UNIVERSITÀ POLITECNICA DELLE MARCHE
Repository ISTITUZIONALE

A review on video-based active and assisted living technologies for automated lifelogging

This is the peer reviewd version of the followng article:

Original
A review on video-based active and assisted living technologies for automated lifelogging / Climent-Pérez, Pau; Spinsante, Susanna; Mihailidis, Alex; Florez-Revuelta, Francisco. - In: EXPERT SYSTEMS WITH APPLICATIONS. - ISSN 0957-4174. - ELETTRONICO. - 139:(2020). [10.1016/j.eswa.2019.112847]

Availability:
This version is available at: 11566/269313 since: 2024-05-12T13:20:51Z

Publisher:

Published
DOI:10.1016/j.eswa.2019.112847

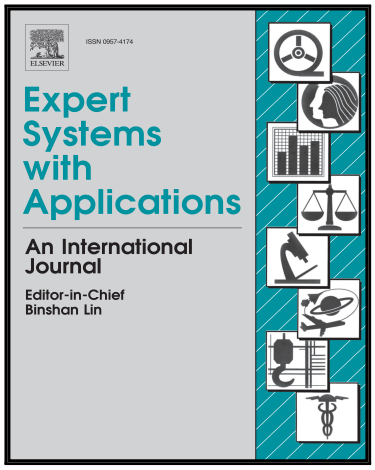
Terms of use:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. The use of copyrighted works requires the consent of the rights' holder (author or publisher). Works made available under a Creative Commons license or a Publisher's custom-made license can be used according to the terms and conditions contained therein. See editor's website for further information and terms and conditions.
This item was downloaded from IRIS Università Politecnica delle Marche (https://iris.univpm.it). When citing, please refer to the published version.

(Article begins on next page)

02 May 2026

Accepted Manuscript



A review on video-based active and assisted living technologies for automated lifelogging

Pau Climent-Pérez, Susanna Spinsante, Alex Michailidis, Francisco Florez-Revuelta

PII: S0957-4174(19)30549-4
DOI: https://doi.org/10.1016/j.eswa.2019.112847
Article Number: 112847
Reference: ESWA 112847

To appear in: *Expert Systems With Applications*

Received date: 4 December 2018
Revised date: 6 July 2019
Accepted date: 26 July 2019

Please cite this article as: Pau Climent-Pérez, Susanna Spinsante, Alex Michailidis, Francisco Florez-Revuelta, A review on video-based active and assisted living technologies for automated lifelogging, *Expert Systems With Applications* (2019), doi: https://doi.org/10.1016/j.eswa.2019.112847

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Highlights

- Explores a topic of social interest: developing technologies for ageing populations
- Emphasises the connection of active assisted living and life-logging, unseen to date
- Reviews literature from two standpoints: technologies used, and application fields
- Covers recent years not covered by others, with an emphasis on 2016-present
- Regards ethical implications of in-home devices, user-centred design and acceptance

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

A review on video-based active and assisted living technologies for automated lifelogging

Pau Climent-Pérez[a,*], Susanna Spinsante[b], Alex Michailidis[c], Francisco Florez-Revuelta[a]

[a]*Department of Computing Technology, University of Alicante, P.O. Box 99, E-03080 Alicante, Spain*
[b]*Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy*
[c]*Department of Occupational Science and Occupational Therapy, University of Toronto, Toronto, Canada*

Abstract

Providing support for ageing and frail populations to extend their personal autonomy is desirable for their well-being as it is for the society at large, since it can ease the economic and social challenges caused by ever-ageing developed societies. Ambient-assisted living (AAL) technologies and services might be a solution to address those challenges. Recent improved capabilities in both ambient and wearable technologies, especially those related with video and lifelogging data, and huge advances in the accuracy of intelligent systems for AAL are leading to more valuable and trustworthy services for older people and their caregivers. These advances have been particularly relevant in the last years due to the appearance of RGB-D devices and the development of deep learning systems. This article reviews these latest developments in the intersection of AAL, intelligent systems, lifelogging, and computer vision. This paper provides a study of previous reviews in these fields, and later analyses newer intelligent techniques employed with different video-based lifelogging technologies in order to offer lifelogging services for AAL. Additionally, privacy and ethical issues associated with these technologies are discussed. This review aims at facilitating

*Corresponding author
Email addresses: pcliment@dtic.ua.es (Pau Climent-Pérez), s.spinsante@staff.univpm.it (Susanna Spinsante), alex.mihailidis@utoronto.ca (Alex Michailidis), francisco.florez@ua.es (Francisco Florez-Revuelta)

ACCEPTED MANUSCRIPT

the understanding of the multiple fields involved.

Keywords: Lifelogging, Computer vision, Human activity recognition, Ambient-assisted living, Quantified self, Telecare, eHealth
2010 MSC: 00-01, 99-00

1. Introduction

The current situation in developed countries with the increase of ageing populations is unsustainable in the long run unless technological and other remedies are put in place. Since age is a factor for the decrease in personal autonomy and the increase in health and social issues, costs associated with these will grow, thus putting pressure on health systems and both professional and informal caregivers, with older people unable to receive assistance and having decreased chances of leading an independent life, and becoming a burden to families and the society at large due to lost working ours by caregivers (absenteeism) and increased expenditures on healthcare providers, as stated by Rashidi & Mihailidis (2013). The European Union recognised the importance of this by funding research directed towards ameliorating this situation and creating new technologies in the field of ambient –or active– assisted living (AAL), see Calvaresi et al. (2017).

AAL systems aim at improving the quality of life and supporting independent and healthy living of older or/and impaired people by using information and communication technologies at home, at the workplace and in public spaces. AAL environments are embedded with a variety of sensors, either located in the environment or worn by the user, that acquire data about the state of both the environment and the individual and/or allow person-environment interaction. These data are processed using expert and intelligent systems in order to provide advanced and personalised healthcare services.

Progress in wearable computing, with a myriad of products in the market (e.g. wearable cameras and smart watches, wristbands and glasses), increased functionality of mobile devices and apps for health and wellbeing, and easier

ACCEPTED MANUSCRIPT

installation of more affordable home automation systems are supporting the design, development, and adoption of healthcare and assisted living services by a larger population. For instance, lifelogging technologies may enable and motivate individuals to pervasively capture data about them, their environment, and the people with whom they interact. Acquisition and processing of physiological signals (e.g. heart rate, respiratory rate, body temperature, and skin conductance), motion, location, performed activities, images seen, and sounds heard, are the basis for the provision of a variety of cutting-edge services to increase peoples' health, wellbeing, and independence. Examples of these services include personalised healthcare, wellness monitoring (physical activity, dietary habits), support for people with memory impairments, social participation, mobility, support to formal and informal caregivers, predictive systems (decline in cognition, aggressive behaviours, fall prevention).

Recently, advances in intelligent systems and computer vision have led to the use of cameras in AAL systems, as they provide richer sensory information than the traditional sensors employed in those systems to monitor people, e.g., magnetic sensors, presence sensors and pressure mats (Nguyen et al., 2016). Video-based AAL systems usually employ conventional "third person" vision systems, where the cameras are located in the environment. An alternative is to mount a camera on the head or the torso of a person and record activities from an egocentric perspective, i.e. from the subject's own point of view.

According to Selke (2016, Ch. 1) *lifelogging* is understood as different types of digital self-tracking and recording of everyday life. The term is often used interchangeably with others such as *self-tracking* or *quantified self* (QS). Yet, normally, the latter is used to refer to the movement of people who monitor themselves or log their lives. More in depth, *lifelogging* means capturing human life in real time by recording physiological as well as behavioural (activity) data and store them for knowledge extraction at a later stage, which allows self-archiving, self-observation and self-reflection. Technologies used tend to be non-intrusive, such as miniature cameras and other sensors (wearable computing, smart watches) with real-time data transfer and ubiquitous access. Another

ACCEPTED MANUSCRIPT

feature of *lifelogging* is that it is a continuous process that requires no user interaction. Data collection is *always on*. In the context of AAL, sensors used for lifelogging can also be *ambient-installed* as opposed to wearable sensors, for instance, video surveillance or other cameras installed in nursing or smart homes to monitor and support older and frail people (Jalal et al., 2014). Furthermore, the data collection performed by users about their habits, shared with other stakeholders (caregivers, medical practitioners) is key to provide assistive means for improved, long-lasting independent living.

Most lifelogging technologies have ethical implications, and may have low user acceptance if the users are not involved in the process. Living labs have been proposed (Bygholm & Kanstrup, 2015; Queirós et al., 2015) as a means to reach a better understanding of user needs, as well as to lower prospective users' resistance that hinder the development and deployment of very much needed technologies for the ageing populations in developed countries. Most existing resistance has to do with ethical concerns of mass surveillance and lack of privacy (Bygholm & Kanstrup, 2015; Arning & Ziefle, 2015; Padilla-López et al., 2015).

This paper presents a literature review of the latest advances in the confluence of these three fields, namely computer vision (CV), AAL, and lifelogging. That is, it explores existing video-based technologies in the context of AAL with a focus on methods whose outputs can be assembled together in order to create a lifelog for the user, who can then share it, at their discretion, with the medical practitioners, social workers, and caregivers of their choice. We have carried out an exhaustive search in *Google Scholar* (GS) of the literature in these areas, analysing previous reviews, and identifying those more recent relevant works. Most of these reviewed works are within the period of 2015 to present, with a focus on 2017–present. Some works are outside of this temporal scope due to their relevance or if they are precursors of current methods. Figure 1 shows the distribution of reviewed papers according to the year they were published. It is worth noting that the GS tool provides both relevant (i.e. peer-reviewed) results from other sites such as *IEEE Explore*, *ScienceDirect* (SD), and *Web of Science*

ACCEPTED MANUSCRIPT

Table 1: Search keywords and inclusion criteria

Topics covered:[†]	(human) action/activity recognition, (human) behaviour understanding/analysis, gait analysis, fall detection, physiological signal monitoring
Keywords used:	`action recognition`[a], `activity recognition`[a], `behaviour understanding`[a], or `analysis`[a,b]; `gait analysis`[b], `fall detection`, or `prevention`; `physiological signal`[c]; `AAL ambient survey` `AAL ambient review`; `CNN`[d], `convolutional`[d], `deep learning`[d], `neural`[d]
Temporal scope:	2015–present (with focus on 2017–present) *exceptions:* precursors or otherwise relevant
Inclusion criteria:	peer-reviewed works (from IEEE, WoS, SD, etc.) *exceptions*: datasets, tools, challenges, or surveys

[†]: all video-based, i.e. using computer vision.

[a]: With and without '`human`', as some authors use variations.

[b]: With and without '`video`' and '`vision`' to find more video-based methods.

[c]: Always with '`computer vision`' or '`from video`' to get relevant results.

[d]: These terms used only in combination to previous ones to find more DL-based methods.

(WoS), among others; as well as non-reviewed or self-archived works. Table 1 provides a summarisation of inclusion criteria, as well as search keywords used, with the aim of search reproducibility.

The remainder of this paper is organised as follows: Section 2 presents and analysis of previous reviews that focus on all, or at least several, of the topics addressed in this paper. Section 3 reviews the different technologies and techniques that are employed in video-based lifelogging for AAL applications, which are presented in Section 4. Section 5 analyses some works dealing with privacy and ethical issues, which hinder user acceptance of these technologies and services. Finally, Section 6 summarises the main outcomes of this review.

ACCEPTED MANUSCRIPT

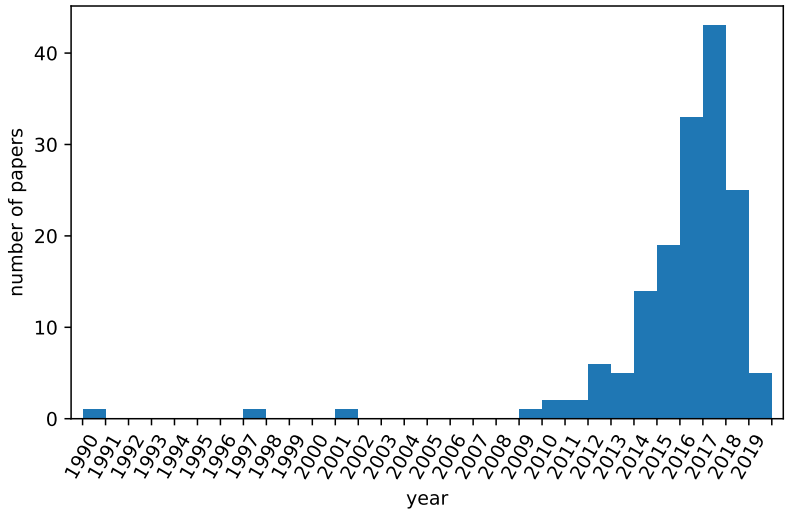


Figure 1: Distribution of all papers reviewed in the present work (all references) according to publication year. Please note year 2019 is ongoing at the time of writing.

2. Analysis of previous reviews

Previous reviews exist, as summarised in Table 2, but some are limited in scope in different ways. For instance, Chaaraoui et al. (2012) is a review on human behaviour analysis for AAL up to 2012, and Aggarwal & Xia (2014) from 2014 is a review of human action recognition from $3D$ data. They are included in this work for the sake of completeness and interest. The survey by Kong & Fu (2018) is much more recent, however it is also limited in scope to action recognition. Another recent survey, by Viana et al. (2019) is limited in scope to bibliometric analysis, that is, by evaluating merely the publication trends on the topic of AAL, by year, country, and other such non-technical dimensions. Conversely, Sathyanarayana et al. (2018) is a very complete, broader scope review, however it covers only works up to 2015. Yet, many advances have occurred since then, like new or renewed efforts in machine learning: sparse coding, deep learning, etc. as well as camera improvements and larger datasets, or the ability to use synthetic data while retaining good generalisation in real-world

ACCEPTED MANUSCRIPT

scenarios. Regarding other more recent reviews: most are vision-based, or have a strong focus on video-based methods additionally to other sensors. There are some exceptions, which are marked accordingly on the table. For instance, Díaz Rodríguez et al. (2014) classify works into either data-driven (inductive learning) or knowledge-based (i.e. using ontologies or other hierarchical structures), but covers the former only in a very broad manner, to then focus on the latter, providing a review of existing ontologies for human activity description. Another example is found in Erden et al. (2016), in which the authors consider ambient-assisted living mostly as fall detection, and thus constrain the problem of action or activity recognition to body pose detection, since most methods in their review consist of the same three classes (i.e. falling, standing, lying). This review aims to be broader in scope, and thus include methods for AAL that can be useful for the purposes of lifelogging.

Examples of broader field reviews in AAL also exist (Rashidi & Mihailidis, 2013; Planinc et al., 2016; Calvaresi et al., 2017; Leo et al., 2017; Prati et al., 2019). These focus more on the assistive technologies, and living tools that AAL can provide. For instance, in Rashidi & Mihailidis (2013), AAL tools for older adults are presented, the focus on *tools* means these are not necessarily methods at the *research* level, but also commercial solutions that can be found in the market. Furthermore, the authors identify the challenges brought forward by an ageing society, namely: increase in diseases, higher health costs, insufficient number of caregivers, more dependency, and larger impacts on society. This last item refers to the economic disruption caused by absenteeism and lost working hours of informal caregivers which are often relatives of the person needing support. Solutions are divided into either 'tools and technology', or 'applications and algorithms'. The tools presented include smart homes, wearables, as well as assistive robotics. On the algorithms, the authors focus on recognition of activities of daily living (ADLs), "one of most important components of AAL." It further divides the task of ADL recognition (or more broadly human activity recognition –HAR–), into methods using wearable sensors, ambient sensors and vision. Furthermore, this review includes a section on cognitive orthotics, i.e.

