

Recognition of Instrumental Activities of Daily Living in Egocentric Video for Activity Monitoring of Patients with Dementia

Iván González-Díaz, Vincent Buso, Jenny Benois-Pineau, Guillaume Bourmaud, Gaele Usseglio, Rémi Mégret, Yann Gaestel, and Jean-François Dartigues

1 Introduction

The task of recognizing human activities in videos has become a fundamental challenge among the computer vision community [15]. In order to face the limited field of view and the difficulty of accessing all relevant information from fixed cameras, an alternative has been found in egocentric videos, recorded by cameras worn by subjects. Indeed, in addition to dealing with the previously listed drawbacks, wearable cameras represent a cheap and effective way to record users activity for scenarios such as telemedicine or life-logging.

In this chapter we focus on the problem of recognizing Instrumental Activities of Daily Living (IADL) for the assessment of the ability of patients suffering from Alzheimer disease and age-related dementia. Indeed, an objective assessment of a patient's capability to perform IADLs is a part of clinical protocol of dementia

I. González-Díaz

Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Leganés, 28911, Madrid, Spain

e-mail: igonzalez@tsc.uc3m.es

V. Buso • J. Benois-Pineau (✉)

LaBRI, UMR 5800 CNRS/University of Bordeaux 1, Talence, France

e-mail: vbuso@labri.fr; benois-p@labri.fr

G. Bourmaud • G. Usseglio • R. Megret

IMS, UMR 5218 CNRS/University of Bordeaux 1, Talence, France

e-mail: guillaume.bourmaud@ims-bordeaux.fr; remi.megret@ims-bordeaux.fr

Y. Gaestel • J.-F. Dartigues

ISPED, U897 INSERM, Bordeaux, France

e-mail: yann.gaestel@isped.u-bordeaux2.fr; dartigues@isped.u-bordeaux2.fr

diagnostics and evaluation of efficacy of therapeutical treatment [1]. Traditional ways of assessment with the help of questionnaires do not bring satisfaction as two kinds of errors have been observed, which do not allow a practitioner to fully trust the responses. The error of the first kind is that one committed by the patients. At the early stage of dementia, they cannot admit that they become less performant in their everyday activities and diminish their difficulties. The error of the second kind is committed by the caregivers. They are permanently stressed watching their relatives' mental capacities to deteriorate. Hence they over estimate the difficulties of patients with dementia [17]. This is why the egocentric video has been first used for the recording of IADLs on patients with dementia in [22]. Later on, first results of recognition of IADLs in such recorded video were reported in [19]. Nevertheless, the recognition problem being very complex, efficient ways of solving it still remain an open research issue.

There has been a fair amount of work on recognizing everyday at home activities by analyzing egocentric videos, many of them based on the fact that manipulated objects represent a significant part of the actions. However, most of the studies were conducted under a constrained scenario, in which all the subjects wearing the cameras perform actions in the same room and, therefore, interact with the same objects: e.g. a hospital scenario in which the medical staff asks patients to perform several activities. Typical constrained scenarios allow to make assumptions on the objects or even to use instance-level visual recognition: the authors in [12] present a model for learning objects and actions with very little supervision, whereas in [28] a dynamic Bayesian network that infer activities from location, objects and interactions is proposed. The problem still open under such a scenario becomes even more complex, if an ecological observation is performed, i.e. at person's home. The individual environment varies, the objects of the same usage, e.g. a tea-pot or a coffee machine, can be of totaly different appearance. We call this scenario "unconstrained". In this case the recognition of activities in a wearable camera video has to be funded on the features of higher abstraction level than simple image and video descriptors computed from pixels.

It is only recently that the more challenging unconstrained scenario has been examined regarding activity recognition, such as in the work of [21], where the authors recognize ego-actions in outdoor environments using a stacked Dirichlet Process Mixture model. Pirsivash and Ramanan [23] propose to train classifiers for activities based on the output of the well-known deformable part model [14] using temporal pyramids. They demonstrate that performances are dramatically increased if one has knowledge of the object being interacted with. The approach making use of these "active" areas for ADL recognition has also been studied by Fathi and al. in [13] under a constrained scenario, where the authors enhanced their performances by defining visual saliency maps.

In the context of medical research on Alzheimer disease the unconstrained scenario means an epidemiological study of performances of patients in an ecological situation at their homes as it was done in [20]. Hence in this paper we model an activity as a combination of a meaningful object the person interacts with and the environment. The rationale here is quite straightforward. Indeed a reasonable

assumption can be made that e.g. if a person is manipulating a tea pot in front of a kitchen table, than the activity consists in “making tea”. If a TV set is observed in the camera view field and the person is in living room, then the activity would be “watching TV”. Therefore, efficient recognition approaches have to be proposed for object recognition and localization of a person in its environment, and, more than that an efficient combination of results of these two detectors have to be designed in the activity recognition framework.

We therefore make the following contributions: (1) we further develop object recognition approach with psycho-visual weighting by saliency maps [5], (2) we show that analyzing the dynamics of a sequence of active objects + context by means of temporal pyramids [23] becomes a suitable paradigm for activity recognition in egocentric videos. However, in this optic we claim that context can be better described by the output of place recognition module rather by the outputs of many non-active object detectors as proposed in [23]. We provide experimental evaluation on a publicly available dataset of activities in egocentric videos.

The remainder of the paper is organized as follows: in Sect. 2 we describe the involved modules in our activity recognition approach. Section 3 assesses our model and compares it to the current state-of-the-art performances and Sect. 4 draws our main conclusions and introduces our further research.

