An approach to provide dynamic, illustrative, video-based guidance within a goal-driven smart home

Joseph Rafferty[1], Chris Nugent[1], Jun Liu[1] and Liming Chen[2]

[1] School of Computing and Mathematics, University of Ulster
Email: {j.rafferty, cd.nugent, j.liu} @ulster.ac.uk

[2] School Computer Science and Informatics, De Montfort University
Email: liming.chen@dmu.ac.uk

Abstract.
The global population is aging in a never-before seen way, introducing an increasing ageing-related cognitive ailments, such as dementia. This aging is coupled with a reduction in the global support ratio, reducing the availability of formal and informal support and therefore capacity to care for those suffering these aging related ailments. Assistive Smart Homes (SH) are a promising form of technology enabling assistance with activities of daily living, providing support of suffers of cognitive ailments and increasing their independence and quality of life. Traditional SH systems have deficiencies that have been partially addressed by through goal-driven SH systems. Goal-driven SHs incorporate flexible activity models, goals, which partially address some of these issues. Goals may be combined to provide assistance with dynamic and variable activities. This paradigm-shift, however, introduces the need to provide dynamic assistance within such SHs. This study presents a novel approach to achieve this through video based content analysis and a mechanism to facilitate matching analysed videos to dynamic activities/goals. The mechanism behind this approach is detailed and followed by the presentation of an evaluation where showing promising results were shown.

Keywords: Smart home, Guidance, Video-processing, Metadata Generation, Activities of Daily Living, Ontology.

1 Introduction

The demographic composition of the global population is changing in an unprecedented manner. This change results in the effect of global aging (United Nations 2010; United Nations 2014). Global aging is where the median age of the global population is rising. Specifically, the global median age is predicted to increase to 38 by 2050, from its 2009 level of 28. This aging has been attributed to many factors; notably among them is a decrease in birth rates in conjunction with reduced levels of adult mortality (United Nations 2010; United Nations 2014). This issue is further compounded in regions which have experienced the baby boom which proceeded World War II, such as Europe and the USA (van de Kaa & Population Reference Bureau Inc. 1987; De Luca et al. 2011).

Such an uneven distribution of the population by age range is predicted to lead to the situation where by 2050 the world will have a demographic distribution where over 20% of the population is aged of 65 or over (United Nations 2010; De Luca et al. 2011). With such a demographic transition it is anticipated that there would be an increase in aging related illness, such as Dementia and Alzheimer's disease. Such an increase in aging related illness will, in turn, lead to an increased demand on healthcare services and informal support.

An ageing population increases demand on healthcare services due to the accompanying decline of physical and cognitive capabilities (McGinnis & Moore 2006; Lowthian et al. 2011). This additional demand is likely to be twinned with a downward trend in the worldwide Potential Support Ratio (PSR). The PSR represents the number of people that are within the working age (15-64) for each person over the age of 65. It is expected that the PSR will globally decrease from 11:1 in 1950, 9:1 in 2001 to 4:1 in 2050 (United Nations 2010).

The reduction in the support capacity of the population will lead to fewer formal and informal resources available for care of the increasing number of elders. As such, the quality of life of older people would, inevitably, be negatively impacted. Additionally, this situation would carry an increased burden on those who provide care in addition to those in need of care (United Nations 2010; United Nations 2014; van de Kaa & Population Reference Bureau Inc. 1987; European Commission n.d.).

Ambient Assistive Living (AAL) is a technological approach to enhancing the life of individuals by integrating assistive, information and communication based, technologies into its user's daily environment. Specifically, this pervasive form of technology when applied to the aging population enables users to remain socially connected, to stay active for longer, and live independently, longer, into old age (European Commission n.d.) .

The assistance that AAL affords can increase the independence of people who are cognitively impaired, such as that which accompanies many aging related illness. The automated assistance provided by AAL can reduce the amount of formal and informal resources needed to care for those suffering aging related impediments while simultaneously reducing the burden on health care providers.

Assistive Smart Homes (SH) represent a promising approach to providing AAL to sufferers of aging-related illnesses (De Luca et al. 2011; Chen, Hoey, et al. 2012; Hwang & Hoey 2012). SHs are residential environments which are augmented with technology enabling AAL to help care for, monitor and enhance the life of its inhabitants. The use of smart home technologies have shown promise in providing care for the aging section of our population (Cook & Das 2007; Chan et al. 2008).

SHs have been used to provide assistance with specific tasks, general Activities of Daily Living (ADL), monitoring inhabitant health signs, providing tele-presence and reminders. SHs providing assistance with ADLs has shown promise in providing assistance with the aging segment of the population, particularly when inhabitants suffer cognitive impairments, such as dementia. ADLs are the activities needed to maintain independence, such as clothing and feeding oneself. In this application, SHs typically monitor activity progresses, detects activities that an inhabitant is having trouble with and provides some guidance to bring this 'stalled' activity to completion.

Such SHs typically employ technology spanning three types of components to realise their functionality; sensing, reasoning and actuation. The reasoning component infers activity from the sensor data and attempts to assist with activities being performed through the use of actuation platforms (e.g. audio prompts delivered through speakers). The reasoning component is largely software based and its implementation greatly affects the performance and utility of SHs.

Typically, SHs provide functionality by leveraging a 'bottom-up' paradigm (De Luca et al. 2011; McGinnis & Moore 2006; Lowthian et al. 2011; Cook & Das 2007; Chan et al. 2008; Acampora et al. 2013), which operates as follows. Multi-modal sensors are deployed within smart homes to monitor and track changes that occur in the environment the inhabitant occupies. The assistive system can observe sensor interactions in conjunction with activity models and discover if an inhabitant is having problems with performing a specific activity. In the event where an inhabitant requires assistance, the assistive system will offer just-in-time, context-aware, help through prompts or actuators. The prompting systems in these environments include use of audio, text or video instruction (Acampora et al. 2013; Chan et al. 2008; Chen, Hoey, et al. 2012; Cook & Das 2007; Mihailidis et al. 2008). Of these types of prompts, video based instruction provides instruction that is both detailed and relatable. Previous studies have shown that video is the most desirable method to provide instruction for tasks such as ADLs (J. Rafferty et al. 2014; Mihailidis et al. 2008; O'Neill et al. 2010; Ballan et al. 2010).

Currently SHs providing such functionality, have a number of deficiencies related to handling variation in activity performance, providing reusable and flexible activity representations, providing dynamic illustrative guidance and some having a reliance on dense sensing (De Luca et al. 2011; McGinnis & Moore 2006; Lowthian et al. 2011; Cook & Das 2007; Chan et al. 2008; Acampora et al. 2013). In order to address the aforementioned issues, a paradigm shift from the current 'bottom-up' approach to a 'top-down', goal-driven, paradigm is necessary (Chen, Nugent, et al. 2012; Rafferty, Chen, et al. 2015).

