An Attention-based Activity Recognition for Egocentric Video

Kenji MATSUO, Kentaro YAMADA, Satoshi UENO, Sei NAITO
KDDI R&D Laboratories Inc.
2-1-15, Ohara, Fujimino-shi, Saitama, Japan
{ke-matsuo, kr-yamada, sa-ueno, sei}@kddilabs.jp

Abstract

In this paper, we propose a human activity recognition method from first-person videos, which provides a supplementary method to improve the recognition accuracy. Conventional methods detect objects and derive a user's behavior based on their taxonomy. One of the recent works has achieved accuracy improvement by determining key objects based on hand manipulation. However, such manipulation-based approach has a restriction on applicable scenes and object types because the user's hands don't always present significant information. In contrast, our proposed attention-based approach provides a solution to detect visually salient objects as key objects in a non-contact manner. Experimental results show that the proposed method classifies first-person actions more accurately than the previous method by 6.4 percentage points and its average accuracy reaches 43.3%.

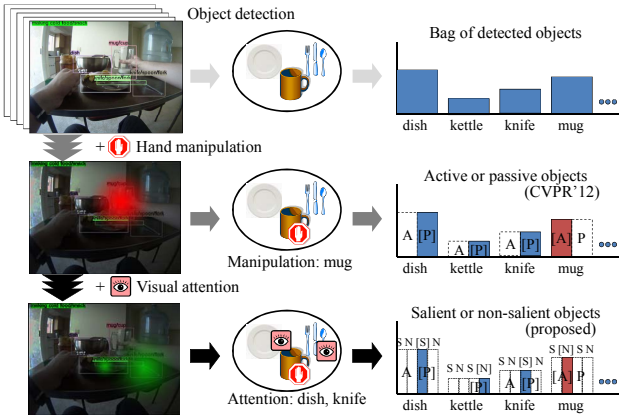

Figure 1. Activity recognition for ego centric video and an enhancement approach based on salient object detection: The user finishes drinking milk and puts the empty mug on the table. The user's next attentions are focused on the knife and bread served in a dish. It suggests that the user would desire spreading peanut butter.

1. Introduction

Egocentric video analysis for a user's own activities has attracted attention. Such a video is suitable for activity analysis since it is captured from a first person view. There has been a consecutive expectation on such video in welfare applications such as behavior support and home monitoring for disabled or elderly persons [1, 2]. In the welfare field, several types of domestic behaviors are defined as ADL - activity in daily living - such as "washing hands" and "drinking water". A home monitoring system achieves one-day analysis on ADL to examine a user's lifestyle. ADL is also used for rehabilitation. In recent years, glass-type camera devices have been provided for consumers, which expand application fields into daily life as well.

One pioneer study on activity analysis in daily life [3] has a unique approach that detects all objects from egocentric video in advance. After that, their taxonomy derives the user's own behavior. For example, if "a TV", "a sofa" and "a remote" appear in a frame, this situation can be estimated as "watching TV"; the previous study has reported another improvement by the hand-based approach. This approach classifies each detected object into an "active object" or "passive object" based on whether the user manipulates the object or not. "Active object" is considered a key object in activity estimation. However, the above approach has an inevitable issue, that is, there is a restriction on applicable scenes and object types. The user's hand does not always appear in all scene and some types of objects are not operated by hand.

In contrast, there have been several recent works on visual attentions [4-14]. These researches try to quantify the user's attentions in still or video images. Such information to identify objects to which the user pays strong attentions could be key factors in self-behavior recognition. Therefore, visual attention could provide a non-contact approach to determine important objects and solve the conventional issues on scene and object type restrictions.

This paper proposes an attention-based approach to achieve precise activity recognition from egocentric videos. In order to improve accuracy, the proposed approach

additionally quantifies the user's focus as a visual attention map and discriminates each object into two groups based on the user's attention. This is a supplemental approach to enforce the conventional approach [3]. Moreover, another contribution in this paper is to optimize visual attention for egocentric video. The conventional saliency map discards the user's own motions, although the ego motions have strong impacts to the user's attention. To this end, the proposed approach complements visual attention in a similar manner to the previous work [14]. Figure 1 summarizes the features of each approach.

