Experiments on an RGB-D Wearable Vision System for Egocentric Activity Recognition

Mohammad Moghimi[1], Pablo Azagra[2], Luis Montesano[2], Ana C. Murillo[1,2] and Serge Belongie[3]

[1]UC San Diego
La Jolla, CA
mohammad@ucsd.edu

[2]DIIS - I3A
University of Zaragoza, Spain
{montesano, acm}@unizar.es

[3]Cornell Tech
New York, NY
tech.cornell.edu

Abstract

This work describes and explores novel steps towards activity recognition from an egocentric point of view. Activity recognition is a broadly studied topic in computer vision, but the unique characteristics of wearable vision systems present new challenges and opportunities. We evaluate a challenging new publicly available dataset that includes trajectories of different users across two indoor environments performing a set of more than 20 different activities. The visual features studied include compact and global image descriptors, including GIST and a novel skin segmentation based histogram signature, and state-of-the art image representations for recognition, including Bag of SIFT words and Convolutional Neural Network (CNN) based features. Our experiments show that simple and compact features provide reasonable accuracy to obtain basic activity information (in our case, manipulation vs. non-manipulation). However, for finer grained categories CNN-based features provide the most promising results. Future steps include integration of depth information with these features and temporal consistency into the pipeline.

1. Introduction

Thanks to advances in consumer electronics, digital cameras are ubiquitous sensors whose presence is constantly growing and offer more and more solutions to real-life problems. Moreover, the miniaturization of camera optics and electronics has facilitated the construction of a wide variety of wearable visual sensors. The ever growing capabilities of local and cloud based computing are pushing the potential for this technology even further. The information captured by wearable or person-mounted cameras presents opportunities for a diverse array of applications. From early prototypes focused on life logging, e.g. [6], to more interactive devices such as Google Glass[1], advancements in the design of these systems are pushing the boundaries of wearable vision applications. In this paper we explore opportunities offered by vision and depth (RGB-D) sensors in this field, in particular for first-person activity recognition. Fig. 1 shows the sensor and images used in this work.

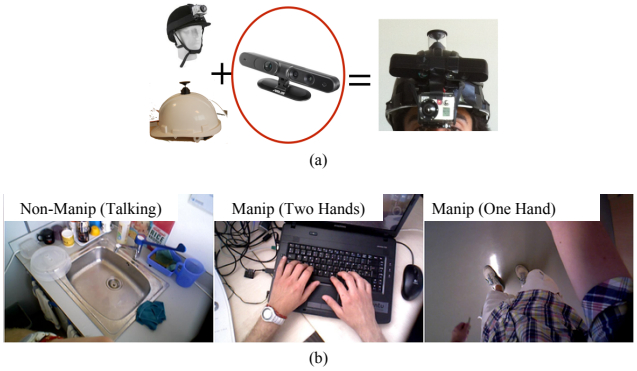


Figure 1. (a) Wearable vision system evaluated in this work (RGB-D camera mounted on a helmet together with other vision sensors not used in this work). (b) Sample images from activities recorded with this system. The second and third images were correctly classified, while the first one was incorrectly labeled, it should belong to the others category.

Distinguishing between fine grained activity labels based solely on still frames (e.g., opening vs. closing a door) is a challenging task. Sequential and temporal information is the key to distinguish such cases, but in many activities still images can nonetheless narrow down the list of possible activities. This work is focused initially on classification of still images (i.e., descriptors computed separately for each frame), but in the future these steps can be integrated with spatiotemporal consistency or additional features to achieve more detailed activity recognition.

The main goals and contributions described in this work

[1]http://glass.google.com

are the following:

- Evaluating a prototype with a helmet-mounted RGB-D camera worn by a set of users performing similar indoor activities. We propose a hierarchy of labels to facilitate the use of contextual information before running fine grained activity classification.

- Analyzing the performance of new image representations for still-frame activity recognition: (a) a set of proposed compact global descriptors built after skin segmentation proposed and (b) Convolutional Neural Network based image representation for activity recognition.