ACCEPTED MANUSCRIPT

Table 2: Previous and recent reviews

Year	Surveys	Topics covered
2012	Chaaraoui et al. (2012)	– Activity recognition (HAR/HBA) for AAL
2013	Rashidi & Mihailidis (2013)	– Living tools
2014	Aggarwal & Xia (2014)	– Activity recognition (HAR) from $3D$ data
	Díaz Rodríguez et al. (2014)[3]	– Ontologies for human activity description
2015	Betancourt et al. (2015)	– Evolution of first person vision methods
	Bygholm & Kanstrup (2015)[1]	– Lack of human-centeredness, acceptance
	Mukhopadhyay (2015)[2]	– Activity monitoring from wearable sensors
	Padilla-López et al. (2015)	– Privacy, user experience, acceptance
	Queirós et al. (2015)[1]	– Usability, accessibility, acceptance
2016	Erden et al. (2016)	– Fall detection (using PIR sensors, or images)
	Hamm et al. (2016)	– Fall prevention, detection, injury reduction
	Nguyen et al. (2016)	– Ego-vision ADL recognition (HAR)
	Planinc et al. (2016)	– Vision-based methods for AAL applications
2017	Calvaresi et al. (2017)	– Systematic review on AAL domain
	Han et al. (2017)	– Space-time skeletal $3D$ representations (for HAR)
	Herath et al. (2017)	– Activity recognition review, including some DL-based
	Khan & Hoey (2017)	– Fall detection (discussion on fall data availability)
	Leo et al. (2017)	– Vision for assistive technology
	Rajagopalan et al. (2017)	– Fall prediction and prevention
	Cippitelli et al. (2017)	– Fall detection from RGB-D and radar
	Wu et al. (2017)	– Activity recognition (using deep learning)
2018	Abdallah et al. (2018)	– Activity recognition with evolving data steams (DL)
	Antunes et al. (2018)	– Activity recognition (of healthcare professionals)
	Faust et al. (2018)[2]	– Physiological signal applications (DL-based)
	Kong & Fu (2018)	– Activity recognition (some DL-based) and prediction
	Sathyanarayana et al. (2018)[4]	– Fall detection, activity, sleep, vital signs, facial cues
	Thevenot et al. (2018)	– Medical diagnosis from faces
2019	Prati et al. (2019)	– Video surveillance (incl. health), wearable sensors
	Viana et al. (2019)	– AAL bibliometric review

[1] Non-technical, from social sciences, medical.

[2] Use other sensors (non-vision).

[3] Knowledge-based, ontologies.

[4] N.B. This review has been **available online since 2015**. Does not cover 2015–2018.

ACCEPTED MANUSCRIPT

tools aimed at helping with cognitive decline. In this section it links with lifelogging using camera-collected pictures, which are useful as a retrospective memory aid. Another such review is that of Planinc et al. (2016), which presents 'computer vision'-based (CV) methods for AAL applications. Depending on the technologies used, it divides video-based (RGB) methods into HAR or human behaviour analysis (HBA), fall detection, tele-rehabilitation, gait analysis (for fall prevention among others), and physiological signal monitoring. In the case of video and depth (using RGB-D devices), applications identified are: fall detection, rehabilitation, serious gaming (also coined as *exergaming* (Hamm et al., 2016; Vaziri et al., 2017)), pose analysis, gesture-based interfaces, and robotics. Another example can be found in Calvaresi et al. (2017), in it the authors criticise the lack of user need-centred reviews, as most are focused on technology. They also insist on the lack of 'need coverage' by solutions, that is, how proposed methods are able to cover, or cater for, a specific need. They attribute it to either lack of interest in need coverage (i.e. most papers are centred around one method), or insufficient need analysis when adapting an existing technology to an AAL scenario, or failing to explicitly analyse need coverage by using general evidence from related fields. Finally, authors raise the need for rigorous evaluation and validation of AAL solutions, and also the need to better understand relationship of users' needs and proposed solutions (i.e. 'need coverage' mentioned above). The most recent, Prati et al. (2019) performs a historical review of intelligent video surveillance (IVS), and continues with wearable sensor networks (WSNs) for activity recognition. Only the last section of this paper presents some recent advancements in the use of IVS for health and care. Namely, three applications are briefly discussed: AAL, patient monitoring, and physiological signal measurement.

In a broader sense, this 'need coverage' is related to user-centred design, which entails other aspects such as privacy and user acceptance (Bygholm & Kanstrup, 2015; Padilla-López et al., 2015; Queirós et al., 2015). In Bygholm & Kanstrup (2015), a broad analysis of the AAL field is presented from the perspective of technologies and applications, but also from the experiences of

ACCEPTED MANUSCRIPT

users, the successes and challenges. One criticism is that existing methods lack real-world applicability due to the complexity of humans and their behaviours, which might be overseen. The paper concludes that research methods comprising a close co-operation among researchers and users is key, and propose the use of living labs for trans-disciplinary work to be carried out among all stakeholders. Similar conclusions are reached in Queirós et al. (2015): usability and accessibility are heavily dependent on a good communication between designers and users, and therefore user-centred design in general, and living laboratories in particular are seen as a promising way to achieve this goals. The authors also point out that interoperability and compatibility among different tools is also important to improve and generate new solutions that provide better usability to final users. Finally, Padilla-López et al. (2015) analyse another aspect of concern for the acceptance of AAL technologies, that is, privacy. The authors present different privacy preservation methods, looking at privacy from different dimensions (enumerated as a list of questions about the data and its processing), methodologies (e.g. the most common being data redaction), and presenting different image filtering, encryption and de-identification, etc. They also discuss privacy at different stages of processing from a data security point of view. Finally, they classify existing methods according to the proposed dimensions.

As identified in broad-scope reviews above, in addition to HAR or ADL recognition, another important field in AAL is fall detection and fall prevention (e.g. via gait analysis) (Hamm et al., 2016; Khan & Hoey, 2017; Rajagopalan et al., 2017; Sathyanarayana et al., 2018). In Hamm et al. (2016), the authors divide interventions depending on whether the patients have already experienced a fall, and therefore have *pre-fall*, and *post-fall* interventions. From the technology point of view, it does not focus on video-based sensors, but discusses about the advantages of re-purposing ambient-installed cameras for fall prevention. Another survey on the field of fall detection is that of Khan & Hoey (2017). They analyse different fall detection techniques from the perspective of data availability, that is, they propose a taxonomy to classify the existing literature

ACCEPTED MANUSCRIPT

as either providing datasets where falls are sufficiently represented, or otherwise being rare or non-existent events in the training data. Methods vary for the three categories: well-represented fall data use multi-class classifiers and similar approaches, whereas unbalanced datasets require sampling and semi-supervised techniques; finally, datasets where falls are not present at all are used in systems that learn a *normal* walking pattern and detect falls as *abnormal* deviations from the common pattern. By contrast, Rajagopalan et al. (2017) propose a review that is more focused on challenges identified in the literature that concern the end-user, namely: performance in real-life conditions, acceptance (e.g. technological intrusiveness), security and privacy concerns, and energy optimisation of sensors (i.e. battery life). The literature reviewed in (Rajagopalan et al., 2017) includes both video- and 'wearable sensor'-based approaches. Finally, although the work in Sathyanarayana et al. (2018) is a general review of patient monitoring techniques using vision, it is worth mentioning here due to the section dedicated to fall detection, including methods from monocular as well as multiple-camera systems, datasets for fall detection and a dedicated discussion on the topic.

Two reviews focus on egocentric vision (Betancourt et al., 2015; Nguyen et al., 2016), which consists in the use of outward-looking cameras worn by the users to identify and track the performance of their ADLs, or analyse their exercise level (e.g active versus sedentary patterns), or walking performance (e.g. irregular gait might indicate deterioration of physical condition, and used for early prevention of falls). Nguyen et al. (2016), presents some of its advantages such as non-occluded view of the ongoing activity, since hand manipulation of objects can be paramount for ADL classification tasks which ambient-installed cameras cannot reach to see due to distance and body occlusion. They also present a review of ADL recognition (subset of HAR) and provide a classification of egocentric vision activity recognition methods as either object-based or motion-based (more on Sec. 3.2.2, wearable or first-person vision). On the other hand, Betancourt et al. (2015) presents a historical evolution of the field of first person vision methods. It explores different camera models, and how

ACCEPTED MANUSCRIPT

applications (i.e. computer vision tasks provided by these devices) have been also evolving, i.e. with more papers showing object and activity recognition in the years closer to this end of their temporal scope (up to 2014). It presents multiple *timelines* with the evolution of different aspects of egocentric vision (e.g. release of devices, key methods, main task of the method).

Although the focus of this review is in purely video-based techniques, several reviews exist that use cameras as an adjunct to, or combined with other sensors. A review showing the diversity of sensors that are available and methods to exploit the data provided by them can be found in Mukhopadhyay (2015), where the authors explore many different types of sensors that can be interesting as these allow to capture patients' temperature, heart rate, brain activity, muscular motion and other data. Sensors explored include: temperature, heart rate monitoring (via photoplethysmography or PPG, sound-based, or based on changes in face brightness (Wu et al., 2012)), accelerometers (mainly for HAR and fall detection), as well as some more *exotic* sensors such as textile patches for the skin that can detect internal activities in the body such as breathing and heart rate, but also hand gesture recognition, swallowing and gait analysis; or sodium ion detectors in the sweat that could reveal electrolyte imbalance or dehydration. In (Faust et al., 2018), the authors focus on four main types of physiological sensors, namely: electromyogram (EMG), electroencephalogram (EEG), electrocardiogram (ECG), and electrooculogram (EOG). A mixture of vision and non-vision sensors with a focus on fall detection using passive infrared (PIR) sensors (constrained vision equating or reducing AAL to only fall detection, though, as said) can be found in Erden et al. (2016). From the reviews that explore video-based methods, it can be seen that most of them explore human activity recognition or behaviour analysis (Aggarwal & Xia, 2014; Chaaraoui et al., 2012; Han et al., 2017; Herath et al., 2017; Kong & Fu, 2018; Wu et al., 2017; Abdallah et al., 2018). This can be justified by the fact that HAR is considered an essential part of AAL (Calvaresi et al., 2017; Rashidi & Mihailidis, 2013) and therefore receives more attention from researchers. Also, human activity recognition requires fine-grained data, as coarser methods for HAR or

ACCEPTED MANUSCRIPT

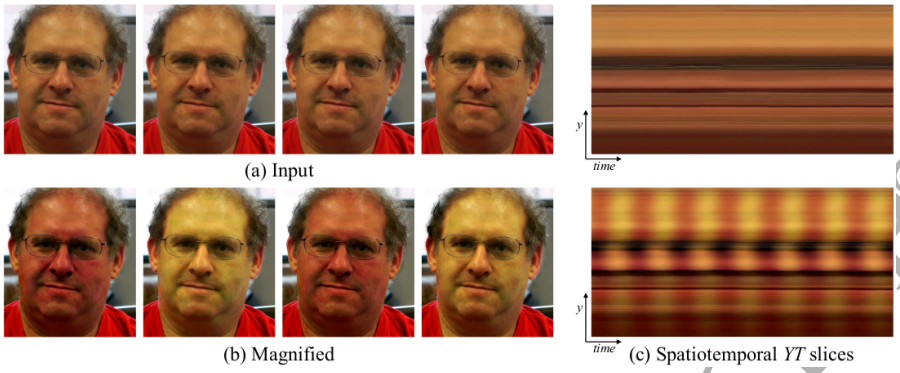


Figure 2: Example of Eulerian video magnification of subtle changes by Wu et al. (2012). The bottom row shows how the method makes heart rate visible to the naked eye (reproduced from (Wu et al., 2012)).

context awareness based on wearable or ambient sensors (contact, radio) can be limited (Antunes et al., 2018). Other fields of AAL such as health monitoring and diagnostics (tele-health) rely mostly on other sensors, as these are considered to be less error-prone and/or have undergone approval from regulatory medical agencies (Faust et al., 2018). The survey in Abdallah et al. (2018) seems to be the most recent one on the field of HAR, and the rapid evolution of this field is made evident by the fact that their survey focuses on evolving data streams, i.e. real-time video with non-delimited markers of activity start or end. However, in recent years, with video magnification of subtle changes by Wu et al. (2012) (see example in Figure 2), and superresolution methods McDuff (2018), as well as deep learning (as seen next), it has been possible to develop purely video-based methods for patient monitoring (Sathyanarayana et al., 2018), including: physiological signal monitoring, diagnostics (Thevenot et al., 2018). A notable commercial example is OxeHealth's OxeCam[1] (Oxford, United Kingdom) to monitor older people in care facilities including heart rate and breath monitoring.

The advent (or rather rebirth) of neural networks and deep learning (DL)

[1] `https://www.oxehealth.com/solution` (accessed: November 2018)

ACCEPTED MANUSCRIPT

have marked the start of a "new era" in many fields including computer vision. As a related field, AAL does not scape this trend either. The review of Herath et al. (2017) seems one of the first to add a review of 'deep learning'-based methods specific to the field of human action recognition (also (Calvaresi et al., 2017) mentioned above). However, the section dedicated to these methods is only one of the many in their review, which in its historical review, goes back to *classical* methods of the 90s and early 00s. Another example is Han et al. (2017), which also devotes a section to 'deep learning'-based representation learning (as a means to avoid manual feature crafting), however only a handful of methods are presented under this section, due to the novelty of the application of such techniques in the field of action recognition. A similar situation is observed in Kong & Fu (2018), which includes DL-based methods for activity recognition, within a review that also explores prediction methods. In that sense, Wu et al. (2017) seem to be the first to have a review that is fully dedicated to DL-based HAR methods. In contrast, there are modern reviews that do not cover works related to DL, such as (Calvaresi et al., 2017), this is due to the period covered at the time of writing (their review only covers years 2007–2013). To this point, most reviews covering DL-based methods are related to activity recognition. This might be due to the fact that DL methods have initially been applied in computer vision and natural language processing tasks, to only later *percolate* into other fields (Faust et al., 2018). Faust et al. (2018) suggest exactly this, and propose a review of DL-based methods for physiological signal applications. Nonetheless, their review does not cover vision-based methods.

Finally, some of the reviews explored are done from a systematic review perspective (Antunes et al., 2018; Bygholm & Kanstrup, 2015; Calvaresi et al., 2017). Following this methodology, one starts by setting some main *research questions*. For the present review, such questions would be the following:

- Which video-based AAL **technologies** can be used for lifelogging?
- How can these technologies translate into lifelogging **applications** for older and frail people?

ACCEPTED MANUSCRIPT

- Are there any **other aspects** of these technologies (such as ethical considerations, privacy issues, legal background, etc.) that are debatable? How can these be countered?

3. Technologies and techniques

This section explores different technologies and techniques that are available in the literature that can help provide the different applications that will be reviewed in Section 4. Machine learning (ML) techniques are at the core of most video-based solutions (with a few exceptions), especially for more complex scenarios such as human activity recognition (HAR). Another very important aspect of developed systems is the number, layout, and type of cameras used. These are two of the most important dimensions in which works can be classified from a technical perspective. Therefore the section is divided into two subsections: first, machine learning techniques commonly used in reviewed papers will be analysed; then, camera arrangements (single, multiple, etc.) and modalities (RGB, depth, etc.) will be explored.

3.1. Machine learning techniques

Within machine learning techniques, it is worth mentioning the trend towards more DL-based methods. This is especially true for activity recognition, and that is why this section will mostly include works using DL, but also others that use different trends in ML techniques. With regards to the former, Herath et al. (2017) classify DL-based methods into four categories: spatio-temporal networks, multiple stream networks, deep generative networks, and temporal coherency networks. For action recognition (a subset of classification tasks) the two first categories are more relevant; also most reviewed works can be classified into either of these two. Spatio-temporal networks include extensions to convolutional neural networks (CNNs, (LeCun et al., 1990; Krizhevsky et al., 2012)) that take into account temporal information: $3D$-CNNs in which convolutional blocks have been augmented to work with $3D$ blocks of XYT

ACCEPTED MANUSCRIPT

pixel colour information using $3D$ convolutions (Ji et al., 2013) (using stacked frames as input), usually with motion information as an additional input channel at the input layer, such as an optical flow (Herath et al., 2017; Rahmani & Mian, 2016). In the work of Tran et al. (2018) the authors explore the idea that full 3D convolutions may be more conveniently approximated by a 2D convolution followed by a 1D convolution, decomposing spatial and temporal modeling into two separate steps. They therefore propose an alternative to $3D$ convolutions named $R(2+1)D$ which they state have the advantages of being able to learn more complex functions (due to additional rectified linear units –ReLUs–), and also easier optimization during training. Temporal extensions to CNNs also include temporal pooling (Yue-Hei Ng et al., 2015). Spatio-temporal networks also include recurrent neural networks (RNNs) using long short-term memory (LSTM) blocks (Hochreiter & Schmidhuber, 1997), as well as hybrid CNN-LSTM networks. Multiple stream networks include those that train colour (RGB) and motion (e.g. optical flow) information in parallel 'subnetworks' that are connected at the decision-making fully-connected layers via their *softmax* scores (Simonyan & Zisserman, 2014), or earlier, which is shown beneficial (Feichtenhofer et al., 2016).

