A Sequential Classifier for Hand Detection in the Framework of Egocentric Vision

Alejandro Betancourt[1,2]
alejandro.betancourt@ginevra.dibe.unige.it
Miriam M. López[1]
miriam.lopez@ginevra.dibe.unige.it
Carlo S. Regazzoni[1]
carlo@dibe.unige.it
Matthias Rauterberg[2]
g.w.m.rauterberg@tue.nl
[1]Department of Naval, Electric, Electronic and Telecommunications Engineering - University of Genoa, Italy
[2]Designed Intelligence Group, Department of Industrial Design - Eindhoven University of Technology, The Netherlands

***Abstract*—Hand detection is one of the most explored areas in Egocentric Vision Video Analysis for wearable devices. Current methods are focused on pixel-by-pixel hand segmentation, with the implicit assumption of hand presence in almost all activities. However, this assumption is false in many applications for wearable cameras. Ignoring this fact could affect the whole performance of the device since hand measurements are usually the starting point for higher level inference, or could lead to inefficient use of computational resources and battery power. In this paper we propose a two-level sequential classifier, in which the first level, a *hand-detector*, deals with the possible presence of hands from a global perspective, and the second level, a *hand-segmentator*, delineates the hand regions at pixel level in the cases indicated by the first block. The performance of the sequential classifier is stated in probabilistic notation as a combination of both, classifiers allowing to test new hand-detectors independently of the type of segmentation and the dataset used in the training stage. Experimental results show a considerable improvement in the detection of true negatives, without compromising the performance of the true positives.**

***Keywords*-Hand Detection; First Person Vision; Egocentric Vision; Video Processing;**

I. Introduction

Hand detection and segmentation for First Person Vision (FPV) applications is gaining more and more attention both in academic and industrial fields due to the fast development of wearable devices. The miniaturization capabilities and the growing interest of companies are closing fast the gap between the academic research interests and the massive consumption of wearable devices. The interest of common users is clear and personal devices like the case of *GoPro*, are becoming very popular. Companies such as Google and Microsoft are putting big efforts to conquer this section of the market. The former already has a commercial prototype, *Project Glass*, which is being tested by a restricted group of users since 2012. In the academic world, the situation is not different, and researchers are proposing methods aimed at exploiting the potential of the videos recorded with this kind of devices, commonly known as egocentric videos. Moreover, the analysis of egocentric videos presents a wide variety of applications in different fields such as health, military, and augmented reality [15, 9, 14, 22, 1].

One of the most explored areas in egocentric vision video analysis is related to the detection and tracking of the user's hands [13, 17]. The potential of this area is high because in many cases it is used as the starting point of high level inference of information about the user, e.g. activity recognition, internal state estimation or user-machine interaction. Efforts of current hand detection methods are focused on segmenting the frame areas belonging to the user's hands while performing activities such as cooking [4, 5, 20] or working with the computer [16]. This problem is usually tackled at pixel level, implicitly assuming an *a priori* hand presence in the frames. This assumption is not always true, particularly for tasks in which hands are not used, such as walking down the street, talking to the office colleagues or speaking on the phone.

Assuming full time presence of hands may lead to important issues: *i)* possible wrong hand measurements, particularly in no-hand frames, would be propagated to other levels of the system and create wrong conclusions or unwanted feedback from the device and *ii)* unnecessary searching for local features in the image, meaning an inefficient use of computational resources and reduction of the battery life. Note that a pixel-by-pixel hand segmentation of a frame with a resolution of 1280×720 pixels involves 921.600 classification tasks. For practical purposes, some authors reduce the resolution of the images without compromising the quality of the results, however the calculation is still (O^2).