2. Related Work

The increasing popularity of wearable cameras has been accompanied by expanded interest in computer vision applications that can be developed for or adapted to the demands of this domain (e.g., first-person viewpoint, low power requirement, and abundance of captured data). We find most of the earlier works to be focused on *life logging* and its applications [14] which naturally motivates the development of automatic summarization techniques such as [9].

Other works in this area focus on shorter-term wearable camera applications, such as user interaction; see for example the hand location and gesture based UI of [13]. Other works in this vein include a system for monitoring a user workspace [1] and an approach to track people moving around a user [7]. Our work similarly explores how to take advantage of RGB-D wearable sensors, but does so in the context of activity recognition.

Some recent works involving activity recognition from an ego-centric point of view include [18] which classifies interactions with the user into categories such as punching or hugging and [11] which leverages the point of view of the camera to estimate the gaze of the user, as the information about where the user is looking is a strong hint for activity recognition.

Action recognition is an important topic in computer vision, not necessarily from an egocentric perspective, so many of the lessons learned in general settings are of interest for our work. In particular, works on skin segmentation [22, 16] play an important role in our study. Skin pixels will contain very important information if we analyze the interactions of a person with the environment, but their location in the image depends heavily on the vantage point of the camera capturing the scene. Therefore, as we will see in later sections, we build our image description starting from pixel-wise skin segmentation rather than from hand and gesture detection libraries, typically designed for a different point of view (frontal).

Also related to our study is the problem of scene understanding. Global scene-level image information can be used to provide context to support finer grained recognition stages [20]. This two stage hierarchical approach has also been used successfully in the context of activity recognition with spatio-temporal features [19].

Another component of our study involves the evaluation of compact image descriptors for basic activity understanding and an exploration of how state of the art features facilitate activity recognition tasks. Convolutional Neural Networks (CNN) provided significant performance gains on computer vision benchmarks including ImageNet[2]. In 2012, a deep CNN model achieved the best results on ILSVRC2012 [8] and we implemented a CNN in their system (DECAF)[3]. Recently, the same model is showed promising results of a range of datasets including PASCAL VOC, Caltech-UCSD Birds and Oxford Buildings [17].

Table 1. Hierarchy of Action Labels used in this work.

Granularity	**Action Labels**	
Level 1	Manipulation	Non-Manipulation
Level 2	One hand Two hands Pick-up Others	Walk Stairs Stand Screen
		Talk Seating

3. Hierarchical Activity Recognition from a Head-mounted camera

As previously mentioned, many recognition systems that need to handle large amounts of data work in a hierarchical manner: first the system prunes the classification options using global information, to either find a set of possible candidates or provide some kind of context information; then, a more detailed analysis is performed to determine which of the possible candidates is the best fit.

In this work we propose how a similar hierarchical process could be useful for the task of activity recognition from a wearable vision system. The following assumes a computer vision system with similar properties to the one we are testing in this work: a head mounted camera, pointing a bit down, to facilitate keeping the user's hands within the field of view. This facilitates the analysis of the actions performed. We can see the described configuration in Fig. 1.

Table 1 shows the labels we are considering at different levels of granularity. At level 1, we perform a binary classification into manipulation and non-manipulation actions. The subsequent levels model progressively fine grained categories within each level 1 group. The non-manipulation labels are assigned to every frame in the sequence while the manipulation labels may not be present in every frame. Frequently, we find frames where several labels occur simultaneously (one from each level 1 group, e.g., we can be standing and opening something at the same time). To simplify this initial study, we consider only the dominant label,

e.g., if the user is seating and reading from the computer screen, we assign the *Screen* label to that frame.

Our dataset also includes a 3rd level with finer categories, but the frequency of these activities is sufficiently low as to make training a classifier impossible. Therefore, manipulation actions have been grouped into similar actions according to how the hands and arms are located (from the first-person perspective) while performing them: *Two-hands*: includes all activities where the user uses/shows both hands while performing the action; *One-hand*: same as before but for one hand; *Pick-up*: includes all activities where the user merely picks up or drops an object but does not otherwise interact with it; *Others*: all activities that imply manipulation but do not fit any of the above.