A goal driven paradigm would instead flexibly model a goal, activities which have actions associated with it. These goals can be represented and combined in a manner that allows variance of goal performance. This dynamic variance of goal performance increases the utility of such a system but introduces issues with providing illustrative, video-based, guidance. To date, video instruction for activity guidance in SHs is provided in a relatively static manner. Specifically, video based instruction is statically mapped to activities or activity elements. Such a static assignment is not compatible with a goal-oriented approach to producing a SH (Acampora et al. 2013; Chan et al. 2008; Chen, Hoey, et al. 2012; Cook & Das 2007; Mihailidis et al. 2008).

In order to address this issue, this study introduces a novel approach of automatically generating metadata and matching this metadata, and therefore guiding video, to dynamically generated goals within a SH. Specifically, in this study, metadata is generated through a process of analysing audio narration and matching semantically compatible utterances of goal actions. These utterances are stored within an ontological store in conjunction with a unique video identifier, a SHA 512-bit hash. This metadata enables videos to be matched to profiles of goals in need of assistance which contain manifests of goal actions. These goal profiles are generated and determined by complementary components of an overall study which aims to realise a goal-driven SH (Rafferty, Chen, et al. 2015).

The remainder of this manuscript is organised as follows. Section 2 introduces related work across video-based metadata generation and matching video metadata to actions. Section 3 describes the novel approach introduced by this study and its place within a goal driven SH. Section 4 evaluates the approach to providing video based assistance. Section 5 details future work and provides a conclusion.

2 Related Work

Video based guidance within a goal driven SH requires a flexible mechanism to determine the most suitable candidate video from a repository for a given goal. In order to identify a suitable match from a repository, accurate and

comprehensive metadata is required. This metadata would contain a description of the actions present in instructional video clips. If exhaustively and completely supplied, such metadata can subsequently be used with a selection mechanism to nominate a suitable video clip.

2.1 Automatic Generation of Metadata for Video Content

Currently, metadata for such video clips are typically provided manually, with some computer vision and audio content efforts being investigated. The use of manually supplied metadata may, through its nature, lead to incomplete or incorrect records. Additionally, production of complete metadata that provides indication of all tasks that are depicted in a video would require a large amount of effort by video reviewers or the original author (Filippova & Hall 2011). Automatic generation of such metadata provides a method to gain consistent, accurate and complete information about the tasks within the depicted in a video.

Automatic video based metadata generation mainly involves analysis of object interactions within a scene, performing analysis on textual elements within videos or analysis of content using statistical analysis (Papadopoulos et al. 2013; Ballan et al. 2010; McCloskey & Davalos 2012; Hentschel et al. 2013; Greco et al. 2016; Gaüzère et al. 2015). Such computer vision based approaches show promising results but require a high amount of computation resources (Yang & Meinel 2014) and training with large data-sets to work efficiently. Additionally, these elements identified are relatively high level and do not support identification of a complex range of tasks as would be performed in goals and ADLs.

Automatic audio based metadata generation mostly focuses on sound analysis, with limited work incorporating Automated Speech Recognition systems (ASR) which convert speech to text (Maratea et al. 2013; Perea-Ortega et al. 2013).

Maratea *et al.* (Maratea et al. 2013) used ASR to provide metadata to provide the basis to search a set of lecturer videos. Speech in these videos were converted to text and automatically summarised. This summary was then used as the basis to provide a searchable corpus, associated with the video files.

Perea-Ortega *et al.*(Perea-Ortega et al. 2013) produced a mechanism of providing semantic labels for video files though use of machine learning. This approach classified documents taken from numerous sources on the web, including Wikipedia and blogs. These documents were classified with a single label which formed part of a multiclass model. This model was then applied to video transcripts produced by ASR. These transcripts where then associated labels from the model. This approach shows promising approach but is unsuitable for use in profiling videos for providing instruction in a goal-driven SH due to the limited application of labels provided.

These audio based approaches do not provide a suitable method of producing activity annotations, as they cannot currently identify a set of goal actions in such video clips nor store their results in a manner that allows matching videos to dynamic goals.

Hybrid computer vision and ASR based approaches have been investigated, of these the following are of note (Yang & Meinel 2014), (Metze et al. 2013).

Yang *et al.* (Yang & Meinel 2014) produced one such study focused on analysis of video based lectures. This combined optical character recognition and ASR to mine content of the lectures to provide a corpus associated with each video. This corpus provides the basis to index and search these videos. This approach does not detect actions depicted within videos and does not provide a suitable data structure that would allow dynamic matching of video based guidance.

Metze *et. al.* (Metze et al. 2013) combines computer vision with ASR and Natural Language Processing (NLP) to produce a machine learning based approach to identify high level concepts, objects depicted in a video, audio signatures and overall topic concepts. This was evaluated against the TRECVID dataset and found that the use of ASR provided the most useful information to reason upon. This approach was able to identify activities such as *"cleaning_an_appliance"*. This is a relatively high level description and was produced through analysis of a large amount of factors and use of a richly annotated training set. This required a large computational effort and a well-annotated pre-existing dataset and so would be unsuitable for use in a goal-driven SH.

The majority of these current approaches to metadata generation for video content requires the specific mechanism to be trained with a dataset beforehand, resulting in a cold start problem (Papadopoulos et al. 2013; Ballan et al. 2010; McCloskey & Davalos 2012; Perea-Ortega et al. 2013). Additionally, computer vision techniques require a significant amount of computation over ASR based analysis, due to the nature of the analysis involved and the greater complexity of compressed video over compressed audio. As the videos to be presented by the goal driven SH to be produced in by this body of work will have clear, concise narration it would be more appropriate to focus on analysis of the audio aspect of such video instruction. This focus on the audio components is the core of the metadata generation technique detailed in Section 3.1.

2.2 Automatic Matching of Video Metadata to Activities

Video based metadata matching approaches have been produced by a number of previous studies (Veltkamp et al. 2013; Chatterjee & Leuski 2015; Choe et al. 2013; Mazloom et al. 2013; Matejka et al. 2014; Jones & Shao 2013; Shabani et al. 2013; Patel & Meshram 2012; Panchal et al. 2012). Some of which are explored in the following paragraphs.