At this point, an intuitive question arises: why to go into detailed pixel-by-pixel classification without knowing first if it is worth it? In order to answer this question, and following the same reasoning of [18] on video analysis, two different tasks should be differentiated, namely *hand-detection* and *hand-segmentation*. The former term has been extensively

2160-7516/14 $31.00 © 2014 IEEE
DOI 10.1109/CVPRW.2014.92

used (and possibly leading to a misunderstanding) for tasks in which the localization of the hands in the scene was required. In this work, this term refers to a step in which a global answer is given to whether hands are present in the scene or not. The latter aims at delineating the hands in a frame at a pixel level. Both problems are closely related, being possible to use *hand-detection* as a pre-filtering stage for *hand-segmentation* under the framework of sequential classification, in which the output of the first classifier is used to decide whether the second one will be used [25]. Furthermore, different features could be applied in each level, being preferable the use of general features for *hand-detection* purposes [19].

The contributions of this article are three-fold: *i)* we propose a sequential classifier, hand-detection/segmentation, which first detects general information and, only if necessary, searches for details, *ii)* we derive the performance of the sequential classifier as a function of subparts, which allows to work separately on each part to subsequently infer the performance of the whole system and *iii)* we analyze the performance of different combinations of feature and classifier as *hand-detectors*. The features used in this work are color histograms (RGB, HSV, LAB), Histograms of Oriented Gradients (HOG) and GIST. The classifiers are Supported Vector Machines (SVM), Decision Trees (DT) and Random Forest (RF). The proposed sequential classifier reduces from 65% to 4% the rate of false positives compared to the state-of-the-art *hand-segmentation* method.

The remainder of this paper is organized as follows: in section II we present existing *hand-detection/segmentation* methods and briefly introduce common features and classification algorithms. Section III states the performance of the sequential classifier. The case of a perfect *hand-detector* is presented along with the performance of state-of-art *hand-segmentation* method when used alone. In section IV, different combinations of features and classifiers are tested as *hand-detectors*, and the best one is analyzed in the sequential classifier. Finally, in section V conclusions are drawn and some lines for future research are proposed.

II. State of the art

Hand-segmentation has recently become one of the most explored objectives in egocentric vision video analysis. It is common to find hierarchical methods for higher inference using as an input the shape and position of the hands. Some example applications lie in the area of activity recognition [4, 5, 20], user-machine interaction [6, 22], or hand posture inference [23].

According to the seminal work proposed in [16], known for being the first public dataset for egocentric object recognition, existing methods for detecting hands in a scene can be divided into two groups: model-driven approaches, which are based on 2D or 3D computerized models of the hands, and data-driven methods, which rely on features extracted from the video frames. Mixtures of both approaches can also be found in the literature [23]. Methods in the first group select the best matching configuration of a hand model to recreate the image of the video frames [21, 23]. These methods are able to infer detailed information of the hands but usually require large computational resources and highly controlled environments. Methods of the second group use extracted features to infer the position of the hands. One of the first methods within this second group was proposed in 1999 by Jones and Rehg [7] to detect skin pixels in images using color histograms. However, egocentric videos introduce additional challenges such as dealing with different skin colors, changes in illumination or camera motion.

Concerning the *hand-detection* task as defined in this paper, no contributions can be easily found in the literature, being still a problem scarcely explored. It could seem that a good hand segmentation method could indirectly solve the hand detection one, however, this is not true in activities with sparse hand presence, leading to considerable drawbacks as already described. In section III we show that the state-of-the-art method for *hand-segmentation* [13] is not able to tackle this problem, and we prove the advantages of facing it as a sequential classifier to improve the overall performance.

A. Features and classifiers for pattern detection

Object detection is a common task in video analysis and have been frequently used to detect vehicles [10], pedestrians [3], or faces [24]. To accomplish this a classifier is trained using large datasets of positive and negative samples and subsequently used for uncontrolled videos. An important aspect of these approaches is that computational time has been considered so that most of them are able to run in real time.

