RECOGNITION OF ACTIVITIES OF DAILY LIVING IN NATURAL "AT HOME" SCENARIO FOR ASSESSMENT OF ALZHEIMER'S DISEASE PATIENTS

Vincent Buso[1], Louise Hopper[2], Jenny Benois-Pineau[1], Pierre-Marie Plans[1], Rémi Mégret[3]

Laboratoire Bordelais de Recherches en Informatique (LaBRI), Bordeaux, France[1]
{vbuso, benois-p, pplans}@labri.fr
Dublin City University (DCU), Dublin, Ireland[2]
louise.hopper@dcu.ie
Laboratoire de l'Intégration du Matériau au Système (IMS), Bordeaux, France[3]
remi.megret@ims-bordeaux.fr

ABSTRACT

In this paper we tackle the problem of Instrumental Activities of Daily Living (IADLs) recognition from wearable videos in a Home Clinical scenario. The aim of this research is to provide an accessible and yet detailed video-based navigation interface of patients with dementia/Alzheimer disease to doctors and caregivers. A joint work between a memory clinic and computer vision scientists enabled studying real-case life scenarios of a dyad couple consisting of a caregiver and patient with Alzheimer. As a result of this collaboration, a new @Home, real-life video dataset was recorded, from which a truly relevant taxonomy of activities was extracted. Following a state of the art Activity Recognition framework we further studied and assessed these IADLs in term of recognition performances with different calibration approaches.

Egocentric vision, Action Recognition, Context, Object Recognition, Place Recognition, Home Clinical Scenario

1. INTRODUCTION AND MOTIVATIONS

Wearable video sensors have gained popularity due to the miniaturization of video devices and their capacity to capture details of person's environment and its instrumental activities, which cannot be captured by ambient video cameras. This is why a strong research activity has been observed since recently in development of video understanding approaches for this specific content [1, 2, 3, 4, 5]. Recognition of Activities of Daily Living is one of the key problems from a computer vision perspective to be specifically adapted for the egocentric video analysis [6].

The authors of [5] were pioneers in developing the first approach for instrumental ADL recognition for dementia /Alzheimer disease diagnosis and evaluation.

Dementia is a progressive condition that can generally be regarded as consisting of three stages – early, middle and late [7]. These stages are qualitatively very different from each other in terms of managing activities of daily living for both the person with dementia (PwD) and his caregiver. The boundaries between these stages are not clearly defined and will vary between individuals. Research suggests however, that looking at performance of activities of daily living in conjunction with existing psychometric dementia-staging measures may improve clinical staging [13]. Intelligent assistive technologies are being developed to monitor and enable certain activities of daily living in people with dementia. Sensors are generally used to monitor a person with dementia's activity in their own home over a period of time [7], and interactive video monitoring has been found to improve their ability to carry out everyday tasks; for example, hand-washing [11], meal preparation [8] and taking medication correctly [14]).

Visual assessment of Person With Dementia (PWD) performances in IADLs is not a fully automatic process. Even indexing of video stream with a taxonomy of predefined ADL for a simple navigation in it, still remains an open research problem. It has been addressed in case of a lab setting [2] and more natural, but still not so much cluttered environment [3]. In this paper we propose an ADL recognition framework in a real-world situation of PwD in his own home for assessment of disease progression. Contrarily to [5], in which low level features were used for instrumental IADL recognition with hierarchical HMM classifiers, we follow a more "high level approach" of [3] and [4] and consider important elements in egocentric visual scenes, such as objects and locations. In this work we study in particular the calibration of classifier outputs. The rest of the paper is organized as follows. In section 2 we present the clinical scenario, the dataset and the taxonomy of activities defined by medical practitioners. In section 3 we overview the activity recognition approach. Experiments and results are presented in Section 4 and Section 5 concludes this work and gives its perspectives.

Activity	Locations	Objects	Videos	Total duration of activities, min
Prepare/eat meals	Kitchen	Bowl, spoon, table, food, Cooker, Bread	6	67.55
Make tea	Kitchen	Kettle, cup, tea bag, milk	5	31.24
Phone call	Various	Phone	5	61.34
Organize/ Take medication	Kitchen	Pills, Medication box, table	4	43.82
Cleaning	Kitchen	Table, Waste Bin	2	15.00
Play a CD	Sitting Room	CD Player, CD,	2	11.30
Water indoor plants	Kitchen, Hall	Jug, Sink, Plants	2	36.56
Feed birds	Outdoors, garden, shed	Bird feeder, Bird Food	3	20.47