Table 3 shows the machine learning (ML) techniques most commonly used in the reviewed works. As can be observed, most recently published methods tend more towards the use of 'deep learning'-based methods. Among these, fully convolutional neural networks, or those with only 1–3 fully-connected (FC) layers on the top for classification are still very widely used (Ding et al., 2017; Elhayek et al., 2015; Fan et al., 2015; Liu et al., 2017a; Ma et al., 2017; Park et al., 2016; Solbach & Tsotsos, 2017; Toshev & Szegedy, 2014; Varol et al., 2017; Wang et al., 2016) (publication years ranging from 2014–present), which is also confirmed by recent reviews (Faust et al., 2018). However, as stated in (Herath et al., 2017), with $3D$ spatio-temporal extensions of CNNs it is difficult to determine which number of frames should be ideal during training. In this sense, hybrid spatio-temporal networks with CNNs connected to RNNs using LSTM blocks seem to be gaining momentum as more methods appear

Table 3: Machine learning techniques

ML technology	Subtype	Cites
Neural networks (deep learning)	CNN	Ding et al. (2017); Elhayek et al. (2015); Fan et al. (2015); Liu et al. (2017a); Ma et al. (2017); Park et al. (2016) Pham et al. (2018); Rahmani & Mian (2016); Rahmani & Bennamoun (2017); Solbach & Tsotsos (2017) Toshev & Szegedy (2014); Varol et al. (2017); Wang et al. (2016); Zhang et al. (2018)
	$3D$-CNN (and $2+1D$)	Carreira & Zisserman (2017); Liu et al. (2016b); Wang et al. (2017b); Tran et al. (2018)
	Multi-stream CNN	Simonyan & Zisserman (2014); Feichtenhofer et al. (2016); Tu et al. (2018); Ma et al. (2018)
	Dynamic images, Temporal representations	Bilen et al. (2016); Xiao et al. (2019); Choutas et al. (2018); Pham et al. (2018)
	RNN, LSTM	Abebe & Cavallaro (2017b,c); Liu et al. (2016a); Nakamura et al. (2017) Núñez et al. (2018); Shahroudy et al. (2016a); Wang et al. (2017b); Zhu et al. (2016)
	CNN+MRF	Tompson et al. (2014)
Other (non-DL)	Feature engineering	Abebe & Cavallaro (2017a)
	Spatial Laplacian pyramids	Ji et al. (2017)
	Sparse representation (ℓ_1/ℓ_2 minimisation)	Shahroudy et al. (2016b); Chen et al. (2016); Rahmani & Mian (2016); Theodorakopoulos et al. (2014)
	SVM	Abebe & Cavallaro (2017a); Ji et al. (2017); Kasturi & Jo (2017); Liu et al. (2016b); Xiao et al. (2019)
	Spectral graph matching	Ardeshir & Borji (2016, 2018)
	Attribute learning	Zhang et al. (2018)

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

that rely on this or similar approaches (Abebe & Cavallaro, 2017b,c; Liu et al., 2016a; Núñez et al., 2018; Shahroudy et al., 2016a; Wang et al., 2017b; Zhu et al., 2016) (publication years ranging 2016–present), although the number of publications is still lower than CNN-based methods. Another trend is the use of *dynamic images*, which are not included in Herath et al. (2017) and are a type of spatio-temporal template from a video, similar to motion history and energy images (MHI/MEI), but using rank pooling as part of a convolutional network (Bilen et al., 2016). The resulting $2D$ action summaries can be observed in Figure 3. Similarly, Pham et al. (2018) propose an action summary image based on temporal $3D$ skeleton information retrieved from RGB-D sensors. If DL-based techniques abound, the opposite can be said of techniques using more *classical* or non-neural approaches. Another recent trend has been to learn *sparse representations* (Wright et al., 2010), in which vocabulary of distinctive object parts is automatically constructed from a set of sample images of object classes. New images are then represented using parts from this vocabulary, together with spatial relations observed among the parts. All this while minimising the number of parts used to describe each new pattern (i.e. which makes the feature vector sparse, hence the name). Yet, not many works were found when performing literature database searches, e.g. (Shahroudy et al., 2016b; Chen et al., 2016; Theodorakopoulos et al., 2014). In Rahmani & Mian (2016) a hybrid approach is presented: the method uses CNNs for feature extraction, and then sparse representations to learn discriminative neuron-sets for each action.

Another interesting trend in machine learning, and specifically in DL-based techniques is that of using synthetic data (Bochinski et al., 2016; Rahmani & Mian, 2016; Varol et al., 2017). Since neural networks generally require larger than usual datasets, or rather, that such dataset provide a much larger benefit in terms of accuracy as the model will generalise better, simulated but *realistic* data is provided to the model during the training stage. Simulated data enables the generation of a large collection of pose variations, as a similar approach to data augmentation, but with almost-infinite possibilities. Also it allows to

ACCEPTED MANUSCRIPT

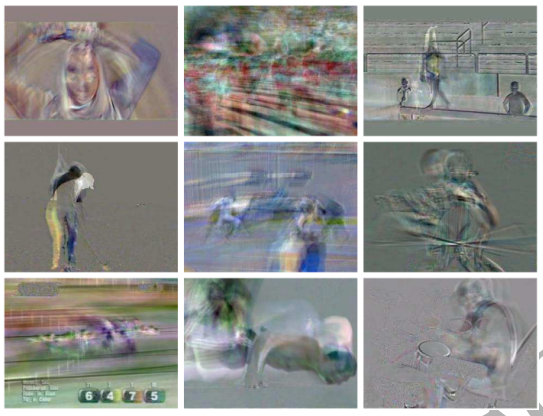


Figure 3: Summarisation of actions by means of dynamic images (reproduced from (Bilen et al., 2016)). From left to right and top to bottom: "blowing hair dry", "band marching", "balancing on beam", "golf swing", "fencing", "playing the cello", "horse racing", "doing push-ups", and "drumming".

collect data samples that might be very unusual (e.g. in fall detection systems positives tend to be a minority case (Khan & Hoey, 2017), for instance 30 instances in 17 months of data (Vlaeyen et al., 2013)), and thus datasets can be more balanced with regards to positive and negative samples. It is also useful to ease the burden (economic, temporal) of large data collections, that need to consider many scenarios and be unbiased. Also, because data has been generated, ground truth is automatically available, therefore it also eases the burden of ground truth labelling. Quality is therefore paramount, as non-realistic or *untransferable* (i.e. unfit for transfer learning) samples will lead to failure of the learnt model when dealing with real-world input during deployment (Baldewijns et al., 2016; Martinez-Gonzalez et al., 2018). For instance, Rahmani & Mian (2016) use $3D$ human models to generate simulated depth data of actors performing different actions. Varol et al. (2017) present SURREAL, a synthetic dataset of human poses for action recognition: it contains realistic images of people, along with synthetic depth and body part segmentation. They prove that a CNN trained on their large-scale dataset is able to provide accurate depth estimation and human part segmentation in real RGB images. Unrelated

ACCEPTED MANUSCRIPT

to AAL, but also worth mention, are Bochinski et al. (2016) who propose to use a realistic videogame engine for the generation of a dataset of humans, vehicles and animals, which is then used on a real-world data classifier installed on a surveillance camera.

A current trend, and one that might be still worth exploring further is that of two-stream or multi-stream networks. Recent works using this type of networks are popular (Ma et al., 2018; Tu et al., 2018; Choutas et al., 2018). Specifically, the current trend is to apply each stream to focus on a small region, body part, or joint, as is done by Ma et al. (2018), where six streams are used to follow relevant body parts. Choutas et al. (2018) propose a pose-time representation based on a temporal joints representation, which is fed to a n-stream CNN. Similarly, Tu et al. (2018) propose a multi-stream network to follow salient (moving) human body regions, as focusing on those yields better results.

To conclude, the results regarding the preference of one type of architectures (i.e. *classical* CNNs, n-stream networks, or residual networks), over the other (i.e. recurrent variants: RNNs, with LSTMs or similar) is not clear. Indeed, to answer this question Ma et al. (2019) perform a series of tests to compare these two families of neural networks. Admitting that multi-stream networks have contributed to a significant progress in human action recognition in recent years, they propose a strong baseline two-stream CNN using a residual network (ResNet-101). Given their results, the authors then propose two different network architectures to further integrate spatio-temporal information: either an extension to RNNs using temporal segments, or an Inception-style temporal convolutional network. Their results show that either solution improves the overall performance, and achieves state-of-the-art results on the standard benchmark datasets used. Finally, it is curious to note that the initial criticism of any feature engineering in the deep learning arena has transitioned slowly to more human-aided deep learning networks, where joints, 'body parts', or other human body information is explicitly provided to a network to facilitate the learning task, or to avoid very deep CNN networks that have trouble working on budget hardware, or RNNs that have trouble with overfitting. It might be

ACCEPTED MANUSCRIPT

worth exploring works from the recent past, just before the last deep learning wave (circa 2012), in order to check which engineered features for body motion description were most successful back then, and be able to *replicate* them using current convolutional neural networks, and then feeding this concurrently with RGB (or RGB-D) information in a multi-stream fashion, as this kind of architectures allows easier integration of human body motion features.

3.2. Camera typology and perspective

As said, the most common camera setup for lifelogging as a memory aid is via a outward-looking camera worn around the neck or as a brooch-like device. This camera perspective is also coined as egocentric vision or *egovision* for short (Nguyen et al., 2016), or *proprioceptive*, as it perceives the wearer's own movements (Abebe & Cavallaro, 2017a). However, as stated in the introduction, cameras can also be installed in the environment as this setup can be less obtrusive. These two camera setups tend to be used more for applications (see Section 4) like human action recognition, or fall prevention and detection: the first (egovision) can detect the wearer's motion patterns with respect to the environment, as well as activities involving the hands and handled objects; whereas the second can recognise motion patterns involving the full body.

In some medical applications of computer vision or methods relying on face analysis, camera setup might need to follow a specific or bespoke setup (so that the sensor is closer to the analysed body part), as in (Huimin et al., 2017; Lewis et al., 2018; Li et al., 2017; Maclaren et al., 2015), or be disguised in an everyday item such as a mirror (Andreu et al., 2016; Colantonio et al., 2015a; Henriquez et al., 2017).

3.2.1. From cameras installed in the environment

Using RGB-only devices. Moved by the scarcity of videos, and the small size of datasets for action recognition, Carreira & Zisserman (2017) propose a new dataset, namely the Kinectics Human Action Video (KHAV or simply *Kinectics*) dataset. They also propose the use of $3D$-CNNs by *inflating* $2D$ filters

ACCEPTED MANUSCRIPT

to $3D$, thus allowing for spatio-temporal feature extraction from video. They demonstrate how current architectures in the state of the art perform when pre-trained on *Kinectics*, and then tested on the much smaller existing action datasets (Hollywood movies –HMDB-51– and University of Central Florida – UCF-101). Aware of the difficulties of creating a dataset as big as *Kinectics*, Ma et al. (2017) propose to crawl the net for action videos, that contain any of the 101 classes in the UCF-101 dataset. They also, as opposed to (Carreira & Zisserman, 2017), use $2D$ convolutions, rather than $3D$ extensions, as they claim spatial networks can perform as well as spatio-temporal, and this enables the usage of single action images for training and starting the process with pre-learnt low-level filters using pre-trained networks (with the ILSVRC[2] subset of the ImageNet dataset). Another option to take temporal information into account for training is to include motion or optical flow data as part of the input (Park et al., 2016; Wang et al., 2016). Finally, Wang et al. (2017b) propose to use both $3D$-CNNs for spatio-temporal feature extraction from adjacent frames, and LSTMs and temporal pooling to explore temporal scales at which different activity instances can be detected.

Another option for analysis of poses for action recognition is using contours or silhouettes. A previous review (Aggarwal & Xia, 2014) introduces works that use contours or silhouettes that can be either retrieved from RGB images using segmentation (which tends to be complex, using background subtraction or similar techniques), or directly from the depth channel of RGB-D devices, in which the segmentation is much easier to perform. Chaaraoui et al. (2013) present a work that uses a bag-of-words (BoW) modelling of features extracted from contours to generate a dictionary of key poses which are then used to learn different actions according to the distributions of learnt words (poses) in video input of performed actions.

[2]Large scale visual recognition challenge: `http://image-net.org/challenges/LSVRC/` (accessed: November 2018)

ACCEPTED MANUSCRIPT

RGB Datasets:. When it comes to RGB-only datasets specific to AAL, it is worth noting than most activity recognition and fall detection datasets in recent years are multi-modal (i.e. with heterogeneous sensor types), and many are recorded from RGB-D sensors rather than classic RGB video cameras. In the field of photoplethysmography (PPG) for monitoring of physiological signals, in spite of the lack of datasets noted in the past by McDuff et al. (2015), two datasets using RGB video stand out as noted by Tulyakov et al. (2016): those are the MAHNOB-HCI[3] by Soleymani et al. (2012), and MMSE-HR, a subset of the MMSE dataset[4] by Zhang et al. (2016b) with annotations for heart rate estimation.

Using depth-based sensors. Using depth data for activity recognition preserves privacy, as people in the images are not recognisable (Padilla-López et al., 2015). Furthermore, depth information is insensitive to changes in lighting, and provides geometric information of the body and handled objects (Liu et al., 2016b; Rahmani & Bennamoun, 2017). An example of recent work in this regard is Rahmani & Mian (2016), in which the authors propose a CNN framework to extract view-invariant features, which are then temporally combined using Fourier Temporal Pyramids (FTPs), and discovering discriminative neuron-sets by solving an ℓ_1/ℓ_2-norm regularised least squares problem, which achieves sparse, discriminative sets per action class. Liu et al. (2016b) learn spatio-temporal features from depth sequences using $3D$-CNNs, decision is made via an SVM which is fed the pre-learnt features as well as skeleton joint information. Ji et al. (2017) also use an SVM classifier as part of their method, but argue that DL-based methods, and data-driven learning in general is bound to require to much computational power and data. Therefore, their features are extracted using a spatial Laplacian and temporal energy pyramid representation. They claim to perform at a similar accuracy level as Shahroudy et al. (2016a), which uses

[3] `https://mahnob-db.eu/hci-tagging/` (accessed: November 2018)

[4] Also referred to as BP4D+: `http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html` (accessed: November 2018)

ACCEPTED MANUSCRIPT

a recurrent network with LSTMs for feature extraction and temporal integration, and at a fraction of the time required per frame. This demonstrates that, although DL-based methods tend to perform better in general, an appropriate mixture of well-picked features can also achieve good accuracies in complex problems. Another example of this is Shahroudy et al. (2016b) in which features are learnt by sparse minimisation of a set of features obtained from depth and skeleton data. Rahmani & Bennamoun (2017) also fuse depth and skeleton data in their method: joint data is pre-processed to obtain a view- and scale-invariant normalised skeleton. Rectangles of interest around the joints in the depth space are also cropped and processed via a CNN to obtain view-invariant joint-context information, which is useful to detect actions involving handled objects. Another idea is to create spatio-temporal templates of action videos either manually (Ijjina & Chalavadi, 2017) or automatically via convolutional networks using rank pooling on the raw images of an action video, which results in *dynamic images* (Xiao et al., 2019). Ijjina & Chalavadi (2017) use *classical* temporal templates (motion history and motion energy images) extracted from the RGB and depth channels independently and then feed these to a CNN for further feature extraction. Since the images (templates) convey spatio-temporal information, the CNN extracts spatio-temporal features that are useful for the classification task. Similarly, Xiao et al. (2019) propose to extend the concept of dynamic images to depth data, by feeding a CNN with the RGB as well as the depth dynamic images. Furthermore, they obtain the depth dynamic images from several simulated viewpoints (by rotating the point cloud accordingly), and finally classify the actions using an SVM classifier. Finally, Zhang et al. (2018) propose to use depth and joint positions in a multi-stream deep convolutional network. Figure 4 shows a diagram of their proposed method. Three CNNs are trained: one with skeleton data ($1D$); another with temporal templates ($2D$); and the last one, a $3D$-CNN with spatio-temporal depth volumes. The activations from the second-last layer from each CNN are then used in an attribute learning framework which uses predefined motion patterns which are discriminative of the different action classes to recognise.