Color histogram is one of the most recurrent features used for image classification [8, 7] due to its straightforward computation and intuitive interpretation. The variety of color spaces such as RGB, HSV, YCbCr or LAB make possible to consider color information while alleviating potential issues with illumination or color skin changes. In particular, HSV is based on the way humans perceive colors, while LAB and YCbCr separate lightness from color components. In egocentric vision, [17] uses a mixture of color histograms and visual flow for *hand-segmentation*, while [22] combined HSV features, a random forest classifier and super-pixels for gesture recognition. Recently, Li and Kitani [13, 12] analyzed the discriminative power of different color features with a random forest regressor for hand segmentation under different illumination configurations.

Dalal and Triggs [2] proposed HOG features in combination with linear Support Vector Machines (SVMs) for pedestrian detection in video streams, achieving relatively high accuracy. HOG captures edge or gradient structure that is very characteristic of the local shape. Information in local

cells is collected into histograms using trilinear interpolation, and overlapping blocks composed of neighbouring cells are normalized with respect to a controllable degree of invariance to translations and rotations.

We also consider GIST [18] as a global scale descriptor of the image, which captures texture information and a coarse spatial layout of the image. GIST can be combined with other local descriptors to accurately detect objects in the scene, and was originally combined with a simple one-level classification tree, as well as with the naive Bayesian classifier. GIST descriptor has been successfully applied for large scale image retrieval and object recognition [19].

Finally, it is worth mentioning the well-known Haar-like features, first introduced by Viola and Jones [24], who built an efficient moving face detector which yielded good results in real time object detection. However, this descriptor is not considered in the rest of this paper since hand shape is highly variable, and therefore, a unique Haar-based detector would not be enough. A modification of this feature for the framework of egocentric vision is mentioned in section V as a promising future research line.

III. Our approach

Given a frame, four possible situations are identified depending on the performance of the *hand-segmentator*: a *true positive* if hand regions are delineated in a frame with hands, a *false negative* if regions are not detected and hands are present, a *false positive* if regions are detected but no hands are present, and finally, a *true negative* for a no-hand frame properly rejected. The *hand-segmentator* is expected to perform successfully in the true positive and the false negative detection rates. Regarding the other two cases, even if the classifier was good at rejecting frames with no hands, it would imply exhaustive scanning over all the pixels of the frame. We address this problem by proposing a *hand-detector* as a trigger of the *hand-segmentator*. In the next section, we state in probabilistic notation the performance of each classifier when used by it own and the performance of the sequential system afterwards.

A. Hand-detection and hand-segmentation as a sequential classification system

Let us assume a dataset D with a considerable mixture of frames with and without hands. Denote $p(y_1) = p(y = 1)$ and $p(y_0) = p(y = 0)$ the probability of hand and no-hand presence, respectively, in any frame k in D. Let $c(i^k, j^k)$ be a *hand-segmentator* given by equation (1), and $gc(k)$ the output induced by $c(i^k, j^k)$ over all the pixels in frame k. According to equation (2), a single positive pixel classification implies a positive label for the whole frame. It is out of the scope of this work to analyze the performance of $c(i^k, j^k)$ at finding real hand regions in k, hence, we assume that if there exists a pixel (i, j) in k such that $c(i^k, j^k) = 1$, then that pixel actually corresponds to a hand region. This assumption works in favor of $gc(k)$ because misdetections of this block are being ignored (see Figure 2).

(a) True positive (source: [13])

(b) False negative (source: [11])

(c) False positive

(d) True negative

Figure 1. Hand detection: possible scenarios.

(a) Example 1

(b) Example 2

(c) Example 3, source: [13]

(d) Example 4, source: [13]

Figure 2. Examples of missdetections.

$$c(i^k, j^k) = \begin{cases} 1 & \text{if pixel } (i,j) \in k, \text{ is part of hand} \\ 0 & \text{if pixel } (i,j) \in k, \text{ is not part of hand.} \end{cases} \quad (1)$$

$$gc(k) = \begin{cases} 1 & \text{if } \sum_{(i,j)\in k} c(i,j) > 0 \\ 0 & \text{if } \sum_{(i,j)\in k} c(i,j) = 0. \end{cases} \quad (2)$$

In order to quantify the performance of $gc(k)$ in a realistic scenario (where $p(y_1) \neq 1$), let us denote by $p_{gc}(\hat{y}|y)$ the probability of labeling the frame as $\hat{y}$ when the correct label (ground truth) is y, under the gc model. Therefore, $p_{gc}(1|1)$ and $p_{gc}(0|0)$ are the probabilities of successful classification of positive and negative cases, respectively, for the gc model.

