User Activity Recognition via Kinect in an Ambient Intelligence Scenario

P. Cottone, G. Maida, M. Morana

In Proceedings of 2013 International Conference on Applied Computing, Computer Science, and Computer Engineering

Publisher: Elsevier

User Activity Recognition via Kinect in an Ambient Intelligence Scenario

Pietro Cottone, Gabriele Maida, Marco Morana*

DICGIM, University of Palermo, Viale delle Scienze, ed. 6, ITALY

Abstract

A great number of sensors are nowadays available, this has stressed the need for new approaches to merge low-level measurements to realize what facts they refer to in the real environment. Ambient Intelligence (AmI) techniques exploit information about the environment state to adapt the environment itself to the users' preferences. Even if traditional sensors allow a rough understanding of the users' preferences, ad-hoc sensors are required to obtain a deeper comprehension of users' habits and activities. In this paper we propose a framework to recognize users' activities via the Microsoft Kinect. The approach proposed here takes advantage of the position of some human body parts estimated by using Kinect depth information. In our system, significant patterns of joints (i.e., postures) are discovered by applying a clustering technique and then classified by means of a multi-class SVM. Each activity is then modeled as a sequence of known postures by using HMMs. A prototype has been implemented by connecting the Kinect to a miniature PC with limited computational resources. Experimental tests have been performed on a dataset we collected at our laboratory and results look very promising.

Selection and peer review under responsibility of Information Engineering Research Institute

Keywords: Ambient Intelligence; User Profiling; Activity Recognition; Kinect

* Corresponding author. Tel.:; fax:.
E-mail address: marco.morana@unipa.it.

1. Introduction

In recent years, the ever-increasing availability of innovative sensors has fostered the production of novel techniques for pervasively monitoring real environments. The goal of such a trend is to make the environment “smart”, that is to provide intelligent mechanisms to connect the user to the environment and vice versa.

Ambient Intelligence (AmI) is a novel research that aims at producing intelligent methods for adapting the environment characteristics to the users’ preferences [1] .

The method proposed here is part of an AmI system [2] designed to monitor an office environment in order to achieve the energy efficiency of the building while satisfying a number of constraints related to the preferences expressed by the user. In our architecture, the sensory component [3] is designed using [4] and consists of a Wireless Sensor and Actuator Network (WSAN) capable of measuring some of the most relevant environmental conditions (e.g., temperature, humidity, light [5]). However, even if traditional sensor nodes allow to capture the environmental characteristic preferred by the user, basic nodes are not sufficient to provide high-level information about the activities the user is performing; for this purpose, high-level vision sensors are a possible solution.

In this work we present a system to perform user activity recognition via the Microsoft Kinect sensor in some offices of our University. In our vision, the activities are described as sequences of different postures defined by the positions of some body joints detected by means of a skeleton tracking algorithm [6] . The most frequent patterns of joints positions (i.e., postures) are discovered by applying a clustering algorithm and then classified using Support Vector Machines. Hidden Markov Models (HMMs) are finally applied to build a representation of each activity as a sequence of known body postures.

In the considered scenario, the sensory system is pervasively deployed in different rooms of the Department; for this reason, the Kinect sensor is also connected to a fanless PC that allows for high levels of mobility and pervasiveness.

The paper is organized as follows: related works are outlined in Section 2, whilst the overall architecture proposed here is illustrated in Section 3. Experimental results are detailed in Section 4, and conclusions are discussed in Section 5.

2. Related Work

Recently, the question of human activity recognition has been analyzed in various works.

Many solutions have been based on the processing of color images captured by traditional cameras. In [7] , the authors used a set of binary images representing human silhouettes as input of a system based on HMMs. The silhouettes were obtained by processing RGB images, thus this method requires a great number of typical image processing steps (e.g., background removal) to produce a reliable output. Another activity recognition technique based on silhouettes and discrete HMMs is presented in [8] . The authors used Fourier analysis to describe the human silhouettes and Support Vector Machines (SVMs) [9] to classify them into different postures. Then, postures were described as the symbols emitted from the HMM in order to recognize different activities. Other works [10] [11] focused on the issue of activity recognition by analyzing the data captured by intrusive sensors, e.g., sensors that can be worn by the user.

