Actions in Context: System for people with Dementia

Àlex Pardo[1], Albert Clapés[1,2], Sergio Escalera[1,2], and Oriol Pujol[1,2]

[1]Dept. Matemàtica Aplicada i Anàlisi, UB, Gran Via 585, 08007 Barcelona.
[2]Centre de Visió per Computador, Campus UAB, Edifici O, 08193 Barcelona.
alexpardo.5@gmail.com, aclapes@cvc.uab.cat, sergio@maia.ub.es, oriol_pujol@ub.edu

Abstract. In the next forty years, the number of people living with dementia is expected to triple. In the last stages, people affected by this disease become dependent. This hinders the autonomy of the patient and has a huge social impact in time, money and effort. Given this scenario, we propose an ubiquitous system capable of recognizing daily specific actions. The system fuses and synchronizes data obtained from two complementary modalities - ambient and egocentric. The ambient approach consists in a fixed RGB-Depth camera for user and object recognition and user-object interaction, whereas the egocentric point of view is given by a personal area network (PAN) formed by a few wearable sensors and a smartphone, used for gesture recognition. The system processes multi-modal data in real-time, performing paralleled task recognition and modality synchronization, showing high performance recognizing subjects, objects, and interactions, showing its reliability to be applied in real case scenarios.

Keywords: Multi-modal Data Fusion, Computer Vision, Wearable Sensors, Gesture Recognition, Dementia

1 Introduction

The number of people living with dementia increases every year. According to The World Alzheimer Report, in 2010 there were 35.6 million dementia affected people and this number is expected to increase to 65.7 million by 2030 and 115.4 million by 2050. The most common type of dementia is Alzheimer's disease. It mainly affects elder people and the most common symptom of this disease is lack of awareness. This hinders the autonomy of the patient because he could not remember where he is, what he had already done, or what he has to do. As a result, he requires constant attention. Intelligent systems help patients to reduce their dependence on carers and improve their quality of life in the early stages of the disease. In this paper, we present an ubiquitous system to assist people with dementia. Our framework can be split into two main modalities, an "ambient" one based on multi-modal visual data for user, object, and user-object

relationship recognition, and an "egocentric" one based on wearable sensors to model user gestures.

From an ambient perspective, different visual-based sensors are placed in the environment and computer vision and pattern recognition techniques are applied. In this field, different ubiquitous systems have been recently proposed in order to assist people with dementia, most of them focused on the detection of falls and risky events. In [2], the authors present a privacy preserving automatic fall detection method. The system is able to recognize five different scenarios (standing, fall from standing, fall from chair, sit on chair, and sit on floor) based on the 3D depth information provided by a 3D camera. Whereas the former system uses quite simple features, more refined approaches to the fall detection have been proposed. In [3], the authors use fuzzy clustering techniques to accurately detect the activity (sitting, being upright, or being on the floor) performed by an elder. In [1], a system for passive fall risk assessment in home environments using Microsoft® Kinect™(RGB-Depth camera) is presented. The risk is evaluated obtaining measurements of temporal and spatial gait parameters. Most of these works are based on background subtraction and people tracking techniques [8]. And, with the increasing adoption of depth sensor data, novel multi-modal RGB and depth object descriptors are being proposed [5, 6].

From an egocentric perspective, previous works on wearable sensors and health-care are based on activity recognition for promoting active lifestyles and prevent related diseases [9,10]. The authors of [12] propose a Smart Reminder of intended activities for people with dementia. In this case, Dynamic Time Warping (DTW) [13] algorithm is applied to perform activity recognitions. DTW is also applied in [14] in order to perform context recognition by means of data acquired from wearable sensors.

Finally, few works have been recently presented combining the power of ambient and egocentric paradigms within the same framework. In [16] a single wearable sensor and a blob-based vision system is used in order to identify daily activities, such as walking, sitting, standing, laying down. The work shows improved recognition results when both modalities are jointly considered. In [17], bicycle maintenance activities are recognized using an ultrasonic hand tracking system and motion sensors placed on the arms. In [18], the authors use a camera worn by the user and an inertial sensor fused data to estimate the pose and create mixed reality scenes. Nevertheless, most of previous works use high complexity algorithms to fuse the data, such as Neural Networks, Hidden Markov Models [19], Gaussian Mixture Model Bayes classifiers [16] and Extended Kalman Filter in [18]. In those cases, the complexity of the algorithms make the reliable and non-invasive implantation of a real-time home-care ubiquitous system for assistance for the elder and people with dementia difficult.