Our approach aims to improve this performance by using a two-level sequential classifier. In the first level, a *hand-detector* $gh(k)$, which serves as a pre-filter to avoid unnecessary executions of the second layer, the *hand-segmentator* $c(i, j)$. Figure 3 shows a simplified diagram of the proposed system. In this case, $gh(k)$ must be trained using positive and negative samples to subsequently be used as a trigger of $c(i, j)$. Following the same notation as for gc, we derive the

performance of the sequential system, denoted by $p_{ss}(\hat{y}|y)$, and its counterparts as expressed in (3). Note that $p_{ss}(\hat{y}|y)$ is a function of $p_{gh}(\hat{y}|y)$ and $p_{gc}(\hat{y}|y)$, making it possible to infer the performance of the whole system for new datasets if similar performance in any of the levels are assumed under a proper training.

$$p_{ss}(\hat{y}|y) = \begin{bmatrix} p_{gh}(1|1)p_{gc}(1|1) & p_{gh}(0|1) + p_{gh}(1|1)p_{gc}(0|1) \\ p_{gh}(1|0)p_{gc}(1|0) & p_{gh}(0|0) + p_{gh}(1|0)p_{gc}(0|0) \end{bmatrix} \tag{3}$$

In order to quantify the improvement introduced by our approach, it is necessary to analyze the conditions that make (3) close to the identity matrix I_2. The better the detector $gh(k)$, the closer $p_{ss}(\hat{y}|y)$ to I_2. For no-hand cases (second row), the improvement is substantial because $p_{gh}(1|0) \leq 1$. For false positive cases, an extra bias is introduced to the system, making evident the importance of avoiding early rejections in $gh(k)$, or in other words, of pushing $p_{gh}(0|1)$ close to 0. Note that, if a perfect hand detector is assumed, then the second row of equation (3) becomes $[0, 1]$ and only the false negatives outputted by the *hand-segmentator* will affect the sequential classifier.

The applicability of the proposed method is assessed by considering as baseline the work of Li and Kitani [13], which proposes a *hand-segmentator* based on random forest and color histograms (using RGV, HSV, LAB color spaces). The original paper reports experiments with other features such as HOG and SIFT, however, the authors conclude that color features performed best. This conclusion implies higher dependence of the trained model on skin characteristics.

In the mentioned work, the classifier is trained only on positive hand samples with different configurations. We train and test their method as described in the original paper, using the EDSH1 database for training and EDSH2 and EDSHK databases for testing. For the estimation of $p_{gc}(\hat{y}|1)$, we count the number of true positives and false negatives. Furthermore, we recorded several videos with no hands in an office environment to complement the dataset, and subsequently counted the number of false positives and true negatives to estimate $p_{gc}(\hat{y}|0)$. Equation (4) shows the performance.

$$p_{gh}(\hat{y}|y) = \begin{bmatrix} 0.98 & 0.02 \\ 0.65 & 0.35 \end{bmatrix} \tag{4}$$

As expected, the *hand-segmentation* level performs satisfactorily when hands are present because the baseline method was proposed for that scenario. However, a very low performance is obtained for the alternative cases. From these results, two conclusions can be drawn: *i)* in the 65% of the no-hand cases, inaccurate information is detected and potentially used for higher inference, and *ii)* in the other 35% of the no-hand cases, exhaustive *hand-segmentation* was performed to finally reject the presence of hands. In the next section we study the performance of different *hand-detectors* and analyze the effects of using them under a sequential classifier structure in conjunction with the method proposed in [13].