2 The Approach

We aim to recognize IADLs by analyzing human-object interactions and as well as the contextual information surrounding them. Hence let us firstly introduce the notion of an ‘*active object*’ (AO). An AO is an object which the subject/patient wearing camera interacts with. Here the interaction is understood as manipulation or observation. We claim that the analysis of this kind of objects becomes the main source of information for the activity recognition, and that the explicit recognition of ‘non-active’ objects as in [20] is not longer needed. We suppose that they can be efficiently encoded in a global descriptor of the scene/context. The activity model is therefore understand as the interaction with specific objects (AO) in a specific environment (context). In this particular work, we have considered that context can be successfully represented by identifying the place in which the user is performing the activity.

We propose a hierarchical approach with two connected processing layers (see Fig. 1). The first layer contains a set of *Active Object detectors* (Sect. 2.1) and a *Place Recognition* system (Sect. 2.2). Hence it allows for identification of the elements of our activity model. The second one addresses the activity recognition task on the basis of identified elements (Sect. 2.3).

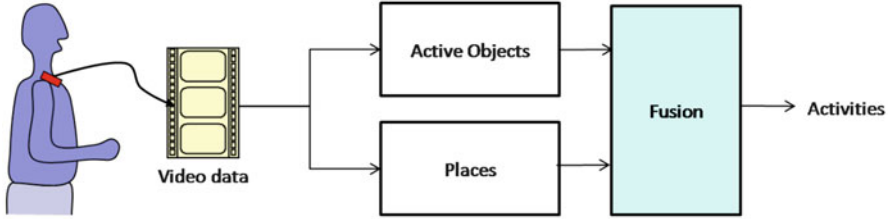


Fig. 1 Processing pipeline for the activity recognition

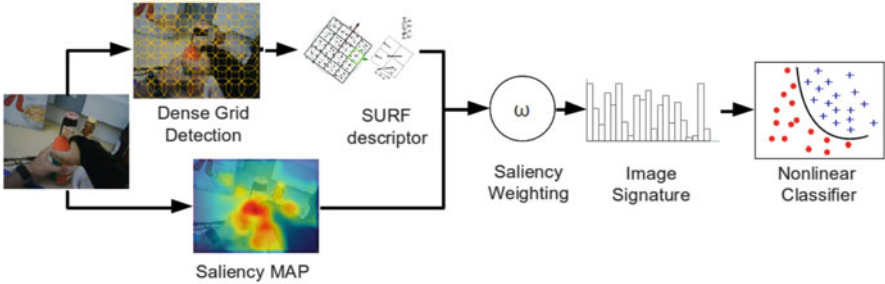


Fig. 2 Processing pipeline for the saliency-based object recognition in first-person camera videos

2.1 Object Recognition

As already mentioned, we aim to recognize activities under an unconstrained scenario in which each video is recorded at a different place. This is therefore a more difficult task than the recognition of specific objects instances. It remains an open problem for the computer vision community.

In general, we consider one individual detector for each object category although, as shown in the processing pipeline presented in Fig. 2, the nonlinear classification stage is the only step that is specific for each category. We have built our model on the well-known Bag-of-Words (BoW) paradigm [9] and proposed to add saliency masks as a way to provide spatial discrimination to the original Bag-of-Words approach. Hence, for each frame in a video sequence, we extract a set of N SURF descriptors d_n [3], using a dense grid of circular local patches. Next, each descriptor d_n is assigned to the most similar word $j = 1..V$ in a visual vocabulary by following a vector-quantization process. The visual vocabulary, computed using a k-means algorithm over a large set of descriptors in the training dataset (about 1M descriptors in our case), has a size of $V = 4,000$ visual words.

In parallel, our system generates a geometric-spatio-temporal saliency map S of the frame with the same dimensions of the image and values in the range $[0,1]$ (the higher the more salient a pixel is, see Fig. 3). Details about the generation of saliency maps can be found in [5]. Here we briefly remind the key components for prediction a salient area in wearable video.



Fig. 3 Illustration of the different saliency cues (geometrical, spatial and temporal) composing the spatio-temporal saliency maps from [5]

First of all the saliency prediction approach we follow is a “bottom-up” one, or stimuli—driven. This means that the local characteristics of video frames are used to model attraction of Human Visual Attention (HVA) by the elements of visual scenes. Since the fundamental work by Itti [18] for stills and later developed models for video [4], bottom-up models for prediction of visual attention incorporate three cues: (i) spatial, (ii) temporal, (iii) geometric. Spatial cue stands for sensitivity of Human Visual System (HVS) to luminance and colour contrasts and orientation in image plane. Temporal cue expresses its sensitivity to motion. Finally, the geometrical cue usually expresses the so-called “central bias” hypothesis put forward by Buswell [6].

As for the spatial cue, various local filtering approaches have been proposed in order to compute local contrast and orientation. In our work as in [5] we used local operators allowing for computation of seven local contrast features in HSI colour domain: Contrast of Saturation, Contrast of Intensity, Contrast of Hue, Contrast of Opponents, Contrast of Warm and Cold Colors, Dominance of Warm Colors, and Dominance of Brightness and Saturation. These features proposed by Aziz and Mertsching [2] proved to be efficient for predicting sensitivity of HVS to colour contrasts. Then the spatial saliency map value for each pixel in a frame is a mean of these features.