Specific Level 3 actions represented in the dataset are as follows. *Two-Hands*: Typing on a keyboard, Using the mouse, Reading a paper, Reading a book, Writing on a paper; *One-Hand*: Hand-shaking, Writing on a board, Open-close door, Open-close window, Open-close fridge, Open-close microwave, Open-close closet; *Pick-up*: Using vending machines, Answering the phone, Drinking, Eating, Picking-up and Droping an object.

4. RGB-D image segmentation and description

As previously described, the focus of this work is on still frame based activity recognition. This section describes the different types of image descriptors studied in this work for our experiments.

4.1. Global image descriptors

The first type of descriptors considered are typical global image descriptors, which give a compact image representation frequently used for scene categorization and place recognition, such as GIST [15], color histograms or global image statistics such as color invariant moments.

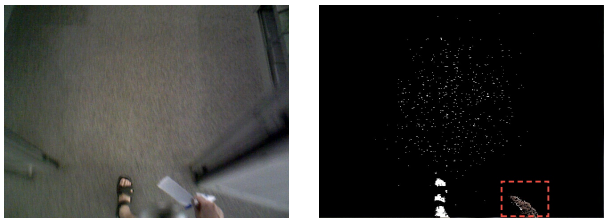
4.2. Skin segmentation based features

We can use our specific settings to design domain-specific features. In particular, we want to achieve egocentric activity recognition from a head mounted camera pointing down. We design an image signature that roughly encodes the distribution and location of skin pixels (arms and hands) in the image.

4.2.1 Skin segmentation

We start by implementing a standard color based skin segmentation step. There is a lot of prior work on how the best ways to segment skin color values. Some are using motion as well as color to perform segmentation (e.g. [10]) while other just focus on static image segmentation. In particular, following the survey in [22], we apply the following filter, (1), to the RGB values of each pixel. Given the R, G and B color values of a pixel p, in a standard range of values from 0 to 255, p is considered a skin pixel if the following expression is true:

$$\begin{aligned}(R > 95) \wedge (G > 40) \wedge (B > 20) \wedge \\ (\max\{R,G,B\} - \min\{R,G,B\} > 15) \wedge \\ (|R - G| > 15) \wedge (R > \max\{B,G\}).\end{aligned} \quad (1)$$



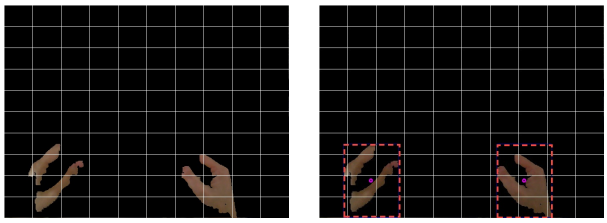
(a) (b)

Figure 2. Skin segmentation. (a) Original raw image (b) Segmentation using color and depth filtering. The white pixels (those that are NOT within the dashed rectangle) were accepted by the color filter but rejected by the depth filter. Depth helps us tp filter out feet and ground reflection pixels within the skin color range.

Including depth information. In addition to the color segmentation, we can use the depth information provided with the dataset (obtained with the RGB-d sensor). Since we aim for a simple signature, we propose to use the depth value as part of the filtering process, so we do not accept skin-color pixels if they are further away than a given threshold. This threshold should correspond to the maximum reach of a user (we set this threshold to 1m after measuring in 5 different users). As we can see in Fig. 2, this helps to get a better skin segmentation, by not only discarding real skin pixels that are further away (e.g., feet), but also rejecting spurious skin-colored regions (e.g., parts of the floor).

The inclusion of depth information is simple but allows us to be less strict with the range of acceptable colors. However, we should note that indeed the skin color segmentation is still very dependent on the users that acquired the dataset, therefore a user-based skin color calibration will be needed in a more general setting.