B. Dataset and models

The dataset used in this work as a ground truth to train the *hand-detector* is composed of 2835 video frames, 1.259 (44%) of which show hands and 1.576 (56%) do not. These frames are the result of sampling a set of videos (24 minutes overall duration) every 0.5 seconds. The ground truth information regarding the presence or absence of hands was created manually, labeling the positive and negative video intervals before the sampling process. The videos were recorded in different places (living room, kitchen, office, auditorium, street, among others) in order to gather variations in lighting conditions, color compositions and background scene configurations. The videos were recorded with a *GoPro hero3+* with 1280×720 *px* resolution and 60 *fps*.

Based on the previous studies, we create a pool of combinations of features and classifiers. HOG features are extracted as shape-based descriptor, GIST is used for quantifying the discriminative power of the global characteristic of the scene and finally, following the line of study of [13], three color spaces (RGB, HSV and LAB) and the concatenation of them as an extra feature were also applied. Note that color histograms are local features (extracted at pixel level). In order to consider a global framework, we average the color values over cells resulting from applying a 3×5 grid to each frame. This grid is proposed to capture color patterns in the areas on which hands tend to appear when are observed from an first person perspective, as shown in Figure 4. All these features were combined with SVMs, DT and RF classification methods.

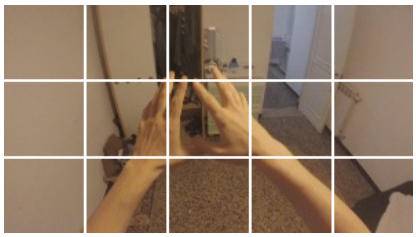

Figure 4. Proposed grid to reduce the color features.

IV. Results

In order to compare the performance of the combinations of features and classifiers proposed in the previous section, we use a 10-fold cross validation for each one. Table IV shows the performance of each approach. Judging by the diagonal of $p_{gh}(\hat{y}|y)$, HOG-SVM yield the best performance, achieving up to 90% and 93% of true positive and true negative detection rates, respectively. These results were

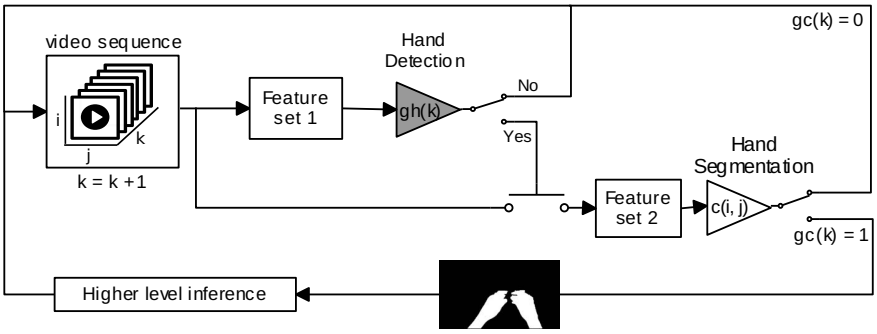

Figure 3. Proposed sequential classifier.

	$p_{gh}(1\|1)$			$p_{gh}(0\|0)$		
	SVM	DT	RF	SVM	DT	RF
HOG	**0.90**	0.72	0.74	**0.93**	0.78	0.93
GIST	0.81	0.76	0.75	0.85	0.80	0.91
RGB	0.56	0.76	0.70	0.89	0.78	0.90
HSV	0.57	0.78	0.75	0.82	0.80	0.92
LAB	0.54	0.80	0.77	0.90	0.82	0.92
(RGB+HSV+LAB)	0.68	0.79	0.77	0.84	0.80	0.94

Table I
PERFORMANCE OF THE PROPOSED *hand-detectors*

expected, considering the high performance of this combination reported previously for the problem of pedestrian detection, for which HOG-SVM was originally proposed [2].