2. Related work

Simplicity: There have been rich historical works on activity recognition in the life logging community [15-19]. Most approaches need to install special devices, such as RFID tags and accelerometers, into surrounding objects. Such preparation is a heavy burden. The proper approach does not require any installations to the user's surroundings. In contrast, activity analysis using egocentric video meets the requirements, since the user just needs to wear a camera to record his/her own activities.

Practicality: Some pioneers have studied activity recognition using a wearable camera. To advance the development under ideal conditions, these primary studies treat noise-less or repeating motions [20-22] and qualify them in a specific scene such as a kitchen and office scenes [23-26]. Therefore, it is difficult for such a method to be applied to practical scenes, including dining, living, and bathroom, as it is. This is because unpredictable fluctuations in the user's behaviors degrade the recognition accuracy. One of the goals in activity recognition is to develop an innovative approach that works well even in practical and in-the-wild environments.

Applicability: To improve recognition accuracy, some works have focused on the interactions between the user's hand and surrounding objects [27-30]. These approaches emphasize manipulated objects as keys to recognize activity. However, the operator's hands often disappear from the egocentric sight. Moreover, the types of applicable object that the user can actually touch are limited. In the previous work, there were only 5 types of active objects in all 21 types [3]. To avoid the above restrictions on scene and object type, the desirable approach should determine key objects in a non-contact manner. The attention-based approach could provide a promising solution for the conventional restrictions, since the visual attention can be measured from egocentric video without limitations on scene and object type.

Suitability: **A** saliency-based method is a modeling approach to extract the user's visual attention from image signals. Recent researches have also attempted to extend the saliency map toward video signals [4-13]. Most of the conventional methods using saliency maps are not available in ego activity analysis, since they are not applied sufficiently in egocentric view. There is a revisable chance that this study may represent a specific visual attention for egocentric video. Our motivation is inspired from a previous work. It also proposed an attention prediction for egocentric video [14]. In our method, the user's visual attention is also calculated not only from static saliency, but also from dynamic ego motions.

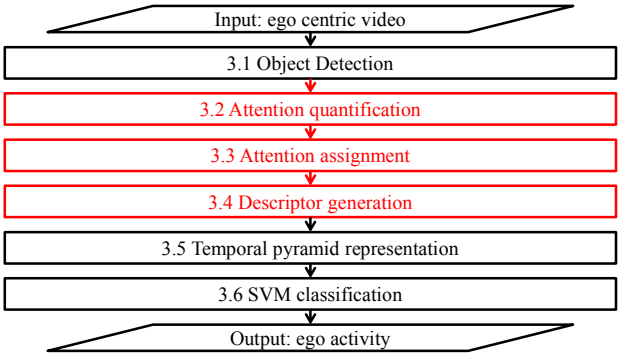

Figure 2. Block diagram of the proposed approach: In comparison to the conventional method [3], Steps 3.2, 3.3, and 3.4 are original and additional steps to achieve attention-based activity recognition.

3. Proposed method

Figure 2 shows the block diagram of the proposed method. Steps 3.1, 3.5, and 3.6 are identical to the steps of the conventional method [3]. Our proposed steps are Steps 3.2, 3.3, and 3.4 as highlighted in red. The concrete descriptions are given below.

3.1. Object detection

The first step is to detect objects from an egocentric video, applying an arbitrary method for object recognition, such as the part-based model [34]. The method provides the region $r(i)$ and the likelihood score $l(i)$ on each detected object i. We assume that the detection algorithm learned M types of objects. Then each detected object is classified into "active object" or "passive object" in the same manner as the hand-based approach [3]. Specifically, whether the user manipulates the object or not causes one of the two attributions. If two objects differ from each other in their types or attributions, this hand-based approach classifies them as different objects. The previous work [3] generates feature descriptor D as a histogram representation (d_1 … d_{2M}) as shown in Figure 1. The dimension is $2M$, since the hand-based approach gives each object two attributions. d_t indicates the cumulative likelihood of the object type t through all objects.

3.2. Attention quantification

Next, Step 3.2 extracts the user's interests from each frame and stores them in a visual attention map. The visual attention maps have the same resolution as the original images. The value at each point means the user's degree of attention on the corresponding pixel. Therefore, the visual attention map is used to estimate the user gaze point on the image. The saliency map is one approach to quantify visual attention and achieves a certain performance in practice. However, it doesn't work well for egocentric video. This is probably because the conventional saliency map discards the user's own motions, although the ego motions have strong impacts on the user's attention. To solve this adverse situation, the proposed approach complements visual attention in a similar manner to that of the previous work [14]. The final goal of this step is to extract the user's own motions as well as saliency from egocentric video, to integrate both observations, and to generate a visual attention map. The first operation measures camera rotation. It is basically equivalent to the user's own motion, that is, ego motion. Camera rotation is obtained through a commonly-used algorithm easily. This algorithm extracts local feature points frame by frame, and identical pairs are found between adjacent frames.