We also explored the use of another step that makes use of depth information: plane segmentation. Our motivation was to remove pixels corresponding to a dominant plane in the scene. This can easily be discovered by fitting a plane to the 3D points. Further, we can determine whether it corresponds to a table where the user is manipulating something, the floor, a wall, and so on. Once we know which pixels correspond to that surface, we can also suppress them from the skin pixel set (avoiding noise for instance when a wooden table has skin colored regions). However, after our initial experiments, we deemed the computational cost increase too high for the improvements obtained, so this filter is not used in the rest of experiments in this work.



(a) (b)

Figure 3. Skin segmentation based descriptors. (a) 10×10 grid on the skin-segmented image used to compute the *SKIN_HIST* descriptor. (b) Bounding boxes obtained in a sample image are represented by the dashed rectangles.

Including superpixel information. Finally, we have also explored superpixel segmentation step to complement the skin segmentation process. We use the fast SEEDS approach of [21] and assign skin or not skin to each superpixel depending on the average RGB color of the superpixel components. This step provides a more robust skin segmentation, and it is used in the rest of the experiments.

Since our final goal is not to achieve an accurate skin segmentation, the quality of this skin segmentation was only visually validated across all the sequences. The goal of this segmentation is to obtain a good starting point to compute the following features.

4.2.2 Skin pixels based features

Given the final skin pixel segmentation, we have explored a variety of descriptors that build on top of it, the most promising of which are as follows.

Skin histogram (*SKIN_HIST*) We divide the image into a 10×10 grid, as shown in Fig. 3(a), and build a histogram that represents the ratio of skin pixels contained in each cell.

Arms-Hands Bounding Box (*BB*) We detect the two largest connected components of skin pixels above a minimum size threshold and compute the bounding box around them as shown in Fig. 3(b). Each bounding box with is described with the following two descriptors:

- The eigenvectors of the scatter matrix of the skin pixels within each bounding box are used to characterize the *shape and orientation* of the set of skin pixels in that bounding box.

- The ratio of pixels in the bounding box that actually are skin-pixels measures the *density* of that bounding box content.

4.3. Convolutional Neural Network (CNN) Based Image Representation

Finally, we explore the performance of CNN based image representation for activity recognition. In this work, we use the implementation of [3] (DECAF). We use a model pre-trained to classify 1000 object categories from the Image Net 2012 challenge. We run a feed-forward pass for all of the video frames of our dataset and use the output of the seventh layer to represent the input frame.

After representing images with a 4096 dimensional feature vector, we trained a linear SVM [5] on labels such as manipulation/non-manipulation or different kinds of hand manipulation. Even though the network is trained to classify objects, the features seem general enough to give the best results on still image activity recognition as well.

5. Experiments

5.1. Dataset

The data we use in this work is part of the *Wearable Computer Vision Systems dataset* recently acquired with the purpose of comparing different wearable cameras for different wearable vision system applications. This dataset is available online[2].

We use the 5 labeled sequences available using an RGB-D camera. These were acquired by 4 different users in 2 different scenarios, where the users performed a set of actions as they were told but without a specified order and at their own pace. These sequences were manually labeled with different granularity level for activity labels as well as with location labels (not used in this work). They present a challenging classification problem with large intra-class variations (due to multiple users and scenarios) and very few labeled instances of less common actions.

5.2. Experimental setup

We have run a leave-one-out cross validation for our experiments. We trained classifiers based on the data from 4 sequences and tested it on the other sequence data. The final performance is reported as the average results across different tests. We use accuracy and confusion matrices to evaluate the different image representations and classification methods proposed in this work as baselines.

5.2.1 Classifiers and features used

The following combinations of descriptors (detailed in Section 4) are considered in our experiments:

GIST descriptor. Typically used for scene categorization, we use the implementation available from [4].

[2] https://i3a.unizar.es/es/content/wearable-computer-vision-systems-dataset

SKIN_HIST (Skin histogram): it is obtained from skin segmentation using color, depth and superpixel filters.

BB (Bounding Box): this descriptor is computed on the bounding boxes around the two largest skin connected components found.

CNN: We run the feed forward path for each frame in the sequence and take the output of the seventh layer as the representation.

