5W1H as a Human Activity Recognition Paradigm in the iSpace

Leon F. Palafox
School of Electric Engineering
The University of Tokyo
Tokyo, Japan
Email: leon@hlab.iis.u-tokyo.ac.jp

Laszlo A. Jeni
School of Electric Engineering
The University of Tokyo
Tokyo, Japan
Email: laszlo@hlab.iis.u-tokyo.ac.jp

Hideki Hashimoto
School of Engineering
The University of Tokyo
Tokyo, Japan
Email: hideki@iis.u-tokyo.ac.jp

***Abstract*—In this paper we propose the use of a widely known paradigm in the forensic area to detect and categorize human activities. 4W1H describes the context in any environment by describing it using 5 base variables, Who, When, What, Where and How. We make use of this description plus an intention variable known as "Why" to be able to predict and react accordingly to the current user's situation. We show the hardware setting required to detect these variables as well as some approaches to sense them from the environment using a minimal hardware setting. We also look into the current work in categorizing the data using Self Organizing Maps and Clustering Techniques.**

I. Introduction

Human Activity Recognition (HAR) is one of the prime problems in the setting of intelligent spaces such as the iSpace [1]. These are rooms equipped with sensors (Fig. 1) that capture information of the users actuating within them to afterwards provide services accordingly to the current situation. Examples of these are self regulated air conditioned systems, or automatic light dimmers. To be able to respond and have an effective interaction with the users of the space, a good machinery, capable of precise recognition of the diverse performed activities, is needed.

There are diverse approaches to the HAR problem, most works have been focused on recognizing human activities based on images retrieved from cameras [2] [3] and using pattern recognition algorithms to match those inputs with previous information stored in a database. There are different problems that affect these approaches; lighting conditions and obstruction possibility among the most critical ones, since in most of the cases these problems make too difficult to do a good detection of the sensed subjects. Recent papers have dealt with these problems by adding sensors, like RFID tags attached to the objects and the users, to have continuous sensing; as well as extracting relevant information from other data sources, like the pitch in a microphone or the input in a keyboard and/or mouse [4].

However, the use of RFID tags (that have accelerometers attached to them) also presents some setbacks, specially since some of the time they may provide unreliable information; for example, in moments of inactivity both the users and the objects present null acceleration, making impossible to track the current activity using only the tag in those specific moments, thus having the need for visual classification using cameras.

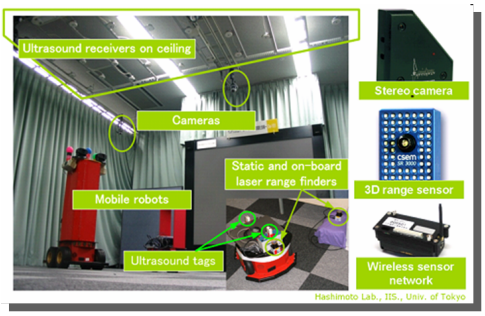


Fig. 1. Intelligent Space with Distributed Sensors and Mobile Robots.

Because of these problems, context, which defines the general state of the system, becomes a crucial variable in the HAR sensing setting. By defining a context, problems such as inactivity or sensing obtrusion may be overcome, and is in defining it that some works such as [4] and [5] have put special emphasis to provide a more comprehensive sensing of the environment, that will allow the overall architecture to drastically improve its classification capabilities.

In this paper we propose a new paradigm to detect context, based on the 4W1H (Who, When, What, Where and How) structure of information proposed before in [6] and [7] for intelligent spaces, in which we organize the sensed data according to specific variables. The new model adds and additional variable to the proposed structure of the sensing known as "Why", which pretends to extend over previous works by also serving as a feedback analysis of the user responses to the environment reaction of his activities, like his feelings concerning the room's heating system being suddenly turned on. Recent research has tried to extract these emotions from cameras, using different paradigms to do a translation from the features sensed in a face to emotions and feelings of the subjects.

The paper is organized as follows, first we will present an overall explanation of the 4W1H paradigm, its advantages and disadvantages against other common sensing schemes, we will describe the kind of hardware necessary to implement it and give a brief review of the proposed solutions to deal with some

of the common troubles this system can present, next, we will show the addition of the new variable "Why", we will show present work being done to asses it, and will briefly analyze the current approach in the iSpace. In the last part, we will present a structure of the overall scheme, and how it is being planned to interact with the users, we will present some of the classification results, to finally present the future directions of the research and the latent problems it presents.