The effects of the above complement are explained as follows. Two objects appear on the opposite side within the user's line of sight. Even if the two objects have almost the same saliency, the above complement using ego motion enables correct estimation. Namely, if the user is turning his/her own head in the right direction, the user would give higher attention to the right object. To generate the saliency map, in this paper, we use the approach proposed by Chen et al. [13].

3.3. Attention assignment

This step finds the salient objects on which the user's sight is strongly focused. The first operation quantifies the magnitude of attention for each object i, which is calculated from the visual attention map based on the object region $r(i)$. It obtains $m(i)$ as the maximum value of the attention map within the region $r(i)$. The next operation finds pixels having a value less than or equal to $m(i)$ all over the visual attention map, and the relative magnitude of attention is defined as $a(i)$. Finally, each object is classified into "salient object" or "non-salient object" based on whether $a(i)$ is greater than the threshold T or not.

3.4. Descriptor generation

This step generates a descriptor but the operation flow is different from the conventional method [3]. The main difference is the reflection of visual attention. To reflect the changes, the detected object is classified into 4 further segments based on whether they are salient or non-salient as well as active or passive. They are treated as different segments in the descriptor. After that, the attention-based approach also changes another operation in measuring cumulative weighted likelihood. The magnitude of attention $a(i)$ weights likelihood $l(i)$ in the following equations before the cumulative operation.

$$S(i) = a(i) \cdot l(i) \quad (1)$$

$$N(i) = \{ 1 - a(i) \} \cdot l(i) \quad (2)$$

Where $S(i)$ and $N(i)$ are weighted likelihoods for salient and non-salient objects, respectively. Finally, descriptor D is represented as a $4M$ dimensional histogram $(d_1, \ldots, d_{4M})$.

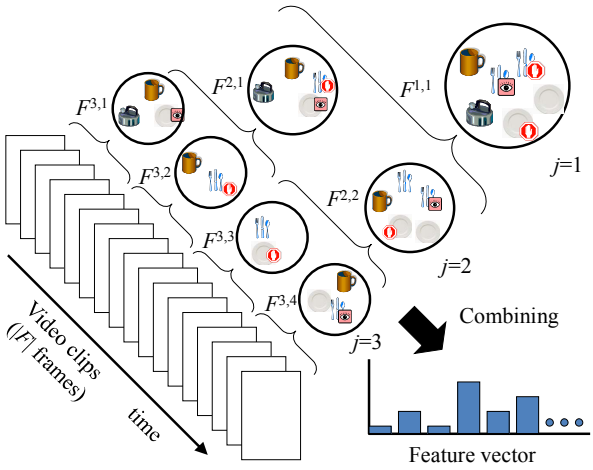

Figure 3. Temporal pyramid representation of feature vector: The target unit consists of $|F|$ frames and the number of layers is set to 3. Step 3.5 generates a feature vector through temporally layered and multiplexed operations of descriptors.

3.5. Temporal pyramid representation

This step finalizes the feature vector generation in a similar manner to that of the conventional temporal pyramid approach [3]. It provides temporal robustness to activity recognition. Naive descriptor D_f in each frame f is not sufficiently robust to combat temporal fluctuations. Therefore, to achieve the robustness against fluctuations, the temporal pyramid approach layers and multiplexes descriptors at a constant temporal interval. Figure 3 shows the conceptual diagram of operational flow. The target unit is a subset F consisting of $|F|$ frames. The number L of total layers is set to 3. The main operation at j-th layer, j=1, 2 … L, divides F equally into 2^{j-1} subsets, k=1, 2 … 2^{j-1}, and computes average descriptor $D^{j,k}$ of each subset.

$$D^{j,k} = \frac{1}{|F^{j,k}|} \sum_{f \in F^{j,k}} D_f \quad (3)$$

The feature vector V is obtained by combining all $D^{j,k}$, that is, $V = [D^{1,1} \ldots D^{j,k} \ldots D^{L,2^{L-1}}]$. Finally the dimension of V is equal to the integer multiple sizes of the original descriptor D_f.

