C. Dobbins, M. Merabti, P. Fergus, and D. Llewellyn-Jones, "The Big Data Obstacle of Lifelogging," in *The 28th IEEE International Conference on Advanced Information Networking and Applications (AINA'14)*, 2014, pp. 922–926

The Big Data Obstacle of Lifelogging

Chelsea Dobbins, Madjid Merabti, Paul Fergus, David Llewellyn-Jones

School of Computing and Mathematical Sciences

Liverpool John Moores University

Liverpool, UK

{C.M.Dobbins, M.Merabti, P.Fergus, D.Llewellyn-Jones}@ljmu.ac.uk

***Abstract*– Living in the digital age has resulted in a data rich society where the ability to log every moment of our lives is now possible. This chronicle is known as a human digital memory and is a heterogeneous record of our lives, which grows alongside its human counterpart. Managing a lifetime of data results in these sets of big data growing to enormous proportions; as these records increase in size the problem of effectively managing them becomes more difficult. This paper explores the challenges of searching such big data sets of human digital memory data and posits a new approach that treats the searching of human digital memory data as a machine learning problem.**

***Index Terms*—Human Digital Memory; Lifelogging; Sensors; Big Data; Clustering; Machine Learning**

I. INTRODUCTION

Time is physically irreversible. The unidirectionality of time is one of nature's most fundamental laws and as long as the universe has existed governs all occurrences; there is no return to yesterday [1]. Although it is impossible to physically go back in time, mental time travel occurs every day. As stated by Tulving [1], "*Time's flow is irreversible. The singular exception is provided by the human ability to remember past happenings. When one thinks today about what one did yesterday, time's arrow is bent into a loop. The rememberer has mentally travelled back into her past and thus violated the law of the irreversibility of the flow of time.*" This unique ability resides within all of us and occurs on a daily basis, without hesitation. As such, human memory is considered to be the most basic and important operation of the brain, with very few cognitive processes (recognition, language, planning, etc.) being able to operate effectively without a contribution from it [2].

However, retaining every aspect of our lives, for example, how we felt or what we did on a specific day is virtually impossible. For example, a birthday party that occurred yesterday is typically remembered in greater detail than a similar event from twenty years ago. As people get older, the ability to remember information declines [3]. Nevertheless, recent advances in technology can alleviate this problem, to a certain extent. Devices are now capable of capturing every moment of daily life. As such, this has led to the phenomenon of 'lifelogging,' which refers to the process of automatically recording aspects of one's life in digital form [4]. As described by Dodge and Kitchin [5], "*A life-log is conceived as a form of pervasive computing consisting of a unified digital record of the totality of an individual's experiences, captured multimodally through digital sensors and stored permanently as a personal multimedia archive*". Such extensive digital collections are often referred to as human digital memories (HDMs). As defined by Kelly [6], "*A HDM is typically a combination of many types of media, audio, video and images*". These personal archives are constructed from a wide range of data sources, across various media types [7]. HDMs are now becoming a reality and reflecting upon those items has become an active part of people's lives [8].

As such, HDMs are becoming richer in content. This is due to the consequences of leading an increasingly digital lifestyle, which results in copious amounts of information being generated. We are living in a data rich society, where the ability to generate and access a number of different data sources is possible. Any object, embedded with a sensor, can provide us with information. Through unique addressing schemes, these pervasive devices are able to interact with each other and cooperate with their neighbours, to reach common goals [9]. This revolution is known as the Internet of Things (IoT) and can be defined as "*a worldwide network of uniquely addressable interconnected objects, based on shared communication protocols*" [10]. These "smart objects" now fit seamlessly into our world, instead of forcing users to enter their environment, a concept first envisioned by Weiser [11]. As it becomes more socially acceptable to continually capture content, whether it is from a wearable camera or sensors, the pool of data that is being amassed is growing rapidly. According to IBM [12], "*Every day, 2.5 quintillion bytes of data are created — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few.*" Using this data, an entire lifetime can be reconstructed, thus creating human digital memories of life experiences.

A powerful source of information that is often used to generate and collect data comes from mobile devices. According to a recent report by Cisco [13], in 2012, global mobile data traffic grew by 70%, compared to 2011. In 2012, this type of traffic reached 885 petabytes per month and was nearly twelve times greater than the total global Internet traffic in 2000 (75 petabytes per month). Reiterating this growth in data is Intel's recent depiction of the exchange of data in a minute (see Fig. 1) [14].

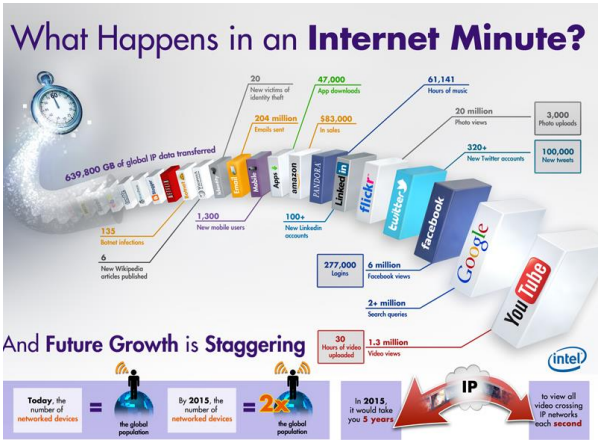

Fig. 1. Data Exchange In One Minute On The Internet [14]

As it can be seen, the generation of data, even in a single minute, is staggering. With all of this data at our fingertips, searching such vast heterogeneous digital archives, in order to find specific moments in time, is a significant challenge. As more data is accumulated, the ability to manage it becomes harder. This interest has led to the task of managing, and using, human digital memories, over a lifetime, being declared a grand challenge in computing [15].

This paper explores the challenges of searching such big data sets of HDM information and posits a new approach that treats the searching of HDM data as a machine learning problem. The preliminary results that have been achieved are interesting and illustrate how a user's HDM information can be successfully searched without the need to define complex queries.

II. Background

Research into capturing and creating human digital memories has received a great deal of attention, from researchers, over the last few decades. Since the *Memex* [16] in 1945, research into how aspects of our lives can be captured and organised, have been investigated. Over time, this vision of storing accumulated items has evolved into digitally capturing and storing information about ourselves and our environment. The culmination of this practise has been to lifelog, i.e. continually capture content with the aid of wearable systems. Lifelogging has many benefits, Sellen and Whittaker [17] summarize these as "*the five Rs*":

1. *Recollecting (mentally re-living specific life experiences),*
2. *Reminiscing (re-living past experiences for emotional or sentimental reasons, either individually or social in groups),*
3. *Retrieving (recovering specific digital information we've encountered over the years, for example, documents, email, and Web pages),*
4. *Reflecting (the reviewing of past experiences that may include examining patterns of about one's behaviour over time),*
5. *Remembering intentions (remembering prospective events in one's life)*

Human digital memories are a digital representation of ourselves that evolve and grow alongside us and are seen as a window into our past. As technology advances and sensors become more prevalent, within our environment, the range of data that we have access to is increasing. This has led to HDMs becoming richer in content. However, as people collect more and more data there is a danger of "information overload" and inadvertently, significant mementos are being lost and forgotten.

One such approach that explores the use of machine learning is *PhotoTOC,* proposed by Platt *et al.* [18]. This application is "*A browser for personal digital photographs that uses a clustering algorithm to automatically generate a table of contents of a user's personal photograph collection*". In this implementation, time-based clustering has been used to choose one photograph from a cluster, which is the most representative of that cluster. These photographs then provide an overview of the entire collection. While Harada *et al.* [19] developed a timeline browser for PDA's that uses a time–based clustering algorithm to organise related photos together. Similarly, Harada *et al.*'s [19] algorithm has been based on previous work by Graham *et al.* [20] in which their original system uses the recursive way in which photographs are taken, in bursts, and represents this using a tree of clusters where photos are stored only at the leaf nodes [20].

Whilst these developments are interesting, in terms of organisation and the way in which data is retrieved, a memory is composed of much more information than just photographs. More data is required so that detailed human digital memories can be created. Information such as physiological changes, temperature, location, etc. would provide more context about such captured times. However, the inclusion of more data creates new challenges in terms of information retrieval. Creating such heterogeneous records requires sophisticated searching methods that can cope with retrieving a variety of data items.