CNN-MULTIWINDOW: this is the same as CNN with the difference that the CNN feature extraction was run on 5 windows of the images (the four corners and the center part). The representations are concatenated to make the full representation.

SIFT - BOW: we extracted SIFT [12] and used 1000 cluster centers to learn 1000 visual words. Input images are represented by a 1000 dimensional feature vector using standard word histogram encoding.

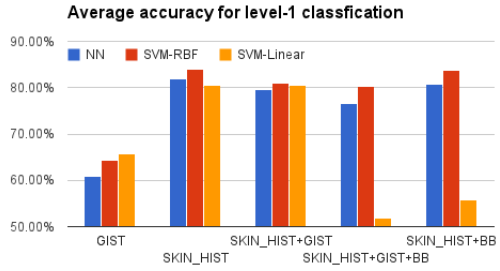
5.3. Performance evaluation

We present the experimental validation of our study organized in two sets of experiments. The first set analyzes the performance of compact and global descriptors and the differences observed when using different classifiers. The second set explores the performance of additional more sophisticated image representations and compares all the baselines proposed.

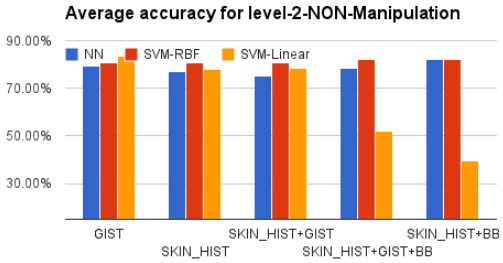
5.3.1 Performance of global image descriptors

Our first set of experiments analyzes different configurations of the global and compact image descriptors described in Sections 4.1 and 4.2. The goal is to evaluate which classification framework would be more suitable for these descriptors and the discriminative power of each of them for the activity recognition. Since these descriptors are compact and efficient to compute (just one descriptor per image), each combination has been used in three different classification frameworks: nearest neighbor (*NN*), linear SVM (*SVM-L*) and SVM with RBF kernel (*SVM-RBF*).

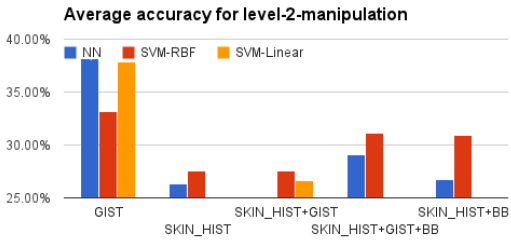
Single step classification As a motivation example for the hierarchical process, we ran a baseline experiment that consisted on training a single classifier for all the eleven level-2 classes at once, with different combinations of descriptors and classifiers. The best results obtained with the different combinations run were a raw accuracy (total number of correctly labeled tests divided by total amount of tests) around 35%. However, if we compute the accuracy normalized per class, it drops to around 15% (slightly better than chance). This means that the discriminative power of those features and the available amount of training data for each of those classes is not enough to directly distinguish among all of them. The classifier ended up assigning almost every test to the dominant test class. Note that the training was done as balanced as possible, but still too many classes did not have enough occurrences in the dataset.



(a)



(b)



(c)*

Figure 4. Each set of columns represents the average performance (correct classification) for all tests using labels from level 1 (a), level 2 - manipulation (b) and level 2 non-manipulation (c). *Note that the percentages are not normalized per class in this plot but per number of tests, therefore results in (b) can be misleading, since they are actually not better than the other levels, as can be seen in the confusion matrix presented in Table 4.

Classification in two steps If we run the two-step classification proposed, we obtain more promising results. Figure 4 shows the classification performance achieved in the two considered levels of the hierarchy of labels defined previously. The different bar plot sets correspond to different combinations of the descriptors detailed in sections 4.1 and 4.2.

The performance shown there is a raw accuracy measure, i.e., number of correct tests normalized by the total amount of tests. This gives us an initial idea of how much

Table 2. Confusion matrix for labels Manipulation vs Non-manipulation, using *Skin_hist* and *SVM_RBF*.