Average operation discards all differential information in temporal order. Therefore, the pyramid representation can suppress temporal fluctuations in activity recognition. Temporal order is actively neglected in the upper layer. On the contrary, the temporal order is flexibly retained in the lower layer. For instance, in the category of "making tea", the user puts a kettle on a stove and pours boiling water into a cup. Temporal information is important. To achieve efficient recognitions even for such situations, the bottom layers maintain supplemental information on the temporal process [31, 32].

3.6. SVM classification

Step 3.6 adopts the support vector machine to learn and recognize the user's own activity from feature vector V. The SVM needs samples for training before the recognition phase. The first preparation in the training phase is to label egocentric video with ground truth activities. Then applying the above Steps 3.1 to 3.5, the feature vectors can be obtained as samples. After training them, a SVM classifier is finally generated.

4. Experiments

This section presents the experiments and results to evaluate the improvement and effectiveness of the proposed approach. All experiments were conducted using the public dataset on egocentric video in daily living activity [33].

4.1. Conditions

A GoPro camera was used to build the dataset. The camera captures high definition quality video corresponding to 1280 x 960 resolution with 170 degrees of viewing angle. The video frame-rate is 30 fps. The dataset consisted of 20 people's egocentric videos recorded in their own apartment. The previous work also provided all video with grand truth labels on M=21 types of object as well as 18 actions of daily activities [3]. As for the video length per video, the average, minimum and maximum were 27 minutes 8 seconds, 8 minutes 12 seconds and 1 hour 3 minutes 19 seconds, respectively. To establish the same condition as the previous approach [3], the part-based model [34] was adopted for object recognition in Step 3.1. In all experiments, 6 people's videos were assigned for training and the remaining 14 people's videos were used for testing. The final preparation detected objects from the testing videos. In evaluation phases, the leave-one-out method for the 14 people's video was applied to obtain reliable precision in activity recognition. In all experiments, the discriminative threshold T for the salient object was fixed to a constant value of 0.964 and the number L of the temporal layer was set to 2.

Table 1. Recognition accuracy and analytical data: For 18 types of domestic activities, this table provides comparable figures on recognition accuracy between the conventional [3] and proposed approach. The "Appearance rate" column means the appearance proportion of positive likelihood in active objects.

Recognition accuracy			
	Conv.	Prop.	Gain
Average	36.9%	43.3%	+6.4
Variance	9.8%	7.1%	-2.7

Average record length
827.7 sec
Appearance rate - active object
26.3%

Figure 4. Experimental results of object detection and visual attention map: The left figure indicates that Step 3.1 provides individual objects and their regions. The right figure is generated in Step 3.2. It suggests that the user focuses strong attention on the central kettle.

4.2. Results

Figure 4 demonstrates the results of the object detection and visual attention map. Step 3.1 provides individual objects and their regions. In this scene, the user is pouring water into a kettle. Then, Step 3.2 generates a visual attention map as shown in the right figure. The map suggests an acceptable result that the user pays high attention to the center of the frame to pour water.

Table 1 shows the comparative results between the proposed attention-based approach and the conventional hand-based approach [3]. The table indicates the average and the variance of recognition accuracy in 18 activity categories. The average appearance proportion of active objects is also shown in the table.

Improvement: The attention-based approach can maintain or increase recognition accuracy in 12 activity categories in comparison with the hand-based approach. In average accuracy, the proposed method achieves superior recognition to the conventional method by 6.4 points, that is, from 36.9% to 43.3%.

Stabilization: The class variance decreases from 9.8% to 7.1%. This suggests that the attention-based approach can achieve more stable recognition toward any activities.

Applicability: Table 1 suggests an interesting fact. The appearance proportions of active objects are not so high in most categories and the average is 26.3%. In such case, since active objects don't contribute to activity determination so much, it could be quite difficult for the hand-based approach to achieve sufficient precision. On the contrary, the attention-based approach indicates a better result even for such situations. It concludes that the visual attention can provide a solution to determine key objects without restrictions on scene and object type.

Suitability: Another experiment was conducted to confirm that the visual attention map is superior to the standard saliency map in activity recognition. In the saliency-based approach, the average accuracy reaches 35.1%. This is 8.2 points lower than the attention-based approach. This result suggests that saliency is not sufficient to estimate the user's attentions in egocentric video. To improve recognition accuracy, ego motion also needs to be reflected in determining important objects.