Table 1 -. Annotated GoPro video taxonomies for @Home Lead Dyad

2. SCENARIO, DATASET AND TAXONOMY

The home-based pilot of the home based study is primarily concerned with assessing the deployment and use of sensor technology to maintain and enable independence of the person with dementia living at home. This is because the home is where the majority of the caring takes place, and it is where people with dementia report that they would rather age [16]. It also affords the opportunity to monitor the individual in their own environment with considerable frequency, to better monitor for change [15]. Although weekly controllable, the home is the most ecologically valid situation in which to evaluate the ability of assistive technology to maintain independence. Monitoring of IADLs with wearable cameras in this setting is intended to both provide direct support to the individual and also to ensure a feedback to the family caregiver and the clinician in order to facilitate personalized enabling support. Indeed making the patient with dementia fulfill IADLs required for detailed clinical assessment stimulates his cognitive load and prevents from digression due to depression.

A multiple case study design was used with particular emphasis on the description of intra-individual, inter-day variability as this is more clinically relevant than absolute values for the person with dementia [15]. Participant dyads comprising of an individual with mild to moderate stage dementia and their family caregiver were recruited through the Memory Works clinic at the University.

1.1. Data collection & taxonomy of IADLs

As previously mentioned, the data presented in this article focus on the monitoring of IADLs in the lead @Home case study. The data was collected on a participant dyad. This dyad comprises of a married couple who lives in their own home. The husband has a diagnosis of dementia and co-morbid epilepsy; his wife works four days a week and is his primary caregiver. At recruitment, he was aged 58 and just post-diagnosis. He was active and independent and both he and his wife have an open and exploratory attitude to technology, and are willing to try anything to see if it will help their circumstances.

A branched semi-structured assessment interview was carried out in order to understand more about the husband's functional status and the current needs of the dyad. Initial interview questions were used to invite participants to discuss their own functioning. Where problems or potential concerns were highlights, a more detailed assessment was completed using psychometric measures previously validated for use with people with dementia. The ADL measures consisted of the Bristol-ADL Scale [9] and the Everyday Competence Questionnaire.

No ADL issues were reported by this dyad in their assessment interview, so no further psychometric measures were completed. Following the interview, the caregiver raised concerns that her husband was having difficulty operating their CD player. He used to be a keen music listener and has a substantial music collection of CDs. She speculated that he had stopped listening to music, as he could no longer operate the CD player. Despite no clinical indicators of ADL difficulties, it was felt that this task could be supported with the use of a GoPro camera. In order to help the person with dementia complete this task, simple operating instructions were created and positioned beside the CD player. The task was then practiced during the researcher's weekly data collection visit and with the caregiver between visits. The introduction of this sensor prompted the dyad to request that other regular daily tasks

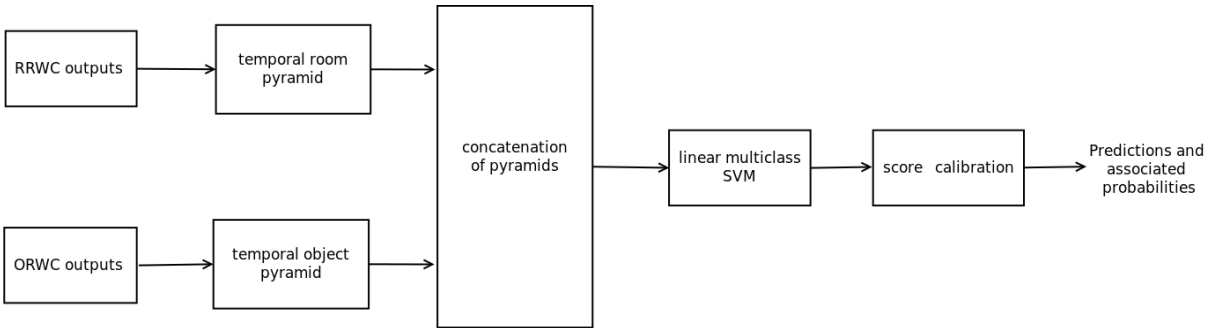

Figure 1 - Activity recognition by fusing location and objects.

would also be monitored. It was agreed that the person with dementia would wear the GoPro camera for one to two hours each morning, as this would capture making and eating breakfast, taking medication, and a variety of natural household chores. On review of the initial four weeks of data, a final taxonomy of eight monitored activities (see Table 1) was agreed upon.

