Activity Recognition in Meetings with One and Two Kinect Sensors

Ramon F. Brena(✉) and Armando Nava

Tecnologico de Monterrey, Monterrey, Mexico
ramon.brena@itesm.mx, armandnavao@gmail.com

Abstract. Knowing the activities that users perform is an essential part of their context, which become more and more important in modern context-aware applications, but determining these activities could be a daunting task. Many sensors have been used as information source for guessing human activity, such as accelerometers, video cameras, etc., but recently the availability of a sophisticated sensor designed specifically for tracking humans, as is the Microsoft Kinect has opened new opportunities. The aim of this paper is to determine some human activities, such as eating, reading, drinking, etc., while a group of persons are seated, using the Kinect skeleton structure as an input. Further, due to occlusion problems, it could be guessed that a combination of two Kinect sensors could give an advantage in activity recognition tasks, especially in meeting settings. In this paper, we are going to compare the performance of a two Kinect system against a single Kinect in order to determine if there is a significant advantage in using two sensors. Also, we compare several classifiers for the activity recognition task, namely Naive Bayes, Support Vector Machines and K-Nearest Neighbor.

Keywords: Kinect · Naive Bayes · Support Vector Machines · K-Nearest Neighbor · KNN · SVM · NB

1 Introduction

In recent years, there has been an increased interest in recognizing human activities [1], as the user activity is an essential element of her/his context: if the system can recognize and understand the activities performed by a human, they can provide assistance to perform those activities [2]. This is indeed the focus of Ambient Intelligence research (AmI). For example, AmI systems can be used to monitor elder people, perceive their needs and preferences, provide then different services to comfort or helping them applying emergency treatment [3] etc.

Many sensors have been used for activity recognition, including accelerometers [4], microphones video-cameras [5] and other. In recent years, the commercial implementation of the Natal project by Microsoft, that is, the Kinect [6], though it was originally intended as a video-game accessory, started to be used as an experimental human-tracking sensor. The Kinect has a camera with other sensors besides the RGB camera (like the infrared camera and the infrared projector).

J.F. Martínez-Trinidad et al. (Eds.): MCPR 2016, LNCS 9703, pp. 219–228, 2016.
DOI: 10.1007/978-3-319-39393-3_22

The Kinect became popular with researchers because it is useful for AmI research, and because it is cheap.

The capability of tracking the human skeleton in real time with the Kinect sensor makes it possible to guess physical human activities, instead of using computer vision algorithms to the raw video image. So, the Kinect sensor takes part of the processing burden, and delivers a high-level data structure representing the human skeleton, which can be further analyzed by activity-recognition algorithms, allowing these to become the main focus of the research.

We focus on tracking the skeletons of seated persons with one and two Kinect sensors (Fig. 1), and recognize what activities are being performed by them, because this corresponds to meeting situations in enterprises, government, schools, etc. The analysis of activities of participants in meetings could give valuable information concerning the level of focus of participants in a presentation, the level of participation in a discussion, etc.

Tracking seated persons, and recognizing what activity is being performed by each person, is a challenging task, because some several different activities look almost the same. Another challenge is the occlusion that sometimes occurs when a part of the virtual skeleton can not be tracked because something is blocking the Kinect vision. So we proposed to use two Kinect sensors instead of one, as a way to reduce the occlusion, because what is hidden to one Kinect could be visible for the other one.

In this paper we propose specialized algorithms for analyzing the skeleton structure of seated persons, so we can differentiate physical activities such as drinking, eating, paying attention, using laptop, checking watch, checking cellphone, writing, using tablet, attending a call, and participating. Also, we are going to compare the performance of two against one Kinect sensor, to see how effective they are recognizing activities, and if there is actually a significant advantage in using two Kinect sensors, and also we are going to experimentally compare several classification algorithms for this activity recognition task.

The rest of the paper is organized as follows: Sect. 2 outlines the related works, Sect. 3 describes the method process used; experiments and results are in Sect. 4, and finally conclusions are presented in Sect. 5.