Our perspective is to consider Kinect as a sensor to transparently gather observations about users’ behavior [12] .

The vision system of the Microsoft Kinect is composed of two cameras (i.e., an RGB camera and an IR camera) with 640x480 resolution, and an IR projector that is responsible of shooting infrared rays toward the environment. The distortion degree of each ray projected against the scene is used to estimate a depth map in which each pixel value represents the distance of a specific 3D point from the Kinect.

The Kinect has already been chosen as input sensor in some other works. A model for human actions representation using the Kinect is presented in [13] . Human bodies are represented in terms of joints and actions are expressed as the interactions that occur between subsets of these joints. Due to the great number of possible features, a data mining algorithm is used to discover the most discriminative features, called Actionlets, so that a specific action can be defined as an Actionlet Ensemble, i.e., a combination of Actionlets. In [14] a fully automatic and robust real-time system for dynamic hand gesture recognition is described. The authors propose an action graph based approach, which shares similar properties with standard HMM but requires less training data since it allows state sharing among different gestures.

3. System Overview

The system proposed here (see Fig. 1(a)) is designed to automatically discover the activity performed by the user by reasoning on a set of known features, i.e., postures. Firstly, the coordinates of a set of 3D points of the human body are detected by means of a skeleton tracker algorithm [4], then the K-means algorithm [15] is applied on this set in order to build a set of known postures. The obtained postures are validated by means of a Support Vector Machines (SVMs) classifier whilst Hidden Markov Models (HMMs) are finally used to express each activity as a sequence of known postures.

The OpenNI/NITE 1.5 skeleton detection method [6] is able to detect (i.e., to find the 3D coordinates) in real time the position of 15 body joints (see Fig. 1(b)). In the considered scenario, to overcome the noise of the IR sensor a subset of 11 joints was selected, discarding those whose detection is less reliable. In particular, some low-mobility joints (i.e., hips and shoulders) have not been considered as suggested by the results obtained in [12] , where Principal Component Analysis was applied to evaluate if and how a reduced feature space affects the system performance.

Since the relative distances of the detected joints depends on a number of factors related to the user (e.g., height, arms length) and its position in the scene, we moved the coordinates of the detected joints to a new coordinate system fixed at the torso (x-axis coinciding with the left-right hip axis) and the features has been scaled with respect to the distance between neck and torso. Reference joints (red dots) are shown in Fig. 1(b), the joints we use for our analysis are depicted in green, while the discarded ones are in gray.

After detecting the joints of interest, we applied the K-means algorithm is applied to reduce to K the number of the observed joint patterns, that is we build a vocabulary of K-words. By adopting this model, we can consider each posture as a specific word of the vocabulary, so that each activity can be modeled as a consistent sequence of words.

A better statistical description of each cluster is obtained by applying a SVM approach. In particular, the K-means produces as output the associations features/cluster that we use to train a multi-class SVM with Gaussian radial basis kernel function. Moreover, since sequences of joint configurations are transformed into a sequence of K-words, we obtain that all the consecutive occurrences of the same posture are merged. In this way a more efficient representation is also obtained allowing for recognizing different instances of activities performed with variable time durations.

Next, we modeled each activity by means of a discrete Hidden Markov Model (HMM) [16] . Each HMM is trained using known posture sequences, then a new (unknown) activity is classified according to the largest posterior probability computed by testing the corresponding posture sequence against the set of HMMs. If a unreliable (i.e., low) probability is measured, the sequence is marked as “unknown”.

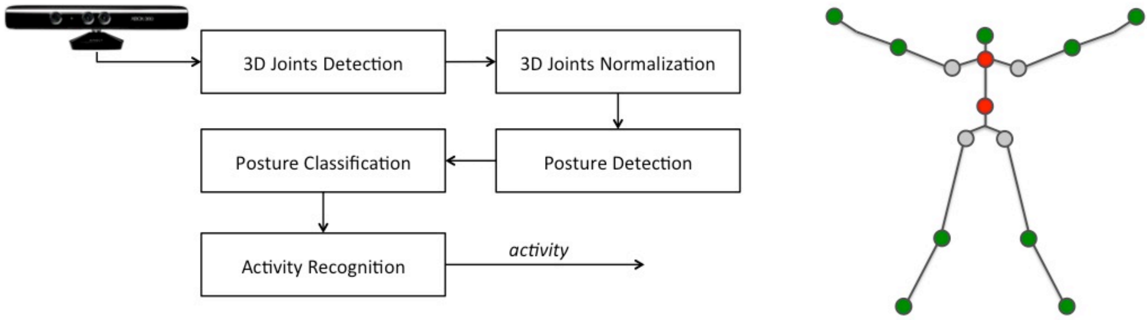

Fig. 1. (a) System Overview; (b) The 15 joints detected by the OpenNI/NITE skeleton tracker. Reference joints (red): neck, torso. Selected joints (green): head, elbows, hands, knees, feet. Discarded joints (grey): shoulders, hips.

4. Results

The method proposed in this paper is part of a bigger system that aims at controlling the actuators installed in a building to create and maintain environmental conditions according to both users preferences and power consumption constraints. In this context, the Kinect device is a node of the sensor network and the output of the activity recognition module is one of the inputs of the AmI system which reasons about different information coming from the sensing infrastructure to infer what the user is doing.

In [12] an early version of the our activity recognition framework has been evaluated on the public MSR Action3D dataset [17] . Since the quality of existing public datasets is often poor, we decided to collect a new dataset which contains 8 activities (*Catch Cap, Toss Paper, Take Umbrella, Walk, Phone Call, Drink, Sit down, Stand up*), each performed 3 times by 10 different subjects.

Several tests have been performed on the 240 captured in order to verify the accuracy and the robustness of the proposed solution.

In particular, the experimental tests started by applying a Grid Search approach to find out the best couple of values for the number of clusters K (i.e., the number of postures) and the number of the HMM states N. The value of each node of the grid has been computed as the mean rate of a Leave One Out Cross Validation (LOOCV) repeated ten times to overcome the randomness of the clustering algorithm. The best recognition rate is obtained with K = 39 and N = 5, with a mean accuracy of 95% and standard deviation of 2.45 between the different runs of the LOOCV.

Motivated by the results obtained over the whole dataset, we focused on investigating how and how much the chosen training set can impact on system performance. For this reason, the whole dataset is divided into subsets and each subset is tested three times in a way similar to the one used in [17] :

- 1/3 Validation: 1/3 of the data captured for each subject is used for training, the remaining part is used for testing;
- 2/3 Validation: 2/3 of the data captured for each subject is used for training, the remaining part is used for testing;
- Cross Subject Validation: 1/2 of the subjects is used for training and the remaining part for testing.

Each of the above tests was repeated ten times, randomly choosing the sequences or subjects of the training and testing sets. The results of the three performed tests are shown in Table 1. The first two rows report accuracy values of 93.75% and 94.87% respectively, which are comparable to the mean accuracy of 95% obtained over the whole dataset. The most significant result is the one obtained by the cross subject test (bottom row) that aimed to measure the ability of the system in recognizing activities performed by new subjects. In fact, the achieved recognition rate of about 91% shows that the method proposed is able to capture a general model of the activity regardless to the user that performed it.

The methods we developed has been implemented and verified using MATLAB and LIBSVM 0. A prototype of the activity recognition module has been realized connecting the Kinect to a miniature PC equipped with Intel Atom Z530 1.6GHz CPU and Linux OS with kernel 2.6.32. This minimal setup allowed to perform real-time processing of the observed scene with minimum levels of obtrusiveness and low power consumptions. In fact, the prototypal version of the system, implemented in JAVA, takes a mean processing time (i.e., consisting of both posture analysis and activity recognition) of about 1 second, whilst the power consumption is about 7W, that is just 1W more than the consumption required during idle.

Table 1. Recognition rates obtained while using three different training sets.

	Accuracy (%)
1/3 Validation	93.75%
2/3 Validation	94.87%
Cross Subject Validation	90.98%

5. Conclusions

In this work we presented a system for the automatic recognition of user activities via the Kinect sensor in an office permeated with small pervasive sensor devices.

We started by estimating the position of some joints, i.e., points at which parts of the human body are joined, by using 3D information provided by the Kinect. The proposed methods recognize the activity performed by the user according to a set of known patterns. Such patterns, called postures, are automatically defined by clustering training data into k clusters and classified by means of SVM-based approach. Each activity is finally modeled by a HMM built on known posture sequences.

Several tests have been performed on a dataset we collected in order to verify the accuracy and the robustness of the system. In particular we focused on investigating both the capacity of the system for distinguishing between similar activities and the scalability of the proposed approach.

Results showed that the system is able to capture a general model of the action regardless to the user that performed it. In particular, we succeed in modeling an activity independently of its time duration or who performs it. The solution proposed resulted to be scalable, that is stable results can be obtained considering a great number of actions, and expandable with new actions not previously recorded. Moreover, the prototypal setup we built demonstrated that the proposed method can be executed on resource-constrained devices providing near real-time processing of the observed scene.

We are currently working on collecting more data for an extended version of our dataset, with additional gestures and subjects, in order to make it publicly available for download and comparisons.

Acknowledgements

This work is supported by the SMARTBUILDINGS project, funded by POR FESR SICILIA 2007-2013.

References

[1] De Paola, M. La Cascia, G. Lo Re, M. Morana, and M. Ortolani. User Detection through Multi-Sensor Fusion in an AmI Scenario. In Proc. of the 15th International Conference on Information Fusion, pages 2502–2509. Published by the IEEE Computer Society, 2012.

[2] A. De Paola, G. Lo Re, M. Morana, and M. Ortolani. An Intelligent System for Energy Efficiency in a Complex of Buildings. In Proc. of the 2nd IFIP Conference on Sustainable Internet and ICT for Sustainability, 2012.

[3] A. De Paola, S. Gaglio, G. Lo Re, and M. Ortolani. Sensor9k: A testbed for designing and experimenting with WSN-based ambient intelligence applications. Pervasive and Mobile Computing. Elsevier, 8(3):448–466, 2012.

[4] A. Lalomia, G. Lo Re, and M. Ortolani. A hybrid framework for soft real-time WSN simulation. In Proceedings of the 13th IEEE/ACM International Symposium on Distributed Simulation and Real Time Applications, 2009. DS-RT '09, pp. 201-207

[5] G. Anastasi, G. Lo Re, and M. Ortolani. WSNs for structural health monitoring of historical buildings. In Human System Interactions, 2009. HSI '09. 2nd Conference on, pages 574 –579, may 2009.

[6] PrimeSense. Openni. http://www.openni.org/.

[7] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In Computer Vision and Pattern Recognition, 1992. Proceedings CVPR '92., 1992 IEEE Computer Society Conference on, pages 379 –385, jun 1992.

[8] M. Pietikainen V. Kellokumpu and J. Heikkila. Human activity recognition using sequences of postures. In Proc IAPR Conf. Machine Vision Applications, pages 570 – 573, 2005.

[9] Bernhard Scholkopf and Alexander J. Smola. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA, 2001.

[10] Stephen J. Preece, John Y. Goulermas, Laurence P. J. Kenney, Dave Howard, Kenneth Meijer, and Robin Crompton. Activity identification using body-mounted sensors – a review of classification techniques. Physiological Measurement, 30(4):R1– R33, 2009.

[11] Ling Bao and Stephen S. Intille. Activity recognition from user-annotated acceleration data. In Pervasive Computing, volume 3001 of Lecture Notes in Computer Science, pages 1–17. Springer Berlin Heidelberg, 2004.

[12] P. Cottone, G. Lo Re, G. Maida, and M. Morana. Motion sensors for activity recognition in an ambient-intelligence scenario. In PerCom Workshops 2013, pages 646–651, 2013.

[13] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 1290 –1297, june 2012.

[14] Kurakin, Z. Zhang, and Z. Liu. A real time system for dynamic hand gesture recognition with a depth sensor. In Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European, pages 1975–1979, 2012.

[15] J. Macqueen. Some methods for classification and analysis of multivariate observations. In 5-th Berkeley Symposium on Mathematical Statistics and Probability, pages 281–297, 1967.

[16] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2):257 –286, feb 1989.

[17] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, pages 9 –14, june 2010.

[18] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011.