4.3. Considerations

Figure 5 shows the recognition accuracy in each activity. The proposed method unfortunately degrades the accuracy in 5 categories. Test videos recorded in three of the categories, that is, "dental floss", "watching TV", and "using a computer", consist mostly of moderate and motionless scenes. For example, there are hardly large motions in "dental floss", since a user stands in front of a mirror constantly and just slightly moves his/her own arm. It is difficult for the attention-based approach to recognize activity, because ego motion is too simple to estimate and feature a user's attention significantly. In such motionless scenes, the proposed method often tends to extract visual attention around the center of the user's line of sight.

In contrast, the remaining categories, that is, "making tea and "making coffee", contain active and rapid scenes. For instance, the user pours water from a kettle into a mug. However, the attention based approach also causes adverse situations in the two categories. This means that salient or non-salient is assigned in succession to a specific object through video sequencing. It is easy to change the attribution of the salient mug into non-salient and vice versa. This situation is unstable in the attention-based approach. For these categories, the above fluctuations cause degradation in performance, because the random assignment of a salient object prevents the specific determining of key objects.

5. Conclusions

To provide an effective algorithm in activity recognition, this paper proposed an attention-based approach for egocentric video. The remarkable features are that it quantifies visual attentions and determines key objects in activity recognition based on the attention map. The experimental results confirmed that the attention-based approach exceeds the conventional approaches in recognition accuracy. One of our future works is to suppress the fluctuations in assigning salient or non-salient, which is mentioned in Section 4.3. To provide robustness against such fluctuations, further study will attempt to apply a temporal complement on visual attention information.

Although the proposed flow requires additional steps and increases computational costs by several times, it could be applicable even in practical scenes. For instance, the offline batch process enables one day's activity analysis during the night. The proposed flow also has future potential to achieve speeding up and real-time processing. The cost reduction in the computations is another future work.

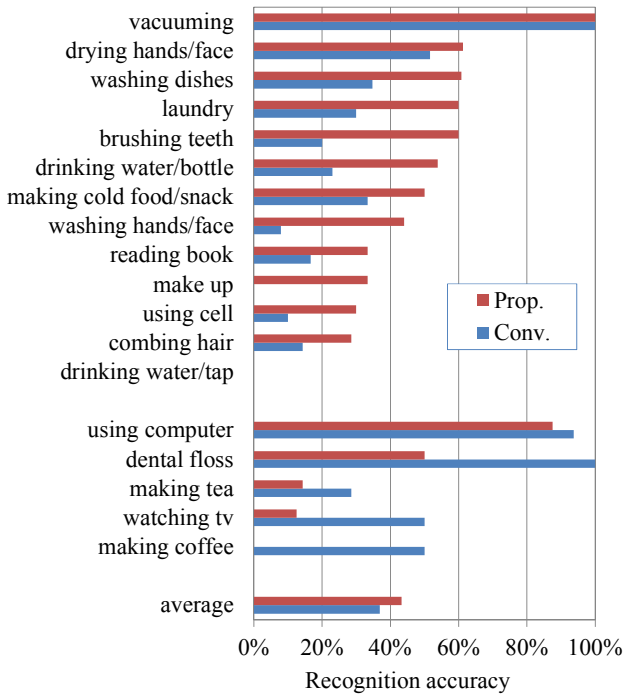

Figure 5. Recognition accuracy: This graph provides comparable figures on recognition accuracy between the conventional [3] and proposed approach for 18 types of domestic activities.

References

[1] A. Catz, M. Itzkovich, E. Agranov, H. Ring, and A. Tamir. SCIM-spinal cord independence measure: a new disability scale for patients with spinal cord lesions. Spinal Cord, 35(12):850–856, 1997.

[2] B. Kopp, A. Kunkel, H. Flor, T. Platz, U. Rose, K. Mauritz, K. Gresser, K. McCulloch, and E. Taub. The Arm Motor Ability Test: reliability, validity, and sensitivity to change of an instrument for assessing disabilities in activities of daily living. Arch. of Physical Medicine and Rehabilitation, 78(6), 1997.