In this paper, we propose a multi-modal fusion pipeline for real-time user detection, object detection, object identification, and user-object relationship recognition using computer vision by analyzing data from a RGB-Depth camera (Microsoft® Kinect™) combined with gesture recognition by means of processing streaming data from a 9-degrees-of-freedom wearable IMUS sensors

(Shimmer®) which measure acceleration, angular speed and magnetic field strength and direction. In order to recognize the subject environment (context), the computer vision part models a RGB-D environment learning a Gaussian distribution for each pixel from a set of initial frames. Then, in each new frame, the foreground is segmented according to the confidence with respect to the learnt environment. From the depth image, subjects are also detected and tracked using robust depth features and a Random Forest classifier. The previously segmented foreground regions, not classified as subjects are considered to be objects. These object regions are described, matched, and recognized using 3D descriptors of normal vectors distributions (FPFH). In the Personal Area Network scenario (PAN), actions are recognized by means of a parallel proposed version of Dynamic Time Warping over the spatio-temporal measurements given by the wearable sensors. Thus, given an interaction between a patient and an object, we can determine which action is done by fusing both modalities. The proposed system is simple yet efficient, runs in real-time and has a high performance rates. The presented ubiquitous system is successfully evaluated on real multi-modal data simulating real use-case scenario for people with dementia taking a medication.

The rest of the paper is structured as follows: Section 2 explains the proposed system. Section 3 describes the data, scenarios, settings and validation measurements and the results. Finally, conclusions are given in Section 4.

2 System

In this section, we present our multi-modal activity recognition system for people with dementia. Actions are meaningful with respect to the context where they took place. Actions in Context means mixing ambient and egocentric features, giving the system a two-sided reality, one from the perspective of the environment and the other from the point of view of the user.

Fig. 1 shows the architecture of the proposed system. In the egocentric computing part, we have an IMUS Sensor connected with an smartphone using a Bluetooth® connection. This mobile device is used as a HUB and its function is to label every sample with a time-stamp and stream it using IP to the server which will process all the data in real-time. This part of the application is relative to subject who is wearing the inertial sensors. The ambient part consists on a RGB-D camera monitoring the scene. This device gives us information about color and depth. Processing this information allows us to be able to perform user and object detection and recognition. Both data streams are synchronized by means of the NIST Internet Time Service. The different modules of our system are described in detail next.

2.1 Egocentric Computing

This part of the system is centered on the gestures the user is performing with his/her dominant arm (left part of block shown in Fig. 1). For this task, we first compute a set of inertial features which are then used to perform gesture recognition based on a parallel version of DTW.

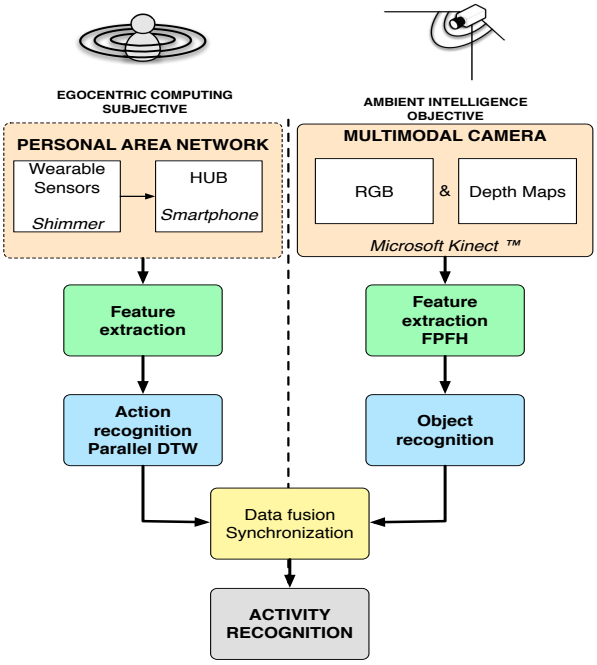


Fig. 1: Block diagram of the overall system.

Egocentric Features. Sensors used in the system have nine degrees of freedom (accelerometer, gyroscope and magnetometer with 3-axis each one). Selected features are raw inputs given from accelerometer and gyroscope and also their acceleration and angular speed energies. Magnetometer data is disregarded since we found it is not useful in gesture recognition.