	Manip	Non-Manip
Manip	**0.84**	0.24
Non-Manip	0.16	**0.76**

Table 3. Confusion matrix for fine grained Manipulation labels considered, using *GIST* descriptor and *NN* classifier.

	two_hands	one_hand	pick_up	others
two_hands	**0.24**	0.06	0.20	0.28
one_hand	0.12	**0.44**	0.18	0.11
pick_up	0.54	0.38	**0.45**	0.35
others	0.11	0.13	0.18	**0.26**

Table 4. Confusion matrix for fine grained NON-Manipulation labels considered, using *GIST* and *SVM-L* classifier. Seat label is not shown because there were no occurrences in this sequence.

	Walk	Stairs	Stand	Screen	Poster	Talk
Walk	0.88	0.00	0.11	0.00	0.01	0.00
Stairs	0.99	0.01	0.01	0.00	0.00	0.00
Stand	0.30	0.00	0.58	0.00	0.12	0.00
Screen	0.00	0.00	1.00	0.00	0.00	0.00
Poster	0.61	0.00	0.28	0.00	0.11	0.00
Talk	0.39	0.00	0.58	0.00	0.03	0.00

of a whole sequence we can understand, but could be misleading because it is not normalized per class, as we saw in the previous example. If we compute the normalized accuracy per class, results in level 1 and level 2-*manip* remain similar, but accuracy normalized per class for level 2-*Non-manip* drops to around 25%. This indicates that this level of classification is not successful with the current image description. We can examine this in more detail using the confusion matrices. in Tables 2, 3 and 4.

We can see that in the last case, the classifier training stage was not successful at all, assigning most of the data into the same dominant class. As we observed in the initial (one step) experiment, the low performance is likely to be due to insufficient training: it was properly balanced for certain actions (level 1 and non-manip), but for others with fewer examples it clearly has insufficient information to obtain a robust classifier.

5.3.2 Performance using state-of-the-art image representations

In this second set of experiments, we explore more sophisticated (and more costly) image representations to determine whether the issues encountered with more compact features arose from a lack of descriptive power and/or scarcity of training examples.

Table 5. Confusion matrix for labels Manipulation vs Non-manipulation, **using *CNN* descriptor**.

	Manip	Non-Manip
Manip	**0.80**	0.20
Non-Manip	0.12	**0.88**

Table 6. Confusion matrix for labels Manipulation vs Non-manipulation, **using *CNN* descriptor**.

	two_hands	one_hand	pick_up	others
two_hands	**0.61**	0.06	0.22	0.10
one_hands	0.06	**0.63**	0.17	0.15
pick_up	0.14	0.11	**0.42**	0.33
others	0.15	0.20	0.27	**0.38**

Table 7. Confusion matrix for fine grained NON-Manipulation labels considered **using *CNN* descriptor**. Seat label is not shown because there were no occurrences in this sequence.

	Walk	Stairs	Stand	Screen	Poster	Talk
Walk	0.99	0.00	0.01	0.00	0.00	0.00
Stairs	0.09	0.91	0.00	0.00	0.00	0.00
Stand	0.87	0.02	0.11	0.00	0.00	0.00
Screen	0.00	0.00	0.00	0.00	0.00	0.00
Poster	0.61	0.38	0.00	0.00	0.02	0.00
Talk	0.92	0.00	0.02	0.00	0.00	0.06

Table 8. Accuracy obtained with the best performing options from all the image representations studied. Top rows are global representation. Bottom rows are the results for more sophisticated image representations.

Descriptor used:	Level 1	Level 2 Manip	Level 2 Non-Manip
SKIN-HIST (SVM-RBF)	**0.84**	0.27	0.80
GIST (SVM-L)	0.65	0.35	0.81
BoW	0.81	0.44	0.77
CNN	**0.84**	0.52	**0.83**
CNN-MULTI	0.83	**0.57**	0.77

We conclude this second set of experiments by comparing the best configurations using the proposed compact image description with more sophisticated features in Table 8. As one could expect, we can observe that for the basic classification, the contextual separation between manipulation and non-manipulation is nicely modeled by our simple description of how the skin (arms-hands) pixels are distributed in the images. However, for more complex and fine grained categorization, the preliminary results we have obtained with the CNN based representation look like a promising new path for activity recognition. Although these features

are more costly to compute, future steps include combining the preliminary results obtained in this work in such a way that the simple per frame classification can be used as a prior or decision step, to select key frames to set the stage for more detailed representations.