It is also remarkable the performance of GIST-SVM, which reaches 81% true positive detection rate, and 80% for true negatives. Regarding decision trees and random forest, the best combination achieves 77% and 94%, with the advantage of being computationally fast, in particular if they are combined with color features. Neither GIST-SVM nor (RGB+HSV+LAB)-random forest are selected because our interest is to minimize the early rejection (see Section III). LAB results in the best individual color-based feature, however a slight improvement is obtained if they are used in conjunction, particularly with random forest classifier.

The performance of the sequential classifier system for HOG-SVM model is given by (5). This result is inferred if is assumed a similar performance of the *hand-segmentator* for the skin colors of our dataset. Note the remarkable reduction from 65% to 4% in the false positive rate, preventing errors from being propagated through the system without a *hand-detector*. Regarding the true negatives, the sequential classifier achieves 96% accuracy rate, yielding a total of 93% of the no-hand cases being filtered by the *hand-detector* global classifier and avoiding the execution of the *hand-segmentator*. Concerning the computational complexity of our approach, assuming real time feature extraction (such as for HOG and color histograms), and given the fast performance of a linear SVM, it can be concluded that in 93% of the no-hand frames the O^2 problem of *hand-segmentation* is avoided.

$$p_{ss}(\hat{y}|y) = \begin{bmatrix} 0.88 & 0.12 \\ \mathbf{0.04} & \mathbf{0.96} \end{bmatrix} \quad (5)$$

In order to visualize the difficult frames of the *hand-detector*, we perform a training over all the frames in the dataset except 100 frames randomly chosen, which are left out. Afterwards, the trained SVM is used to estimate the probability of hand presence in those 100 frames. Figure 5 shows this probability in the x-axis for some of these frames: the more on the left, the higher the probability of being classified as no-hand, and vice-versa for the hand case. Therefore, the 'difficult' cases are located in the middle of the cloud. Some misclasifications, shown on the top of the figure, are found to be classified as hand-frames when diagonal structures appear in the frame. On the other side, we found some rejections when the hands barely appear. Another important aspect to highlight is the capability of the classifier to reject frames with extra people on it. In our case, we use a linear SVM with the same weight for positive and negative frames (that is, no *a priori* knowledge was introduced in the system), however, a reduction of the early rejections of hand frames could be achieved by assigning higher weights to positive frames in the training process.

V. CONCLUSIONS AND FUTURE RESEARCH

In this paper, the necessity for differentiation between *hand-detection* and *hand-segmentation* is stated and a theoretic approach for both cases is presented. Based on this idea, a sequential classification system is proposed aimed at improving the performance and reliability of hand inference in the framework of wearable devices. Our approach is validated by quantifying the maximum improvement that could be reached if a *hand-detector* was used as a trigger for a *hand-segmentator*. The insertion of the proposed *hand-detector* block avoids unnecessary executions of the pixel-by-pixel classifier in the 93% of the no-hand cases. A total of 6 features and 3 classifiers were tested, concluding that the best one is the HOG-SVM combination.

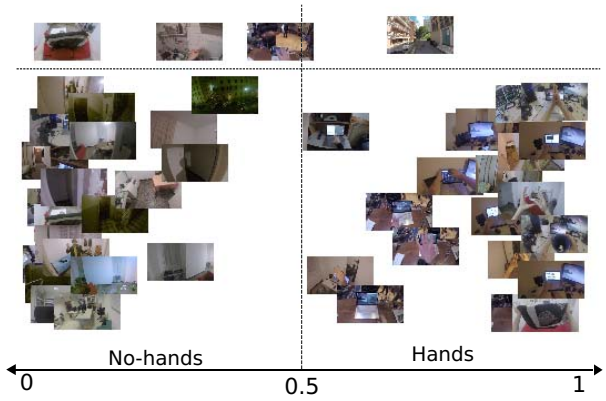

Figure 5. Probability of being a hand-frame inferred by the SVM.

This paper shows the potential of the traditional features by solving the *hand-detection* problem, however it is still open the study of egocentric inspired image features, such as possible adaptations of Haar-like features to deal with hand detection.

VI. Acknowledgement

This work was supported in part 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.