Parallel Dynamic Time Warping. The problem of processing the information in real-time is solved using a variant of Dynamic Programming. The system has to process a large amount of input data (50 samples per second with 8 values and a time-stamp). In this sense, we use a parallel version of Dynamic Time Warping (DTW) [15] in order to speed up the activity recognition process. DTW algorithm provides a simple and fast way of aligning sequences. It computes the minimum path reconstruction of the input given a pattern. It is simply parallelized by taking each new sample as a possible beginning of a gesture. For each sample there will be a new thread processing it. We propose to store only the last two iterations and a reference to the beginning of the sample (i.e. the time-stamp). Then, when a pattern is accepted, we have the ending sample (with its time-stamp) and the reference we previously stored at the creation of the thread. This effectively identifies the gesture duration. Squared weighted Euclidean distance between the two eight-dimensional vectors is used as a composed metric.

The update rule is as follows,

$$M(i,j) = min(M(i-1,j), M(i,j-1), M(i-1,j-1)) + d(p_i, s_j),$$
$$d(p_i, s_i) = (p_i - s_i)^T(\omega^T I)(p_i - s_i),$$

where s_i is a sample vector, p_i is a pattern vector, M is the dynamic programming cost matrix, and I is the identity matrix. The weight vector ω is used to balance the importance of the accelerometer energy and gyroscope energy and to account for the different dynamic ranges of each feature. In order to speed up the performance of the algorithm, all threads computing a partial result higher than the acceptance threshold are finished at once.

2.2 Ambient Intelligence

The ambient module of the system is capable of detecting subjects that appear in the scene as well as to detect and recognize new objects appearing in the environment. This part is composed of 4 main submodules: environment modeling, user detection, object detection and recognition, and user-object interaction analysis.

Environment modeling A background subtraction strategy is applied. This allows to learn an adaptive model of the scene [6]. Given an initial set of multimodal frames, each one composed of a RGB image and a range image obtained from the Kinect™ device, a gaussian-distribution for each pixel is modeled[1]. Once the background has been modeled at pixel level, the apparition/removal of objects in the scene is detected whenever the pixels in a region show an absolute difference bigger than their learnt confidence values, i.e. greater than δ standard deviations.

User detection At each new frame, we perform user detection by segmentation using Random Forest approach on depth data. For this task, each cloud voxel captured from the scene is evaluated by a forest of trees trained on offsets of depth features. As a result, we obtain a user pseudo-probability for each point in the scene. From this, the user can be detected and an skeletal model as a spatial configuration of body limbs defined [4].

Object recognition Let the segmented image contain 1 at positions detected as foreground "objects" by the background subtraction ambient module. Each connected component of the segmented image which has not been classified as user is considered as a new object whenever its distance to the camera is smaller than the one obtained for the modeled background. In that case we compute a

[1] Note that due to the enrichment given by range data a single Gaussian model suffices for modeling background. Very small improvements have been observed using Gaussian Mixture Models in the studied environments.

normalized description of that particular 3D object view using the Fast Point Feature Histogram (FPFH) [7]. The FPFH is a point-wise 3D descriptor which describes the relative orientation of each point surface normal vector with respect to the average normal vector of the points in the k-neighborhood and encodes the information in a one-dimensional histogram. Those relative orientations are, basically, the roll, pitch, and yaw rotations discretized into 11 bins each one, summing up to a 33-bin descriptor. When a new object is detected in the scene, the descriptor for each point is computed to describe that particular object view. This description is compared to the data set of object descriptions using k-NN to classify the object.

2.3 Fusing Egocentric and Ambient Intelligence

Outputs from egocentric computing and ambient modules are synchronized in order combine the user, object, and activity recognition performed by both modalities. The synchronization is made using time-stamps adjusted with the NIST (National Institute of Standards) Internet time. Since an activity is defined as an action performed with a specific object, the intersection between object interaction and gesture recognition defines the activity the user is performing.

3 Results

In order to present the results, first, we describe the hardware, data and settings of the system in order to perform the experiments.

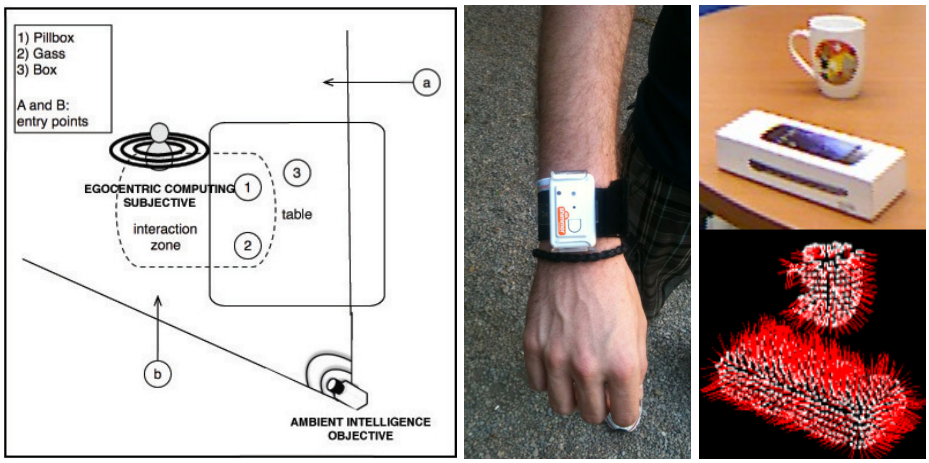
3.1 Hardware

The system is composed by a RGB-Depth camera (i.e. Microsoft® Kinect™) and a 9-degrees-of-freedom wearable Shimmer® device (Fig. 2), that includes triaxial accelerometer, gyroscope and magnetometer, Bluetooth® communication, up to 50Hz sampling rate and the possibility to include additional modules (ECG, GPS, etc.). Also we used a Samsung GT-I9250 smartphone to stream the data to the PC running the server on a quad-core processor equipped with 8GB of RAM memory.

3.2 Data

We defined a dataset containing 13 different multi-modal synchronized recordings in 4 different types of scenes, with a length between 30 and 60 seconds. As a case of study for people with dementia, we include a scenario where the patient is taking his/her medication (a pill from a pillbox). This enables the evaluation of different objects, users and interactions in several scenarios.
In this case, the scenario has a table with three objects on it: a pillbox, a glass of water and a box. The camera is pointing towards the pillbox and the glass so the user will appear from one of the sides (see Fig. 2(a)).

We have defined four possible scenarios for the validation: An unknown subject leaves the pillbox and a glass on the table. The patient appears in the scene. Then,



(a) (b)

Fig. 2: (a) Setup of the ubiquitous system in a test scenario (b) On the left Shimmer® device tied to the arm of the user. On the right, the point clouds with the estimated surface normals of the objects (used in the FPFH object descriptior computation).

1. He takes the pillbox standing sideways from the camera, puts the pill on the mouth, leaves the pillbox on the table, drinks water, and leaves the room.
2. He takes the pillbox facing the camera, turns so that the camera cannot see what is doing, puts the pill on the mouth, leaves the pillbox on the table, drinks a bit of water, and leaves the room.
3. He takes the pillbox facing the camera, turns so that the camera cannot see what is doing, puts the pill on the mouth, leaves the pillbox on the table, drinks a bit of water, and leaves the room with the glass on the hand.
4. He takes the pillbox facing the camera, puts the pill on the mouth, leaves the pillbox on the table, drinks a bit of water, and leaves the room.

Although there are more than one scenario, the system only recognizes a single gesture[2]. Two important constraints of these recordings are: first, the camera has to see the patient taking the pillbox and leaving it to be able to detect the interaction. And second, the pill has to be taken by using the dominant arm, i.e. the one wearing the sensor. All objects, users and actions have been manually annotated in order to validate the performance of the system. The scenarios were performed by a single user without dementia.

3.3 Settings

Parallel Dynamic Time Warping threshold is set using leave-one-out cross-validation to value 5.7 10^7, weights w used to compute the distance in DTW are experimentally set to 0.8 for raw data and 0.2 for the energy measures. 400 frames

[2] The extension to a small set of gestures of interest can be easily achieved without a significant loss in performance [11].

are used to learn a background model. Pixels segmented in each frame are those with change in its value with respect to the learnt model greater $\delta = 1.15$ standard deviations. A set of objects of interest is defined offline. Once a new object appears in the scene, it is classified as the nearest object minimizing he distance among the corresponding FPFH descriptors (see Fig. 2 (b)) if this value is lower than 0.5 (otherwise it will be considered an unknown object).

3.4 Validation

The system is aimed to recognize activities. Recognizing an activity involves two main processes, knowing the object which the user is interacting with and also understanding the gesture performed during this interaction. In order to validate the system, we considered an overlapping measure, the Jaccard Index ($J = \frac{A \cap B}{A \cup B} = \frac{\mathrm{TP}}{\mathrm{TP+FP+FN}}$).

The different parts of the system work as follows:

- The interaction begins with the change of depth in the region of the object of interest. That is, moving the object out of its bounding box of the static position. The detection may take some frames from the beginning of the interaction (pick up event).
- In order to detect a new object, it has to be dropped and remain static for a few frames. This delay constraint in the object detection is set for the sake of robustness.
- Gesture detection is performed when the hand of the user starts a motion towards the mouth and returns to the first position.[3]

Fig. 3 shows the average Jaccard index split in the four different scenarios, compared to the average system performance. Three bar sets show the individual performance measurements using only ambient data, egocentric data and the fusion of both. Observe that ambient analysis (multi-modal vision) achieve lower results. That is because in the vision system the taking-a-pill action is defined by the pick-up and release events but the action could be shorter if the users keep holding the object in his/her hands. This decreases the accuracy of the recognition since the detection could be done before the action begins or could exceed the action end. Additionally wearable sensor data usually performs well but not as well as the fused version. Notice the performance difference between scenarios 1-4 and 2-3 in the accuracy of the sensor. This happens when the subject is turning, that increases the length of the gesture of taking the pill. The fusion of the system maintains its robustness thanks to the help of the camera in the elimination of these outliers. Fig. 4(a)-(b) shows a real example of the system working. The interface of the application shows the depth map, the detected user, the detected objects in the scene, and detects the action of taking the medication (red square in (b)) shown by the inertial sensor features on the bottom of the images. Fig. 4(c) shows an example of how fusion decreases the number of false positives. The patient touching his mouth is similar to taking

[3] Notice that the drinking action is not detected because the system is sensitive to the hand orientation.

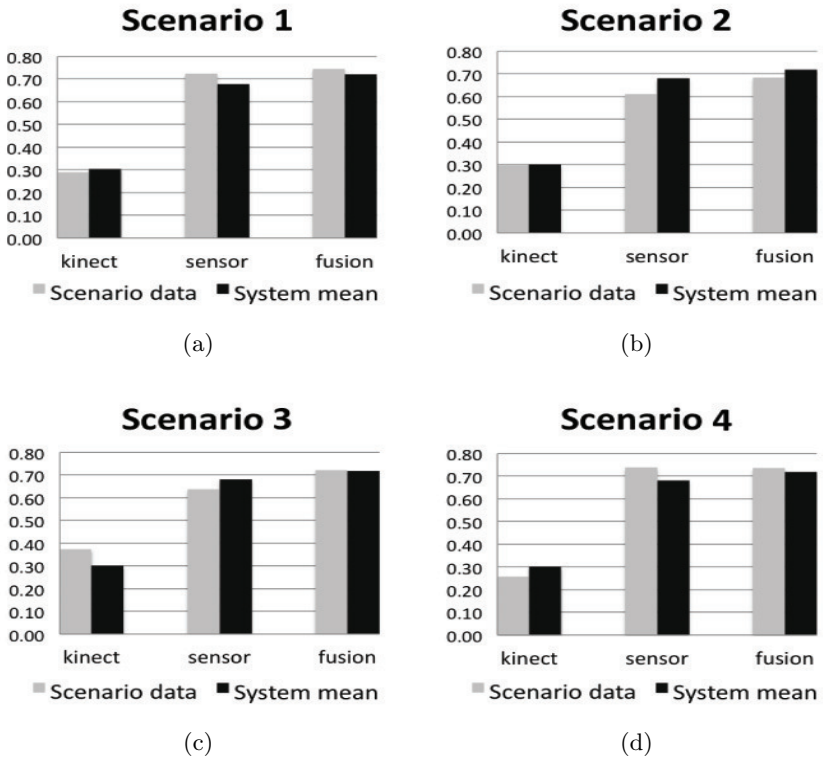


Fig. 3: Mean Jaccard value for each of the scenarios.

a pill, thus it would be recognized by the sensor but not by the camera; in the fused data it would be a negative response.

Next, we include a discussion about the performance and reliability of the proposed system performing subject and object detection, object identification, and interaction and gesture recognition.

– The **subject detection** in RGB-D data by means of Random Forest (RF) has become a standard approach, being also exploited in commercial entertainment systems as Microsoft® XBOX360™. This technique provides very accurate results, and in our case it is reinforced by the usage of a background subtraction technique, since those regions not subtracted from the background are not considered people even if the RF detector gives a false positive, and thus discarded. From this, we cope with almost all the possible subject detection false positives except for the ones in non-human segmented foreground which actually never occur since we deal with small objects that differ from the random depth features learnt by the algorithm. Although the method can still give very few false negatives, it turns out to be very fast and accurate approach.

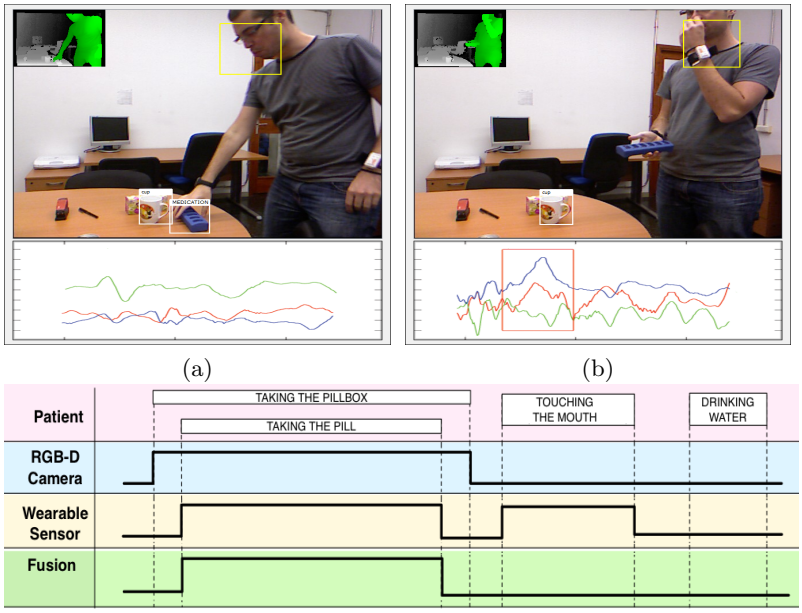


Fig. 4: (a) and (b) show object and gesture recognition.(c) Is an example of the fusion.

- The **object detection** performance is highly dependent on the capability of the system to accurately model the environment. Even though the infrared sensor of the Kinect™ device provides quite noisy measures, its price makes it a very good device to be considered in a wide range of applications. Because of these hardware limitations, the depth measures in a given pixel, even with the device fixed, range considerably. However, the variance in a certain pixel follows the same distribution throughout time, which we assumed to be normal. Then, to correctly model the pixels' variability a considerably large set of frames (400) is needed. Due to the noisy acquisition, very small objects can not be correctly modeled. Despite this, we have seen the system is able to detect objects with high precision if their area greater than 150 pixels in the dense depth map.
- The **object recognition** part presents the typical difficulties when performing in noisy and low-resolution images, as the case of dense depth images. One of the main limitations of the presented system is scalability, i.e. considerably increasing the number of objects to discriminate. In many applications, such as the presented one, we are interested in being able to classify among a small set of objects of interest. In this scenario, the method has shown a high classification accuracy performance.
- From the point of view of ambient intelligence, the **subject interaction with objects** task depends on the goodness of both the background subtraction and the subject detection. Since both of them perform correctly,

and the interaction is straightforward to detect, the results are also accurate and precise.
- In the egocentric computing area, the **action recognition** is very accurate and delimits the gesture with high precision. However, in a future work we plan to include additional interaction and gesture recognition categories in order to analyze the scalability of generalization capability of both multimodal vision and egocentric features.

4 Conclusion

We proposed to fuse egocentric and ambient features to define an integrated ubiquitous system to model daily actions of people with dementia. From the point of view of ambient intelligence, we learned a pixel-based Gaussian distribution of the background. Foreground segmentation is used to detect and recognize both user and objects based on 3D descriptor and statistical classifiers. From the egocentric point of view, we used a set of movement features computed from wearable sensors to test a parallelized Dynamic Time Warping gesture recognition method. We showed that fusing ambient and egocentric modalities provides a fast and robust system with high application potential, capable of recognizing different everyday activities involving different objects under the natural conditions of indoor scenes.

Acknowledgments

This work has been partly supported by RECERCAIXA 2011 Ref. REMEDI and TIN2009-14404-C02.

References

[1] E. E. Stone and M. Skubic. Pervasive Computing Technologies for Healthcare (PervasiveHealth): Evaluation of an inexpensive depth camera for passive in-home fall risk assessment. pp. 71-11, 2011.
[2] C. Zhang, Y. Tian, and E. Capezuti. Proceedings of the 13th international conference on Computers Helping People with Special Needs: Privacy preserving automatic fall detection for elderly using RGBD cameras. 2012.
[3] Banerjee, T. Banerjee, J. Keller, J. Skubic, and E. E. Stone. Fuzzy Systems, IEEE Transactions: Day or Nigh Activity Recognition from Video Using Fuzzy Clustering Techniques. pp. 1-1, 2013.
[4] J. Shotton, A. Fitzgibbon, M. Cook, et. al. CVPR: Real-Time Human Pose Recognition in Parts from Single Depth Images. pp. 1297-1304, 2011.
[5] S. Escalera, Human Behavior Analysis From Depth Maps: Articulated Motion and Deformable Objects 2012. pp. 282-292, 2012.
[6] A. Clapés, M. Reyes, and S. Escalera, Pattern Recognition Letters 34(7): Multi-modal user identification and object recognition surveillance system. pp. 799-808, 2013.

[7] R.B. Rusu, N. Blodow, and M. Beetz, Fast Point Feature Histograms (FPFH) for 3D Registration. The IEEE International Conference on Robotics and Automation (ICRA). 2009. Kobe, Japan.
[8] P. F. Felzenszwalb, D. A. McAllester, and D. Ramanan, CVPR: A discriminatively trained, multiscale, deformable part model. pp. 1-8, 2008.
[9] M. Ermes, J. Pärkkä, J. Mäntyjärvi, and I. Korhonen, TITB: Detection of Daily Activities and Sports With Wearable Sensors in Controlled and Uncontrolled Conditions. Vol. 12, no. 1, 2008.
[10] K. Ouchi, T. Suzuki, and M. Doi, Distributed Computing Systems, A Wearable Healthcare Support System Using User's Context. pp. 791-792, 2002.
[11] J. Lichtenauer, E. Hendriks, M. Reinders, IEEE Trans on Pattern Analysis and Machine Intelligence, Sign Language Recognition by Combining Statistical DTW and Independent Classification, pp. 2040-2046, 2008.
[12] S. Jiang, Y. Cao, S. Iyengar, et. al., Body Area Networks: CareNet: An Integrated Wireless Sensor Networking Environment for Remote Healthcare. pp. 9:1-9:3, 2010.
[13] T. K. Vintsyuk, Kibernetika: Speech discrimination by dynamic programming. Vol. 4, pp. 81-88, 1968.
[14] K. Ming Hsiao, G. West, S. Venkatesh, and M. Kumar, ISNIPC: Online context recognition in multisensor systems using Dynamic Time Warping. pp. 283-288, 2005.
[15] H. Sakoe and S. Chiba, IEEE Transactions on Acoustics, Speech and Signal Processing, Dynamic programming algorithm optimization for spoken word recognition. Vol. 26(1) pp. 43- 49, 1978.
[16] J. Pansiot, D. Stoyanov, et. al., 4th Int W. on Wearable and Implantable Body Sensor Networks, Ambient and Wearable Sensor Fusion for Activity Recognition in Healthcare Monitoring Systems. pp. 208-212, 2007.
[17] T. Stiefmeier, G. Ogris, H. Junker, P. Lukowicz and G. Troster, Wearable Computers, 2006 10th IEEE International Symposium: Combining Motion Sensors and Ultrasonic Hands Tracking for Continuous Activity Recognition in a Maintenance Scenario. pp. 97-104, 2006.
[18] S. You and U. Neumann, Virtual Reality: Fusion of vision and gyro tracking for robust augmented reality registration. pp. 71-78, 2001.
[19] C. Zhu, W. Sheng, Pervasive and Mobile Computing 7: Motion- and location-based online human daily activity recognition. pp. 256-269, 2011.