Motion saliency map was built on the basis of non-linear sensitivity of HVS to motion magnitude, proposed by Daly [10] and expressing the fact that HVS is not sensitive to very low motion magnitude and to a very high motion magnitude neither. As the measure of motion magnitude we took the “residual motion” which is a local motion observed in image plane after compensation of camera motion according to affine motion model [5].

Finally, we devoted a specific study to the geometrical cue, which also is a non-trivial question in case of body-worn cameras. Indeed Buswell’s hypothesis of central bias is not hold when the body-worn camera is not fixed on “symmetry axis” of human head as it was the case in the study [5]. Nevertheless, a reasonable assumption that the gaze direction coincides with the camera optical axis orientation can be made when the camera is fixed on the body such as on glaces or in a central position on the chest. (Note this is the case in the dataset we use in this chapter for experiments). In this case the geometric saliency map can be modelled by an isotropic gaussian with a spread $\sigma = 5$ visual degrees centered on image center.

All three maps: spatial $S_{sp}(x, y)$, temporal $S_t(x, y)$ and geometric $S_g(x, y)$ are normalized by their respective maximum, which is called “saliency peak”. An illustration of these three cues is given in Fig. 3.

The resultant saliency map is obtained from the three normalized saliency maps by a linear combination

$$S(x, y) = \alpha * S_{sp}(x, y) + \beta * S_t(x, y) + \gamma * S_g(x, y) \quad (1)$$

The coefficients α, β, γ are estimated by linear regression with regard to ideal maps from a training set that is gaze fixation maps of subjects.

Coming back to the visual signature of a video frame, we use the resultant saliency map to weight the influence of each descriptor in the final image signature, so that each bin j of the BoW histogram H is computed following the next equation:

$$H_j = \sum_{n=1}^N \alpha_n w_{nj} \quad (2)$$

where the term $w_{nj} = 1$ if the descriptor or region n is quantized to the visual word j in the vocabulary and the weight α_n is defined as the maximum saliency value S found in the circular local region of the dense grid. Finally, the histogram H is L1-normalized in order to produce the final image signature.

Once each image is represented by its weighted histogram of visual words, we use a SVM classifier [8] with a nonlinear χ^2 kernel, which has shown good performance in visual recognition tasks working with normalized histograms as those ones used in the BoW paradigm [27]. Using the Platt approximation [24], we finally produce posterior probabilistic estimates O_k^t for the occurrence of the object of class k in the frame t .

2.2 Place Recognition

In this section we detail the place recognition module. Place recognition plays a role of *context recognition* in our overall approach for IADLs modeling and recognition.

The general framework can be decomposed into three steps. First of all, for each image, a global image descriptor is extracted. We choose the Composed Receptive Field Histograms (CRFH) [25] since it was proven to perform well for indoor localization estimation [11]. Then a non-linear dimensionality reduction method is employed. In our case, we use a Kernel Principal Component Analysis (KPCA) [26]. The purpose of this step is twofold: it reduces the size of the image descriptor which alleviates the computational burden of the rest of the framework, and it provides descriptors on which linear operations can be performed. Finally, based on these features, a linear Support Vector Machine (SVM) [8] is applied to perform the place recognition, and the result is regularized using temporal accumulation [11].

For the application considered in this paper, each video is taken in a different environment. Consequently, our module has to learn generic concepts instead of specific ones as it is usually the case [11]. In this context, we need to define concepts both relevant for action recognition and as constrained as possible to obtain better performances. Indeed, for example the concept ‘stove’ has probably less variability and may be more meaningful for action recognition than the concept ‘kitchen’. This will be discussed in detail in Sect. 3.3.

Again, following the Platt approximation [24], the output of this module is then a vector P_j^t with the probability of a frame t representing the place j .

2.3 Activity Recognition

Our activity recognition module uses the temporal pyramid of features presented in [23], which allows to exploit the dynamics of user’s behaviour in egocentric videos. However, rather than combining features for active/non-active objects, we represent activities as sequences of AOs and places (context). For instance, cooking may involve user’s interaction with various utensils whereas cleaning the house might require a user to move around various places of the house.

In particular, for each frame t being analyzed, we consider a temporal neighborhood Ω_t corresponding to the interval $[t - \Delta/2, t + \Delta/2]$. This interval is then iteratively partitioned into two subsegments following a pyramid approach, so that at each level $l = 0 \dots L - 1$ the pyramid contains 2^l subsegments. Hence, the final feature of a pyramid with L levels is defined as:

$$F_t = \left[F_t^{0,1} \dots F_t^{l,1} \dots F_t^{l,2^l} \dots F_t^{L-1,2^{L-1}} \right] \quad (3)$$

where $F_t^{l,m}$ represents the feature associated to the subsegment m in the level l of the pyramid and is computed as:

$$F_t^{l,m} = \frac{2^l}{\Delta} \sum_{s \in \Omega_{tm}^l} f_s \quad (4)$$

where Ω_{tm}^l represents the m temporal neighborhood of the frame t in the level l of the pyramid and f_s is the feature computed at frame s in the video. In the experimental section, we will assess the performance of our approach using the outputs of K object detectors $[O_1^s \dots O_K^s]$, the outputs of J place detectors $[P_1^s \dots P_J^s]$, or the concatenation of both, as features f_s .

In this work, we have used a sliding window method with a fixed window of size Δ , parameter that is later studied in the Sect. 3, and a pyramid with $L = 2$. Finally, the temporal feature pyramid has been used as input for a linear multiclass SVM in charge of deciding the most likely action for each frame.

The complexity of the classifier system, being layered, precludes the easy interpretation of the results as probabilistic elements, as they are defined on an arbitrary axis that is suitable for deciding of a best class, but not to associate a probabilistic interpretation to it. Since automatic activity recognition from wearable camera is a difficult problem, it is very important to be able to assign confidence measures to these predictions, in order to monitor their validity and uncertainty for higher level inference. This problem corresponds to a calibration problem [16]. Even though the automatic detection of all possible events is not possible in all cases, computing confidences can mitigate this, by trusting the prediction only when the system is confident.

For a two-class classifier, each observation x_k (in our case the input features belonging to a multi-dimensional space) is associated a predicted binary label y_n in $\{0, 1\}$. In practice the prediction is based on the thresholding of the classifier score s_n , which is produced by the decision function as $s_n = f(x_n)$. The calibration problem consists in finding a transformation $p_n = g(s_n)$ of these scores into a value in the interval $[0, 1]$ such that the result can be interpreted as the probability $p_k = P(y_k = 1|x_k)$ that of a true positive conditioned on the observed sample. The calibrated values have then reasonable properties to be used in a fusion approach with other sources of information.

In our work, we used the Platt approach [24], generalized to the one-to-one multi-class classification [29] and detailed in [7]. Each test sample is therefore associated with probabilistic confidence value $p_{kc} = P(L_k = c|x_k)$ that it belongs to class c , such that it is normalized by $\sum_c p_{kc}$.

The experimental part will evaluate both the raw recognition performance, using the classification strategy that assigns a sample to the class with higher probability, as well as the reliability of the estimated confidence value.

3 Experimental Section

3.1 Experimental Set-up

We have assessed our model in the ADL dataset, proposed by the authors of [23], that contains videos captured by a chest-mounted GoPro camera on 20 users performing various daily activities at their homes. This dataset was already annotated for 44 object-categories and 18 activities of interest (see Fig. 4) and we have additionally labeled 5 rooms and 7 places of interest.

This dataset is very challenging since both the environment and the object instances are completely different for each user, thus leading to an unconstrained scenario. Hence, and due to the hierarchical nature of the activity recognition process, we have trained every module following a leave-k-out procedure ($k = 4$ in our approach). This approach allows us to provide real testing results in object and place recognition for every user, so that the whole set can be later used for activity



Fig. 4 Overview of the 18 activities annotated in the ADL dataset

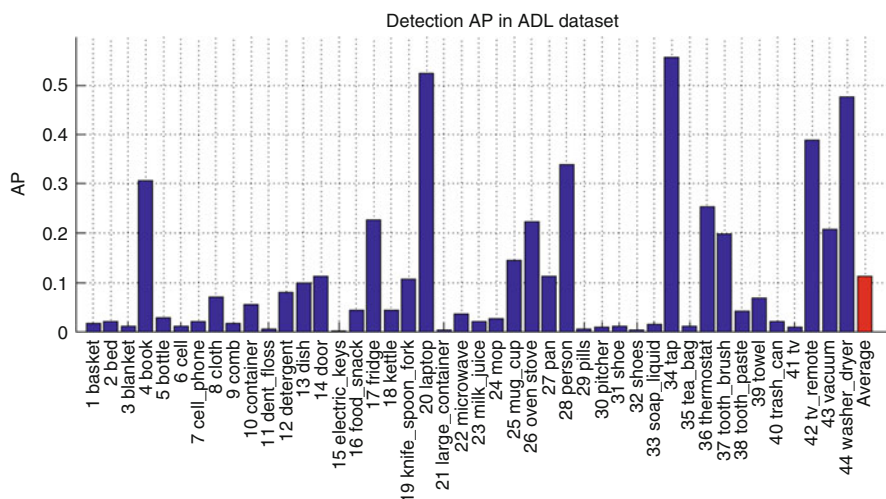


Fig. 5 Results in object detection

recognition. Furthermore, for activity recognition, the first six users have been taken to cross-validate the parameters of a linear SVM [8], whereas the remainder ones (7–20) have been used to train and test the models following a leave-1-out approach. The library libSVM [7] was used for the classification.

3.2 Object Recognition Results

Figure 5 shows the per-category and average results achieved by our active object detection approach in terms of Average Precision (AP). We have used this quality measure rather than accuracy due to the nature of the dataset, which is highly unbalanced for every category. The mean AP of our approach is 0.11 but, as can be

noticed from the figure, the performance notably differs from one class to another. Main errors in classification are due to various reasons: (a) a high degree of intra-class variation between instances of objects found at different homes, what leads to poor recognition rates (e.g. bed clothes or shoes show large variations in their appearance), (b) some objects are too small to be correctly detected (dent floss, pills, etc.), and (c) for some objects that theoretically show a lower degree of intra-class variation (TV, microwave), performance is lower than expected since it is very hard for a detector to distinguish when they can be considered as ‘active’ in the scene (e.g. a user just faces a ‘tv remote’ or a ‘laptop’ when using them, whereas the TV or the microwave are more likely to appear in the field of view even when they are not ‘active’ for the user).

3.3 Place Recognition Results

In this section, we report the results obtained on the ADL dataset for the place recognition module. We use a χ^2 kernel and retain 500 dimensions for the KPCA. We compared two different types of annotation of the environment: a room based annotation compound of five classes (bathroom, bedroom, kitchen, living room, outside) and a place based annotation compound of seven classes (in front of the bathroom sink, in front of the washing machine, in front of the kitchen sink, in front of the television, in front of the stove, in front of the fridge and outside).