6. Conclusions and Future Work

In this paper, we presented results of our quantitative analysis of different feature extraction methods for the task of activity recognition using an RGB-D wearable vision system. Toward this end we make use of a novel and challenging public dataset and propose a hierarchy of labels for the included activities.

Our experiments show that classification in still frames with compact features can give good priors for more sophisticated classifiers/descriptors. Based on our experiments, CNN-based image features provide the best representation for finer grained activity recognition steps, compared to other baselines including bag of words representation or ad-hoc skin based descriptors.

There is still plenty of room for improvement based on the use of temporal consistency and increased leveraging of depth information within the image representation. Besides, the dataset used includes data from additional wearable cameras that recorded user activities simultaneously. In our future work we will pursue continued analysis on all of the camera/sensor streams to compare their strengths and weaknesses for the different classification tasks.

Acknowledgements

The authors would like to thank the colleagues in the University of Zaragoza, in special Luis Riazuelo and Javier Civera, for their help processing the dataset. This work is partially funded by a Google Focused Research Award and Spanish projects DPI2012-31781 and DPI2012-32100.

References

[1] D. Damen, A. Gee, W. Mayol-Cuevas, and A. Calway. Egocentric real-time workspace monitoring using an RGD-D camera. In *IROS*, 2012. 2

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[3] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013. 2, 4

[4] M. Douze, H. Jegou, H. Sandhawalia, L. Amsaleg, and C. Schmidt. Evaluation of GIST descriptors for web-scale image search. In *Int. Conf. on Image and Video Retrieval*, 2009. 4

[5] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *J. of Machine Learning Research*, 9:1871–1874, 2008. 4

[6] S. Hodges, E. Berry, and K. Wood. SenseCam: A wearable camera that stimulates and rehabilitates autobiographical memory. *Memory*, 19(7):685–696, 2011. 1

[7] O. H. Jafari, D. Mitzel, and B. Leibe. Real-time RGB-D based people detection and tracking for mobile robots and head-worn cameras. In *ICRA*, 2014. 2

[8] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012. 2

[9] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012. 2

[10] C. Li and K. M. Kitani. Pixel-level hand detection in egocentric videos. In *CVPR*, 2013. 3

[11] Y. Li, A. Fathi, and J. M. Rehg. Learning to predict gaze in egocentric video. In *ICCV*, 2013. 2

[12] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999. 5

[13] W. W. Mayol, A. J. Davison, B. J. Tordoff, N. Molton, and D. W. Murray. Interaction between hand and wearable camera in 2D and 3D environments. In *Proc. British Machine Vision Conference*, 2004. 2

[14] K. O'Hara, M. M. Tuffield, and N. Shadbolt. Lifelogging: Privacy and empowerment with memories for life. *Identity in the Information Society*, 1(1):155–172, 2008. 2

[15] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. Journal of Computer Vision*, 42(3):145–175, 2001. 3

[16] S. L. Phung, A. Bouzerdoum, and D. Chai Sr. Skin segmentation using color pixel classification: analysis and comparison. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(1):148–154, 2005. 2

[17] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. *arXiv preprint arXiv:1403.6382*, 2014. 2

[18] M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *CVPR*, 2013. 2

[19] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li. Hierarchical spatio-temporal context modeling for action recognition. In *CVPR*, 2009. 2

[20] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *ICCV*, 2003. 2

[21] M. Van den Bergh, X. Boix, G. Roig, B. de Capitani, and L. Van Gool. SEEDS: Superpixels extracted via energy-driven sampling. In *ECCV*. 2012. 4

[22] V. Vezhnevets, V. Sazonov, and A. Andreeva. A survey on pixel-based skin color detection techniques. In *Proc. Graphicon*, 2003. 2, 3