ACCEPTED MANUSCRIPT

RGB-D Datasets:. As mentioned, most recent activity recognition and fall detection datasets are multi-modal, that is, captured from networks equipped with different sensor types: RGB-D cameras, but also wearable, and binary devices (i.e. contact sensors for cabinet and house doors, electric switches, etc.). Two examples of large datasets captured in recent years amounting days and even months of data are: a) those compiled by Twomey et al. (2016) for the SPHERE project as part of their challenge[5] which consisted of RGB-D, accelerometer and PIR sensor data to detect (classify) 20 different motions, postures, or posture transitions; b) the NTU dataset for activity recognition introduced by Shahroudy et al. (2016a), which includes 56 thousand video samples from 40 different subjects performing a total of 60 labelled action classes. Other interesting datasets (including fall detection) can be found in the specific reviews by Zhang et al. (2016a) and Cai et al. (2017). Physiological signal monitoring from RGB-D sensors and heart rate monitors for ground truth labelling also exist, an example is SWELL stress dataset[6] by Koldijk et al. (2014). For fall detection, Zhang et al. (2015) and to a minor extent Cai et al. (2017) provide good reviews of available datasets.

Some authors also use top-view depth imagery (Liu et al., 2017a; Kasturi & Jo, 2017; Cippitelli et al., 2015, 2016). It is the case of Liu et al. (2017a), which propose to perform transfer learning along with cross-layer inheriting feature fusion (CLIFF). In their scheme the lower layers of a VGG16 model (7 convolutions) are kept frozen, with an additional block of convolution added. The result of the pooling block after that additional convolution is concatenated with features coming directly from the pooling after the fourth convolution. The authors claim this allows to train on a small dataset such as the one used, avoiding vanishing gradients or over-fitting. Like Liu et al. (2017a), Kasturi & Jo (2017) also use top-view imagery, but, as opposed to all works reviewed so far using depth-based sensors, their aim is to provide a fall detection system.

[5] `https://www.irc-sphere.ac.uk/sphere-challenge/home` (accessed: November 2018)
[6] `http://cs.ru.nl/~skoldijk/SWELL-KW/Dataset.html` (accessed: November 2018)

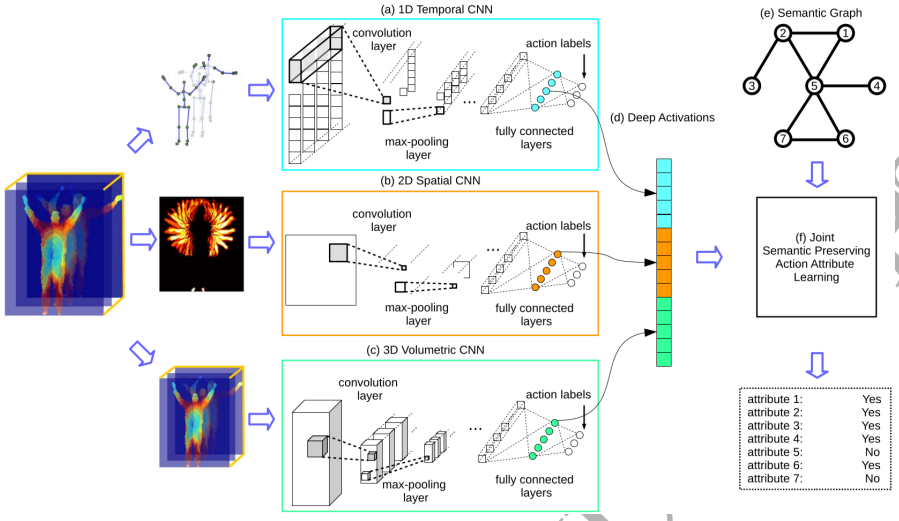
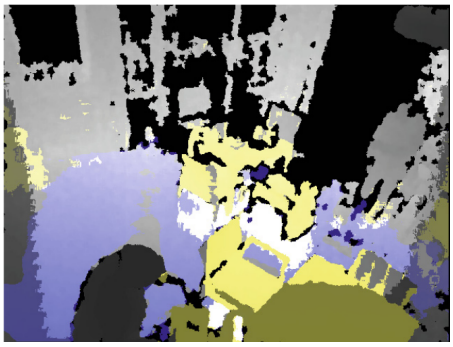


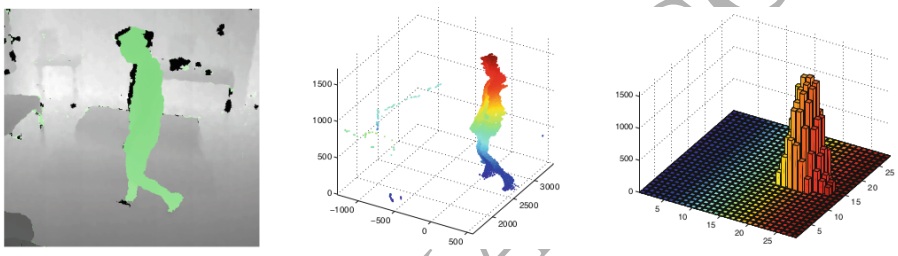
Figure 4: Example of multi-stream CNN using skeletal data, temporal templates, and spatio-temporal depth volumes in a single learning scheme. Features from the second-last layer as input for an attribute learning framework (reproduced from (Zhang et al., 2018)).

They extract shape-based features from the silhouettes, and classify them into fall or non-fall instances using an SVM classifier. It is worth noting here, that unlike HAR, fall detection tends to use simpler machine learning, and rarely uses DL-based methods. Another option is to extrapolate a worldtop view from the point cloud as done by Pramerdorfer et al. (2016). In their proposal they first identify the ground plane by an iterative RANSAC plane fitting. Objects are then detected filtering points with a height of 30 to 90 cm above the ground plane, horizontal planes are then analysed to determine whether they can be fall areas. Motion detection and tracking is performed, and therefore falls when multiple people are in a room are supported. Height and occupation maps are derived from the data, and subsequently used for tracking, finally falls detected from body poses and a rules-based reasoning for exclusion criteria.

Using skeleton data. Some methods using depth along with skeleton data have been reviewed above. However, methods exist using only skeleton data or data derived from joint information only. To recognise activities from skeleton data,



(a)



(b)

Figure 5: Pramerdorfer et al. propose an advanced fall detection framework: a) Frame showing colourisation of detected floor (blue) and furniture (yellow); b) from left to right: depth map, point cloud in plan-view space, and height map of moving objects (reproduced from (Pramerdorfer et al., 2016)).

first skeleton construction or recovery methods need to be run. These methods provide information on the location of the joints of a person, either in $2D$ (image) or $3D$ (real world) coordinates. These methods can be run from single RGB cameras, multiple camera views (can be positioned orthogonally, or not), and more recently with cameras providing depth information (i.e. RGB-D devices like Kinect and similar). Application programming interfaces (APIs) for RGB-D devices often offer a way to obtain the skeleton information. For instance Microsoft Kinect SDK offers an implementation of Shotton et al. (2011).

Before (Shotton et al., 2011), there were other attempts at using body-part detectors to find limb positions in images of people: either full body, e.g. Eich-

ACCEPTED MANUSCRIPT

ner et al. (2012); or depending on application, upper body, e.g. Cippitelli et al. (2016). With the advent of 'deep learning' and the success of convolutional neural networks at solving certain visual tasks, other researchers have tried to relax the constraints of using depth-enabled devices. That is, being able to extract skeleton information from RGB-only devices. An example of this is *OpenPose* by Wei et al. (2016), a CNN-based real-time system that can jointly detect human body, hand, facial and foot keypoints from single images. $2D$ capability is available for multiple people, whereas $3D$ point resolution is only available for a single detected person with triangulation from multiple views (Cao et al., 2017; Simon et al., 2017; Wei et al., 2016).

However, if the methods are based on a single RGB image, pose is normally estimated as a $2D$ skeleton (that is, on the image plane) as in Cherian et al. (2014), rather than a $3D$ skeleton in real world coordinates, although attempts at this exist, e.g. Andriluka et al. (2010). Several methods using DL-based techniques for $2D$ pose estimation via skeleton joint localisation exist (Fan et al., 2015; Tompson et al., 2014; Toshev & Szegedy, 2014). For instance, in Tompson et al. (2014) a CNN is trained to find the locations of joints in a fashion similar to that of Eichner et al. (2012), then a spatial model is learnt via Markov Random Fields, to constrain the pose to a *plausible* one. The CNN component looks at patches at different resolutions, in order to fine-tune the location of the joint. Similarly, in Fan et al. (2015) a *dual-source* CNN is used. This type of CNN receives both a general image and a close-up of the joint region, in order to better train for accurate joint position estimation. Finally, Toshev & Szegedy (2014) present *DeepPose*, which consists of a multi-stage process. The first stage (consisting of a CNN) is given a general view of the image, to estimate locations of joints. This first network's receptive field is limited in pixel size (ca. 220×220 image), therefore estimations tend to be coarse. All subsequent stages consist of a second CNN which receives a patch around the joint location of the first network and is trained to refine the joint locations (i.e. using real-valued regressors).

Formation of $3D$ skeletons tends to use systems with multiple cameras. For

ACCEPTED MANUSCRIPT

instance, Elhayek et al. (2015) propose to use a CNN as a 'body part' detector (estimating joint probabilities) which is then used with a model-based generative approach for skeleton fitting and skeletal motion tracking. They obtain $3D$ skeletons by aggregating joint information from multiple views. This is not novel, as reported in their literature review, but they are able to achieve minimal number of cameras to obtain $3D$ skeletons, stating that they can use as little as 2–3 cameras.

Once the skeleton is obtained, or constructed from the data, many works exist for action recognition (Ding et al., 2017; Liu et al., 2016a; Núñez et al., 2018; Zhu et al., 2016) using different skeletal data representations as reviewed by Han et al. (2017). Zhu et al. (2016) propose a regularisation to learn a joint co-occurrence feature of skeleton joints using skeletal data as input to a RNN with LSTM blocks. However, Ding et al. (2017) criticise RNN networks due to how these overemphasize temporal information, and therefore decide to encode temporal information via texture images. They compare different skeleton-based features (joint-to-joint distances, orientations, vectors; joint-line distances; line-line angles). Each feature is represented as a texture colour image, i.e. where columns represent spatial features in a frame, and rows encode the sequence of a specific feature. These features are then given to separate CNNs in a multi-stream fashion. Results are provided for all features combined, as well as for subsets of features (using feature selection). Another option to counter overemphasis in temporal information is provided by Liu et al. (2016a), who propose to use an RNN with LSTMs in a different way. Aware of the importance of spatial joint arrangement for action discrimination, the LSTMs in their network encode both spatial and temporal relationships. They do so by adding contextual information about other joint positions as well as the position of the joint in the previous time step. Furthermore, due to the nature of the sensors, which might include noisy inputs, trust gates are used as a mechanism to accept or ignore new data that might distort the spatio-temporal joint model learnt so far. Similarly, Núñez et al. (2018) use a combined neural network, consisting of a CNN and a RNN using LSTM blocks. The CNN is trained separately to learn

ACCEPTED MANUSCRIPT

representations from spatio-temporal skeletal data. The LSTM-RNN is then used to determine the action based on the underlying (input) representation. This method of training two 'deep learning'-based networks that are combined to form one single inference engine is the *de facto* standard in DL-based work in recent years (Ren & Zemel, 2017), although Núñez et al. train the components separately, instead of in an end-to-end fashion as is common.

3.2.2. From wearable or first-person vision

As stated in Nguyen et al. (2016), recognising activities where objects are manipulated in front of the hands (which includes many ADLs) can be hard from ambient-installed cameras, since the head and torso or a cluttered environment could occlude the activity. Furthermore, with cameras installed on the forehead (or disguised into smart glasses), or the chest, actions tend to take place (and objects tend to be) in the centre of the image, where camera focus is also better. As with ADLs detected from ambient-cameras, methods can be classified according to the semantic level of the behaviours being analysed: from motion, to actions, and activities, or long-term behaviours (Chaaraoui et al., 2012). Furthermore, the authors also differentiate between object-based activity recognition (using detected objects to infer activities being performed), as opposed to motion-based methodologies, which use physical features (magnitude, angle, frequency) of detected motions to recognise what the person is doing. The former have the challenge of detecting an activity with a set of missing detections, whereas the latter is a holistic approach better apt for coarser types of activities, which involve bigger motion patterns (e.g. a motion-based method will not be able to differentiate actions involving small object manipulation in the hands). In another work, Nguyen et al. (2018) propose using a CNN-based hand detection for improved activity recognition from egocentric vision, using the *EgoHands* dataset of Bambach et al. (2015). Related to motion-based first-person vision methods, *proprioceptive HAR* is another line of research that consists on recognising activities from the perception of the wearer's movements. One could use well established methods such as optical flow or interest point correspondence

ACCEPTED MANUSCRIPT

to detect variation of the scene as seen by the camera between consecutive or near frames. The detected motion can then be characterised via its strength, periodicity (or lack thereof) to recognise different activities. In this field, the works by Abebe & Cavallaro stand out (Abebe & Cavallaro, 2017b,c,a). Two of the methods are DL-based, whereas the other is not. In (Abebe & Cavallaro, 2017a) motion features are extracted by interest point detection and matching. These include magnitude, direction, as well as point descriptor changes. These are then temporally accumulated to create higher level features. In (Abebe & Cavallaro, 2017c), stacked spectrograms of motion patterns extracted from optical flow vectors and the displacement vectors of the intensity centroid are used in a CNN with LSTMs to encode temporal dependency. Stacking of spectrograms allows for the usage of $2D$ convolutional filters, which are much more common in off-the-shelf DL-based architectures. Temporal information is, according to the authors, the most important characteristic for the recognition of proprioceptive activities, and the LSTM component in the network is in charge precisely of this.

Egovision datasets:. Section 5 of the review by Nguyen et al. (2016) contains a summary of relevant datasets on the field of egocentric vision for AAL, specifically for ADL recognition. Most relevant datasets mentioned are the Activities of Daily Living (ADL) dataset of Pirsiavash & Ramanan (2012) from 2012; the Georgia Tech Egocentric Activities (GTEA) dataset[7] by Fathi et al. (2011), and GTEA Gaze+[8]. However, since the publication of (Nguyen et al., 2016) other datasets have appeared such as some datasets for manipulated object detection such as the EMMI dataset by Wang et al. (2017a). In their paper they also explore other manipulated object recognition datasets, which could be interesting to the reader. Also there have been extensions to previously existing datasets, such as the *extended* GTEA Gaze+ which subsumes the orig-

[7] `http://www.cbi.gatech.edu/fpv/` (accessed: November 2018)

[8] `http://ai.stanford.edu/~alireza/GTEA_Gaze_Website/GTEA_Gaze+.html` (accessed: November 2018)

ACCEPTED MANUSCRIPT

inal Gaze+ dataset, and features 28 hours (de-identified) of cooking activities from 86 unique sessions of 32 subjects. These videos come with audio and gaze tracking. The authors have further provided human annotations of actions (human-object interactions) and hand masks. The CMU Multi-Modal Activity Database[9] (CMU-MMAC) by De la Torre Frade et al. (2009) is mentioned in (Nguyen et al., 2016), but due to the lack of publicly available annotations, it has seldom been used. This is likely to change with the recent publication of the semantic annotation done by Yordanova et al. (2018). Another recent dataset is the EPIC-KITCHENS dataset by Damen et al. (2018): a large-scale egocentric video benchmark recorded by 32 participants in their native kitchen environments. Their videos depict non-scripted daily activities: they simply asked each participant to start recording every time they entered their kitchen. Recording took place in 4 cities (in North America and Europe) by participants belonging to 10 different nationalities, resulting in highly diverse cooking styles. The dataset features 55 hours of video consisting of 11.5M frames, which were densely labelled for a total of 39.6K action segments and 454.3K object bounding boxes. The resulting annotation is unique in that the participants narrated their own videos (after recording), thus reflecting true intention, the authors then crowd-sourced ground-truths based on these narrations.