The initial phase of the @Home pilot was, with the participants' consent, to provide training data for the location, activity, and object recognition models. During this 12 weeks period, 134 recordings were captured, which amounted to 33.3 hours of data. Representative samples of each activity were identified for annotation, and the development of associated taxonomies (see Table 1).

3. ACTIVITY RECOGNITION MODEL

In this section is presented the framework for activity recognition from wearable video content. It consists in a fusion between object and location detectors. It follows a hierarchical approach with two connected processing layers (see Figure 1). The first one contains a set of object detectors and place detectors referred in the rest of the paper as ORWC (object recognition with wearable camera) and RRWC (room recognition with wearable camera). The second layer uses the outputs of the first layer to perform the activity recognition task. The full pipeline is depicted in Figure 1.

Firstly, we recover for a frame t, a vector of probabilities $O_t = (o_t^1, \dots, o_t^K)$, consisting of K object detectors outputs (resp. $P_t = (p_t^1, \dots, p_t^J)$ consisting of J place detectors outputs) from both object and room recognition modules (outputs of ORWC and RRWC, see [6], [17]). These probabilities are obtained after an SVM classification using Platt approximation (see [17]). Once the outputs O_t, P_t of detectors have been recovered for each analysis frame t temporal pyramids are built from them and used as frame signatures. Here, for each analysis frame t, 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 sub-segments following a pyramid approach, so that at each level $l=0...L-1$ the pyramid contains 2^l sub-segments. Hence, the final feature of a pyramid with L levels is defined as:

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

Where $F_t^{l,m}$ represents the feature associated to the sub-segments m in the level l of the pyramid and is computed as:

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

Where Ω_{tm}^l represents the m-frames temporal neighborhood of the frame t at the level l of the pyramid and f_g is the feature computed at the frame g in the video. In this work, we have used a sliding window method with a fixed window of size $\Delta = 1200$ frames and a pyramid with $L = 2$ (according to experiments performed in [17]).

Hence equation (1) is applied to both O_t and P_t features. Finally Object and Place related pyramids are concatenated into a unique frame signature.

The temporal feature pyramid has then been used as input for a linear multiclass SVM in charge to predict the most likely action for each frame. The multiclass SVM was trained in 1-vs-1 fashion using libsvm software[20].

Finally, in order to map the SVM prediction scores to posterior probabilities, a calibration has been applied on each binary score. Three kinds of calibration have been implemented:

- The simplest one consists in normalizing the scores to [0, 1] using a sigmoid function.

$$g(s) = \frac{1}{1 + \exp(s)} \quad (3)$$

- The second one uses a direct calibration approach with the Platt calibration method [18].

$$g(s) = \frac{1}{1 + \exp(As + B)} \quad (4)$$

The coefficients A and B are estimated by fitting the sigmoid $g(s)$ to modified targets t_i:

$$t_i = \begin{cases} \dfrac{N_+ + 1}{N_+ + 2} & \text{if } y_i = +1 \\ \dfrac{1}{N_- + 2} & \text{if } y_i = 0 \end{cases} \quad (5)$$

where N_+ is the number of positive samples and N_- is the number of negative samples. This is done by minimizing:

$$-\sum_{i=1}^{N} t_i \log(g(s_i)) + (1 - t_i)\log(1 - g(s_i)) \quad (6)$$

A more general calibration function is given by monotonic functions. Their shape is not parameterized, as they only satisfy:

$$r < s \Rightarrow g(r) \leq g(s) \quad (7)$$

The underlying assumption is that two-class classifier ranks the sample correctly. Hence calibrating the scores consists of finding the monotonic mapping from score space to probability space. This can be done with isotonic regression and implemented by using the efficient Pairwise Adjacent Violators algorithm (PAV)[19].

Choosing a different calibration measure for each individual location detector has an effect on both the quality of the final confidence, but also on the relative ranking of the detectors when dealing with multi-class classification. Therefore, a decision based on first ranked class can be influenced by the calibration.

Hence, in the following section, we will assess our activity recognition module and will compare these three score calibration approaches. We will evaluate the reliability of the probabilistic scores and decide which calibration method is best suited in the specific context of activity recognition from wearable camera videos.

4. EXPERIMENTS AND RESULTS

Our model was assessed and different score calibration methods presented above were compared on the @Home dataset described in Section 2.