We have obtained average accuracies of 58.6 and 68.4 %, for the room and place recognition, respectively. We will consider both features as contextual information for the recognition of activities.

3.4 IADL Recognition Results

In this section we show our results in IADL recognition in egocentric videos. As already mentioned, our system identifies the activity at every frame of the video using a sliding window. The performance is evaluated using the accuracy at frame level, which is defined as the number of correctly estimated frames divided by the total number of frames. For that end, we have also included a new class ‘no activity’ associated to frames that are not showing any activity of interest. It is also worth noting that the global performance is computed by averaging the particular accuracies for each class (rather than simply counting the number of correct decisions) and, thus, adapts better to highly unbalanced sets as the one being used (where most of the time there is no activity of interest).

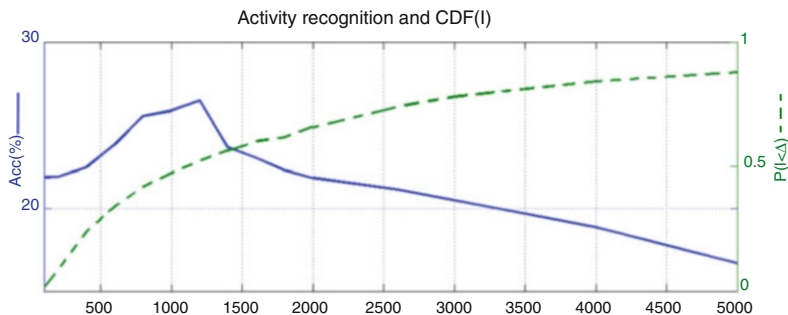


Fig. 6 Activity recognition accuracy with respect to the window size Δ (blue solid line) and cumulative distribution of activity lengths (green dotted line) (Color figure online)

3.4.1 Window Size

In our first experiment, we have studied the influence of the window size Δ defined in Sect. 2.3. Based on the results shown in Fig. 6 (blue line), we can draw interesting conclusions: on the one hand, too short windows do not model the dynamics of an activity, understood in our case as sequences of different active objects or places. Oppositely, too long windows may contain video segments showing various activities. Although, from our point of view, this fact might help to detect several strongly related activities by reinforcing the knowledge about one activity by the presence of the other (e.g. washing hands/face and drying hands/hair are activities that usually occur following the same temporal sequence), it might also lead to features containing too many active objects and places. These features would therefore make these frames difficult to assign to a particular activity. In our case, the value that best fits the activities in ADL dataset is $\Delta = 1,200$ frames, which corresponds to approximately 47 s of video footage. In fact, looking at the cumulative distribution of the activities length in the dataset (green line in Fig. 6), we have found this value is close to the median value which yields approximately 1,100 frames, thereby being consistent with the intuition that the window size should be chosen to be representative of typical activities length.

3.4.2 Recognition Performance

In the first column of Table 1, we show the results of our approach using either just active object or place detectors, and using an early combination of both of them by feature concatenation. As one can notice from the results, the active objects using sliency alone achieves slightly better performance than the approach of [23]. The place and room information alone yield lower performance, possibly being less informative to discriminate the activities. Combining objects and their context (the place where they are located) notably improves the performance achieved by simply

Table 1 Activity recognition accuracy for our approach computed at frame and segment level, respectively

Approach	Avg Fr. Acc (%)	Avg Seg. Acc (%)
Active objects (AO)	24.0	37.4
Places	18.5	6.1
Places + Rooms (early)	20.0	11.1
AO + Places (early)	27.3	38.5
AO + Places + Rooms (early)	26.3	36.5
AO + Places (late)	25.0	40.0
AO + Places + Rooms (late)	24.8	39.3
Pirsiavash et al. [23]	23.0	36.9

Bold values state the best performing approach in each scenario.

using the object detectors. Let us note that we have also tested several late fusion schemes (linear combinations, multiplicative, logarithmic, etc.) that did not lead to improvements in the system performance.

Furthermore, for comparison, we also include the results obtained with the software provided by the authors of [23]. This approach uses the outputs of various detectors of active and non-active objects implemented using the Deformable Part Models (DPM) [14]. Let us note that, as mentioned by the authors in the software, results differ from the ones reported in [23] due to changes in the dataset. From the results, and due to the similar classification pipeline of both methods, we can conclude that our features are more suitable for the activity recognition problem.

Finally, as made in [23], we additionally include results of a segment based evaluation in which ground truth time segmentations of the video are available in both training and testing steps. Hence, this case simplifies the activity recognition from a category segmentation problem to a simple classification problem for each segment. This case lacks the ‘no activity’ class, so that only video intervals showing activities of interest are taken into account. Combining objects and context provides the best performance, which is again superior to the one obtained by Pirsiavash and Ramanan [23].

In order to analyze these results in more details, Fig. 7 shows the Average Precision for each class separately. Performance is shown for several approaches, either using each mid-level feature alone, or using early or late fusion. It is clear from the results that several profiles appear for different kind of activities: some activities (Watching TV, Using the computer) are better recognized from object detection alone, while others (laundry, washing dishes, making coffee) are more linked to places. Their fusion tend to improve the mean performance, although the best way to do so depends on the activity category.

Overall, these results lead us to conclude that, recognizing activities in egocentric video does not require identifying every object in a scene, but simply detect the presence of ‘active’ objects and provide a compact representation of the object context. This context has been implemented in this work by means of a global classifier of the place. Future work could consider additional complementary features.

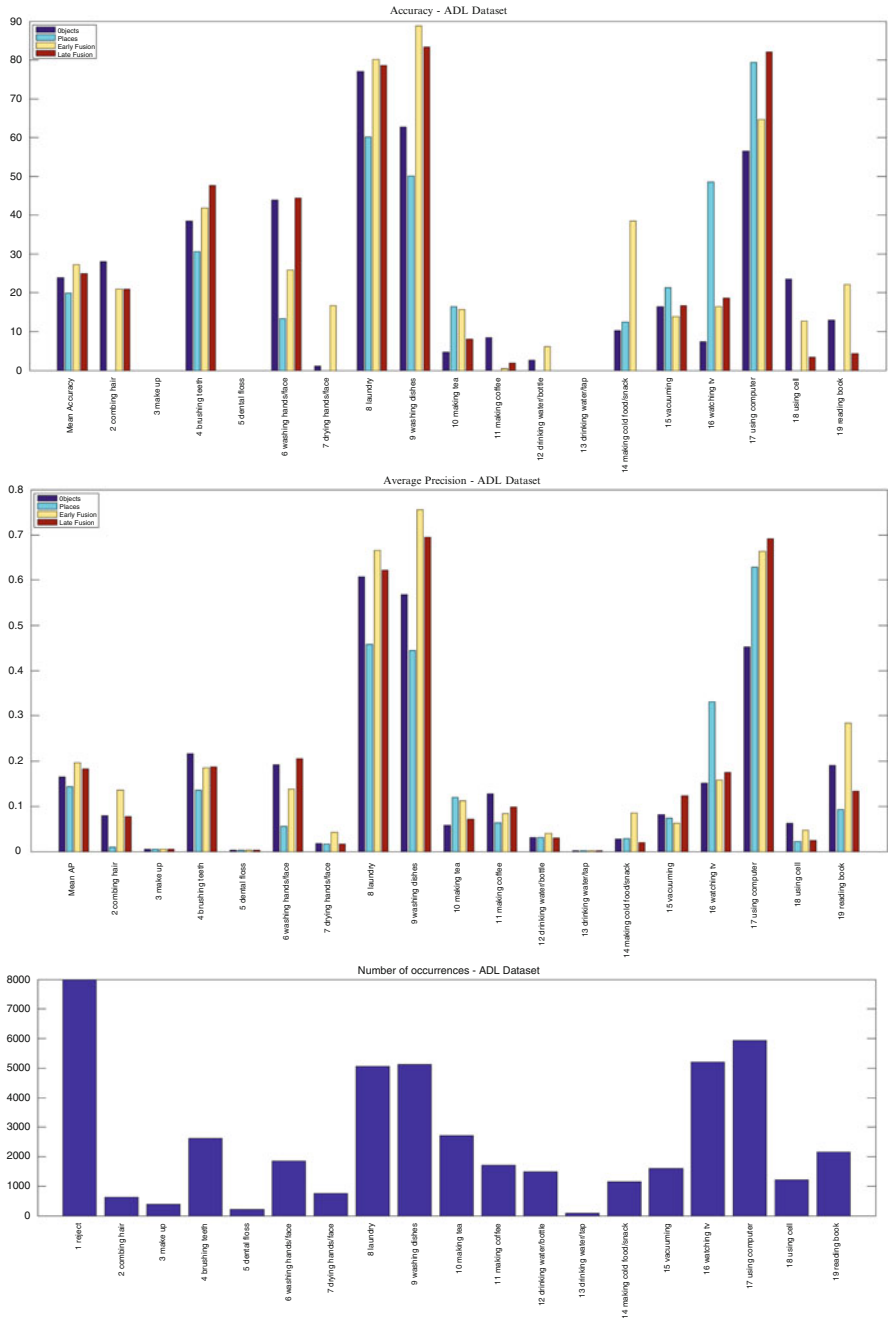


Fig. 7 (Top) Accuracy and (Middle) Average precision for activity recognition for various strategies: active object alone, places alone, early fusion, late fusion of active objects and places. (Bottom) Number of occurrences in the dataset

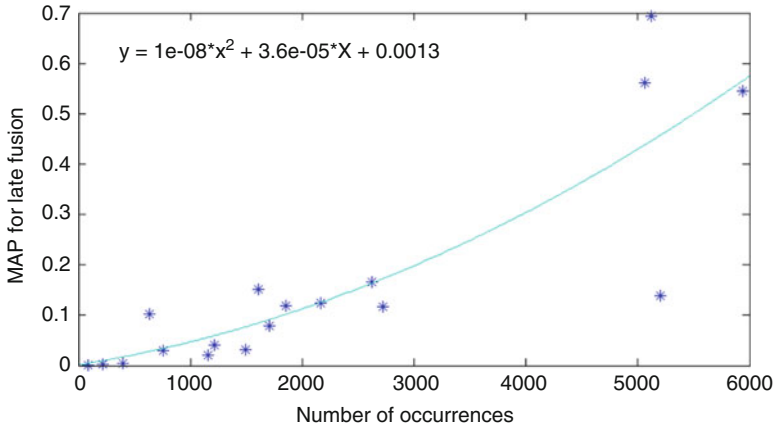


Fig. 8 MAP vs number of occurrences in dataset. Each star of the scatter plot represents one category. The best quadratic fitting is shown

3.4.3 Amount of Training Data

It is very interesting to note that the performance seems to be positively correlated with the amount of training data available, as illustrated in Fig. 8. There is indeed a sharp difference of average performance between the categories with a larger amount of training data and the others. Therefore, one main bottleneck of the recognition for this type of data remains the availability of sufficient training data, in order for the training to be representative of the test data. Although acquiring a large corpus of relevant data is an actual challenge when dealing with the monitoring of patients, these results suggest that ongoing and future efforts to obtain larger amount of training data in wearable camera setups is needed. Do they deal with control or patient subjects or not, they will likely contribute in notable improving the quality of the developed systems in terms of correct recognition of the activities.

3.5 Reliability of Confidence Values

Figure 9 shows the reliability plot of the main activity recognition approaches, based on Objects and Places features. The confidence values were computed as explained in the theoretical section.

The x -axis represents the predicted confidence in the $[0, 1]$ range. All frame based predicted confidence values are quantized into ten intervals over the $[0, 1]$ range. For each confidence interval, the value on the y -axis represents the empirical probability that the samples that have confidence within the interval are correctly classified. Therefore, an unperfect classifier with perfect calibration is called reliable and should ideally produce confidence values that match the empirical

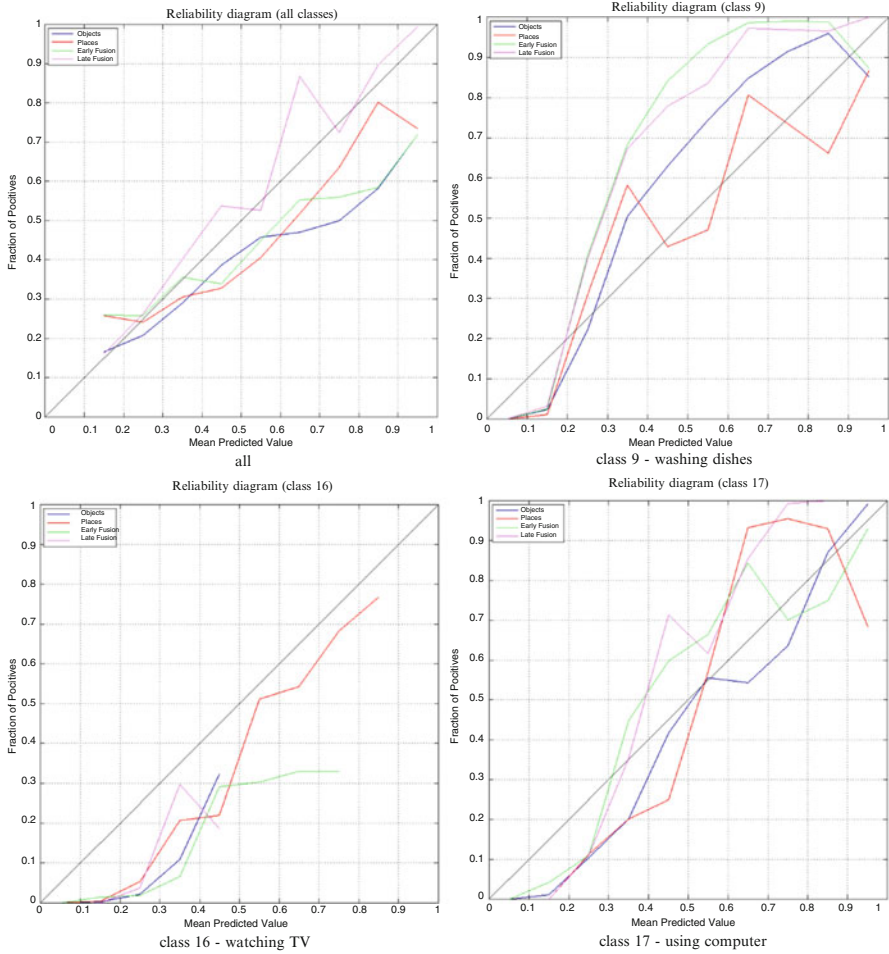


Fig. 9 Reliability diagrams of estimated confidence values

probability: points on the $(0, 0) - (1, 1)$ diagonal. Points over the diagonal show an underestimation of the quality of the classifier; point under the diagonal are overconfident.

Overall, for all classes, the reliability of all shown approaches follows approximately the diagonal. We have also shown the reliability plots of the categories with the three largest amounts of training data. Class 9, which has the best AP performance overall, has a reliability that is slightly under-confident: the estimates are actually better than predicted. Class 16 has a quite low AP performance, although a large amount of training data is available; this is reflected in the truncated curves, since no predictions on the test data were produced with high confidence.

Although the plot shows a slight over-confidence, it is interesting to see that this self-assessment of the algorithm avoids labeling samples with high probability in error.

These results show that on average, the confidence tend to be overestimated, as the effective accuracy for the samples belonging to each interval of confidence is lower than the predicted confidence.

4 Conclusion

In this chapter we have shown how activity recognition in egocentric video can be successfully addressed by the combination of two sources of information: (a) active objects either manipulated or observed by the user provide very strong cues about the action, and (b) context also contributes with complementary information to the active objects, by identifying the place in which the action is being made.

For that end, an activity recognition method that models activities as sequences of actives objects and places have been used on a challenging egocentric video dataset showing daily living scenarios for various users. We have demonstrated how the combination of both objects+context provides notable improvements in the performance, and outperforms state-of-the-art methods using active+passive objects representations.

The results also show that activity recognition in unconstrained scenarios is still a challenging task, that requires the fusion of complementary sources of information. Future research directions may consider the use of additional complementary features such as motion, hand positions, presence of faces for social activities, and continue the very important task of collecting significant amount of wearable video data in order to improve the representativity of training datasets for the target tasks. This is actually the case in the first prototype of Dem@care system which is under tests with volunteers patients with dementia.

Acknowledgements This research is supported by the EU FP7 PI Dem@Care project under grant agreement #288199. The authors would like to thank Olalla Rodríguez López for her valuable help.

References

1. Amieva, H., Goff, M. L., Millet, X., Orgogozo, J. M. M., Pérès, K., Barberger-Gateau, P., et al. (2008). Prodromal alzheimer's disease: Successive emergence of the clinical symptoms. *Annals of Neurology*, 64(5), 492–498.
2. Aziz, M. Z., & Mertsching, B. (2008). Fast and robust generation of feature maps for region-based visual attention. *IEEE Transactions on Image Processing*, 17(5), 633–644.
3. Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110, 346–359.

4. Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1), 185–207.
5. Boujut, H., Benois-Pineau, J., & Mégret, R. (2012). Fusion of multiple visual cues for visual saliency extraction from wearable camera settings with strong motion. In *European Conference on Computer Vision - Workshops, ECCV'12*, (pp. 436–445).
6. Buswell, G. T. (1935). *How people look at pictures*. Chicago, IL: The University of Chicago Press.
7. Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2, 27:1–27:27. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
8. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297.
9. Csurka, G., Dance, C. R., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV* (pp. 1–22).
10. Daly, S. J. (1998). Engineering observations from spatiovelocity and spatiotemporal visual models. In *IS&T/SPIE Conference on Human Vision and Electronic Imaging III*, (Vol. 1).
11. Dovgalecs, V., Mégret, R., & Berthoumieu, Y. (2013). Multiple feature fusion based on co-training approach and time regularization for place classification in wearable video. *Advances in Multimedia, 2013*, 22 pp. doi:10.1155/2013/175064. Article ID 175064.
12. Fathi, A., Farhadi, A., & Rehg, J. M. (2011). Understanding egocentric activities. In *International Conference on Computer Vision, 2011, ICCV '11* (pp. 407–414). Washington, DC, USA.
13. Fathi, A., Li, Y., & Rehg, J. M. (2012). Learning to recognize daily actions using gaze. In *Proceedings of the 12th European conference on Computer Vision - Volume Part I, ECCV'12* (pp. 314–327). Berlin/Heidelberg: Springer.
14. Felzenszwalb, P. F., Girshick, R. B., McAllester, D. A., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9), 1627–1645.
15. Gaidon, A., Marszalek, M., & Schmid, C. (2009). Mining visual actions from movies. In A. Cavallaro, S. Prince & D. Alexander (Eds.), *British machine vision conference* (pp. 125.1–125.11). Londres, United Kingdom: British Machine Vision Association BMVA Press. Page web de l'article : <http://lear.inrialpes.fr/pubs/2009/GMS09/>.
16. Gebel, M. (2009). *Multivariate calibration of classifier scores into probability space*. Saarbrücken, Germany: VDM Publishing.
17. Helmer, C., Peres, K., Letenneur, L., Gutierrez-Robledo, L. M., Ramarosan, H., Barberger-Gateau, P., et al. (2006). Measuring the objectness of image windows. *Geriatric Cognitive Disorders*, 1(22), 87–94.
18. Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11), 1254–1259.
19. Karaman, S., Benois-Pineau, J., Dovgalecs, V., Mégret, R., Piquier, J., André-Obrecht, R., et al. (2014). Hierarchical hidden markov model in detecting activities of daily living in wearable videos for studies of dementia. *Multimedia Tools and Applications*, 69(3), 743–771.
20. Karaman, S., Benois-Pineau, J., Mégret, R., Dovgalecs, V., Dartigues, J.-F., & Gaëstel, Y. (2010). Human daily activities indexing in videos from wearable cameras for monitoring of patients with dementia diseases. In *International Conference on Pattern Recognition (ICPR), 2010* (pp. 4113–4116).
21. Kitani, K. M., Okabe, T., Sato, Y., & Sugimoto, A. (2011). Fast unsupervised ego-action learning for first-person sports videos. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3241–3248).
22. Mégret, R., Dovgalecs, V., Wannous, H., Karaman, S., Benois-Pineau, J., Khoury, E. E., et al. (2010). The immed project: Wearable video monitoring of people with age dementia. In *Proceedings of the International Conference on Multimedia, MM '10* (pp. 1299–1302). New York: ACM.

23. Pirsiavash, H., & Ramanan, D. (2012). Detecting activities of daily living in first-person camera views. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.
24. Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers* (pp. 61–74). Cambridge: MIT Press.
25. Pronobis, A., Mozos, O. M., Caputo, B., & Jensfelt, P. (2010). Multi-modal semantic place classification. *The International Journal of Robotics Research (IJRR)*, 29(2–3), 298–320.
26. Schölkopf, B., Smola, A., & Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computing*, 10(5), 1299–1319.
27. Sreekanth, V., Vedaldi, A., Jawahar, C. V., & Zisserman, A. (2010). Generalized RBF feature maps for efficient detection. In *Proceedings of the British Machine Vision Conference (BMVC)*.
28. Sundaram, S., & Cuevas, W. W. M. (2009). High level activity recognition using low resolution wearable vision. In *Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2009* (pp. 25–32).
29. Wu, T.-F., Lin, C.-J., & Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5, 975–1005.