Chatterjee *et al* (Chatterjee & Leuski 2015) used active learning techniques to rank candidate videos based upon queries. Active learning increases the performance of an approach by incorporating human opinion and feedback. This active learning technique incorporates statistical metrics such as Sample Uncertainty, Density and Diversity. This works requires the exhaustively annotated TRECVID and USC SmartBody video sets to learn from and perform queries. This is a limitation of the approach and reduces its use within goal-driven SHs due to the unrealistic levels of annotation that are provided within this dataset. Additionally, this dataset does not have a focus on ADLs/IADLs or inhabitant goals.

Choe *et al* (Choe et al. 2013) proposed an approach to facilitate querying grouped events from a large video database. Unlike previous approaches, such as matching and nomination upon a search for similar scenes, this approach proposes matching based on semantic events. In this study, these semantic events were essentially actions, such as parking a vehicle or exchanging a package. This approach was based upon using subgraph matching and topic learning algorithms to form search based upon vector comparison, reducing computational complexity for this search process once indexing has been performed. This proposed approach has a number of caveats, these are related to high computational complexity during indexing and availability of a suitable training set. During initial video indexing categorisation of actions and events required scene element extraction using super/sub pixel analysis in conjunction colour segmentations applied to a Markov Random Field and Swanson Cut based machine learning system. Once indexed search using graph structures, converted to vectors, proved to be an efficient mechanism to query the mix of actions and events detected through this visual information. This work leverages TRECVID introducing the limitations covered previously.

Jones *et al* (Jones & Shao 2013) produced a mechanism to retrieve content depicting human actions from realistic video databases. This approach intended to address limitations of querying video databases by incorporating the ability to index and search the content of video, opposed to relying on user provided metadata or text-based context. Specifically, this study focused on improving upon contemporary works that retrieving videos that depict human actions by reducing their requirement for very clean and well annotated data. These activities are coarse grained and

per scene, such as horse riding or playing tennis. This approach uses metadata generated by scene detection using a number of computer vision techniques. This metadata is then incorporated into a bag-of-words and is queries through a vocabulary-guided pyramid match and spatio-temoporal pyramid match. This approach uses active learning to further apply weight to these pyramids, though a number of SVM and naïve-Bayes based algorithms. This approach relies extensive computer vision techniques, active learning and only has low resolution activity definitions reducing scope to used its query mechanism in a goal driven SH.

These previous works, while promising, are not been able to match video metadata about actions to activities of daily living or goals. As such, there was the need to produce a novel mechanism to do this.

3 A Mechanism to Provide Dynamic, Illustrative, Video-Based Guidance in a Goal-Driven Smart Home

In this study, an annotation method capable of generating rich metadata for video clips and subsequently matching this metadata to goal descriptions has been devised, implemented and tested. This was implemented in a platform called ABSEIL (Audio BaSEd Instruction profiLer) which is in turn a component of a larger assistive SH system, named INSigHt (Inteligent ageNt Smart Home). In order to provide dynamic, illustrative video-based instructions a two stage mechanism is required. Stage one of this mechanism involves automatic production of metadata form candidate video clips, a novel process to achieve this is shown in Section 3.1. Stage two involves using this metadata to match video files to a specific goal, this is shown in Section 3.2. The mechanisms presented in this study are an extension of previous work by the authors (J. Rafferty, Nugent, et al. 2015). This previous work has been modified to provide assistance within a goal driven SH within this study and is been evaluated in Section 4 with a focus on use within a SH.

3.1 A mechanism to automatically generate metadata for video files

This novel metadata generation process is designed to work in conjunction with the video repository. Specifically, the video repository generated by the Personal IADL Assistant (PIA) project has been chosen with access negotiated. PIA was an EU AAL funded project which provided assistance with Instructional Activities of Daily Living (IADL) (Lawton & Brody 1988) through providing narrated instructional videos, in an on-demand manner (J. Rafferty et al. 2014). As such, the PIA repository provides a source of high quality, narrated, video-based instruction, suitable for guiding inhabitants of a smart home.

In the devised approach, videos are converted to audio clips. The audio clips are then sent to a black-box ASR which returns a transcription. This transcription is processed to identify goal actions extracted from the INSigHt goal repository in order to determine if any of these actions present within the narration. Due to individual personalisation, terms used in narration and modelling of the goal actions may differ in specific terms used but still refer to semantically compatible variations. As such, compatible, alternative, variations of these actions are catered for using a semantic lexicon and homophone dictionary. A depiction of this method is presented in Figure 1 and explained in greater detail in the proceeding paragraphs in this section.

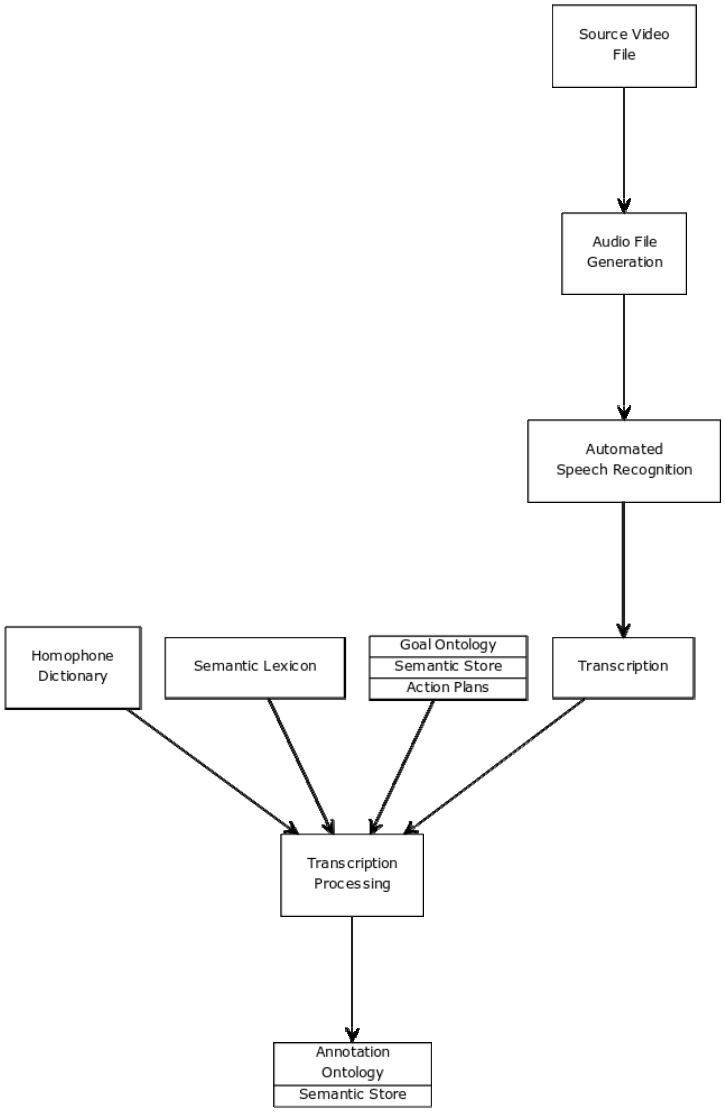

Fig 1. A depiction of the automatic metadata generation process devised within this study.