We have used 26 videos in this experiment. All recordings were performed using a GoPro camera at a frame rate of 30 frames per second with a resolution of 1280*960. Each video lasts fifteen minutes on average and contains around 27000 frames. We split the dataset by videos into 5 subsets of near-equal size (5-5-5-5-6) in order to perform a cross validation. Table 2 displays the overall accuracies obtained by our three types of calibration (Normalized, Platt and PAV).

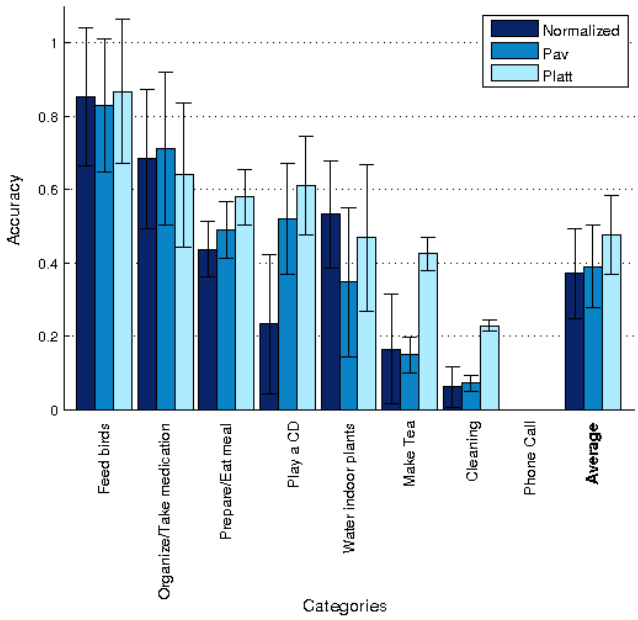

Figure 2 - Accuracies class per class

In terms of overall accuracy, the Platt approach performs better, followed by PAV, then Normalized.

To refine this analysis, accuracies for each class are displayed in and associated by the number of occurrences class per class. For the sake of comparison, chance gives an accuracy of 0.125.

For the accuracies class per class, none of the three methods (eq. (3), (4), (7)) is an absolute winner. Indeed, the best classification changes according to the classes. However it is worth noting that for the classes with a weak number of occurrences ("Make tea" or "Cleaning"), only the Platt approach seems to work.

Overall, performances are good for categories presenting small intra-class variability. Indeed the best performances correspond to activities performed either in characteristic locations ("Feed birds") or with a specific small set of manipulated objects ("Organize/take medication", "Play a CD", "Water Indoor plants"). Activities such as "Cleaning" or "Prepare/Eat meal" however present a much larger intra-class variability of manipulated objects or locations. Here the difference of performances is explained by a weak number of occurrences in annotated data (see "Cleaning" and the time of activities' recording in Table 1).

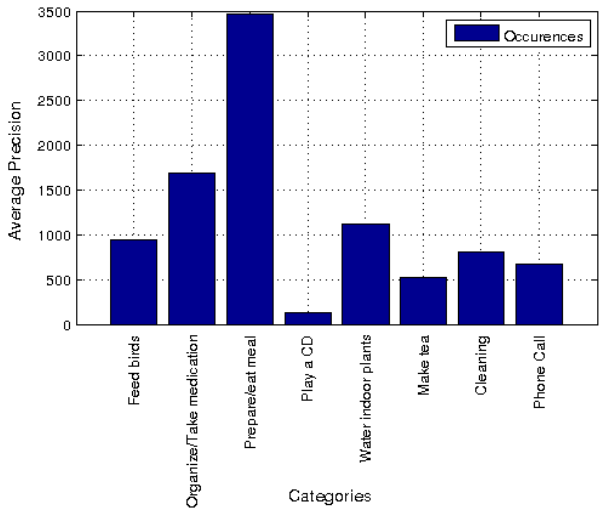

Figure 3 – Average precision per class

	Normalized	**PAV**	**Platt**
Mean accuracy (%)	37±12.4	39±11.2	47.7±10.8

Table 2 - Overall accuracies

We can also observe that the performance for the "Phone call" class is very low which makes sense since it is the most difficult activity to recognize. The drop in performance in this specific category exposes the limits of activity detection with wearable cameras. Indeed, the benefit given by the point of view of the wearable device no longer exists in the present scenario since the phone cannot be seen (close to the ear) and the room does not provide information since a phone call could happen anywhere in the house.

5. CONCLUSIONS AND PERSPECTIVES

Hence in this paper, we further study the problem of instrumental activity recognition in wearable videos in a real life @Home scenario for observations of patients with dementia. From the application point of view, this study helped stimulation of persons with dementia. From a methodological point of view we studied different calibrations of the SVM output as to give a probabilistic response. To our best knowledge our work still remains pioneering in what concerns application domain. As far as the complexity of the data set is concerned, the obtained scores are encouraging and can be improved incorporating more sensor modalities.