II. 4W1H

First proposed in the forensic field [8] to perform a correct assessment of every important variable concerning a specific situation, 4W1H was first used in intelligent environments in [7], and it describes a paradigm in which we fragment every activity within the space into 5 elemental variables:

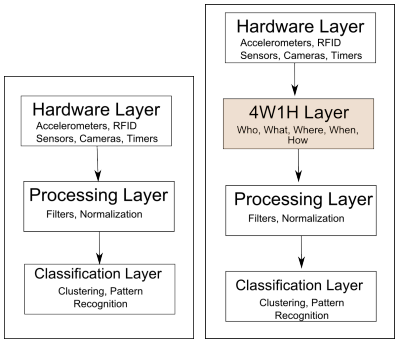
1) *Who*: The current user of the space and objects (John, Mary, Citizens, Country Population)
2) *Where*: The actual location where the activity is taking place (Home, Kitchen, Bedroom, Tokyo, Indonesia)
3) *When*: The moment in time for the action (2 p.m., afternoon, 1999, 5th Century)
4) *What*: The objects the user is interacting with (car, cup, book, ship)
5) *How*: The way the user is handling himself or the objects (standing, pouring, turning pages)

We can clearly see that this kind of codification allows us to analyze not only individual humans activities, but global situations as well, the model is fairly flexible, and allows us to asses the context of any possible environment for further analysis.

To extract the information for each variables, we need different sensors, each of them devoted to a specific element. It is important to notice that when using 4W1H is not necessary to have the information regarding all the variables at every sensing time. Every new variable that we have increases the level of certainty in the sensing for further recognition, but by being capable of obtaining only a subset of elements in the 4W1H (Who and What, for example) we still can make a fair assumption regarding the current state of the space.

In the 4W1H scheme, elements can present redundancy and have a high codependency among them, for example, while sleeping, the object with the tag {bed} will be highly dependent on the time of the day, which will usually have the tag {night} [9]; that means that while knowing the current time of the day to be {night}, we can make a fair assumption that the object associated will be the bed, and the overall activity of the person will be sleeping, thus making the tags {night} and {bed} highly correlated from a statistical point of view and making them adequate to perform a correct classification. This kind of correlation is present in almost every activity a normal human performs in a daily basis, the tag {office} for example and the tag {keyboard} are also correlated, as most of the time a person spends in the office he or she is always interacting with a computer.

Is this redundancy what makes the 4W1H a highly flexible sensing paradigm, in which problems like camera obtrusion or lighting conditions that ruins a sensor's performance may be overcome by simply obtaining the information from the other sensors in the environment. If a camera is suddenly unable to sense "who" the current user is, the sensing setting can still infer who he is given his current location or interacting objects.



(a) Normal (b) 4W1H

Fig. 2. Different Sensing Settings

A. Implementation Scheme

In Figure 2 we show the difference between sensing schemes in a layered architecture, where each layer transmits information to the next one along different transmission means, e.g. cables, wireless data, etc.

The most common approach is to have at most 3 layers (Fig. 2a), the top layer will be in charge of retrieving all the information from the sensors located in the physical space, it may be composed of cameras, RFID tags, accelerometers, etc. This layer is also known as the "Phyiscal Layer" and it is composed by the hardware elements of our sensing system.

In the next layer, we do a buffering of all the data from the sensors and perform different pre-processing algorithms. When sensing audio waves, for example, and the objective is to extract human sounds and speech, a band pass filtering needs to be performed over the common human speech band, (about 200 to 9000 Hz) to work only with relevant signals; as well, some algorithms require normalization and zero-mean transformation in order to work correctly and make different sources of information comparable for later classification. Some sensors, have this preprocessing algorithms already included in the hardware layer, but it is usually performed in a separate one to allow tunning and personalization of the parameters.

The final layer is formed by the different classification, clustering and pattern recognition algorithms used to detect important features in the sensed data. It also outputs the most fitted response for the current situation in the environment, if the sensed data, for example, tells the system someone is searching the light switch and the lights are off, the most fitted response will be to turn the lights on.

The classification is the most critical structure, since it often determines the setting and parameters for the previous layers in the system, if the classification is done via a clustering

algorithm over the different sensors, we need to process the variance of the data in the pre-processing layer to obtain comparable data. The classification scheme, as well, defines the sampling frequencies in the hardware layer to have synchronization among the different sensors to obtain coherent data in a given time lapse.

In the 4W1H sensing scheme (Fig. 2b), we add a new layer to the architecture, the 4W1H layer, the specific function of this middleware layer is to perform a deterministic clustering, in which we allocate the raw information from each of the sensors in specific variable modules (4W1H elements), and then, perform the classification based on the 5 variables, doing then, a synthetic space reduction of the order in our data. We will reduce N dimensional data to a 5 dimension data vector {Who, What, When, Where, How}.

It is important to notice that the principal advantage that the 4W1H setting offers is that the layer does not need to receive all the information concerning the sensors, since it will work solely with the data it has at hand, thus making it robust against the failure of any sensor. Further layers like preprocessing and classification would be directed influenced by the 4W1H layer, since the nature of the data will change, considerably reducing the classification space dimension.

B. Hardware Setting

To implement the 4W1H paradigm, the hardware setting differs from that of conventional HAR schemes. The most fundamental difference is that while in most settings the overall sensing is performed mainly by one of a set of sensors (camera, RFID, Accelerometers), in the 4W1H setting we retrieve the data from different sensors in a synchronized manner (Fig. 3), it can be clearly seen that by sensing in this way, even if we lose communication from one of the sensors (time t), we still can perform certain classification techniques that do not require extensive data from all of the sensors at once. And that in perfect conditions (time t+n), we can use the data of all the sensors to perform a classification.

To do the implementation of the scheme, the main variables are allocated to each of the elements in 4W1H from the sensors that capture the specific data of specific variables:

1) *Who*: It is usually retrieved from RFID tags in the users wrists (can be performed by cameras as well)
2) *Where*: Depending on the level of location, it can be sensed from a GPS (global location), ZPS (Local sensing) or even the IP of a stationary computer. It is also worth to notice that other approaches, like cameras, can be used.
3) *When*: Each of the sensors are synchronized by a master clock that is also synchronized with the global time.
4) *What*: Objects have RFID tags attached to them, each of them with accelerometers incorporated to them, so we can sense the specific moment each object is being interacted with.
5) *How*: We retrieve this information by doing classification over the data set of an accelerometer, gyroscope and magnetometer sensor that is attached to the user at every moment, to do an assessment of his/her current activities.

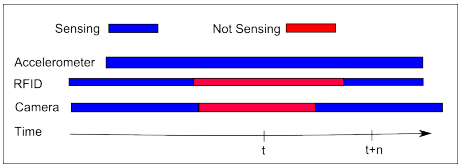


Fig. 3. 4W1H Temporal Scheme.

C. 5W1H

Commonly known as 5W1H, also known as 4W1H+W. In this specific paradigm, we add an additional variable to the sensing architecture, ”Why“; this variable has been studied before in [10] by being tracked and analyzed using a neural network to classify the facial features for recognizing emotions in humans. This emotions are directly related to the willingness or disposition of humans regarding the current task, by inferring this variable, we can gain some insight in the ulterior motives of the users, providing a good feedback support for further classification in the current space. As well, it provides us with richer information regarding the user's state of mind in the space, for example, is the user currently happy by doing work or not? How likely is he to finish the task at hand given his current mood; if he is typing and very excited, how likely is that he is writing a personal email rather than a work related email?

III. Classification and Recognition in the 5W1H scheme

Classification in the 5W1H scheme has to be done in different steps, since some of the variables like ”Where“ and ”When“ are easily extractable from the sensors, but more subtle variables, such as ”How“ and ”Why“ need a pre-classification of the raw data coming from the sensors. Yet, depending on the selected setting, the data processing step might be ignored or simply done by a deterministic data allocation. The selected setting will be determined by the application we wish to do with our sensing environment, hospitals and houses have different requirements in sensing for accuracy and flexibility, hospitals for example, are highly controlled environments, where the presence of new objects and users is constantly being monitored and the amount of possible actions for a patient in a room is also limited by the disease he or she might have, while a house is a more flexible environment in which the possible combination of activities and objects increases exponentially.

A. Who

While in the current setting, we use simply the information received from RFIDS associated with the users in the environment, further approaches can be done for a less invasive hardware setting, there has been extensive work in the use of cameras to perform user recognition [11], [12], [13]. This approach of using cameras offers a less invasive setting, albeit

to the sacrifice of processing speed in the overall setting, while in the RFID approach we only need to sense serial data from the sensors, using cameras may slow our system due to the heavy image processing associated to each captured frame.

B. Where

There is a number of different works regarding localization of humans in controlled spaces [14] using different sensors, as well as with the 'Who" variable, the specific setting used so far only relies in the deterministic assignment of the IP assigned to the sensing computer in a structured network (where each IP is accounted for, physically). Less invasive and more flexible algorithms are available as well for place recognition using cameras, like the work in [15], the trade-off between processing time and flexibility is now present, and it will depend in our sensing application and desired scalability on which approach should we chose.

C. When

Time is a variable that is easily extractable from a computer hardware, and the most common approach will be to sample it from the computer or the sensor clock each time an action is presented in the environment. There may be the possibility of extracting time of the day from a scenery, but is a lose-lose relation, since our precision won't be as high as with a clock, and the processing time will suffer greatly due to the recognition algorithms.

D. What

Extensive research has been done in deciding which is the optimal way to sense individual objects in an environment, in our setting, we choose RFID sensors, due to the data and the processing of the information being direct. Making further classification or data mining unnecessary, since every piece of data we retrieve from an RFID antenna comes with a tag that associates itself with an object. One current problem is the necessity of an object database on which objects and tags are being paired, and each new object in the environment will require a new tag to be created in order to allow its recognition in the system.

Other approach currently being used is the use of cameras to detect objects performing training algorithms [16], [17] based on previous presentation of the objects, and thus being able to create dynamic databases using only images detected from cameras.

For a setting with a controlled amount of objects and with the requirement of high sensing accuracy, like a hospital, the use of cameras wouldn't be advisable since a single mistake recognizing a different kind of medicine or object may have grave consequences, yet in more flexible environments, where the amount of objects grows exponentially as well as the amount of replacement also happens quite often, like a normal house, RFID would present a suboptimal solution, due to the need of updating the database each time a new object enters the environment, thus making an automatic tagging system, like a camera based, the best option.

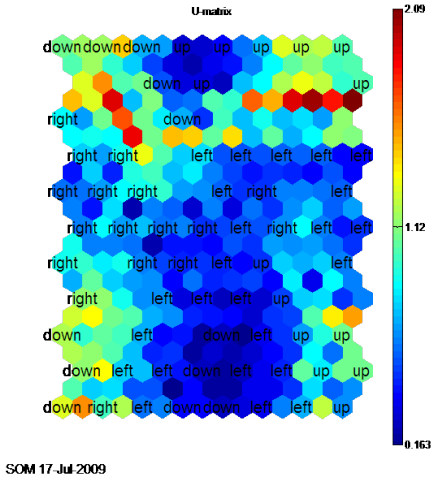


Fig. 4. Self Organizing Map classification of simple movements.

E. How

This variable has been commonly the most critical in the environment as well as the only one that under no sensing setting can be classified in a deterministic approach, there is usually the necessity of doing a previous classification, that will depend directly on the hardware equipment used to sense it.

In our setting, we used sensors from the Xsens family, [18], which has an accelerometer, a gyroscope and a magnetometer that allow for precise information related to the user's current movements and activities. Previous works have proven that the use of Self Organized Maps (SOM), Wavelets and Compressed Sensing, [19] greatly improves the sensing times as well as the accuracy.

Using SOM, we can obtain a good clustering of the sensor's data (Fig. 4) that will allow us to later do a good assessment of the overall situation, if a user is holding a cup, and moving his hand up, his most likely activity will be drinking.

Other approaches to sense this specific variables are to infer the current movements of the people using cameras [3], albeit all the problems using cameras to sense this kind of situation may bring (Precision, Processing time, lightning conditions).

F. Why

This variable is used both for later classification, and as a feedback system as well. As shown in [10] this system consists in the use of a camera to track facial expressions, that afterwards will be interpreted into a coding system developed by Paul Ekman [20]. (Fig. 5)

The graphical representation on how the "Why" variables allows us to do both classification and feedback, is shown in Figure 6, here we can appreciate how the variable is evaluated both at the sensing part, but again at the reaction part of a fully interactive system. An example of this application would be a system turning on the light by itself given that the user seems to feel uncomfortable with the current lighting setting, if the

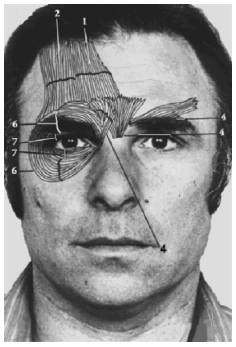


Fig. 5. Action Units in Head

user react happily, the action was correct, and thus reinforces the system's training, yet, if the user becomes upset or angry, the system may evaluate that as a negative feedback, and thus helping the classification system to improve its performance next time. The cameras used for the experiments are capable

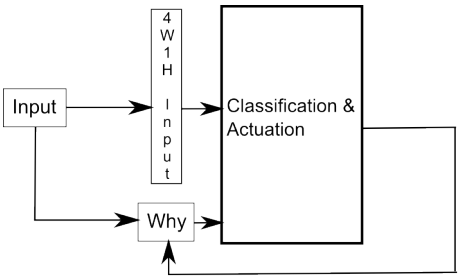


Fig. 6. Role of the "Why" Variable

of sampling 30 fps, and after a normalization and a feature recognition described in [10], a recognition algorithm will track and decode specific patterns. Which are described by Ekman as the basic patterns that comprise every possible human emotion, and by being able to actively track them, we can infer with a high precision level the current feelings of the system's user.

G. Classification algorithms for 5W1H

Once every variable from the 5W1H has been stored in a database, we need to undergo pattern recognition algorithms that allows us to perform a correct assumption of the current situation of the user. The classification algorithms will be classified in 2 possible learning settings according to their applications; for systems with a small set of possible movements or users and a limited set of actions, like one person's desk or a hospital room, we use a supervised setting, where mostly of all the situations are accounted for, and new situations are not expect to happen with a high frequency. On the other hand, for highly dynamic settings like an office building, or a classroom, or someone's house where the set of actions is exponentially large, an unsupervised approach needs to be taken, where the system creates the possible set of states by itself without prior knowledge of it. It is important to know

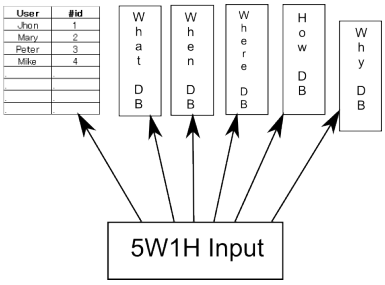


Fig. 7. Allocation of new inputs in the 5W1HDB's.

this definitions are set apart from the traditional definition of Supervised and Unsupervised Learning algorithms.

1) Supervised setting: A hospital room, usually, has a fixed number of possible interactions between the user and its environment, mainly because the user is restricted by any number of conditions to the room. A set of possible interactions is possible to be inferred by asking experts that interact with the system in a daily basis, in this specific scenario the nurses and the doctors would be the most likely experts as well as some patients who are currently residing in the hospital.

This kind of situation, where we are aware of the number of possible states in our system is called supervised setting, and for the specific problem of 5W1H, we need to perform a classification over a vector $S \in \mathbb{R}^{5\times 1}$, whose elements are the real valued numerical tags associated with each of the variables in the 5W1H, [Who, Where, When, What, Why, How] are related to numerical tags stored in dedicated databases for each of the variables, and each time we have a new action, or user, a new tag is allocated (Fig. 7). The task of creating this database of users and activities for the firsts set of sensing is done by the experts.

Among some of the algorithms used for unsupervised settings, we have clustering algorithms, in which we group variables according to their similitude with each other, in the 5W1H setting for example, we can group sets of actions, places, users accordingly with how alike their interaction with the environment is (a cup of tea and a glass have very similar interactions), among some of the clustering we could make and their objectives, we find:

Clustering around users

In this setting, we can classify actions, places, objects happening around users, we can do clusters in which different users with their 5W1H elements are the variables to be grouped, and in this way, we may relate users with the places, objects and activities they do often in the space.

Clustering around places

By grouping users, activities and objects around places, we can know in certain settings, which are the most required or crowded places, in a hospital room, we can now how often someone used the bathroom to keep a health monitoring system, or how often that person was off the bed so we can make some assesement regarding his situation.

Clustering around objects

This way, we are able to know the most used objects in a controlled setting, this way, we can now how a person is interacting with the TV, or for example, if someone is reading a certain book, we could create a recommendation system that is able to offer a similar book to the person.

It is important that, for most clustering algorithms to have precise and accurate results, previous information about the space, concerning the total number of possible objects, users, and overall classes should be known beforehand, while there are some heuristics to implement clustering algorithms in unsupervised settings, it is a difficult task, given that the total number classes is usually obtained empirically from the total energy of the data.

2) Unsupervised Setting: In this setting, while we have the same way of allocation as in Figure 7 we now have a constantly increasing number of entries in each of the variables, with new users and new activities being performed on a daily basis, this setting often found in households or office buildings, where the interaction dynamics as well as the degrees of freedom lack any sense of restriction from an external agent.

In this setting, it is difficult to find or point out a single expert capable of providing all the information regarding the possibilities of the environment, the number of possible classes to analyze can grow exponentially as the possible number of users and objects increases with it.

There is a number of different approaches to solve this kind of problem, like model the whole environment as a probabilistic scheme, where each new action or user only adds some level of complexity to the system and thus making inference over latent variables like context.

To train this kind of systems, we usually have to use freely available labeled databases from the Internet, in which third parties have done the extensive and expensive work of observing the largest possible sets of users and interactions in a confined space.

IV. Preliminary Results

In some of the recent work, we have obtained good results for this kind of systems, albeit limited, the current plots show how this system has a good capability to create clusters in a controlled setting.

In figure 8 we present the results of an analysis of clustering using Particle Swarm Organization clustering methods on the set of data retrieved from the 4W1H [21]. We can clearly see how for a monitoring system design, this way to arrange data becomes highly efficient and convenient, since we can transform activities and users into a plotting of a space representing the current physical areas where activities are happening. The plotting results are fairly efficient, having a high percentage of accuracy recognizing the places, since for this experiments the space variable was of a deterministic nature, it is one of the more reliable clusterings we can make regarding any of the variables.

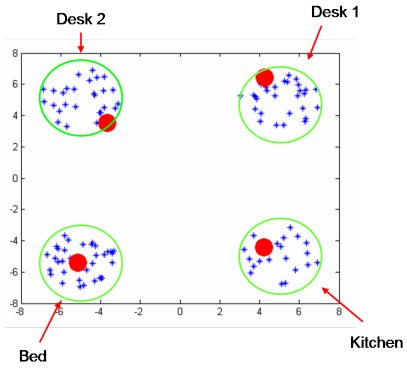


Fig. 8. Clusters around different places

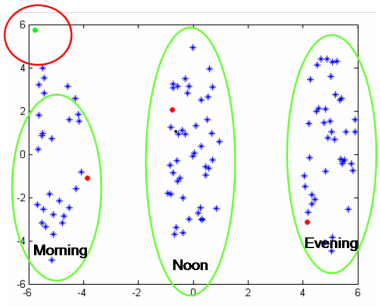


Fig. 9. Clusters around time variables

We give in Figure 9 similar clustering results for time variables, although equally convenient as the one we presented before, it exemplifies one of the limitations of this model. Unless the clustering is done with previous knowledge of the number of variables within the environment, we can have problems of over-clustering, shown in Figure 9 as a red area, which may introduce noise in the final recognition task of the user for this space since for most of the examples the output clusters are previously unlabeled.

We show in Figure 10 a final representation of the monitoring system, in which for a given space and user, we have a plot, that in the x-axis presents the time and in the y-axis the amount of involvement. It is seen in this way how different users interact with their environment at certain times, building in this way a system that to certain extent is aware of the context in which activities are happening. The plotting can be compared to that of a fuzzy setting, in which the value in the Y-axis represents a degree of membership, in this case it represents a degree of usability.

V. Conclusion

We have defined the human activity recognition problem, and have described a new scheme 5W1H in which we allocate all the information we receive from the sensors to a specific set of variables. We have analyzed how each of the variables interact with each other and have seen the impact or possibilities in different settings of human activity.

We presented as well some plots of results we obtained using the 4W1H setting, that effectively prove that a clas-

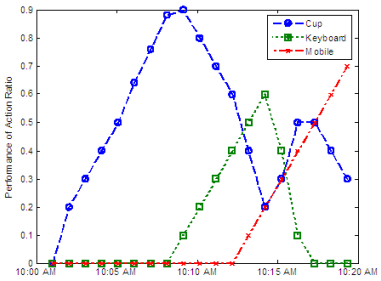


Fig. 10. Fuzzy Representation of the activity in the space

sification based on the discretization of data into specific variables (4W1H - 5W1H) allows us to have a deeper and more comprehensive analysis of the current situation within the space.