ABSEIL uses software to transcode a candidate video file into a set of audio files. Once the audio files have been produced, a version is then streamed to the Google Speech API (GSAPI) (Google n.d.), an ASR system. GSAPI was chosen over alternative, available, ASR solutions (Apple, Dragon, Windows Speech) due to its better performance during evaluation of systems for this application (Joseph Rafferty et al. 2014; Joseph Rafferty, Nugent, et al. 2015).

In this approach, a repository of ontologically modelled goals for the INSigHt SH exists. This repository may contain inhabitant goals such as *MakeTea,* which may also reference sub-goals, such as *GetMug*. Goals have an associated action plan which contains a number of atomic actions, these are steps required to achieve the goal, such as *OpenTea*. This goal ontology is presented as a hierarchy of concepts in Figure 2.

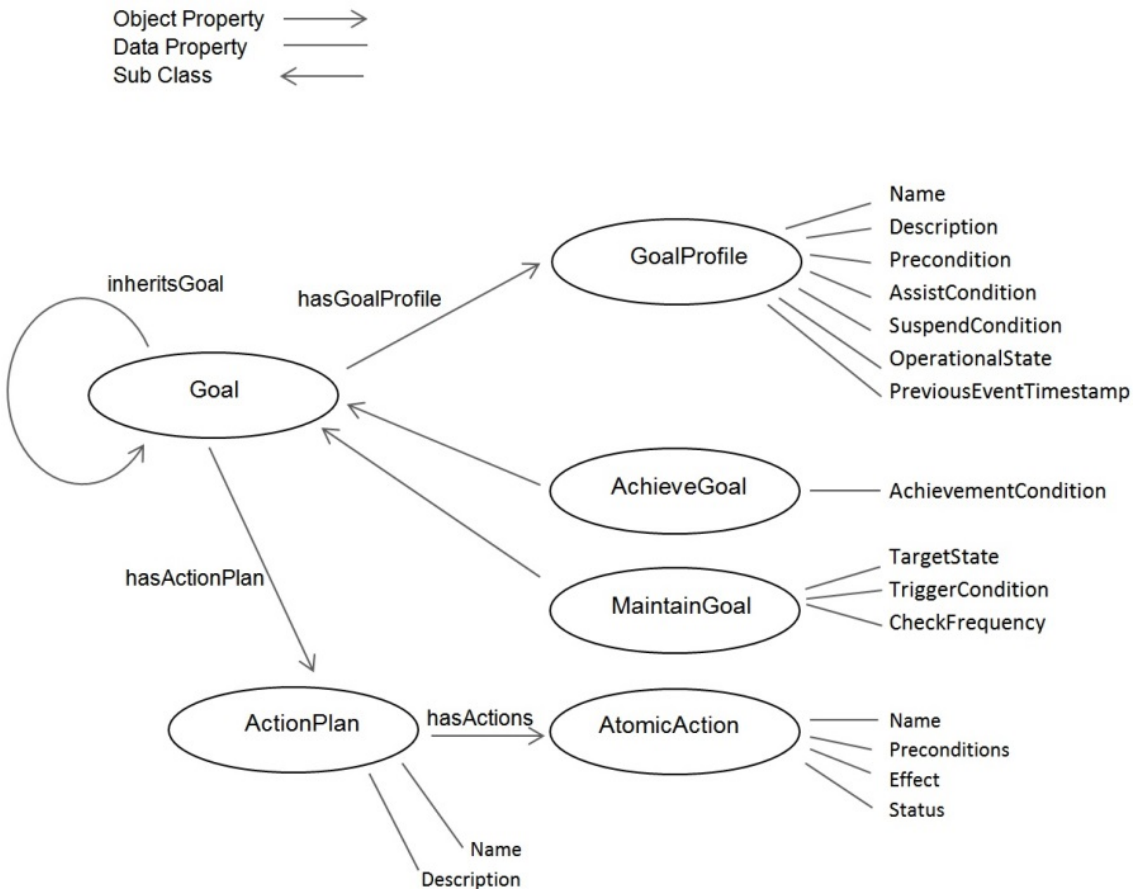

Fig. 2. The classes, object properties and data properties of the proposed goal ontology.

Actions in the goal repository are used to search transcriptions of video files. These atomic actions are used to generate a four search sets, which are applied to the transcription; these sets are *{direct | homophone substitution | synonym substitution | homophone/synonym substitution}*.

Contemporary ASR systems are black-box mechanisms which are based upon machine learning. As such, they are unable to offer insight into, or influence over, their transcription process. Such hidden complexity introduces issues when nuance involving homophones and synonyms is required.

Homophones are words which may be phonetically confused, such as 'ate' and 'eight'. Contemporary works within ASR correct for these homophones within the black-box core of the ASR process (Chen & Ananthakrishnan 2013; Mehla & Aggarwal 2014). As such, an alternative method of substitution needs be implemented. In this approach, substitution produces combinations of words from a pre-existing homophone dictionary (SIL n.d.). These combinations are placed in a *homophone substitution* search set. For example, a subset generated from the atomic action *PourFlour* could contain {*pore flour | pour flour | poor flour | pore flower | pour flower | poor flower*}.

Additionally, exact words used in modelling actions are specified by whoever modelled that goal and so are subject to personalisation. This introduces an issue when terms used in narration use alternative, semantically similar, words, known as synonyms, to those expressed in the modelled atomic actions. An example of this is as follows, in place of the word '*cup'* some of its synonyms could be used, some of these are *{cup| mug| teacup}.* This can cause issues when narration and the modelled atomic actions differ. For example, an utterance of "Place the tea in the mug" in the narration of a video could be used to describe an action which is compatible with the *PlaceTeaInCup* atomic action.

To cater for this variance, a semantic knowledgebase was incorporated to produce a *synonym substitution* search sets. In this evaluation platform, the chosen semantic lexicon was WordNet 3.1 (Princeton University 2010). WordNet is maintained by Princeton University and is one of the foremost lexical databases for the English language.

In order to correct for instances where both homophone errors and use of semantically compatible words occur, it is necessary to create a search set that contains these combinations, the *homophone/synonym substitution* search set. The combinations in this set provide useful alternatives to atomic actions that are present in such narrated videos. Once the search sets have been generated they are used to search the body of the transcription to identify utterance of their terms. During the search process, terms in search sets are given a four-word window between them. This allows atomic actions such as *TakeCup* to be found when alternative variations are uttered, for example "Take the closest cup".