6. REFERENCES

[1] K. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3241–3248, 2011.

[2] Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In International Conference on Computer Vision, 2011, ICCV '11, pages 407–414, Washington, DC, USA, 2011.

[3] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2012.

[4] Lee, Y.,J., Graumann, K. Predicting Important Objects for Egocentric Video Summarization, Int. Journal of Computer Vision, January 2015, DOI 10.1007/s11263-014-0794-5

[5] S.Karaman, J.Benois-Pineau, V. Dovgalecs, R.Mégret, J.Pinquier, R.André-Obrecht, Y.Gaëstel and J.-F. Dartigues,"Hierarchical Hidden Markov Model in Detecting Activities of Daily Living in Wearable Videos for Studies of Dementia", Multimedia Tools and Applications,69(3), 2014, pp. 743-771

[6] González-Díaz,I, Benois-Pineau, J. Buso, V, Fusion of Multiple Visual Cues for Object Recognition in Videos. Fusion in Computer Vision 2014: 79-107, Springer

[7] Alzheimer Society of Ireland *The stages of Alzheimer's disease* [Online]. Available from: http://www.alzheimer.ie/about-dementia/stages-progression.aspx [Accessed 4/19/2012 2012].

[8] Bharucha, A.J., Anand, V., Forlizzi, J., Dew, M.A., Reynolds,Charles F., I.,II, Stevens, S. and Wactlar, H. 2009. Intelligent assistive technology applications to dementia care: Current capabilities, limitations, and future challenges. *The American Journal of Geriatric Psychiatry,* 17(2), pp.88-104.

[9] Bucks, R. S., Ashworth, D. L., Wilcock, G. K., & Siegfried, K. (1996). Assessment of activities of daily living in dementia: development of the Bristol Activities of Daily Living Scale. *Age and ageing*, *25*(2), 113-120.

[10] Kalisch, T., Richter, J., Lenz, M., Kattenstroth, J. C., Kolankowska, I., Tegenthoff, M., & Dinse, H. R. (2011). Questionnaire-based evaluation of everyday competence in older adults. *Clinical interventions in aging*, *6*, 37.

[11] Mihailidis, A., Boger, J.N., Craig, T. and Hoey, J. 2008. The COACH prompting system to assist older adults with dementia through handwashing: An efficacy study *BMC Geriatrics,* 8(1), pp.28.

[12] Rosenberg, L., Nygård, L., & Kottorp, A. (2009). Everyday technology use questionnaire: Psychometric

evaluation of a new assessment of competence in technology use. *OTJR: Occupation, Participation and Health*, *29*(2), 52-62.

[13] Shay, K.A., Duke, L.W., Conboy, T., Harrell, L.E., Callaway, R. and Folks, D.G. 1991. The clinical validity of the Mattis Dementia Rating Scale in staging Alzheimer's dementia. *Journal of Geriatric Psychiatry and Neurology,* 4(1), pp.18-25.

[14] Smith, G.E., Lunde, A.M., Hathaway, J.C. and Vickers, K.S. 2007. Telehealth home monitoring of solitary persons with mild dementia. *American Journal of Alzheimer's Disease & Other Dementias,* 22pp.20-26.

[15] Kaye, J. (2008). Home-based technologies: A new paradigm for conducting dementia prevention trials. Alzheimer's and Dementia, 4, 60-66.

[16] Frank, J.B. (2002). The paradox of aging in place in assisted living. London: Bergin & Garvey.

[17] Gonzalez Diaz, V. Buso, J. Benois-Pineau, G. Bourmaud, R. Megret, Modeling instrumental activities of daily living in egocentric vision as sequences of active objects and context for Alzheimer disease research, in: *Proceedings of the 1st ACM International Workshop on Multimedia Indexing and Information Retrieval for Healthcare*, MIIRH '13, ACM, 2013

[18] Platt, John (1999). "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods". *Advances in large margin classifiers* 10 (3): 61–74.

[19] Jan de Leeuw, Kurt Hornik, Patrick Mair (2009). Isotone Optimization in R: Pool-Adjacent-Violators Algorithm (PAVA) and Active Set Methods. *Journal of Statistical Software*, 32(5), 1-24. URL http://www.jstatsoft.org/v32/i05/.

[20] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm