Journal of Visual Communication and Image Representation 51 (2018) 169–190

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

journal homepage: www.elsevier.com/locate/jvci

Multi-modal activity recognition from egocentric vision, semantic enrichment and lifelogging applications for the care of dementia☆

Georgios Meditskos[a,*], Pierre-Marie Plans[b], Thanos G. Stavropoulos[a], Jenny Benois-Pineau[b], Vincent Buso[b], Ioannis Kompatsiaris[a]

[a] *Information Technologies Institute, Centre for Research and Technologies - Hellas, Greece*
[b] *LABRI, University of Bordeaux, France*

ARTICLE INFO

Keywords:
Instrumental activity recognition
Egocentric camera
Mechanical measurements
Visual cues
Ontologies
Semantic knowledge graphs

ABSTRACT

We describe a framework for lifelogging monitoring in the scope of dementia care, based on activity recognition from egocentric vision and semantic context-enrichment. As pure vision-based approaches appear to be already saturating in terms of recognition accuracy, we propose their enhancement with wearable bracelet accelerometer information. For that purpose, we design and study appropriate early and late fusion schemes to increase accuracy. The incorporation of mechanical variables, such as jerk, improves the recognition accuracy of activities that require fine motion. In addition, we describe a framework for semantic activity representation and interpretation, using Semantic Web technologies for building interoperable activity graphs. The system is personalized, as deployment-specific activity models are authored, while problems related to the disease are detected by rules. Complemented by lifelogging applications, the system is able to support interventions by clinicians, and endorse a feeling of safety and inclusion for end-users and their carers.

1. Introduction

In recent years, automatic human activity recognition has drawn much attention in the field of video analysis technology due to the growing demands from many applications, such as surveillance environments, entertainment environments and healthcare systems [1,2]. Among others, wearable cameras have already proved to be useful for lifestyle monitoring, object recognition from video summarisation [3] and fall detection [4]. In addition, medical studies conducted with individuals with Alzheimer's disease [5,6] showed that egocentric camera can give a more accurate view to clinicians on the difficulties patients face while performing Instrumental Activities of Daily Living (IADLs), such as cooking and signing a check. Research also shows that the close-up egocentric view allows for gathering more elements related to IADLs than ambient cameras alone. At the presence of such opportunities, traditional methods can be re-purposed, extended and combined to provide advanced monitoring qualities, such as social interactions [7].

Although it is traditionally a uni-modal task (e.g. image-based), recent developments show that video-based human activity recognition can benefit from complementary sources of information [8]. For instance, Internet of Things (IoT [9]) presents new opportunities, enabling the infusion of additional context in the activity recognition process. As such, activity recognition from multi-modal data has already been applied in various domains, such as in healthcare where numerous clinical studies exist that assess the cognitive state of individuals according to their performance on key IADLs [10]. In this direction, multi-modal activity recognition has the potential to greatly benefit clinical practice, automating the clinical assessment through carefully designed trials and recordings, enabling at the same time the subjective monitoring of people behaviour in order to drive clinical interventions and care [11].

As already mentioned, traditional applications use exclusively visual data for activity recognition and the healthcare domain is no exception. Meanwhile, the visual modality can be considered insufficient, especially in the egocentric scenario, i.e. when the camera is worn by a person shooting first person video [12]. At the same time, wearable devices are flourishing [13]: many wearable devices, in the form of bracelets, are flooding the retail market, providing the affordable and reliable means to measure movement and vitals for fitness and lifestyle applications.

Moreover, healthcare and assisted living applications require a broader understanding of human behaviour in order to assess and

☆ This paper has been recommended for acceptance by Mariella Dimiccoli.

[*] Corresponding author.

E-mail addresses: gmeditsk@iti.gr (G. Meditskos), pplans@labri.fr (P.-M. Plans), athstavr@iti.gr (T.G. Stavropoulos), jenny.benois@labri.fr (J. Benois-Pineau), vincent.buso@labri.fr (V. Buso), ikom@iti.gr (I. Kompatsiaris).

https://doi.org/10.1016/j.jvcir.2018.01.009
Received 15 March 2017; Received in revised form 28 September 2017; Accepted 14 January 2018
Available online 31 January 2018

promote well-being. This knowledge extends beyond activity detection events to idiosyncratic and habitual knowledge, such as the manner in which an activity is performed and its recurrent patterns (e.g. regarding a bed time routine). To accomplish this, applications need to model and couple activity recognition with models of clinical knowledge including deviations from typical behaviours, incoherences and problematic situations.

In this paper we propose a framework that employs multi-sensory data analysis, including egocentric camera and sensing from a wearable bracelet, in order to recognize IADLs useful for clinical purposes. In detail, the framework entails egocentric vision and wearable accelerometer data analysis, followed by Semantic Web technologies [14,15] to model activities, fuse information and classify activities and problematic situations. In detail, semantic enrichment entails a knowledge-driven layer that encapsulates clinical domain and user profile knowledge to monitor behavioural aspects. Semantic Web ontologies, reasoning and querying technologies are used for the declarative definition of formal, interoperable and semantically rich interpretation rules to highlight problematic situations.