When search set terms are discovered within a transcription they are stored in a video action ontology. In addition to search set terms, metadata about a video is stored; a unique identifier, an optional title and an optional description. The unique identifier is a 512-bit hash digest of the original video file, facilitating a means to associate the metadata to the file without recording an association. By default, the optional video title and descriptions are retrieved from the source repository, when possible. The video action annotation ontology contains four classes to hold matched terms from the four search sets. These are the *DirectTerms, HomophoneTerms, SynonymTerms* and *HomophoneSynonymTerms* classes. All these classes contain the same set of data properties: *DepictedAction, TimeStamp and Duration*. The video action annotation ontology is depicted as a hierarchy of concepts in Figure 3.

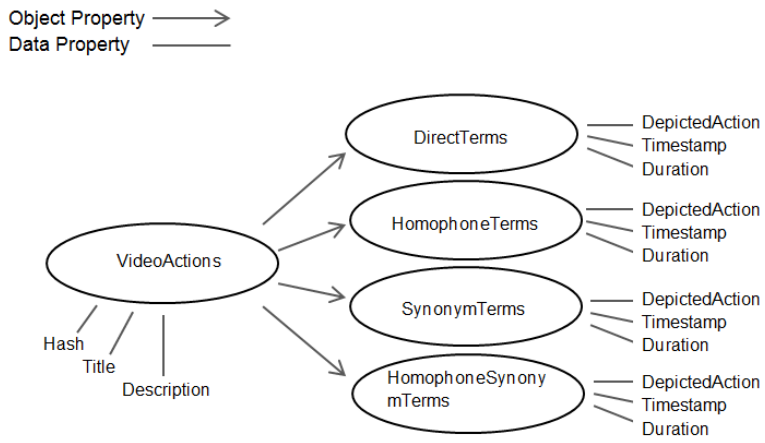

Fig. 3. A hierarchy of concepts showing the classes and object/data properties of the video action annotation ontology.

Once produced, this metadata may be used to provide assistance for given inhabitant goals. In order to achieve this a mechanism for matching videos to the needs of inhabitant goals need to be incorporated. A novel mechanism to achieve this has been produced and is present in Section 3.2, below.

3.2 A mechanism to automatically match video files to nominated inhabitant goals

In this study, a mechanism capable of nominating suitable video clips to provide relevant assistance for SH inhabitant goals and IADLs has been produced. This method is intended to select appropriate instructional videos for dynamically constructed inhabitant goals and IADLs, as specified by a goal-driven SH.

This approach uses semantic rules and queries to reasoning across a number of ontologies, including the goal library of a goal-based SH system and the video action annotation ontology previously detailed in Section 3.1. The goal ontology of this SH records the goal names and atomic actions required for nominating appropriate video clips.

This mechanism was integrated into the ABSEIL system and was made accessible through a REST based web service endpoint. This allows consuming applications (in this case, INSigHt) to issue a request to a query service.

In the case of the SH system, it can issue a query in the form of a goal profile or a specific goal name. A goal profile contains a manifest of actions that an instructive video would be matched to. These actions will be matched to goals that are present in the INSigHt goal ontology. A goal profile based request may not necessarily contain actions that are present within action plans stored in the INSigHt goal ontology, further information about this goal ontology and its lifecycle can be obtained in (Rafferty, Chen, et al. 2015).

Typically, a goal name request will be from a goal that is present in the INSigHt goal ontology. This type of request will access the INSigHt goal ontology to obtain or generate a goal profile. In the case of generation of a goal profile from a given name, semantic reasoning based atomic action is performed through the rules shown in (1) and (2). This generated profile is then used to match metadata as shown in this chapter and therefore obtain video-based instruction.

These queries are then processed against a list of goal names from the semantic store, using the Lucene-based technique detailed in (Joseph Rafferty et al. 2014), to determine if any were uttered. If any goal matches are identified, the goal name is used in the video nomination process used directly by the SH.

Once the nomination process has been completed, it will respond to the client applications. If a video is identified, its SHA512 hash is returned to the client. The client can then use that hash within a URI to obtain the video from a cloud-based repository via a HTTP GET command. In the event that a suitable match is not nominated, a null value will be returned to the client, the client will then handle this as appropriately.

The video nomination process takes a goal name and then applies semantic reasoning, in combination with some traditional programming in order to determine the most appropriate video from a number of candidates, if any suitable candidate is available.

The reasoning behind this video nomination mechanism was implemented through a range of semantic technologies, specifically a combination of SWRL and SPARQL. These semantic technologies are used in conjunction traditional logic programming, where appropriate. The SPARQL and SWRL rules used in this process shown in (1) and (2). The logic behind this process is formally depicted in Table I.

SWRL rule, shown in (1), reasons about goal inheritance, facilitating querying actions from inherited goals. This enables generation of complete action plans from dynamically created activity structures, Goals. In this SWRL rule, *?g* represents a goal under consideration, *?gp* represents the goal profile of *?g*. Goal profiles contain the goal name and action plan. *?g* inherits goals as though the *inheritsGoal* property. Inherited goals are represented by *?ig* and the profiles of inherited goals are represented as *?igp*. these inherited goals, and their action plans are then applied to the original goal though asserting the *hasGoalProfile* property.

```
hasGoalProfile(?g, ?gp),
hasGoalProfile(?ig, ?igp),
inheritsGoal(?g, ?ig),                                          (1)
Name(?gp, ?gn)
    -> hasGoalProfile(?g, ?igp)
```

SPARQL query (2) combines with SWRL rule (1) to extract a list of actions from a goal ontology. This will produce a list of actions for a given goal, as represented by *?gn* in this query

```
SELECT ?acts
  WHERE {
     {                                                          (2)
        ?p ug:Name ?gn.
        ?goal ug:hasGoalProfile ?p.
```

```
        ?goal ug:hasGoalProfile ?pn.
        ?allGoals ug:hasGoalProfile ?pn.
    }.{
        ?allGoals ug:hasActionPlan ?ap.
        ?ap ug:hasActions ?aa .
        ?aa ug:Name ?acts
    }
}
```

Table I. Logical rules that form the core reasoning in the nomination mechanism.

