Starner, J Weaver, and A Pentland, "Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1371–1375, 1998.

[14] T Kurata, T Okuma, M Kourogi, and K Sakaue, "The Hand-mouse: A Human Interface Suitable for Augmented Reality Environments Enabled by Visual Wearables," in *Symposium on Mixed Reality*, Yokohama, 2000, pp. 188–189.

[15] M. Kolsch and M Turk, "Robust hand detection," in *International Conference on Automatic Face and Gesture Recognition*. 2004, pp. 614–619, IEEE.

[16] Dima Damen, Andrew Gee, Walterio Mayol-Cuevas, and Andrew Calway, "Egocentric Real-time Workspace Monitoring using an RGB-D camera," in *RSJ International Conference on Intelligent Robots and Systems*. Oct. 2012, pp. 1029–1036, IEEE.

[17] C Li and K Kitani, "Pixel-Level Hand Detection in Ego-centric Videos," in *Computer Vision and Pattern Recognition*. June 2013, pp. 3570–3577, Ieee.

[18] C Li and K Kitani, "Model Recommendation with Virtual Probes for Egocentric Hand Detection," in *ICCV 2013*, Sydney, 2013, IEEE Computer Society.

[19] K Kitani, "Ego-Action Analysis for First-Person Sports Videos," *Pervasive Computing*, vol. 11, no. 2, pp. 92–95, 2012.

[20] Steve Mann, Jason Huang, Ryan Janzen, Raymond Lo, Valmiki Rampersad, Alexander Chen, and Taqveer Doha, "Blind navigation with a wearable range camera and vibrotactile helmet," in *International Conference on Multimedia*, New York, New York, USA, 2011, p. 1325, ACM Press.

[21] C de Boer, J van der Steen, R J Schol, and J J M Pel, "Repeatability of the timing of eye-hand coordinated movements across different cognitive tasks.," *Journal of neuroscience methods*, vol. 218, no. 1, pp. 131–8, Aug. 2013.

[22] Seungyeop Han, Rajalakshmi Nandakumar, Matthai Philipose, Arvind Krishnamurthy, and David Wetherall, "GlimpseData: Towards Continuous Vision-based Personal Analytics," in *Workshop on physical analytics*, New York, New York, USA, 2014, vol. 40, pp. 31–36, ACM Press.

[23] P Morerio, L Marcenaro, and C Regazzoni, "Hand Detection in First Person Vision," in *Information Fusion*, Istanbul, 2013, University of Genoa, pp. 1502 – 1507.

[24] Jingtao Wang and Chunxuan Yu, "Finger-fist Detection in First-person View Based on Monocular Vision Using Haar-like Features," in *Chinese Control Conference*. July 2014, pp. 4920–4923, Ieee.

[25] Gregory Rogez and JS Supancic III, "3D Hand Pose Detection in Egocentric RGB-D Images," in *ECCV Workshop on Consumer Depth Camera for Computer Vision*, Zurich, Switzerland, Nov. 2014, vol. Sep, pp. 1–14, Springer.

[26] Gregory Rogez, James S. Supancic, and Deva Ramanan, "Egocentric Pose Recognition in Four Lines of Code," in *Computer Vision and Pattern Recognition*, Nov. 2015, vol. Jun, pp. 1–9.

[27] S Lee, S Bambach, D Crandall, J Franchak, and C Yu, "This Hand Is My Hand: A Probabilistic Approach to Hand Disambiguation in Egocentric Video," in *Computer Vision and Pattern Recognition*, Columbus, Ohio, 2014, pp. 1–8, IEEE Computer Society.

[28] Simone Chiappino, Pietro Morerio, Lucio Marcenaro, and Carlo S. Regazzoni, "Bio-inspired relevant interaction modelling in cognitive crowd management," *Journal of Ambient Intelligence and Humanized Computing*, vol. 6, no. 2, pp. 171–192, Feb. 2014.

[29] Vladimir Pavlovic, JM Rehg, and J MacCormick, "Learning switching linear models of human motion," *NIPS*, 2000.

[30] Radhakrishna Achanta, Appu Shaji, and Kevin Smith, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2281, 2012.

[31] Pietro Morerio, Gabriel Claudiu Georgiu, Lucio Marcenaro, and Carlo Regazzoni, "Optimizing Superpixel Clustering for Real-Time Egocentric-Vision Applications," *IEEE Signal Processing Letters*, vol. 22, no. 4, pp. 469–473, Apr. 2015.

[32] VP Bhuvana, "Distributed object tracking based on square root cubature H-infinity information filter," in *Information Fusion*, Salamanca, 2014, pp. 1 – 6, IEEE Signal Processing.

[33] M Philipose, "Egocentric Recognition of Handled Objects: Benchmark and Analysis," in *Computer Vision and Pattern Recognition*, Miami, FL, June 2009, pp. 1–8, IEEE.