End-user lifelogging applications complement the framework, providing the means for end-users, their relatives and clinicians to view meaningful information, feel included, assess progress and drive care further. The application for clinicians, the clinician interface, shows aggregated information of total activity duration and problems per day, week or month to enable pattern detection, trends and progress in time. It also shows in-depth events in a timeline to investigate causes and detailed insights. On the other hand end-users themselves and their relatives are shown, via the end-user interface, digested, simple graphs of selected activities and mild problem warnings, in order to keep them engaged while feeling confident, secure and encouraged. The clinical value of the presented framework is presented via pilot applications in real-world homes of elders.

The paper is organized as follows. Section 2 presents related work in the fields of activity recognition from egocentric videos and ontology-based behaviour modelling. Section 3 presents our multi-modal approach for IADLs recognition in egocentric perspective. Section 4 elaborates on the representation, enrichment and high-level interpretation of the detected activities, as well as for the recognition of problems through the execution of SPARQL queries over the activity graphs. Section 5 gives an overview of the lifelogging applications that have been developed to support clinical experts. Experimental results, as well as rule examples for problem detection and visualisation solutions to assist clinical staff, summarizing individuals' performance and highlighting abnormal situations are presented in Section 6. Finally, Section 7 concludes our work.

2. Related work

2.1. Egocentric video analytics and fusion

In videos recorded with egocentric video cameras [16], the recognition of IADLs is achieved using local features, probabilistic classification via Hidden Markov Model (HMM) [17] or Bayesian networks [18]. Nevertheless, more semantic cues, such as objects [19,20], have been also considered for recognition of (Instrumental) Activities of Daily Living ((I) ADLs). While in earlier work [19], the objects were identified with Radio-Frequency Identification (RFID) tags, advances in object recognition from visual data allowed the authors in [20] to use visual objects detectors. Recently, the use of more high-level models has been proposed that convey semantic views, such as presence of activity-related objects and localisation of the person in certain areas [8]. Both "views" were obtained using egocentric video data exclusively. In all these research works, performances of different approaches are measured in terms of accuracy metrics with regard to the annotated ground truth. Hence a specific annotation interface was developed for ground truthing of continuous egocentric video when Alzheimer's Disease (AD) patients were performing IADLs from a taxonomy prescribed by medical doctors. The activities were performed in an arbitrary order suitable for patients in ecological, at-home environment [17]. Obviously, small annotation errors can occur, specifically in delimitation of activities in time. This is why an auxiliary class "other" is often added to the activity classification process.

Still, due to the complexity and variability of human motion, the recognition of IADLs remains a challenging task. The method proposed in the original paper [16] for IADL recognition in a rather constrained environment (GTEA dataset[1]) gives only 26% of average accuracy, which has been slightly improved to 29% in [12]. The method proposed in [20] gives 23% of average accuracy on a more challenging dataset,[2] while [8] reaches a slightly higher figure of 26.9%. Also, in our previous work [17] the best combination of colour, motion and audio features from video allows to attain 30% of peak accuracy. Hence, we argue that uni-modal egocentric vision-based approaches attain saturation, hampering their applicability in real-world use cases.

At the same time, intensive research has been conducted for activity recognition from inertial sensor data [21,22]. The majority of well-recognised activities are the so-called "ambulation activities", such as walking, running, and lying [21]. Triaxial accelerometers are the most widely used sensors as reported in [21]. Frequency domain-features extracted from accelerometer signals prove to be efficient for this task [23], but for the instrumental activities, such as eating and cooking, using only inertial sensors remains insufficient. Furthermore, studies such as [23,21], were not focused on the fact that an instrumental activity is composed of fine hand movements. The research in analysis of human actions in biomechanics put forward the so-called "jerk" [24]. Jerk is mathematically defined as the rate of change of acceleration. The authors in [24] report on successful use of the 2D jerk for the description of fine human hand movements. The easy availability of 3D accelerometers allows for its use in 3D space.

Capitalising on the fact that individual modalities give no satisfactory recognition rates in real-world applications, we propose their fusion in an Instrumental Activity recognition task, investigating optimal fusion schemes of the available multi-modal information. As far as visual information is concerned, the problem of fusion of different features to improve recognition scores is being explored in various contexts [25,26]. In [27] we studied the optimal combination of HMM classifiers in early, mid-level and late fusion frameworks to improve the recognition scores of activities using low level visual features. The problem of fusion of dynamic features and visual cues in Instrumental Activity recognition tasks is even more challenging, since IADLs represent a complex combination of elementary moves. Hence, we hypothesize that the combination of two views, namely, fine dynamic and holistic visual, could increase recognition performance. In this paper we study optimal fusion strategies for recognition of IADLs in a complex description space comprising visual and dynamic information.

2.2. Semantic web and ontologies

The inherent ability of ontologies to formally represent knowledge with machine-understandable and explicitly defined semantics has proved particularly appealing in various application domains. Out of the numerous domains of interest, the recognition of human activities is a notable case where ontologies provide unique solutions for the contextualized modelling of profile and behavioural information at different levels of granularity and abstraction. Within this vision, ontologies and in particular the Web Ontology Language (OWL/OWL 2 [28]), aim to bring to the table the ability to formally capture intended semantics and to support automated reasoning, sharing, integration and management of knowledge. The core idea is to capture everyday

[1] http://ai.stanford.edu/alireza/GTEA/.

[2] http://people.csail.mit.edu/hpirsiav/codes/ADLdataset/adl.html.