Rule	Description
G	A goal (G) that an inhabitant may pursue
$G = \langle A \rangle$	Goals have actions (A) that need to be performed to complete the goal tasks
$G = \{G\boldsymbol{x}\ ;\ x = 1, \ldots, N\}$	A set of goals (G_x) a inhabitant may pursue
$MD = \{MD\boldsymbol{x}\ ;\ x = 1, \ldots, N\}$	A set of metadata generated by the ABSEIL system. These are DM, SM, HM and SHM; representing the direct, synonym, homophone and synonym/homophone sets.
$MD = \langle A, H \rangle$	Generated metadata sets have an associated a set of actions (A) and a hash record (H). H is used to identify videos within a repository.
$\boldsymbol{CMD} = \{\boldsymbol{x} \in (\boldsymbol{A_{DM_x}} \cap \boldsymbol{A_{SH_x}} \cap \boldsymbol{A_{HM_x}} \cap \boldsymbol{A_{SHM_x}});\ \boldsymbol{x} \in \boldsymbol{A_{G_x}}\}$	The CSM, DM, CHM and CSHM sets are transformed into to a candidate metadata set. This set only contains metadata entries that contain all the actions present in the nominated goal.
$\boldsymbol{NV} = \{\boldsymbol{x} \in \boldsymbol{CMD_x} \Leftrightarrow \boldsymbol{x} = \boldsymbol{min}\{\boldsymbol{A_{CMD_x}}\}\}$	NV is the metadata record of the nominated video. The nominated video is the metadata record in the CMD set that has the lowest number of actions. The hash record of NV is returned to a client, allowing location and playback of the video.

In addition, measures need to be taken to handle the case of a request for instruction made by the goal driven SH, where there are no suitable videos available. In this particular scenario, two different *null* responses need to be catered for. These *null* responses are if there are no suitable videos or no matching goal in the repository.

If there are no goals in the goal ontology that match the request, then a completely *null* response will be sent. This will contain no video identifier or action plan. Typically, this should never normally be the case as the goal request sent by an SH would generally be present in the goal ontology. However, this condition may be encountered when a spurious SH request is sent to the REST endpoint presented by the extension made to the ABSEIL system.

In the event a goal name is presented by a SH system and there are no suitable videos are matched but a goal profile may be generated, then a partial *null* response is sent. In a partial *null* response, the video flag in the response will be sent to an empty string but an additional action plan will be returned to the SH client. This action plan may then be used by the SH client to display stepwise instruction, or alternatively, use text-to-speech to synthesize audio guidance that guides an SH inhabitant. This is not catered for in the current iteration of the developed goal driven SH.

4 Evaluation

To evaluate the performance of this automatic annotation generation method it was implemented in the form of the ABSEIL system which was integrated as a component of the INSigHt SH system. Due to the two stage function of this system it was necessary to evaluate both of these elements. Section 4.1 evaluates the metadata generation process and Section 4.2 provides an evaluation of the metadata matching process. These components were implemented in a Java based service which incorporated the Pellet reasoner, which used Apache Jena to incorporate into a Virtuoso-based semantic/ontological store.

4.1 An evaluation of the metadata generation process

During evaluation, eighteen instructional videos were evaluated. Eight of which were from the PIA project and ten were narrated instructional videos following the guidelines used within the PIA project.

These videos contained a narrated depiction of instruction on how to perform various ADLs. These range of ADLs depicted included making coffee, making tea, how to take medications, setting a timer, use of a DVD player, use of a television, adjusting a climate control system and use of a coffee maker. Narration was performed in a clear, detailed and succinct manner, in order to relay the highest quality guidance for the target audience. This target audience is primarily individuals suffering aging related mild cognitive impairment.

In this evaluation process, a set of manually generated metadata was produced for these videos using the atomic action set extracted from a coffee making goal, and therefore its sub goals, as a reference for target terms. These terms were extracted through the use of APIs and tools produced for the INSigHt SH system.

Videos were subsequently analysed with the ABSEIL platform and where automatically generated metadata was compared to the manually generated metadata. The results of this evaluation are presented in Table II. Each of the four search sets were assessed and averaged, with erroneously identified actions being noted as false positives.

The set of narrated videos covered a range of beverage making tasks. The set of videos taken from the PIA project contained a single video that involved the steps of making coffee. This single video was used to evaluate incorrect profiling of video clips, where metadata would be generated without any relevant content.

Table II. The results of evaluating the accuracy of metadata produced by the automatic metadata method compared to manually generated metadata. False positives percentages are indicated in brackets.

Video source	Accuracy of generated annotation			
	Direct Terms	**Homophone Terms**	**Synonym Terms**	**HomophoneSynonym Terms**
PIA (8 videos)	87.5 % (0)	87.5 % (0)	87.5 % (0)	87.5 % (0)
Independently Narrated (10 videos)	66.86% (0)	73.97 % (0)	79.93% (2.9)	82.59 % (2.7)

As shown in this evaluation, the devised method shows promise for use in automatically generating metadata to enable guidance provision for a goal driven SH.

On occasion, semantically compatible words were used in place of those specified in the action plan and were not discovered by the processing. To address this deficiency, an additional source of synonyms may be incorporated into the system.

The false positives that arose were all nonsensical phrases. These typically incorporated additional prepositions or unusual word combinations, such as generation of the phrase "*Impoverished Water In Vessel*" generated from the atomic action *PourWaterInCup*. However, such nonsensical phrases should never be present in videos that are analysed as these videos are intended to provide clear instruction.

Additional issues were encountered when a narrator referred to a previous object as "it". Such utterances require modelling context and nuance within language and can be handled by the incorporation of a more advanced NLP toolkit such as the Natural Language Toolkit (NLTK Project n.d.).

4.2 An evaluation of the goal to video matching process

The video to goal matching component of this approach, this was implemented within a component of the ABSEIL system. This component was accessible via a REST-based web service, facilitating testing of this mechanism from two different perspectives; a goal nomination tool that emulates an assistance request from a goal-driven SH and assistance requests generated from a goal-driven SH. The requests generated by the goal-driven SH were dynamically generated though monitoring real world sensor-object interactions during inhabitant performance of goals. These sensor-object interactions were subsequently processed using an intention recognition process. During the performance of these goals a number were interrupted and so the intention recognition process nominated them for assistance.

Prior to the evaluation of the matching component, 20 videos were profiled and metadata was generated. This metadata was used as the basis video to goal matching in this evaluation. Additionally, the goal repository used in this evaluation incorporated a range of goals covering a variety IADLs, all bar two of these were represented in at least one of the videos previously processed by ABSEIL.

During this evaluation, 58 total requests were sent to the nomination mechanism using the voice application, goal nomination tool and requests generated by the goal-driven SH.

A set of 38 requests were expected to produce a positive result where a match was found. This match was compared to manual assessment to determine if it was the most suitable video for that goal. This set is referred to as set I.