III. The Research Challenges of Searching a Lifetime of Data

The vision of the *Memories for Life: managing information over a human lifetime* grand challenge is to help people manage and use their digital memories across their *entire* lifetime [15]. Collecting data over this extensive period of time yields a phenomenal amount of information. When human digital memories are created this enormous amount of data needs to be intelligently searched and the associated information succinctly brought together. As stated by Ranpura [21], "*Memories are rich because they are formed through associations. When we experience an event, our brains tie the sights, smells, sounds, and our own impressions together into a relationship. That relationship itself is the memory of the event*". Whilst humans can do this type of processing, subconsciously, in a matter of nanoseconds, creating these associations, digitally, poses a greater challenge. The complex and heterogeneous nature of a human digital memory means that the simple ranked retrieval of information is unlikely to support many of the user's information searching tasks [22]. Furthermore, queries that require sophisticated interpretation need to be

efficiently handled [15]. For example, queries such as, "When have I spoken to Joe?" or "Find all of my happy memories?" requires an intelligent method of data analysis that enables multi–dimensional queries to be executed across a vast amount of data. Consequently, the system needs to learn about its user and understand their data.

Machine learning techniques are seen as a way to overcome some of these challenge. Intelligent search, instead of keyword matching, and query answering is facilitated and provides a way to search data from distributed sources, irrespective of its format [23]. Using a matrix representation of the data, allows the searching of this information to be searched based on the similarities in a vector object. Consequently, a wider range of information can be included in the memory; the user is not limited by needing to have a pre-existing knowledge of the information. For example, structuring queries requires the user to define exactly what they are looking for and the location of this data. This approach is limited because as more data is amassed managing this information becomes harder. However, clustering enables this data to be explored without the user necessarily knowing what they are looking for; instead similar pieces of information are automatically retrieved.

IV. Evaluation

In exploring this idea, this section presents the preliminary results that have been achieved from searching human digital memory data, using the well-known *k-means* clustering technique. *K-means* was chosen because of its simplicity, and because it is the most widely used clustering algorithm in practice, which has been used in a variety of application domains [24]–[26].

In order to demonstrate this idea, the user undertook a variety of activities, over seven days. This included lying down, sitting, standing, walking, running, ascending and descending stairs, vacuum cleaning and ironing. As a result, a sample of photographs, location, heart rate and data from three accelerometers, which were worn on the ankle, chest and hand have been collected. This information has been pre-processed and features have been extracted. The feature set comprises of a variety of features from the *time*, *frequency* and *geographic* domains, for example, *mean*, *median*, *standard deviation*, *energy*, *entropy*, *peak frequency*, and *geographic mean*, to name but a few.

Using this set of features, the results from the *k-means* algorithm have been analysed. In order to explore this idea, the user first logs into the DigMem [27] system and chooses from a set of pre-defined questions (see Fig. 2) how they would like the system to query their data. Questions enable the user to gather more of an insight into their behaviours, without defining search queries. Using questions one and four as an example, these are used to explore times of high and low energy expenditure. In this instance, 'high-energy' activity refers to walking, running, ascending and descending stairs and vacuum cleaning, whilst 'low-energy' activities denotes lying down, sitting, standing and ironing.

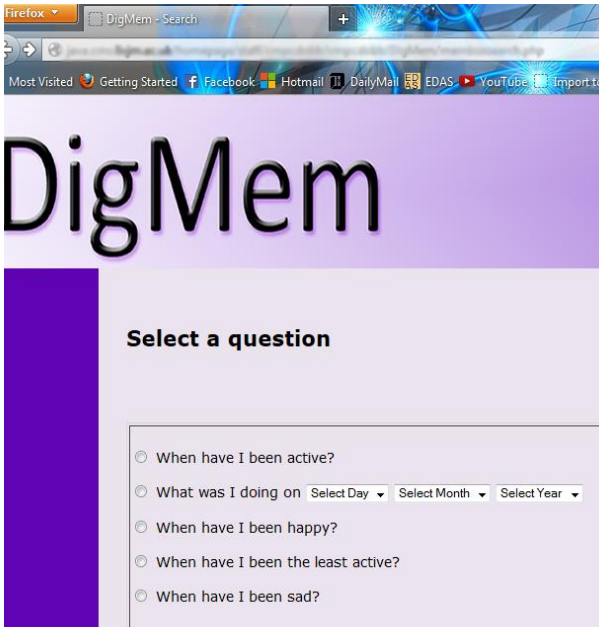

Fig. 2. DigMem Questions Web Page

In order to demonstrate this idea, the first question that was selected was, "When have I been active?" Fig. 3 illustrates these results. As it can be seen, two clearly defined clusters are present. Cluster 1 is composed of 45% of the data, whilst cluster 2 is 55%. As it can be seen, there is a clear divide in the data. This illustrates that the majority of the high-energy activities, which were being performed, had a higher irregular motion pattern, as the entropy levels are quite high. Entropy characterizes the consistency in an activity, and helps to differentiate between signals that have similar energy values but correspond to different activity patterns [28], [29]. As more energy is used, the activities become more repetitive in their frequency. In particular, it can be seen that there are three periods of time that exhibit particularly high energy but lower entropy levels. Therefore, less time was spent doing high-energy activities that involved a high repeated frequency, such as running, since a lesser portion of the data is in cluster 1.

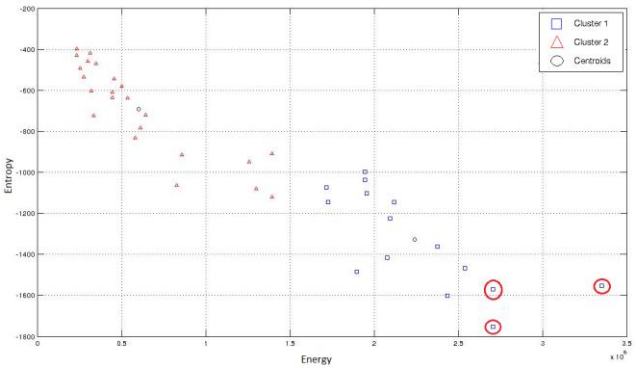

Fig. 3. *K-means* Analysis – When Was I Most Active?

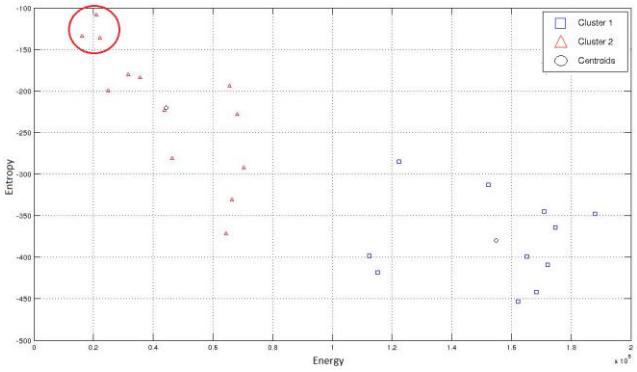

Fig. 4. *K-means* Analysis – When Was I Least Active?

The second question "When have I been the least active?" has also been asked. Fig. 4 illustrates the results from the question. As it can be seen, the range of energy has decreased. This illustrates that fewer intensive activities have been performed. Lower entropy also suggests that those activities were also repetitive in nature, such as sitting down. Cluster 1 is composed of 46% of the data, whilst cluster 2 is 54%. Since Cluster 1 has higher energy but lower entropy is can be deduced that during those times the user was walking, as this has a high repetitive frequency.

As it can be seen, these preliminary results support the idea that clustering data is a viable method of searching HDM data. They clearly illustrate the periods of time that the user has spent being active and sedentary. A direct correlation between energy and entropy is also visible. Simply searching this data with specific queries, or keywords, is cumbersome and the potential for human-error to omit information in the search queries is greater. As demonstrated, unsupervised machine learning is able to treat the challenge of searching this data as a clustering problem and to retrieve information based on features. The user simply selects a query from a set of pre-defined questions and the clustering algorithm retrieves this information automatically, thus eliminating the need for the user to define their search criteria.

By treating the searching of HDM data as a clustering problem removes the need to label the feature vectors. By letting the algorithm cluster similar pieces of data together, removes the need to have a pre-existing knowledge about the data. Furthermore, the system is not 'learning' about the user's memories; therefore, testing and training sets are not required, as is the case in classification. Additionally, this method is beneficial as it overcomes the limitations of searching data with complex query languages, such as the SPARQL Protocol for RDF (SPARQL) [30]. SPARQL is a complicated language that relies on the user understanding the domain before queries can be constructed. If the user is unfamiliar with the underlying data, then finding information can almost be impossible. In addition, navigating SPARQL's complex labyrinth of syntax is a difficult task entirely. However, by transforming the raw data into HDM vectors, and treating the searching of this data as a clustering problem eliminates the need to have a pre-existing knowledge of the dataset. Furthermore, these vectors can become extremely large, especially if a lifetime of data is being recorded. In spite of this, clustering algorithms are able to deal with these sets of big data quite easily. By transforming extremely large datasets, of raw data, into features enables the HDM vectors to be rich with information. The bigger the feature space is the more detailed a memory is.

V. Summary and Future Work

As more and more data is being generated, a great deal of information can be gathered about ourselves and the environment. This information can then be used to re-create any time throughout our lives. However, there is a danger of information overload. As we accumulate more data, the difficulty in managing it becomes apparent. For example, finding key moments in twenty years' worth of data can seem almost impossible. Advanced searching techniques are required, which can find information with minimal user involvement. As posited in this paper, clustering techniques aim to address this challenge. Using pre-defined questions, the algorithms group data based on similarity. The user does not need to define their search criteria, thus limiting the possibility of overlooking data items.

One limitation of the system is that the questions approach only considers data that is straightforward to measure (location, accelerometer and heartbeat). Machine learning algorithms can easily classify this type of data. Future work would consider expanding the range of questions so that photographic data could be queried, instead of being linked in at a later time. Executing such queries requires sophisticated interpretation, such as "*Find a picture of me playing with Peter when he was a toddler*" [15]. This type of query places considerable focus on computer vision and image understanding [31]–[33]. In order to execute this query, an innate understanding of who the people in the picture are and activity recognition are required. Incorporating this type of question is an exciting avenue for the research, as is the idea of allowing the user to create customised questions. This would enable the system to fully understand its user.

References

[1] E. Tulving, "Episodic Memory: From Mind to Brain," *Annu. Rev. Psychol.*, vol. 53, pp. 1–25, Jan. 2002.

[2] D. Tranel and A. R. Damasio, "Chapter 2 Neurobiological Foundations of Human Memory," in in *The Handbook of Memory Disorders*, John Wiley & Sons, 2003, p. 17.

[3] M. W. Prull, L. L. C. Dawes, A. M. Martin, H. F. Rosenberg, and L. L. Light, "Recollection and familiarity in recognition memory: adult age differences and neuropsychological test correlates.," *Psychol. Aging*, vol. 21, no. 1, pp. 107–18, Mar. 2006.

[4] A. R. Doherty, N. Caprani, C. Ó. Conaire, V. Kalnikaite, C. Gurrin, A. F. Smeaton, and N. E. O'Connor, "Passively Recognising Human Activities Through Lifelogging," *Comput. Human Behav.*, vol. 27, no. 5, pp. 1948–1958, Sep. 2011.

[5] M. Dodge and R. Kitchin, "'Outlines of a world coming into existence': pervasive computing and the ethics of forgetting," *Environ. Plan. B Plan. Des.*, vol. 34, no. 3, pp. 431–445, 2007.

[6] L. Kelly, “The Information Retrieval Challenge of Human Digital Memories,” in *BCS IRSG Symposium: Future Directions in Information*, 2007.

[7] C. Gurrin, D. Byrne, N. O’Connor, G. J. F. Jones, and A. F. Smeaton, “Architecture and Challenges of Maintaining a Large-scale, Context-aware Human Digital Memory,” in *5th International Conference on Visual Information Engineering*, 2008, pp. 158–163.

[8] V. Kalnikaite and S. Whittaker, “A Saunter Down Memory Lane: Digital Reflection on Personal Mementos,” *Int. J. Hum. Comput. Stud.*, vol. 69, no. 5, pp. 298–310, Jan. 2011.

[9] L. Atzori, A. Iera, and G. Morabito, “The Internet of Things: A survey,” *Comput. Networks*, vol. 54, no. 15, pp. 2787–2805, Oct. 2010.

[10] L. Mainetti, L. Patrono, and A. Vilei, “Evolution of wireless sensor networks towards the Internet of Things: A survey,” in *19th International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, 2011, pp. 1–6.

[11] M. Weiser, “The Computer for the 21st Century,” *ACM SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 3, no. 3, pp. 3–11, Jul. 1999.

[12] IBM.com, “What is big data?,” 2013. [Online]. Available: http://www-01.ibm.com/software/data/bigdata/. [Accessed: 19-Mar-2013].

[13] Cisco, “Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012–2017.” pp. 1–34, 2013.