3.2.3. From bespoke camera installations

As mentioned in the introduction to this section on camera perspectives, some methods need to have special conditions for a good-quality analysis of the signals to be processed from the images. Most of these entail physiological signal monitoring, such as breath and cardiac activity sensing via images (Chen & McDuff, 2018; Maclaren et al., 2015; Colantonio et al., 2015b; Andreu et al., 2016), affective status and well-being detection from faces (Colantonio et al., 2015a; Andreu-Cabedo et al., 2015; Henriquez et al., 2017), wound healing monitoring (Huimin et al., 2017), or automatic food journalling (Sen, 2017; Cippitelli

[9] `http://kitchen.cs.cmu.edu/` (accessed: Nov, 2018)

ACCEPTED MANUSCRIPT

et al., 2015, 2016). Huimin et al. (2017) use close-up images of wounds to analyse their healing process and determine whether the patient might require further care. For this purpose, they use a CNN to learn to segment the mask of the wound in the picture, and therefore determine its size. Maclaren et al. (2015) propose a system for measuring respiratory and cardiac information from a camera during a magnetic resonance (MR). The patient is lying during the test, and the camera is mounted above the forehead of the patient. Their method measures colour changes for heartbeat detection (based on ideas similar to Wu et al. (2012)), and motion in the head-foot direction for breath pattern detection. European Union's FP7 project SEMEOTICONS (Colantonio et al., 2015b,a) (2013–2017) led to a number of publications regarding the use of face analysis (semeiotics) for the diagnostic of cardio-metabolic syndrome. The project's main tool is a smart mirror (the *wize mirror* in project's terms (Andreu et al., 2016; Andreu-Cabedo et al., 2015; Henriquez et al., 2017)) equipped with multiple cameras and depth sensors, that along with a gas sniffer (*wize sniffer* (Germanese et al., 2017)) is able to detect most risk factors for cardiac and metabolic (type-2 diabetes) diseases, namely: the amount of face fat (indicator of overweight and obesity), its location near the eyelids (hypercholesterolemia), lack of skin micro-circulation (visible after local heating in healthy individuals), noxious habits (smoke and alcohol byproducts detected by sniffer), anxiety issues (via expression analysis of face), etc. Chen & McDuff (2018) propose a DL-based method for heart and breath rate detection from imagery consisting of close-up videos of the head and upper torso of individuals. Their *DeepPhys* framework consists of a convolutional attention network (CAN), which is a type of network that, based on knowledge about the human eye, gives more attention to a central area of the image (fovea), and less to the surrounding (context). This can also translate to just focusing more on a subset of features, and less on another group. The authors claim that in the fields of physiological measurements using computer vision, use of CNNs in the past was limited to feature extraction from images, but not to the calculation of the physiological metrics themselves. They claim to be the first to propose an end-to-end system that can simultaneously

ACCEPTED MANUSCRIPT

learn the spatial mask to detect the appropriate regions of interest (RoIs) and recovers the blood volume pulse (BVP) and respiration signals.

Finally, and on a different topic, in the thesis of Sen (2017, Ch. 4) the author presents an automated food journalling application using images captured from a smart watch. An accelerometer-based approach triggers the camera when eating-like motions are present in the wrist. Pictures are taken at the point of the motion where the biggest portion of the plate can be seen, these are then sent to a server for analysis and to determine food presence in pictures to filter uninteresting pictures. By doing this, an automated food journal can be created, to for instance, check adherence to a diet plan, or to calculate general well-being indices from food intake. Food intake analysis is also explored by Cippitelli et al. (2015), where a top-view RGB-D camera with improved matching of colour and depth is used to detect the individual, plates, cutlery, and contents of the plates. This is further extended in Cippitelli et al. (2016), where $3D$ localisation of upper limbs and head from a top-view RGB-D sensor is used in a system to analyse food type and intake behaviours.

4. Applications of lifelogging for AAL

A good review on video-based monitoring of patients and older people can be found in the work by Sathyanarayana et al. (2018). However their review is much more focused at *institutionalised* patients, and is much more dedicated to detection of medical conditions, and action recognition within the hospital. The authors look at different solutions focusing on the application. Specifically, they cover seven possible application fields, namely detection and/or monitoring of: falls, activities, sleep, apnoea, epilepsy, vital signs, and facial expression. All of these fields could be of interest to a person trying to monitor their own overall health and independence status, and to establish an early diagnosis of some conditions such as sleep disorders, apnoea, etc. Therefore, all applications presented in (Sathyanarayana et al., 2018) would be useful for lifelogging in an AAL scenario. Furthermore, the review is focused in vision techniques, i.e.

ACCEPTED MANUSCRIPT

using video (RGB), depth (RGB+D), infrared (IR) and time-of-flight (TOF) cameras; although some multi-sensor methods (including cameras in most cases but not all) are also presented. As stated in the motivation, Planinc et al. (2016) also propose a division of computer vision methods for AAL based on technologies and applications. These methods are part of the key enabling technologies laid out in the work by Moschetti et al. (2014), which describes the AALIANCE2 project, and explores the technologies for AAL currently existing throughout Europe and other parts of the world, identifies stakeholders and analyses their needs, and then propose a series of key enabling technologies that need to be present to achieve the goals of greater independence for older people, among others. It can be observed, from these two reviews, that common themes appear: activity recognition, fall prevention and detection, and physiological signs monitoring (including affective status, sleep quality, indicators of apnoea or epilepsy, and other niche areas). Other applications are not as useful for data collected into a personal lifelog (i.e. tele-rehabilitation, serious gaming, gesture-based interfaces, and assistive robotics). We will therefore focus on the former group. Table 4 shows how different methods reviewed are used or could potentially be used for different applications.

4.1. Human activity recognition

One of the tasks regarded as essential to AAL is human activity monitoring or human activity recognition (HAR) (Calvaresi et al., 2017; Rashidi & Mihailidis, 2013), sometimes mentioned under the umbrella of a wider "context awareness" concept (Queirós et al., 2015). In Aggarwal & Xia (2014) these methods are classified according to the type of feature that is used for recognition, namely: depth data, contours and silhouettes, and skeleton information. This same division has been used for methods using cameras installed in the environment under Sec. 3.2. The reader is referred to that section for an in-depth analysis of methods aiming at this application.

ACCEPTED MANUSCRIPT

Table 4: Applications provided by works surveyed in this review

Application	Reviewed literature	
	Real provision	Potential provision
HAR and HBA	Carreira & Zisserman (2017); Ding et al. (2017); Ijjina & Chalavadi (2017); Ji et al. (2017); Liu et al. (2016b,a, 2017a); Ma et al. (2017); Park et al. (2016); Rahmani & Bennamoun (2017); Shahroudy et al. (2016b); Wang et al. (2016, 2017b); Zhang et al. (2018); Zhu et al. (2016) Abebe & Cavallaro (2017b,c,a); Nakamura et al. (2017, 2016)	Eichner et al. (2012); Elhayek et al. (2015); Fan et al. (2015); Tompson et al. (2014); Toshev & Szegedy (2014); Varol et al. (2017)
Fall detection	Kasturi & Jo (2017); Mastorakis et al. (2018); Solbach & Tsotsos (2017)	(reviews: Khan & Hoey (2017))
Gait analysis and Fall prevention	Dubois & Charpillet (2014); Vaziri et al. (2017)	(reviews: Hamm et al. (2016); Rajagopalan et al. (2017); Cippitelli et al. (2017))
Physiological signal monitoring and well-being assessment	Andreu-Cabedo et al. (2015); Andreu et al. (2016); Chen & McDuff (2018); Colantonio et al. (2015a,b); Coppini et al. (2017); Germanese et al. (2017); Henriquez et al. (2017); Hurter & McDuff (2017); Irani (2017); Lewis et al. (2018); Li et al. (2017); Lopez-Martinez & Picard (2017); Huimin et al. (2017); Maclaren et al. (2015); Picard et al. (2001); Sen (2017)	Wu et al. (2012) (reviews: Betancourt et al. (2015); Faust et al. (2018); Sathyanarayana et al. (2018))

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

4.2. Gait analysis, fall detection and prevention

Several reviewed works focus on fall detection, or gait analysis for fall prevention (Dubois & Charpillet, 2014; Mastorakis et al., 2018; Solbach & Tsotsos, 2017; Stavropoulos et al., 2016; Vaziri et al., 2017; Kasturi & Jo, 2017; Cippitelli et al., 2017; Pramerdorfer et al., 2016). Cippitelli et al. (2017) is a review on fall detection methods from depth and radar sensors. Kasturi & Jo (2017) has already been mentioned under depth camera-based methods in Sec. 3.2.1. Most works on fall detection do not rely on advanced machine learning algorithms for decision, but rather use threshold-based methods. Solbach & Tsotsos (2017) propose to use stereo camera information to estimate the human pose and the ground plane in $3D$. Once this is achieved, they propose a number of measures to determine whether a person is fallen. Even if the human pose is calculated using a CNN, the reasoning behind, i.e. to detect a fall, is based on simple hand-crafted features, since detecting a fall can be derived from a knowledge-based reasoning, e.g. using a distance from ground calculated as the distance of the centre of gravity to the floor (ground plane). One of the problems with fall detection techniques is the lack of big datasets that are representative of a wide variety of fall instances as exposed in the review by Khan & Hoey (2017). To tackle this problem, Mastorakis et al. (2018) propose to use a physics-based simulated approach. They claim that fall recordings are unnecessary for modelling falls, since the simulation engine employed can produce a variety of fall events that can mimic an individual's physical conditions using myoskeletal models.

Focusing now on the gait analysis and fall prevention, Dubois & Charpillet (2014) propose to track three different gait parameters: length of steps, their duration, and speed of the gait. They compare the data measured in different situations (e.g. walking normally, or with actors wearing a skirt that impedes normal gait) to a *ground truth* consisting of the same gait parameters obtained by an actimetric carpet.

Finally, Vaziri et al. (2017) shows a quantitative and qualitative analysis of a fall prevention intervention, named *iStoppFalls* which is a video game-based system (exercise gaming, or *exergaming*) for older adults which aims to

ACCEPTED MANUSCRIPT

improve balance and strengthen key muscles which are frail in high risk fallers. Since adherence to a exercising routine is key to success in fall prevention, they quantitatively monitor patient progress and/or failure using several metrics. Furthermore, because of other factors beyond technical, they also propose a qualitative assessment to discover how older people regard the system, and what do they think could improve their likeliness to use the system for longer periods of time.

4.3. Physiological signal monitoring

Computer vision methods for physiological signal monitoring can be seen as an alternative to invasive systems requiring patch sensors on the skin (Irani, 2017; Li et al., 2017). Also, alternatives based on radar and laser can also be very costly. As said above, a good review on physiological signal monitoring is that of Sathyanarayana et al. (2018), which is focused on video devices. Also the reviews commented in the introduction about non-vision sensors, as much of physiological signal monitoring happens with other types of (mostly medical-grade) sensors (the review by Faust et al. (2018) is entirely on methods using these devices). One of the earliest works in this field is Picard et al. (2001), which estimates affective state of a patient by using four different physiological signals. Although vision is not used, it demonstrates the ability of physiological signals to provide valuable information for affective state recognition.

Hurter & McDuff (2017) present *Cardiolens* to provide a visual aid to perform remote physiological monitoring of heart rate. The idea is to integrate their algorithm in smart glasses to monitor subjects in front of the wearer, but it could well be adapted to other uses (e.g. a mirror as in (Colantonio et al., 2015a)). They propose a photoplethysmography algorithm using RGB information along with frequency filtering to obtain heart rate as per a previously validated method by the same authors.

The PhD thesis of Irani (2017) explores techniques for the analysis of human facial videos to provide contact-less (non-intrusive) methods for physiological signal recovery, including: heartbeat estimation, muscle fatigue detection, and

ACCEPTED MANUSCRIPT

pain/stress recognition. They propose a new method for heart estimation, that unlike others is not colour-based, but rather motion-based (i.e. performing tracking of facial landmarks). For pain, the author proposes a spatio-temporal technique based on energy changes of the facial muscles due to discomfort. For stress detection, they use a combination of RGB and thermal information, along with features from super-pixels, rather than directly from pixels as reported in the literature, achieving state-of-the-art performance.

Li et al. (2017) present a means for non-contact vision-based cardiopulmonary monitoring in different sleeping positions. Their method is aimed at apnoea detection during sleep and aims to be robust against postural change while sleeping. Their method is motion-based (tracking of distinctive points) and uses infra-red (IR) imagery (since presence of light would impede sleep in patients). They compare their results against a ground truth based on a polysomnography recording and report low mean percentage errors for heart and respiratory rates ($< 5.0\%$ and $< 3.4\%$ respectively).

Lewis et al. (2018) present a system for continuous cardiac activity monitoring combining an RGB-D device with a video camera (RGB). They claim that methods which can run on real-time have the potential to be embedded on a device, and call for better on-line methods, as opposed to existing methods which tend to evaluate data post-hoc, i.e. off-line. The RGB-D device is used to monitor the patient's face, whereas the features for cardiac activity monitoring are extracted from a video camera with better resolution. They also compare their results against ground truth ECG data.

At the convergence of first-person video and physiological monitoring, is the work by Nakamura et al. (Nakamura et al., 2016, 2017). They collected a dataset consisting of egocentric video augmented with heart rate and acceleration signals with more than 30 hours of video (Nakamura et al., 2016). Furthermore they propose a method for energy expenditure calculation and activity recognition using video and acceleration data, but using heart rate data during the training stage as a soft labelling of real energy expenditure. Their regression works on a recurrent network (using CNNs for feature extraction form the video

ACCEPTED MANUSCRIPT

and engineered acceleration features, with early fusion, then fed to LSTMs to consider the temporal dimension too). Also analysing energy expenditure, yet from ambient cameras is Tao et al. (2018). The proposed method uses a combination of visual (RGB-D) and inertial sensors to calculate energy expenditure. The proposed framework is individual-independent and fuses information from both modalities leading to improved estimates beyond the accuracy of each single modality and manual methods based on "metabolic equivalents of task" (MET) energy expenditure lookup tables, which are currently commonly used by professionals. In another work, Tao et al. (2017) compare calorific expenditure estimated from RGB-D data against physical gas exchange measurements in a domestic environment. From their experiments, the authors conclude that the proposed vision pipeline is suitable for home monitoring in a controlled environment.

5. Privacy and user acceptance

User views and preferences are important in the design and marketing of AAL solutions, as collected in several reviews on the topic (Arning & Ziefle, 2015; Bygholm & Kanstrup, 2015; Queirós et al., 2015). The works by Bygholm & Kanstrup (2015) and Queirós et al. (2015) have already been presented in the introduction (Section 1). However, Arning & Ziefle (2015) focus more on the user acceptance of AAL solutions based on not only the medical effectiveness of the proposed systems, but the combination of this factor with others such as camera typology and perceived privacy. Their conclusions are that acceptance has a lot to do with effectiveness of the proposed monitoring method (i.e. *medical safety* as is worded by the authors), and privacy is a concern in private spaces a lot more than it is in public spaces. Privacy concerns were mostly related to being recognisable, and less related to data privacy (e.g. storage of video for medical purposes). In fact, there are some completely unacceptable technologies according to the individuals interviewed: face recognition in the private scenario, storage in some cases, and seamless integration (i.e. cameras integrated in the

ACCEPTED MANUSCRIPT

home in an invisible manner).

Padilla-López et al. (2015) offer a review of different privacy preservation methods, first defining a series of dimensions of privacy (enumerated as a list of questions about the data and its processing) that systems deemed secure need to consider. The review shows an emphasis on different methodologies that can be followed (intervention, blind vision, secure processing, redaction and data hiding). Redaction methods are the most common, according to the authors, and they present different image filtering, encryption, de-identification of faces, object removal (via impainting), and visual abstraction (e.g. the use of avatars to hide the person's identity). They also discuss privacy at all levels of processing (acquisition, processing, storage, and retrieval), with advantages and drawbacks from the data security point of view. Furthermore, it also enumerates some of the video surveillance systems that take privacy into account, and to what extent they take into account all dimensions of privacy preservation proposed. Following this review, (Padilla-López et al., 2014) propose a series of image filters for privacy-aware real-time video redaction, with different levels of access depending on the person accessing the secure video channel (privacy-by-context). For instance, close relatives might be able to see the video with a filter that shows the face and pose of the person, whereas other stakeholders might only be able to see a more redacted output that still allows them to interpret what is happening in the scene without privacy-revealing details.

Ribaric et al. (2016) present the concept of de-identification in multimedia for privacy protection. They present a taxonomy of features that can identify a person (both biometric and non-biometric, such as textual information) and review existing methods to overcome identification (i.e. by detecting and replacing or scrambling the identifying data). In the biometric de-identification methods they include: face, fingerprints, voice, ear, gait and gestures; as well as soft identifiers such as height, body silhouette, gender, age, ethnicity, scars and tattoos, etc.

Another possibility for de-identification is *cartooning* (Erdélyi et al., 2013, 2014; Hassan et al., 2017). Erdélyi et al. (2013) propose a MeanShift-based

ACCEPTED MANUSCRIPT

method for cartooning (i.e. reducing the total number of colours and simplifying texture based on pixel property neighbourhood) with edge recovery to preserve sharp edges. This is done to obscure the identity of people while preserving video intelligibility. As part of their algorithm they also recolour personal items (such as scarves and carried bags) by shifting the hue, and perform further blurring of faces. In a later work, Erdélyi et al. (2014) propose to have an adaptive filter, i.e. where an operator can determine the level of obscuring performed. They also provide comparison to intelligibility, privacy, and appropriateness with pixelation and simple blurring. Finally, Hassan et al. (2017) use a similar cartooning method, and propose a deep learning based approach (using region-based convolutional networks, or R-CNNs for short) to replace personal identifying items (e.g. toothbrushes, TV and computer screen contents, etc.) with clip-art images from a pre-selected collection. They also apply this method on first-person videos (where such personal items are much more visible, especially screen contents), and claim to be the first to do so.

More particular to the field of lifelogging (LL), Gurrin et al. (2014) introduce a proposal for a privacy by design framework for LL. They introduce the stakeholders of a lifelogging system, namely: the individual, subjects the individual interacts with, passive bystanders (recorded unintentionally), and a host or hosts (people given access to the lifelog by the individual). They then analyse which aspects of lifelogging (devices used, stakeholders) have a potential for privacy breaches, and propose measures to counter them. For instance, they state the use of video logging is much more likely to cause breaches of privacy of bystanders, whereas pedometer data and other wearable data (e.g. temperature, heart rate, breathing) might not pose such privacy concerns. Among the measures are secure transmission and storage, as well as the right of anyone to chose whether to be in or out of someone else's lifelog.

Some reviewed works cover user acceptance studies of specific projects or finalised systems (Coppini et al., 2017; Stavropoulos et al., 2016; Vaziri et al., 2017). Coppini et al. (2017) provides a user acceptance and usability study regarding the *wize mirror* proposed in (Colantonio et al., 2015a; Henriquez

ACCEPTED MANUSCRIPT

et al., 2017). Another example is Stavropoulos et al. (2016) present the results of a system called `Dem@care`, which combines multiple types of sensors, including video and audio, but also wearable physiological signal devices. The system has undergone clinical trials in different countries and therefore it has been validated under several jurisdictions. One of the things the authors note, for instance, is how different national-level regulations allow or prevent the use of certain types of sensors in different environments due to how privacy issues are perceived in each society. These examples show the recent trend and effort in involving final users, to counter the issues perceived in reviews like (Bygholm & Kanstrup, 2015; Queirós et al., 2015) as shown in the introduction.

Finally, in the context of privacy and data security, it is worth mentioning recent developments in 'deep learning'-based methods (Abadi et al., 2016; Malekzadeh et al., 2018; Phan et al., 2016). With the advent of generative adversarial networks (GANs), it has been possible to extract information about the training data and/or to fool systems (Malekzadeh et al., 2018). The lack of studies about privacy preservation in DL-based methods has also been pointed out (Phan et al., 2016). For instance in (Abadi et al., 2016), privacy leaks from the perspective of the training data are analysed. If a training dataset contains real-world sensitive data, it could be possible to create attacks that target DL-based systems to retrieve training examples (Fredrikson et al., 2015) via *model inversion*. To counter this, Abadi et al. (2016) propose a framework for using *differential privacy* within the context of DL neural networks. They achieve this by using modified version of the stochastic gradient descent (SGD) algorithm: a differentially private SGD. It is also worth mentioning auto-encoders (AE), which can be used to preserve privacy when dealing with sensory data, such as in Malekzadeh et al. (2018), where a *replacement AE* (rAE) is proposed. This type of auto-encoder can retain accuracy while preserving the privacy of sensitive information. To do so, the rAE learns how to transform discriminative features that correspond to the inference of sensitive instances into a set of features that have been observed more often in non-sensitive data; all this while preserving the important features of desired inferences unchanged to allow for

ACCEPTED MANUSCRIPT

data sharing through public networks (e.g. usage of cloud services). To exemplify a usage scenario in an AAL environment, Malekzadeh et al.'s method would consider e.g. 'bathroom usage' as *sensitive* (and therefore substituted in the activity log with a *faked* non-sensitive event), 'reading' as *non-sensitive*, and an eldery falling as an *important event* (thus preserved in the activity log available to caregivers or medical practitioners). They demonstrate that GANs cannot deduce or find which non-sensitive inferences are actual ones, and which are *substitutions* of uninteresting but privacy-sensitive events. In a fashion similar to (Abadi et al., 2016), Phan et al. (2016) propose a *deep private auto-encoder* (dPA), which also uses principles based on ϵ-differential privacy.

6. Conclusions

The most recent advances in video-based intelligent lifelogging systems for AAL applications have been reviewed. Common technologies and techniques used across a number of applications in the field have been introduced. These applications have also been commented, especially those which can serve the purpose of *feeding* a lifelog that can be useful as a retrospective memory aid for patients, but also for caregivers and medical practitioners to know more about the day to day performance of the lifelogger, as well as their overall health status.

After analysing previous reviews in these areas, carried out until two years ago, it is clear that in the field of intelligent systems, deep learning techniques seem to have swept the board, at least for activity recognition and most other applications requiring advanced machine learning techniques (i.e. an exception to this is fall detection, which can still be successfully detected by using other methods). Among DL, it is interesting to note how CNN methods should still be preferred as the first architecture even for problems dealing with sequential data, in light of a recent systematic review by Bai et al. (2018) evaluating performance on tasks commonly used to benchmark recurrent neural networks (RNNs using LSTMs), in which results showed better performance for CNNs and even longer

ACCEPTED MANUSCRIPT

effective memory capabilities. This has also been noted in the reviewed works where multi-stream CNNs coding temporal data as $2D$ distributions and feeding them to a $2D$-filter network outperform other more complex methods (Bilen et al., 2016; Ma et al., 2017; Xiao et al., 2019).

Other interesting techniques for future work are those mixing dictionary-based approaches that had great acceptance in the past (bag-of-words modelling, Fisher vector encoding) with features obtained from convolutional neural networks trained with current means, as introduced by Liu et al. (2017b); Xie et al. (2017), instead of using handcrafted features. Alternatively, there is also the proposal of using decision trees as combinations of features extracted from CNNs, as explored by Tanno et al. (2019).

Newer video-based technologies, as RGB-D devices, which capture not only images but also depth and human pose information, and the application of deep learning models, are considerably improving the accuracy and reliability of lifelogging AAL services. However, their deployment in real environments is still far from being a reality, as systems need to deal with cluttered and changing environments, with differences in the way individual people perform their daily activities, and with the changes in the behaviour of a particular user along time. There are a couple of issues that need to be addressed in order to improve the results:

1. the lack of massive amounts of video data related to AAL applications, which are necessary to train modern intelligent systems; and
2. the necessity to involve older and frail people from the inception, and into the design, development and deployment of new technologies (Bygholm & Kanstrup, 2015; Queirós et al., 2015). The literature seems to indicate living labs are the best solution, as they allow an iterative trial and error approach with users, thus assuring their needs are met. Furthermore, proper testing and validation of proposed technologies is a must if these technologies are to be considered more than mere futuristic prototypes (Calvaresi et al., 2017).

ACCEPTED MANUSCRIPT

Privacy is also a concern, since technologies at the intersection of the fields mentioned are usually installed in private environments, where people develop their personal lives and have high expectations of privacy (Arning & Ziefle, 2015). Further studies regarding perceived privacy especially with regards to proposed image filtering approaches are needed, as well as to establish which measures could be taken to improve user acceptance, since benefits of the proposed technologies could potentially serve people most at need, and assist them in living on their own, preserving health for longer, and reassuring their caregivers and families.

Acknowledgements

This work is part of the PAAL – "Privacy-Aware and Acceptable Lifelogging services for older and frail people" project[10]: The support of the Joint Programme Initiative "More Years, Better Lives" (JPI MYBL, award number: PAAL_JTC2017) and the Canadian Institutes of Health Research, the German Bundesministeriums für Bildung und Forschung (grant no: 16SV7955), the Italian Ministero dell'Istruzione dell'Università e della Ricerca, the Spanish Agencia Estatal de Investigación (grant no: PCIN-2017-114) and the Swedish Research Council for Health, Working Life, and Welfare (grant no: 2017-02302) is gratefully acknowledged.

References

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In *2016 ACM SIGSAC Conference on Computer and Communications Security* CCS '16 (pp. 308–318). New York, NY, USA: ACM. doi:doi:10.1145/2976749.2978318.

[10] `www.paal-project.eu` (accessed: November 2018)

ACCEPTED MANUSCRIPT

Abdallah, Z. S., Gaber, M. M., Srinivasan, B., & Krishnaswamy, S. (2018). Activity recognition with evolving data streams: A review. *ACM Computing Surveys (CSUR)*, *51*, 71.

Abebe, G., & Cavallaro, A. (2017a). Hierarchical modeling for first-person vision activity recognition. *Neurocomputing*, *267*, 362 – 377. doi:doi:10.1016/j.neucom.2017.06.015.

Abebe, G., & Cavallaro, A. (2017b). Inertial-vision: Cross-domain knowledge transfer for wearable sensors. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)* (pp. 1392–1400). doi:doi:10.1109/ICCVW.2017.165.

Abebe, G., & Cavallaro, A. (2017c). A long short-term memory convolutional neural network for first-person vision activity recognition. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)* (pp. 1339–1346). doi:doi:10.1109/ICCVW.2017.159.

Aggarwal, J., & Xia, L. (2014). Human activity recognition from 3d data: A review. *Pattern Recognition Letters*, *48*, 70 – 80. doi:doi:10.1016/j.patrec.2014.04.011. Celebrating the life and work of Maria Petrou.

Andreu, Y., Chiarugi, F., Colantonio, S., Giannakakis, G., Giorgi, D., Henriquez, P., Kazantzaki, E., Manousos, D., Marias, K., Matuszewski, B. J., Pascali, M. A., Pediaditis, M., Raccichini, G., & Tsiknakis, M. (2016). Wize mirror - a smart, multisensory cardio-metabolic risk monitoring system. *Computer Vision and Image Understanding*, *148*, 3 – 22. doi:doi:10.1016/j.cviu.2016.03.018.

Andreu-Cabedo, Y., Castellano, P., Colantonio, S., Coppini, G., Favilla, R., Germanese, D., Giannakakis, G., Giorgi, D., Larsson, M., Marraccini, P., Martinelli, M., Matuszewski, B., Milanic, M., Pascali, M., Pediaditis, M., Raccichini, G., Randeberg, L., Salvetti, O., & Stromberg, T. (2015). Mirror

ACCEPTED MANUSCRIPT

mirror on the wall ... an intelligent multisensory mirror for well-being self-assessment. In *2015 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1–6). doi:doi:10.1109/ICME.2015.7177468.

Andriluka, M., Roth, S., & Schiele, B. (2010). Monocular 3d pose estimation and tracking by detection. In *2010 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 623–630). IEEE.

Antunes, R. S., Seewald, L. A., Rodrigues, V. F., Costa, C. A. D., Jr., L. G., Righi, R. R., Maier, A., Eskofier, B., Ollenschläger, M., Naderi, F., Fahrig, R., Bauer, S., Klein, S., & Campanatti, G. (2018). A survey of sensors in healthcare workflow monitoring. *ACM Comput. Surv., 51*, 42:1–42:37. doi:doi:10.1145/3177852.

Ardeshir, S., & Borji, A. (2016). Ego2top: Matching viewers in egocentric andtop-view videos. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *European Conference on Computer Vision (ECCV)* (pp. 253–268). Cham: Springer International Publishing.

Ardeshir, S., & Borji, A. (2018). Egocentric meets top-view. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, (pp. 1–1). doi:doi:10.1109/TPAMI.2018.2832121.

Arning, K., & Ziefle, M. (2015). "get that camera out of my house!" conjoint measurement of preferences for video-based healthcare monitoring systems in private and public places. In *International Conference on Smart Homes and Health Telematics* (pp. 152–164). Springer.

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, .

Baldewijns, G., Debard, G., Mertes, G., Vanrumste, B., & Croonenborghs, T. (2016). Bridging the gap between real-life data and simulated data by pro-

ACCEPTED MANUSCRIPT

viding a highly realistic fall dataset for evaluating camera-based fall detection algorithms. *Healthcare Technology Letters*, *3*, 6–11(5).

Bambach, S., Lee, S., Crandall, D. J., & Yu, C. (2015). Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *2015 IEEE International Conference on Computer Vision (ICCV)* (pp. 1949–1957). doi:doi:10.1109/ICCV.2015.226.

Betancourt, A., Morerio, P., Regazzoni, C. S., & Rauterberg, M. (2015). The evolution of first person vision methods: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, *25*, 744–760. doi:doi:10.1109/TCSVT.2015.2409731.

Bilen, H., Fernando, B., Gavves, E., Vedaldi, A., & Gould, S. (2016). Dynamic image networks for action recognition. In *2016 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3034–3042).

Bochinski, E., Eiselein, V., & Sikora, T. (2016). Training a convolutional neural network for multi-class object detection using solely virtual world data. In *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 278–285). doi:doi:10.1109/AVSS.2016.7738056.

Bygholm, A., & Kanstrup, A. M. (2015). The living challenge of ambient assisted living - a literature review. In *SHI 2015, Proceedings from The 13th Scandinavien Conference on Health Informatics, June 15-17, 2015, Tromsø, Norway* 115 (pp. 89–92). Linköping University Electronic Press, Linköpings universitet.

Cai, Z., Han, J., Liu, L., & Shao, L. (2017). Rgb-d datasets using microsoft kinect or similar sensors: a survey. *Multimedia Tools and Applications*, *76*, 4313–4355. doi:doi:10.1007/s11042-016-3374-6.

Calvaresi, D., Cesarini, D., Sernani, P., Marinoni, M., Dragoni, A. F., & Sturm,

ACCEPTED MANUSCRIPT

A. (2017). Exploring the ambient assisted living domain: a systematic review. *Journal of Ambient Intelligence and Humanized Computing*, *8*, 239–257.

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *2017 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7291–7299).

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4724–4733). IEEE.

Chaaraoui, A. A., Climent-Pérez, P., & Flórez-Revuelta, F. (2013). Silhouette-based human action recognition using sequences of key poses. *Pattern Recognition Letters*, *34*, 1799–1807.

Chaaraoui, A. A., Climent-Pérez, P., & Flórez-Revuelta, F. (2012). A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert Systems with Applications*, *39*, 10873 – 10888. doi:doi:10.1016/j.eswa.2012.03.005.

Chen, C., Liu, K., & Kehtarnavaz, N. (2016). Real-time human action recognition based on depth motion maps. *Journal of Real-Time Image Processing*, *12*, 155–163. doi:doi:10.1007/s11554-013-0370-1.

Chen, W., & McDuff, D. (2018). DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 349–365).

Cherian, A., Mairal, J., Alahari, K., & Schmid, C. (2014). Mixing body-part sequences for human pose estimation. In *2014 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2361–2368). doi:doi:10.1109/CVPR.2014.302.

Choutas, V., Weinzaepfel, P., Revaud, J., & Schmid, C. (2018). Potion: Pose motion representation for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7024–7033).

ACCEPTED MANUSCRIPT

Cippitelli, E., Fioranelli, F., Gambi, E., & Spinsante, S. (2017). Radar and rgb-depth sensors for fall detection: a review. *IEEE Sensors Journal*, *17*, 3585–3604.

Cippitelli, E., Gasparrini, S., De Santis, A., Montanini, L., Raffaeli, L., Gambi, E., & Spinsante, S. (2015). Comparison of rgb-d mapping solutions for application to food intake monitoring. In B. Andò, P. Siciliano, V. Marletta, & A. Monteriù (Eds.), *Ambient Assisted Living: Italian Forum 2014* (pp. 295–305). Cham: Springer International Publishing. doi:doi:10.1007/978-3-319-18374-9_28.

Cippitelli, E., Gasparrini, S., Gambi, E., & Spinsante, S. (2016). Unobtrusive intake actions monitoring through rgb and depth information fusion. In *2016 IEEE 12th International Conference on Intelligent Computer Communication and Processing (ICCP)* (pp. 19–26). doi:doi:10.1109/ICCP.2016.7737116.

Colantonio, S., Coppini, G., Germanese, D., Giorgi, D., Magrini, M., Marraccini, P., Martinelli, M., Morales, M. A., Pascali, M. A., Raccichini, G., Righi, M., & Salvetti, O. (2015a). A smart mirror to promote a healthy lifestyle. *Biosystems Engineering*, *138*, 33 – 43. doi:doi:10.1016/j.biosystemseng.2015.06.008.

Colantonio, S., Germanese, D., Moroni, D., Giorgi, D., Pascali, M., Righi, M., Coppini, G., Morales, M. A., Chiarugi, F., Pediaditis, M., Larsson, M., Stromberg, T., Henriquez, P., Matuszewski, B., Milanic, M., & Randeberg, L. (2015b). Semeoticons - reading the face code of cardio-metabolic risk. In *2015 International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM)* (pp. 1–5). doi:doi:10.1109/IWCIM.2015.7347092.

Coppini, G., Zuccaia, V. C., Marià, R. D., Nazare, J. A., Morales, M. A., & Colantonio, S. (2017). User acceptance of self-monitoring technology to prevent cardio-metabolic diseases: The wize mirror. In *2017 IEEE 13th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)* (pp. 265–271). doi:doi:10.1109/WiMOB.2017.8115837.

ACCEPTED MANUSCRIPT

Damen, D., Doughty, H., Maria Farinella, G., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W. et al. (2018). Scaling egocentric vision: The EPIC-KITCHENS dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 720–736).

De la Torre Frade, F., Hodgins, J. K., Bargteil, A. W., Artal, X. M., Macey, J. C., Castells, A. C. I., & Beltran, J. (2009). *Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database*. Technical Report CMU-RI-TR-08-22 Carnegie Mellon University Pittsburgh, PA.

Díaz Rodríguez, N., Cuéllar, M. P., Lilius, J., & Calvo-Flores, M. D. (2014). A survey on ontologies for human behavior recognition. *ACM Computing Surveys*, *46*, 43:1–43:33. doi:doi:10.1145/2523819.

Ding, Z., Wang, P., Ogunbona, P. O., & Li, W. (2017). Investigation of different skeleton features for cnn-based 3d action recognition. In *2017 IEEE International Conference on Multimedia Expo Workshops (ICMEW)* (pp. 617–622). doi:doi:10.1109/ICMEW.2017.8026286.

Dubois, A., & Charpillet, F. (2014). A gait analysis method based on a depth camera for fall prevention. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 4515–4518). doi:doi:10.1109/EMBC.2014.6944627.

Eichner, M., Marin-Jimenez, M., Zisserman, A., & Ferrari, V. (2012). 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. *International journal of Computer Vision*, *99*, 190–214.

Elhayek, A., de Aguiar, E., Jain, A., Tompson, J., Pishchulin, L., Andriluka, M., Bregler, C., Schiele, B., & Theobalt, C. (2015). Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. In *2015 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3810–3818). IEEE.

ACCEPTED MANUSCRIPT

Erdélyi, Á., Barát, T., Valet, P., Winkler, T., & Rinner, B. (2014). Adaptive cartooning for privacy protection in camera networks. In *Advanced Video and Signal Based Surveillance (AVSS), 2014 11th IEEE International Conference on* (pp. 44–49). IEEE.

Erdélyi, Á., Winkler, T., & Rinner, B. (2013). Serious fun: Cartooning for privacy protection. In *MediaEval workshop* (pp. 1–2).

Erden, F., Velipasalar, S., Alkar, A. Z., & Cetin, A. E. (2016). Sensors in assisted living: A survey of signal and image processing methods. *IEEE Signal Processing Magazine*, *33*, 36–44. doi:doi:10.1109/MSP.2015.2489978.

Fan, X., Zheng, K., Lin, Y., & Wang, S. (2015). Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In *2015 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1347–1355).

Fathi, A., Ren, X., & Rehg, J. M. (2011). Learning to recognize objects in egocentric activities. In *2011 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3281–3288). IEEE.

Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S., & Acharya, U. R. (2018). Deep learning for healthcare applications based on physiological signals: A review. *Computer Methods and Programs in Biomedicine*, *161*, 1 – 13. doi:doi:10.1016/j.cmpb.2018.04.005.

Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In *2016 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1933–1941).

Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In *22Nd ACM SIGSAC Conference on Computer and Communications Security* CCS '15 (pp. 1322–1333). New York, NY, USA: ACM. doi:doi:10.1145/2810103.2813677.

ACCEPTED MANUSCRIPT

Germanese, D., Righi, M., Benassi, A., D'Acunto, M., Leone, R., Magrini, M., Paradisi, P., Puppi, D., & Salvetti, O. (2017). A low cost, portable device for breath analysis and self-monitoring, the wize sniffer. *Applications in Electronics Pervading Industry, Environment and Society*, (p. 51).

Gurrin, C., Albatal, R., Joho, H., & Ishii, K. (2014). A privacy by design approach to lifelogging. In K. O'Hara, C. Nguyen, & P. Haynes (Eds.), *Digital Enlightenment Yearbook 2014* (pp. 49–73). The Netherlands: IOS Press.

Hamm, J., Money, A. G., Atwal, A., & Paraskevopoulos, I. (2016). Fall prevention intervention technologies: A conceptual framework and survey of the state of the art. *Journal of Biomedical Informatics*, *59*, 319 – 345. doi:doi:10.1016/j.jbi.2015.12.013.

Han, F., Reily, B., Hoff, W., & Zhang, H. (2017). Space-time representation of people based on 3d skeletal data: A review. *Computer Vision and Image Understanding*, *158*, 85–105.

Hassan, E. T., Hasan, R., Shaffer, P., Crandall, D. J., & Kapadia, A. (2017). Cartooning for enhanced privacy in lifelogging and streaming videos. In *2017 IEEE conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)* (pp. 1333–1342).

Henriquez, P., Matuszewski, B. J., Andreu-Cabedo, Y., Bastiani, L., Colantonio, S., Coppini, G., D'Acunto, M., Favilla, R., Germanese, D., Giorgi, D., Marraccini, P., Martinelli, M., Morales, M. A., Pascali, M. A., Righi, M., Salvetti, O., Larsson, M., Strömberg, T., Randeberg, L., Bjorgan, A., Giannakakis, G., Pediaditis, M., Chiarugi, F., Christinaki, E., Marias, K., & Tsiknakis, M. (2017). Mirror mirror on the wall... an unobtrusive intelligent multisensory mirror for well-being status self-assessment and visualization. *IEEE Transactions on Multimedia*, *19*, 1467–1481. doi:doi:10.1109/TMM.2017.2666545.

Herath, S., Harandi, M., & Porikli, F. (2017). Going deeper into action recognition: A survey. *Image and Vision Computing*, *60*, 4 – 21. doi:doi:10.1016/

ACCEPTED MANUSCRIPT

j.imavis.2017.01.010. Regularization Techniques for High-Dimensional Data Analysis.

Hochreiter, S., & Schmidhuber, J. (1997). Lstm can solve hard long time lag problems. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 473–479).

Huimin, L., Bin, L., Junwu, Z., Yujie, L., Yun, L., Xing, X., Li, H., Xin, L., Jianru, L., & Seiichi, S. (2017). Wound intensity correction and segmentation with convolutional neural networks. *Concurrency and Computation: Practice and Experience*, *29*, e3927. doi:doi:10.1002/cpe.3927.

Hurter, C., & McDuff, D. (2017). Cardiolens: Remote physiological monitoring in a mixed reality environment. In *ACM SIGGRAPH 2017 Emerging Technologies* SIGGRAPH '17 (pp. 6:1–6:2). New York, NY, USA: ACM. doi:doi:10.1145/3084822.3084834.

Ijjina, E. P., & Chalavadi, K. M. (2017). Human action recognition in rgb-d videos using motion sequence information and deep learning. *Pattern Recognition*, *72*, 504 – 516. doi:doi:10.1016/j.patcog.2017.07.013.

Irani, R. (2017). *Computer Vision Based Methods for Detection and Measurement of Psychophysiological Indicators*. Ph.D. thesis Technical Faculty for IT, Aalborg University.

Jalal, A., Kamal, S., & Kim, D. (2014). A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments. *Sensors*, *14*, 11735–11759. doi:doi:10.3390/s140711735.

Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3d convolutional neural networks for human action recognition. *IEEE transactions on Pattern Analysis and Machine Intelligence*, *35*, 221–231.

Ji, X., Cheng, J., Tao, D., Wu, X., & Feng, W. (2017). The spatial laplacian and temporal energy pyramid representation for human action recognition

ACCEPTED MANUSCRIPT

using depth sequences. *Knowledge-Based Systems*, *122*, 64 – 74. doi:doi:10.1016/j.knosys.2017.01.035.

Kasturi, S., & Jo, K. H. (2017). Human fall classification system for ceiling-mounted kinect depth images. In *2017 17th International Conference on Control, Automation and Systems (ICCAS)* (pp. 1346–1349). doi:doi:10.23919/ICCAS.2017.8204202.

Khan, S. S., & Hoey, J. (2017). Review of fall detection techniques: A data availability perspective. *Medical Engineering and Physics*, *39*, 12 – 22. doi:doi:10.1016/j.medengphy.2016.10.014.

Koldijk, S., Sappelli, M., Verberne, S., Neerincx, M. A., & Kraaij, W. (2014). The swell knowledge work dataset for stress and user modeling research. In *16th International Conference on Multimodal Interaction* (pp. 291–298). ACM.

Kong, Y., & Fu, Y. (2018). Human action recognition and prediction: A survey. *arXiv preprint arXiv:1806.11230*, .

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 1097–1105).

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., & Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 396–404).

Leo, M., Medioni, G., Trivedi, M., Kanade, T., & Farinella, G. (2017). Computer vision for assistive technologies. *Computer Vision and Image Understanding*, *154*, 1–15. doi:doi:10.1016/j.cviu.2016.09.001.

Lewis, G. F., Davila, M. I., & Porges, S. W. (2018). Novel algorithms to monitor continuous cardiac activity with a video camera. In *2018 IEEE conference*

ACCEPTED MANUSCRIPT

on Computer Vision and Pattern Recognition Workshops (CVPR-W) (pp. 1395–1403).

Li, M. H., Yadollahi, A., & Taati, B. (2017). Noncontact vision-based cardiopulmonary monitoring in different sleeping positions. *IEEE Journal of Biomedical and Health Informatics*, *21*, 1367–1375. doi:doi:10.1109/JBHI.2016.2567298.

Liu, A. S., Li, Z. J., Yeh, T. H., Yang, Y. H., & Fu, L. C. (2017a). Partially transferred convolution neural network with cross-layer inheriting for posture recognition from top-view depth camera. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 4139–4143). doi:doi:10.1109/IROS.2017.8206273.

Liu, J., Shahroudy, A., Xu, D., & Wang, G. (2016a). Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision (ECCV)* (pp. 816–833). Springer.

Liu, X., Zhang, S., Huang, T., & Tian, Q. (2017b). E^2BoWs: An end-to-end bag-of-words model via deep convolutional neural network. *Neurocomputing (accepted)*, .

Liu, Z., Zhang, C., & Tian, Y. (2016b). 3d-based deep convolutional neural network for action recognition with depth sequences. *Image and Vision Computing*, *55*, 93 – 100. doi:doi:10.1016/j.imavis.2016.04.004.

Lopez-Martinez, D., & Picard, R. W. (2017). Multi-task neural networks for personalized pain recognition from physiological signals. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* (pp. 181–184).

Ma, C.-Y., Chen, M.-H., Kira, Z., & AlRegib, G. (2019). TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. *Signal Processing: Image Communication*, *71*, 76–87.

ACCEPTED MANUSCRIPT

Ma, M., Marturi, N., Li, Y., Leonardis, A., & Stolkin, R. (2018). Region-sequence based six-stream cnn features for general and fine-grained human action recognition in videos. *Pattern Recognition*, *76*, 506–521.

Ma, S., Bargal, S. A., Zhang, J., Sigal, L., & Sclaroff, S. (2017). Do less and achieve more: Training cnns for action recognition utilizing action images from the web. *Pattern Recognition*, *68*, 334 – 345. doi:doi:10.1016/j.patcog.2017.01.027.

Maclaren, J., Aksoy, M., & Bammer, R. (2015). Contact-free physiological monitoring using a markerless optical system. *Magnetic resonance in medicine*, *74 2*, 571–7.

Malekzadeh, M., Clegg, R. G., & Haddadi, H. (2018). Replacement autoencoder: A privacy-preserving algorithm for sensory data analysis. In *2018 IEEE/ACM Third International Conference on Internet-of-Things Design and Implementation (IoTDI)* (pp. 165–176). doi:doi:10.1109/IoTDI.2018.00025.

Martinez-Gonzalez, P., Oprea, S., Garcia-Garcia, A., Jover-Alvarez, A., Orts-Escolano, S., & Garcia-Rodriguez, J. (2018). Unrealrox: An extremely photorealistic virtual reality environment for robotics simulations and synthetic data generation. *arXiv preprint arXiv:1810.06936*, .

Mastorakis, G., Ellis, T., & Makris, D. (2018). Fall detection without people: A simulation approach tackling video data scarcity. *Expert Systems with Applications*, *112*, 125 – 137. doi:doi:10.1016/j.eswa.2018.06.019.

McDuff, D. (2018). Deep super resolution for recovering physiological information from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 1367–1374).

McDuff, D. J., Estepp, J. R., Piasecki, A. M., & Blackford, E. B. (2015). A survey of remote optical photoplethysmographic imaging methods. In *2015 37th Annual International Conference of the IEEE on Engineering in Medicine and Biology Society (EMBC)* (pp. 6398–6404). IEEE.

ACCEPTED MANUSCRIPT

Moschetti, A., Fiorini, L., Aquilano, M., Cavallo, F., & Dario, P. (2014). Preliminary findings of the aaliance2 ambient assisted living roadmap. In S. Longhi, P. Siciliano, M. Germani, & A. Monteriù (Eds.), *Ambient Assisted Living* (pp. 335–342). Cham: Springer International Publishing.

Mukhopadhyay, S. C. (2015). Wearable sensors for human activity monitoring: A review. *IEEE Sensors Journal*, *15*, 1321–1330. doi:doi:10.1109/JSEN.2014.2370945.

Nakamura, K., Alahi, A., Yeung, S., & Fei-Fei, L. (2016). Egocentric multimodal dataset with visual and physiological signals. In *2017 IEEE conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)* (pp. 1–2).

Nakamura, K., Yeung, S., Alahi, A., & Fei-Fei, L. (2017). Jointly learning energy expenditures and activities using egocentric multimodal signals. In *2017 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1868–1877).

Nguyen, T.-H.-C., Nebel, J.-C., & Florez-Revuelta, F. (2016). Recognition of activities of daily living with egocentric vision: A review. *Sensors*, *16*. doi:doi:10.3390/s16010072.

Nguyen, T.-H.-C., Nebel, J.-C., & Florez-Revuelta, F. (2018). Recognition of activities of daily living from egocentric videos using hands detected by a deep convolutional network. In A. Campilho, F. Karray, & B. ter Haar Romeny (Eds.), *Image Analysis and Recognition* (pp. 390–398). Cham: Springer International Publishing.

Núñez, J. C., Cabido, R., Pantrigo, J. J., Montemayor, A. S., & Vélez, J. F. (2018). Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. *Pattern Recognition*, *76*, 80 – 94. doi:doi:10.1016/j.patcog.2017.10.033.

ACCEPTED MANUSCRIPT

Padilla-López, J. R., Chaaraoui, A. A., & Flórez-Revuelta, F. (2014). Visual privacy by context: A level-based visualisation scheme. In R. Hervás, S. Lee, C. Nugent, & J. Bravo (Eds.), *Ubiquitous Computing and Ambient Intelligence. Personalisation and User Adapted Services* (pp. 333–336). Cham: Springer International Publishing.

Padilla-López, J. R., Chaaraoui, A. A., & Flórez-Revuelta, F. (2015). Visual privacy protection methods: A survey. *Expert Systems with Applications*, *42*, 4177 – 4195. doi:doi:10.1016/j.eswa.2015.01.041.

Park, E., Han, X., Berg, T. L., & Berg, A. C. (2016). Combining multiple sources of knowledge in deep cnns for action recognition. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1–8). doi:doi:10.1109/WACV.2016.7477589.

Pham, H.-H., Khoudour, L., Crouzil, A., Zegers, P., & Velastin, S. A. (2018). Exploiting deep residual networks for human action recognition from skeletal data. *Computer Vision and Image Understanding*, *170*, 51–66.

Phan, N., Wang, Y., Wu, X., & Dou, D. (2016). Differential privacy preservation for deep auto-encoders: an application of human behavior prediction. In *AAAI* (pp. 1309–1316). volume 16.

Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *23*, 1175–1191. doi:doi:10.1109/34.954607.

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)* (pp. 2847–2854). doi:doi:10.1109/CVPR.2012.6248010.

Planinc, R., Chaaraoui, A. A., Kampel, M., & Florez-Revuelta, F. (2016). Computer vision for active and assisted living. In *Active and Assisted Living:*

ACCEPTED MANUSCRIPT

Technologies and Applications Healthcare Technologies (pp. 57–79). Institution of Engineering and Technology. doi:doi:10.1049/PBHE006Ech4.

Pramerdorfer, C., Planinc, R., Van Loock, M., Fankhauser, D., Kampel, M., & Brandstötter, M. (2016). Fall detection based on depth-data in practice. In G. Hua, & H. Jégou (Eds.), *European Conference on Computer Vision Workshops (ECCV-W)* (pp. 195–208). Cham: Springer International Publishing.

Prati, A., Shan, C., & Wang, K. I.-K. (2019). Sensors, vision and networks: From video surveillance to activity recognition and health monitoring. *Journal of Ambient Intelligence and Smart Environments*, *11*, 5–22.

Queirós, A., Silva, A., Alvarelhão, J., Rocha, N. P., & Teixeira, A. (2015). Usability, accessibility and ambient-assisted living: a systematic literature review. *Universal Access in the Information Society*, *14*, 57–66. doi:doi:10.1007/s10209-013-0328-x.

Rahmani, H., & Bennamoun, M. (2017). Learning action recognition model from depth and skeleton videos. In *The IEEE International Conference on Computer Vision (ICCV)* (pp. 5832–5841).

Rahmani, H., & Mian, A. (2016). 3d action recognition from novel viewpoints. In *2016 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1506–1515).

Rajagopalan, R., Litvan, I., & Jung, T.-P. (2017). Fall prediction and prevention systems: Recent trends, challenges, and future research directions. *Sensors*, *17*, 2509. doi:doi:10.3390/s17112509.

Rashidi, P., & Mihailidis, A. (2013). A survey on ambient-assisted living tools for older adults. *IEEE Journal of Biomedical and Health Informatics*, *17*, 579–590.

Ren, M., & Zemel, R. S. (2017). End-to-end instance segmentation with recurrent attention. In *2017 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6656–6664).

ACCEPTED MANUSCRIPT

Ribaric, S., Ariyaeeinia, A., & Pavesic, N. (2016). De-identification for privacy protection in multimedia content: A survey. *Signal Processing: Image Communication*, *47*, 131 – 151. doi:doi:https://doi.org/10.1016/j.image.2016.05.020.

Sathyanarayana, S., Satzoda, R. K., Sathyanarayana, S., & Thambipillai, S. (2018). Vision-based patient monitoring: a comprehensive review of algorithms and technologies. *Journal of Ambient Intelligence and Humanized Computing*, *9*, 225–251. doi:doi:10.1007/s12652-015-0328-1.

Selke, S. (2016). *Lifelogging: Digital self-tracking and Lifelogging-between disruptive technology and cultural transformation*. Springer.

Sen, S. (2017). *Fusing Mobile, Wearable and Infrastructure Sensing for Immersive Daily Lifestyle Analytics*. Ph.D. thesis Singapore Management University.

Shahroudy, A., Liu, J., Ng, T.-T., & Wang, G. (2016a). Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *2016 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1010–1019).

Shahroudy, A., Ng, T. T., Yang, Q., & Wang, G. (2016b). Multimodal multipart learning for action recognition in depth videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *38*, 2123–2129. doi:doi:10.1109/TPAMI.2015.2505295.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., & Blake, A. (2011). Real-time human pose recognition in parts from single depth images. In *2011 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1297–1304). IEEE.

Simon, T., Joo, H., Matthews, I., & Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. In *2017 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1145–1153).

ACCEPTED MANUSCRIPT

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 568–576).

Solbach, M. D., & Tsotsos, J. K. (2017). Vision-based fallen person detection for the elderly. In *The IEEE International Conference on Computer Vision (ICCV) Workshops* (pp. 1433–1442).

Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M. (2012). A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, *3*, 42–55.

Stavropoulos, T. G., Meditskos, G., Briassouli, A., & Kompatsiaris, I. (2016). Multimodal sensing and intelligent fusion for remote dementia care and support. In *2016 ACM Workshop on Multimedia for Personal Health and Health Care* MMHealth '16 (pp. 35–39). New York, NY, USA: ACM. doi:doi: 10.1145/2985766.2985776.

Tanno, R., Arulkumaran, K., Alexander, D., Criminisi, A., & Nori, A. (2019). Adaptive neural trees. In K. Chaudhuri, & R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning* (pp. 6166–6175). Long Beach, California, USA: PMLR volume 97 of *Proceedings of Machine Learning Research*. URL: `http://proceedings.mlr.press/v97/tanno19a.html`.

Tao, L., Burghardt, T., Mirmehdi, M., Damen, D., Cooper, A., Camplani, M., Hannuna, S., Paiement, A., & Craddock, I. (2018). Energy expenditure estimation using visual and inertial sensors. *IET Computer Vision*, *12*, 36–47(11).

Tao, L., Burghardt, T., Mirmehdi, M., Damen, D., Cooper, A., Hannuna, S., Camplani, M., Paiement, A., & Craddock, I. (2017). Calorie counter: Rgb-depth visual estimation of energy expenditure at home. In C.-S. Chen, J. Lu, & K.-K. Ma (Eds.), *2016 Asian Conference on Computer Vision Workshops (ACCV-W)* (pp. 239–251). Cham: Springer International Publishing.

ACCEPTED MANUSCRIPT

Theodorakopoulos, I., Kastaniotis, D., Economou, G., & Fotopoulos, S. (2014). Pose-based human action recognition via sparse representation in dissimilarity space. *Journal of Visual Communication and Image Representation*, *25*, 12 – 23. doi:doi:https://doi.org/10.1016/j.jvcir.2013.03.008. Visual Understanding and Applications with RGB-D Cameras.

Thevenot, J., López, M. B., & Hadid, A. (2018). A survey on computer vision for assistive medical diagnosis from faces. *IEEE Journal of Biomedical and Health Informatics*, (pp. 1–1). doi:doi:10.1109/JBHI.2017.2754861.

Tompson, J. J., Jain, A., LeCun, Y., & Bregler, C. (2014). Joint training of a convolutional network and a graphical model for human pose estimation. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 27 (NIPS)* (pp. 1799–1807). Curran Associates, Inc.

Toshev, A., & Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. In *2014 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1653–1660).

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 6450–6459).

Tu, Z., Xie, W., Qin, Q., Poppe, R., Veltkamp, R. C., Li, B., & Yuan, J. (2018). Multi-stream cnn: Learning representations based on human-related regions for action recognition. *Pattern Recognition*, *79*, 32–43.

Tulyakov, S., Alameda-Pineda, X., Ricci, E., Yin, L., Cohn, J. F., & Sebe, N. (2016). Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. In *2016 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2396–2404).

ACCEPTED MANUSCRIPT

Twomey, N., Diethe, T., Kull, M., Song, H., Camplani, M., Hannuna, S., Fafoutis, X., Zhu, N., Woznowski, P., Flach, P., & Craddock, I. (2016). The SPHERE challenge: Activity recognition with multimodal sensor data. *arXiv preprint arXiv:1603.00797*, .

Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M. J., Laptev, I., & Schmid, C. (2017). Learning from synthetic humans. In *2017 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4627–4635). IEEE.

Vaziri, D. D., Aal, K., Gschwind, Y. J., Delbaere, K., Weibert, A., Annegarn, J., de Rosario, H., Wieching, R., Randall, D., & Wulf, V. (2017). Analysis of effects and usage indicators for a ict-based fall prevention system in community dwelling older adults. *International Journal of Human-Computer Studies*, *106*, 10 – 25. doi:doi:10.1016/j.ijhcs.2017.05.004.

Viana, J., Ramalho, A., Valente, J., & Freitas, A. (2019). Ambient assisted living – a bibliometric analysis. In Á. Rocha, H. Adeli, L. P. Reis, & S. Costanzo (Eds.), *New Knowledge in Information Systems and Technologies* (pp. 218–228). Cham: Springer International Publishing.

Vlaeyen, E., Deschodt, M., Debard, G., Dejaeger, E., Boonen, S., Goedemé, T., Vanrumste, B., & Milisen, K. (2013). Fall incidents unraveled: a series of 26 video-based real-life fall events in three frail older persons. *BMC Geriatrics*, *13*, 103. doi:doi:10.1186/1471-2318-13-103.

Wang, P., Li, W., Gao, Z., Zhang, J., Tang, C., & Ogunbona, P. O. (2016). Action recognition from depth maps using deep convolutional neural networks. *IEEE Transactions on Human-Machine Systems*, *46*, 498–509. doi:doi:10.1109/THMS.2015.2504550.

Wang, X., Eliott, F. M., Ainooson, J., Palmer, J. H., & Kunda, M. (2017a). An object is worth six thousand pictures: The egocentric, manual, multi-image (emmi) dataset. In *ICCV Workshops* (pp. 2364–2372).

ACCEPTED MANUSCRIPT

Wang, X., Gao, L., Song, J., & Shen, H. (2017b). Beyond frame-level cnn: Saliency-aware 3-d cnn with lstm for video action recognition. *IEEE Signal Processing Letters*, *24*, 510–514. doi:doi:10.1109/LSP.2016.2611485.

Wei, S.-E., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional pose machines. In *2016 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4724–4732).

Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T. S., & Yan, S. (2010). Sparse representation for cmputer vision and pattern recognition. *IEEE*, *98*, 1031–1044. doi:doi:10.1109/JPROC.2010.2044470.

Wu, D., Sharma, N., & Blumenstein, M. (2017). Recent advances in video-based human action recognition using deep learning: A review. In *2017 International Joint Conference on Neural Networks (IJCNN)* (pp. 2865–2872). doi:doi:10.1109/IJCNN.2017.7966210.

Wu, H.-Y., Rubinstein, M., Shih, E., Guttag, J., Durand, F., & Freeman, W. (2012). Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics*, *31*, 65:1–65:8. doi:doi:10.1145/2185520.2185561.

Xiao, Y., Chen, J., Wang, Y., Cao, Z., Zhou, J. T., & Bai, X. (2019). Action recognition for depth video using multi-view dynamic images. *Information Sciences*, *480*, 287 – 304. doi:doi:https://doi.org/10.1016/j.ins.2018.12.050.

Xie, G., Zhang, X., Yan, S., & Liu, C. (2017). Hybrid cnn and dictionary-based models for scene recognition and domain adaptation. *IEEE Transactions on Circuits and Systems for Video Technology*, *27*, 1263–1274. doi:doi:10.1109/TCSVT.2015.2511543.

Yordanova, K., Krüger, F., & Kirste, T. (2018). Providing semantic annotation for the cmu grand challenge dataset. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)* (pp. 579–584). IEEE.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In *2015 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4694–4702).

Zhang, C., Tian, Y., Guo, X., & Liu, J. (2018). Daal: Deep activation-based attribute learning for action recognition in depth videos. *Computer Vision and Image Understanding*, *167*, 37–49. doi:doi:10.1016/j.cviu.2017.11.008.

Zhang, J., Li, W., Ogunbona, P. O., Wang, P., & Tang, C. (2016a). Rgb-d-based action recognition datasets: A survey. *Pattern Recognition*, *60*, 86 – 105. doi:doi:https://doi.org/10.1016/j.patcog.2016.05.019.

Zhang, Z., Conly, C., & Athitsos, V. (2015). A survey on vision-based fall detection. In *8th ACM international conference on Pervasive technologies related to assistive environments* (p. 46). ACM.

Zhang, Z., Girard, J. M., Wu, Y., Zhang, X., Liu, P., Ciftci, U., Canavan, S., Reale, M., Horowitz, A., Yang, H., Cohn, J. F., Ji, Q., & Yin, L. (2016b). Multimodal spontaneous emotion corpus for human behavior analysis. In *2016 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3438–3446). doi:doi:10.1109/CVPR.2016.374.

Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., & Xie, X. (2016). Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *Thirtieth AAAI Conference on Artificial Intelligence* AAAI'16 (pp. 3697–3703). AAAI Press.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

None

Signed, Pau Climent-Pérez (corresponding author):

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Credit Author Statement

Author	***Role(s)***
Pau Climent-Pérez (corresponding)	Writing – original draft; Investigation; Visualization
Susana Spinsante	Funding acquisition; Investigation; Writing – review & editing
Alex Mihailidis	Funding acquisition; Investigation; Writing – review & editing
Francisco Flórez-Revuelta	Funding acquisition; Project administration; Supervision; Investigation; Writing – review & editing

Possible Roles: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Roles/Writing - original draft; Writing - review & editing.

ACCEPTED MANUSCRIPT