A set of 20 requests were expected to produce a negative result where a match should not be found. These requests nominated goals that should have no matching video instruction. This is referred to as set II. A summarization of this evaluation is presented in Table III.

Table III. The accuracy of the video nomination process as determined by queries issued by a goal driven SH and a goal nomination tool. Set I are queries that are expected to produce a reply containing a video. Set II are queries that are expected to produce a null reply.

		Goal nomination tool	**Goals nominated from Goal-Driven SH**
Set I	Request Count	30	8
	Correct Response Count	25	8
	Accuracy	83%	100%
Set II	Request Count	10	10
	Correct Response Count	10	10
	Accuracy	100%	100%

In this evaluation a number of incorrect responses were received which were related to wrong nomination and incorrect production of transcriptions. Both the goal nomination tool and mobile application were susceptible to wrongly nominated videos when two specific goals were requested. This was due to nomination of a large, mostly unrelated, video, which held some utterances of goal activity as part of a lengthier instructional video. These actions were depicted within the video but were a small element of the instruction being shown. This wrongful nomination could lead to confusion if the person in need of assistance did not watch the entire video. A future revision of this nomination mechanism will make matching more conservative in order to factor out videos with too many unrelated actions.

Finally, all goals that were nominated by the goal-driven SH had a high success rate. This success is due to the detection of tasks that were modelled across all semantic stores and were depicted in videos processed by ABSEIL. Errors could be introduced where modelled goals have not been depicted in videos produced by ABSEIL.

This evaluation shows that this approach has promise in matching relevant instructional videos to dynamically produced goal profiles, such as those generated within a goal driven SH. Specifically, user's interaction with sensorised objects may be monitored by the intention recognition core of a goal driven SH. This intention recognition process can identify a dynamic goal that an inhabitant is performing. If an identified goal encounters some difficulty, such as a stall, a goal profile containing steps needed to complete the goal may be produced. The presented approach can match these goal profiles to suitable videos. These videos can subsequently be presented to end users in order to guide their completion of the task.

5 Conclusion and Future Work

This study presents a novel approach to providing illustrative, video-based, assistance within a goal-driven SH. Specifically, this approach analyses the narration of video files to identify actions that are present within the activity repository of goal-driven SH. Once identified, these video files can be matched to goal profile or goal names through use of semantic web technologies, specifically; ontologies, reasoners and rules in conjunction with traditional programming logic.

The development and evaluation of this approach was separated into elements, testing each stage of the workflow of this approach. Stage one profiles videos and stage two matches these video to goals through use of the metadata generated in stage one.

Through evaluation it is evident that the overall approach shows promise, with good results being generated in both the video profiling and metadata generation and the video to goal matching aspects of this approach. It should be noted that this system is limited by requiring narration being to be present in videos.

Future work would investigate the utility of this system across a larger range of videos and activities. In order to increase accuracy of the metadata generation process, incorporation of a NLP toolkit should be explored. This may allow greater identification of activities depicted within videos, especially when object references are being made in context. Additionally, a more accurate ASR system could be identified and integrated in order to better profile videos.

Additionally, a subsequent study will integrate this system into a goal driven SH which will be deployed to a test environment. During this testing accuracy of goal recognition and matching of goals in need of assistance to suitable videos will be evaluated.

Finally, this approach could provide utility within a range of application beyond that of a goal-driven SH system, such as use within on-demand assistive systems.

6 References

AAcampora, G. et al., 2013. A Survey on Ambient Intelligence in Health Care. *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers*, 101(12), pp.2470–2494. Available at:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3890262&tool=pmcentrez&rendertype=abstract.

Ballan, L. et al., 2010. Event detection and recognition for semantic annotation of video. *Multimedia Tools and Applications*, 51(1), pp.279–302. Available at: http://link.springer.com/10.1007/s11042-010-0643-7 [Accessed June 2, 2014].

Chan, M. et al., 2008. A review of smart homes- present state and future challenges. *Computer methods and programs in biomedicine*, 91(1), pp.55–81. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18367286 [Accessed July 20, 2012].

Chatterjee, M. & Leuski, A., 2015. A Novel Statistical Approach for Image and Video Retrieval and Its Adaption for Active Learning. In *Proceedings of the 23rd ACM International Conference on Multimedia*. Brisbane: ACM, pp. 935–938.

Chen, L., Hoey, J., et al., 2012. Sensor-Based Activity Recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, pp.1–19. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6208895.

Chen, L., Nugent, C.D. & Wang, H., 2012. A Knowledge-Driven Approach to Activity Recognition in Smart Homes. *IEEE Transactions on Knowledge and Data Engineering*, 24(6), pp.961–974. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5710936.

Chen, W. & Ananthakrishnan, S., 2013. ASR error detection in a conversational spoken language translation system. *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp.7418 – 7422. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6639104 [Accessed May 28, 2014].

Choe, T.E. et al., 2013. Semantic video-to-video search using sub-graph grouping and matching. *Proceedings of the IEEE International Conference on Computer Vision*, (1), pp.787–794.

Cook, D.J. & Das, S.K., 2007. How smart are our environments? An updated look at the state of the art. *Pervasive and Mobile Computing*, 3(2), pp.53–73. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1574119206000642 [Accessed October 5, 2012].

European Commission, The Ambient Assisted Living (AAL) Joint Programme. Available at: http://ec.europa.eu/information_society/activities/einclusion/docs/ageing/aal_overview.pdf.

Filippova, K. & Hall, K., 2011. Improved video categorization from text metadata and user comments. *SIGIR '11 Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pp.835–842. Available at: http://dl.acm.org/citation.cfm?id=2010028 [Accessed June 10, 2014].

Gaüzère, B. et al., 2015. Semantic web technologies for object tracking and video analytics. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9475, pp.574–585. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84952801857&partnerID=40&md5=36ddc1d67dedd953b3f41a0abf4fcf1c.

Google, Google Speech API. Available at: http://www.google.com/speech-api/v1/recognize.

Greco, L. et al., 2016. Abnormal Event Recognition: A Hybrid Approach Using Semantic Web. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. pp. 58–65.

Hentschel, C., Blümel, I. & Sack, H., 2013. Automatic Annotation of Scientific Video Material based on Visual Concept Detection. *Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies - i-Know '13*, pp.1–8. Available at: http://dl.acm.org/citation.cfm?doid=2494188.2494213.

Hwang, A. & Hoey, J., 2012. Smart home, the next generation: Closing the gap between users and technology. *AAAI Fall Symposium on Gerontechnology. Arlington, VA*, pp.14–21. Available at: http://www.aaai.org/ocs/index.php/FSS/FSS12/paper/viewPDFInterstitial/5549/5784 [Accessed March 6, 2014].

Jones, S. & Shao, L., 2013. Content-based retrieval of human actions from realistic video databases. *Information Sciences*, 236, pp.56–65. Available at: http://dx.doi.org/10.1016/j.ins.2013.02.018.

van de Kaa, D.J. & Population Reference Bureau Inc., W.D.C., 1987. *Europe's Second Demographic Transition*, Distributed by ERIC Clearinghouse, [Washington, D.C.] :

Lawton, M. & Brody, E., 1988. Instrumental Activities of Daily Living Scale (IADL). Available at: https://www.abramsoncenter.org/PRI/documents/IADL.pdf [Accessed March 5, 2014].

Lowthian, J. a et al., 2011. The challenges of population ageing: accelerating demand for emergency ambulance services by older patients, 1995-2015. *The Medical journal of Australia*, 194(11), pp.574–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21644869.

De Luca, d'Alessandro E., Bonacci, S. & Giraldi, G., 2011. Aging populations: the health and quality of life of the elderly. *La*

Clinica Terapeutica, 162(1), p.e13.

Maratea, A., Petrosino, A. & Manzo, M., 2013. Generation of description metadata for video files. *Proceedings of the 14th International Conference on Computer Systems and Technologies - CompSysTech '13*, 767, pp.262–269. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84889596966&partnerID=tZOtx3y1.

Matejka, J., Grossman, T. & Fitzmaurice, G., 2014. Video Lens: Rapid Playback and Exploration of Large Video Collections and Associated Metadata. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*. UIST '14. New York, NY, USA: ACM, pp. 541–550. Available at: http://doi.acm.org/10.1145/2642918.2647366.

Mazloom, M. et al., 2013. Querying for Video Events by Semantic Signatures from Few Examples. In *Proceedings of the 21st ACM International Conference on Multimedia*. pp. 609–612.

McCloskey, S. & Davalos, P., 2012. Activity detection in the wild using video metadata. *Pattern Recognition (ICPR)*, pp.3140–3143. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6460830 [Accessed June 2, 2014].

McGinnis, S. & Moore, J., 2006. The impact of the aging population on the health workforce in the United States: Summary of key findings. *Cahiers de sociologie et de démographie …*, (March 2006). Available at: http://cat.inist.fr/?aModele=afficheN&cpsidt=17965740 [Accessed May 20, 2013].

Mehla, R. & Aggarwal, R., 2014. Automatic Speech Recognition: A Survey. *International Journal of Advanced Research in Computer Science and Electronics Engineering*, 3(1), pp.45–53. Available at: http://www.ijarcsee.org/index.php/IJARCSEE/article/view/440 [Accessed May 28, 2014].

Metze, F. et al., 2013. Beyond audio and video retrieval: topic-oriented multimedia summarization. *International Journal of Multimedia Information Retrieval*, 2(2), pp.131–144. Available at: http://link.springer.com/10.1007/s13735-012-0028-y [Accessed November 13, 2013].

Mihailidis, A. et al., 2008. The COACH prompting system to assist older adults with dementia through handwashing: an efficacy study. *BMC geriatrics*, 8, p.28. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2588599&tool=pmcentrez&rendertype=abstract [Accessed November 6, 2012].

NLTK Project, Natural Language Toolkit. Available at: http://www.nltk.org/.

O'Neill, S. a. et al., 2010. Video Reminders as Cognitive Prosthetics for People with Dementia. *Ageing International*, 36(2), pp.267–282. Available at: http://link.springer.com/10.1007/s12126-010-9089-5 [Accessed March 6, 2014].

Panchal, P., Merchant, S. & Patel, N., 2012. Scene detection and retrieval of video using motion vector and occurrence rate of shot boundaries. In *2012 Nirma University International Conference on Engineering (NUiCONE)*. pp. 1–6.

Papadopoulos, D.P. et al., 2013. Automatic summarization and annotation of videos with lack of metadata information. *Expert Systems with Applications*, 40(14), pp.5765–5778. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0957417413001322.

Patel, B. V & Meshram, B.B., 2012. Content Based Video Retrieval Systems. *International Journal of UbiComp*, 3(2), pp.13–30.

Perea-Ortega, J.M. et al., 2013. Semantic tagging of video ASR transcripts using the web as a source of knowledge. *Computer Standards & Interfaces*, 35(5), pp.519–528. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0920548912000888 [Accessed June 2, 2014].

Princeton University, 2010. About WordNet. *WordNet. Princeton University*. Available at: http://wordnet.princeton.edu.

Rafferty, J., Nugent, C., et al., 2015. *A mechanism for nominating video clips to provide assistance for instrumental activities of daily living*,

Rafferty, J., Nugent, C., et al., 2015. Automatic Metadata Generation Through Analysis of Narration Within Instructional Videos. *Journal of Medical Systems*, 39(9), pp.1–7. Available at: http://dx.doi.org/10.1007/s10916-015-0295-2.

Rafferty, J. et al., 2014. Automatic Summarization of Activities Depicted in Instructional Videos by Use of Speech Analysis. In L. Pecchia et al., eds. *Ambient Assisted Living and Daily Activities*. Lecture Notes in Computer Science. Springer International Publishing, pp. 123–130. Available at: http://dx.doi.org/10.1007/978-3-319-13105-4_20.

Rafferty, J., Chen, L., et al., 2015. Goal Lifecycles and Ontological Models for Intention Based Assistive Living within Smart Environments. *Computer Systems Science and Engineering*, 30(1), pp.7–18.

Rafferty, J. et al., 2014. NFC based provisioning of instructional videos to assist with instrumental activities of daily living. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014*. pp. 4131–4134.

Shabani, A.H., Zelek, J.S. & Clausi, D. a., 2013. Multiple scale-specific representations for improved human action recognition. *Pattern Recognition Letters*, 34(15), pp.1771–1779. Available at: http://dx.doi.org/10.1016/j.patrec.2012.12.013.

SIL, American English Homophones. Available at: http://www-01.sil.org/linguistics/wordlists/english/.

United Nations, 2014. Concise Report on the World Population Situation in 2014.

United Nations, 2010. *World Population Ageing 2009 (Population Studies Series)* Pap/Cdr Ed.,

Veltkamp, R., Burkhardt, H. & Kriegel, H.-P., 2013. *State-of-the-art in content-based image and video retrieval*, Springer Science & Business Media.

Yang, H. & Meinel, C., 2014. Content based lecture video retrieval using speech and video text information. *IEEE Transactions on Learning Technologies*, 7(2), pp.142–154.