References

[1] A. Borji, D. Sihite, and L. Itti. Probabilistic Learning of Task-Specific Visual Attention. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 470–477, June 2012.

[2] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 886–893, 2005.

[3] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian Detection: A Benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 304–311, 2009.

[4] A. Fathi, Y. Li, and J. M. Rehg. Learning to Recognize Daily Actions Using Gaze. In *12th European Conference on Computer Vision*, pages 314–327, Florence, Itaty, 2012.

[5] A. Fathi, X. Ren, and J. M. Rehg. Learning to recognize objects in egocentric activities. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3281–3288, June 2011.

[6] R. Grasset, T. Langlotz, and D. Kalkofen. Image-driven view management for augmented reality browsers. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 177–186, 2012.

[7] M. J. Jones and J. M. Rehg. Statistical Color Models with Application to Skin Detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–23, 1999.

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

[9] S. Karaman, J. Benois-Pineau, R. Megret, V. Dovgalecs, J.-F. Dartigues, and Y. Gaestel. Human Daily Activities Indexing in Videos from Wearable Cameras for Monitoring of Patients with Dementia Diseases. In *International Conference on Pattern Recognition (ICPR)*, pages 4113–4116, 2010.

[10] Z. Kim and J. Malik. Fast vehicle detection with probabilistic feature grouping and its application to vehicle tracking. pages 524–531, 2003.

[11] K. Kitani. Ego-Action Analysis for First-Person Sports Videos. *Pervasive Computing*, 11(2):92–95, 2012.

[12] C. Li and K. Kitani. Model Recommendation with Virtual Probes for Egocentric Hand Detection. In *International Conference in Coputer Vision (ICCV)*, pages 2624–2631, 2013.

[13] C. Li and K. M. Kitani. Pixel-Level Hand Detection in Ego-centric Videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3570–3577, 2013.

[14] S. Mann, M. Ali, R. Lo, and H. Wu. FreeGlass for developers, haccessibility, and AR Glass+ LifeGlogging Research in a (Sur/Sous) Veillance Society. In *Information Society (i-Society)*, pages 51–56, 2013.

[15] S. Mann, J. Nolan, and B. Wellman. Sousveillance: Inventing and Using Wearable Computing Devices for Data Collection in Surveillance Environments. *Surveillance & Society*, 1(3):331–355, 2003.

[16] W. W. Mayol and D. W. Murray. Wearable Hand Activity Recognition for Event Summarization. In *Ninth IEEE International Symposium Wearable Computers*, pages 122–129, 2005.

[17] P. Morerio, L. Marcenaro, and C. Regazzoni. Hand Detection in First Person Vision. In *16th International Conference on Information Fusion (FUSION)*, pages 1502–1507, 2013.

[18] K. Murphy, A. Torralba, D. Eaton, and W. Freeman. Object detection and localization using local and global features. In *Lecture Notes in Computer Science*, volume 4170, pages 382–400, 2006.

[19] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[20] M. Philipose. Egocentric recognition of handled objects: Benchmark and analysis. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, June 2009.

[21] M. Schlattman and R. Klein. Simultaneous 4 Gestures 6 DOF Real-time Two-hand Tracking Without Any Markers. In *Proceedings of the Symposium on Virtual reality software and technology*, pages 39–42, 2007.

[22] G. Serra, M. Camurri, and L. Baraldi. Hand Segmentation for Gesture Recognition in Ego-vision. In *Proceedings of the 3rd ACM International Workshop on Interactive Multimedia on Mobile & Portable Devices*, pages 31–36, 2013.

[23] L. Sun, U. Klank, and M. Beetz. EYEWATCHME–3D Hand and object tracking for inside out activity analysis. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, pages 9–16, 2009.

[24] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, 1:I–511–I–518, 2001.

[25] Y. Zhu and S. Schwartz. Efficient face detection with multiscale sequential classification. In *International Conference on Image Processing (ICIP)*, volume 2, pages II–121–II–124, 2002.