A. Future Work

In the future, we will use the 5W1H implementation to train and classify a probabilistic model that will take care of use all the available information, and have as its output the probability of the current main state given the current context, where that probability will be the best possible assessment for the current situation in the environment. To do this work, extensive studies in the areas of Machine Learning and Data Mining have to be undergone to asses which algorithms best fit the problem at hand.

REFERENCES

[1] J. H. Lee and H. Hashimoto, "Intelligent space concept and contents," *Advanced Robotics*, vol. 16, no. 3, pp. 265–280, 2002.

[2] D. Tran and A. Sorokin, "Human activity recognition with metric learning," *Computer Vision–ECCV 2008*, pp. 548–561, 2008.

[3] A. Madabhushi and J. Aggarwal, "A bayesian approach to human activity recognition," in *Visual Surveillance, 1999. Second IEEE Workshop on,(VS'99)*. IEEE, 2002, pp. 25–32.

[4] N. Oliver, E. Horvitz, and A. Garg, "Layered Representations for Human Activity Recognition," in *In Fourth IEEE Int. Conf. on Multimodal Interfaces*, 2002.

[5] C. Sminchisescu, A. Kanaujia, and D. Metaxas, "Conditional models for contextual human motion recognition," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 210–220, 2006.

[6] L. Palafox and H. Hashimoto, "A Movement Profile Detection System Using Self Organized Maps in the Intelligent Space," in *Tokyo, Japan: IEEE Workshop on Advanced Robotics and its Social Impacts*, 2009, p. 114.

[7] M. Niitsuma, K. Yokoi, and H. Hashimoto, "Describing human-object interaction in intelligent space," in *HSI'09: Proceedings of the 2nd conference on Human System Interactions*. Piscataway, NJ, USA: IEEE Press, 2009, pp. 392–396.

[8] K. Cheung, "Developing the interview protocol for video-recorded child sexual abuse investigations: A training experience with police officers, social workers, and clinical psychologists in Hong Kong," *Child abuse & neglect*, vol. 21, no. 3, pp. 273–284, 1997.

[9] J. Pardey, S. Roberts, L. Tarassenko, and J. Stradling, "A new approach to the analysis of the human sleep/wakefulness continuum," *Journal of Sleep Research*, vol. 5, no. 4, pp. 201–210, 1996.

[10] L. Palafox, L. Jeni, H. Hashimoto, and B. Lee, "Recognizing Facial Expressions in the Intelligent Space," in *Proc. on the 2010 International Symposium on Intelligent Systems*, 2010.

[11] L. Wiskott, J. Fellous, N. Kuiger, and C. Von der Malsburg, "Face recognition by elastic bunch graph matching," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 7, pp. 775–779, 2002.

[12] M. Turk and A. Pentland, "Face recognition using eigenfaces," in *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*. IEEE, 2002, pp. 586–591.

[13] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*. IEEE, 2002, pp. 84–91.

[14] D. Brscic and H. Hashimoto, "Tracking of humans inside intelligent space using static and mobile sensors," in *Proc. of the 33th Annual Conference of the IEEE Industrial Electronics Society (IECON'07)*, 2007, pp. 10–15.

[15] A. Pronobis, B. Caputo, P. Jensfelt, and H. Christensen, "A discriminative approach to robust visual place recognition," in *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*. IEEE, 2007, pp. 3829–3836.

[16] J. MacCormick and A. Blake, "A probabilistic exclusion principle for tracking multiple objects," *International Journal of Computer Vision*, vol. 39, no. 1, pp. 57–71, 2000.

[17] P. Duygulu, K. Barnard, J. De Freitas, and D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," *Computer VisionECCV 2002*, pp. 349–354, 2006.

[18] Xsens, "Specification sheet," World Wide Web electronic publication, 2006. [Online]. Available: http:/www.xsens.com

[19] L. Palafox and H. Hashimoto, "A compressive sensing approach to the 4w1h architecture," in *Industrial Technology (ICIT), 2010 IEEE International Conference on*, 3 2010, pp. 1599 –1604.

[20] P. Ekman, W. Friesen, and J. Hager, *Facial action coding system*. Consulting Psychologists Press Palo Alto, CA, 1978, vol. 160.

[21] L. Palafox and H. Hashimoto, "Human action recognition using 4W1H and Particle Swarm Optimization Clustering," in *Human System Interactions (HSI), 2010 3rd Conference on*. IEEE, 2010, pp. 369–373.