[3] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.
[4] M. Cerf, J. Harel, W. Einhauser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. Advances in Neural Information Processing Systems (NIPS), Vol. 20, pp. 241-248, 2007.
[5] L. Itti, N. Dhavale, F. Pighin, et al. Realistic avatar eye and head animation using a neurobiological model of visual attention. In SPIE 48th Annual International Symposium on Optical Science and Technology, Vol. 5200, pp. 64-78, 2003.
[6] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. Advances in Neural Information Processing Systems (NIPS), Vol. 19, pp. 545-552, 2006.
[7] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology, Vol. 4, No. 4, pp. 219-227, 1985.
[8] L. Itti, C. Koch, and E. Neibur. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 20, No. 11, pp. 1254-1259, 1998.
[9] T. Avraham and M. Lindenbaum. Esaliency (extended saliency); Meaningful attention using stochastic image modeling. IEEE Transactions on Pattern Analysis and Machine intelligence (PAMI), Vol. 32, No. 4, pp. 693-708, 2010.
[10] L. F. Coasta. Visual saliency and attention as random walks on complex networks. ArXiv Physics e-prints, 2006.
[11] W. Wang, Y. Wang, Q. Huang, and W. Gao. Measuring visual saliency by site entropy rate. In Computer Vision and Pattern Recognition (CVPR), pp. 2368-2375, IEEE, 2010.
[12] T. Foulsham and G. Underwood. What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. Journal of Vision, Vol. 8, No. 2;6, pp. 1-17, 2008.
[13] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu. Global contrast based salient region detection. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.
[14] K. Yamada, Y. Sugano, T. Okabe, Y. Sato, A. Sugimoto, and K. Hiraki. "Attention prediction in egocentric video using motion and visual saliency," in Proc. 5th Pacific-Rim Symposium on Image and Video Technology (PSIVT) 2011, vol.1, pp. 277–288, Nov. 2011.
[15] M. Blum, A. Pentland, and G. Troster. Insense: Interest based life logging. Multimedia, IEEE, 13(4):40–48, 2006.
[16] D. Patterson, D. Fox, H. Kautz, and M. Philipose. Fine-grained activity recognition by aggregating abstract object usage. In IEEE Int. Symp. On Wearable Computers, 2005.
[17] A. Pentland. Looking at people: Sensing for ubiquitous and wearable computing. IEEE TPAMI, 22(1):107–119, 2002.
[18] M. Philipose, K. Fishkin, M. Perkowitz, D. Patterson, D. Fox, H. Kautz, and D. Hahnel. Inferring activities from interactions with objects. IEEE Pervasive Computing, 2004.
[19] E. Tapia, S. Intille, and K. Larson. Activity recognition in the home using simple and ubiquitous sensors. Pervasive Computing, pages 158–175, 2004.
[20] I. Laptev and P. Perez. Retrieving actions in movies. In ICCV, 2007.
[21] E. Spriggs, F. De La Torre, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In IEEE Workshop on Egocentric Vision, 2009.
[22] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. Rehg. A scalable approach to activity recognition based on object use. In ICCV, pages 1–8. IEEE, 2007.
[23] A. Fathi, A. Farhadi, and J. Rehg. Understanding egocentric activities. In ICCV, 2011.
[24] M. Hanheide, N. Hofemann, and G. Sagerer. Action recognition in a wearable assistance system. In ICPR, 2006.
[25] L. Sun, U. Klank, and M. Beetz. Eyewatchme3d hand and object tracking for inside out activity analysis. In IEEE Workshop on Egocentric Vision, 2009.
[26] S. Sundaram and W. Cuevas. High level activity recognition using low resolution wearable vision. In IEEE Workshop on Egocentric Vision, 2009.
[27] M. Argyle and B. Foss. The psychology of interpersonal behaviour. Penguin Books Middlesex. England, 1967.
[28] A. Fathi, X. Ren, and J. Rehg. Learning to recognize objects in egocentric activities. In CVPR, 2011.
[29] M. Land, N. Mennie, and J. Rusted. The roles of vision and eye movements in the control of activities of daily living. Perception, 28(11), 1999.
[30] W. Mayol and D. Murray. Wearable hand activity recognition for event summarization. In International Symposium on Wearable Computers. IEEE, 2005.
[31] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In CVPR, 2006.
[32] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders. Kernel codebooks for scene categorization. In ECCV, 2008.
[33] http://deepthought.ics.uci.edu/ADLdataset/adl.html, access date 2013/07/05.
[34] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. IEEE TPAMI, 32(9), 2010.