[14] intel, “What Happens in an Internet Minute?,” 2013. [Online]. Available: http://www.intel.com/content/www/us/en/communications/internet-minute-infographic.html. [Accessed: 19-Mar-2013].

[15] A. Fitzgibbon and E. Reiter, “Grand Challenges in Computing Research: GC3 Memories for life: managing information over a human lifetime,” in *Conference on Grand Challenges in Computing Research*, 2005, pp. 13–16.

[16] V. Bush, “As We May Think,” *The Atlantic Monthly*, no. JULY 1945, 1945.

[17] A. J. Sellen and S. Whittaker, “Beyond Total Capture: A Constructive Critique of Lifelogging,” *Commun. ACM*, vol. 53, no. 5, pp. 70–77, May 2010.

[18] J. C. Platt, M. Czerwinski, and B. A. Field, “PhotoTOC: automatic clustering for browsing personal photographs,” in *Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint*, 2003, pp. 6–10.

[19] S. Harada, M. Naaman, Y. J. Song, Q. Wang, and A. Paepcke, “Lost in Memories: Interacting With Photo Collections On PDAs,” in *Proceedings of the 2004 joint ACM/IEEE conference on Digital libraries - JCDL ’04*, 2004, p. 325.

[20] A. Graham, H. Garcia-Molina, A. Paepcke, and T. Winograd, “Time as Essence for Photo Browsing Through Personal Digital Libraries,” in *Proceedings of the second ACM/IEEE-CS joint conference on Digital libraries - JCDL ’02*, 2002, p. 326.

[21] A. Ranpura, “How We Remember, and Why We Forget,” *BrainConnection.com*, 2000. [Online]. Available: http://brainconnection.positscience.com/how-we-remember-and-why-we-forget/.

[22] L. Kelly and G. J. F. Jones, “Venturing into the labyrinth: the information retrieval challenge of human digital memories,” in *Workshop on Supporting Human Memory with Interactive Systems, Lancaster, UK*, 2007, pp. 37–40.

[23] D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster, *Semantic Web: Why, What, and How?* Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1–653.

[24] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl, “Constrained K-means Clustering with Background Knowledge,” in *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, 2001, pp. 577–584.

[25] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, “Top 10 algorithms in data mining,” *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, Dec. 2007.

[26] M. G. Forero, F. Sroubek, and G. Cristóbal, “Identification of tuberculosis bacteria based on shape and color,” *Real-Time Imaging*, vol. 10, no. 4, pp. 251–262, Aug. 2004.

[27] C. Dobbins, M. Merabti, P. Fergus, and D. Llewellyn-Jones, “Creating Human Digital Memories for a Richer Recall of Life Experiences,” in *Proceedings of 10th IEEE International Conference on Networking, Sensing and Control (ICNSC’13)*, 2013, pp. 246–251.

[28] D. Figo, P. C. Diniz, D. R. Ferreira, and J. M. P. Cardoso, “Preprocessing Techniques for Context Recognition from Accelerometer Data,” *Pers. Ubiquitous Comput.*, vol. 14, no. 7, pp. 645–662, Mar. 2010.

[29] W. Song, C. Ade, R. Broxterman, T. Barstow, T. Nelson, and S. Warren, “Activity Recognition in Planetary Navigation Field Tests Using Classification Algorithms Applied to Accelerometer Data,” in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2012, vol. 2012, pp. 1586–1589.

[30] W3C, “SPARQL Protocol for RDF,” 2008. [Online]. Available: http://www.w3.org/TR/rdf-sparql-protocol/. [Accessed: 27-Jun-2013].

[31] V. Delaitre, I. Laptev, and J. Sivic, “Recognizing human actions in still images: a study of bag-of-features and part-based representations,” *Procedings Br. Mach. Vis. Conf.*, vol. 2, no. 5, p. 7, 2010.

[32] N. Bicocchi, M. Lasagni, and F. Zambonelli, “Bridging Vision and commonsense for Multimodal Situation Recognition in Pervasive Systems,” in *IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2012, pp. 48–56.

[33] C. Nakajima, M. Pontil, B. Heisele, and T. Poggio, “People Recognition in Image Sequences by Supervised Learning,” in *Report for Massachusetts Institute of Technology (MIT) Artificial Intelligence Laboratory and Center for Biological and Computational Learning a Department Of Brain and Cognitive Sciences*, 2000, no. 1688, pp. 1–12.