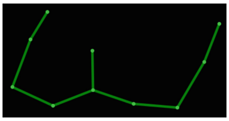

Fig. 1. Seated skeleton

2 Related Work

Human activity analysis is made with data coming from a single sensor or a group of sensors. These sensors can be embedded in the environment or can be attached to the person body, so a basic distinction can be made between infrastructure-based approaches, which use fixed devices such as vision cameras, sensitive floor, infrared sensors, etc., and wearable sensors, in which the person about who the activities are going to be recognized is actually wearing the sensor, like a smart watch with accelerometers [4]. Of course wearable approaches favor user privacy, but they also impose requirements on the equipment the user must carry. In this paper we are assuming infrastructure sensors, in particular a set of Kinect sensors. The use of two Kinect sensors was also proposed by Mazurek [7] as well as Sthone et al. [8], who combines data from multiple Kinect sensors to create pose estimates of a human.

2.1 Classifiers for Activity Recognition

Several classification algorithms have been used for activity recognition system, in particular Naive Bayes, Support Vector Machines (SVM) and k-Nearest Neighbors algorithm (k-NN); we review the use of them in the following. The comparison of classification algorithms is one of the goals of the present paper.

From the outset there is no particular reason for using one classifier instead of another when it comes to activity recognition from sensors, but in our experiments (Sect. 4) we found that there could be significant differences in their performance.

Naive Bayes [9] is a simple probabilistic classifier, which uses the Bayes theorem with a naive independent assumptions between the features. It requires a small amount of data to be trained.

Works using the Naive Bayes classifier with infrastructure sensors include Mazurek et al. [10], and others, while Ravi et al. [11] uses Naive Bayes to recognize activities from the data collected from a wearable sensor. A Bayesian classifier is also used by Song et al. [12], who uses two image-related sensors with the depth information from two cameras to track upper body movements.

Support Vector Machines (SVM) can classify linear and nonlinear data. Linear mapping searches for a lineal optimal line to separate "hyperplanes" [9].

Support Vector Machine classifiers are used by Cottone et al. [13], Megavannan et al. [14] and others.

Multiple cameras and a SVM classifier are used by Cohen et al. [15], who uses four cameras to capture 3D human body shapes and infers body postures using SVM to identify the postures and the activity performed by the user.

The specific use of Kinect sensors and SVM is reported by Zhang et al. [16].

K-Nearest Neighbor (k-NN) finds for a given data point the nearest k points in a dataset and assigns it to the most frequent class among those k points, so it is a kind of voting algorithm [9].

The k-NN classifier has been used, together with image sensors, by Gordon et al. [17], Ofli et al. [18] and others.

Wearable sensors with k-NN is used by Abdullah et al. [19], who uses a smartphone attached to a person to recollect data.

The reviewed works fall short in the two aspects that we emphasize in the present work, namely evaluating the advantage of using several Kinect sensors, and comparing several very different classifiers, for assessing their relative strengths.

3 Solution Method

The approach we follow for activity recognition is entirely data-driven: after collecting data from the Kinect sensors during a meeting, we extract features and then train classifiers. Validation experiments use the trained classifiers against test data, and precision assessment is made using ground-truth made by humans who identify the activity of each participant in every frame of a meeting session (Fig. 3).

So the steps of our approach are the following:

1. Collect data from meeting sessions
2. Apply geometric transformation and consolidate skeletons from two Kinect sensors
3. Establish ground truth from video with the activities of each user
4. Extract features
5. Apply data to classifier (Bayes, SVM, k-NN) and obtain predicted activities for each user
6. Check the result against the ground truth and calculate precision

In the following we will present the details of these steps.

3.1 Data Collection

Activities of participants in a meeting are going to be captured by a 60-degree separated arrangement of Kinect sensors connected to a computer and specialized software (see details in Sect. 4).

The activities that we are going to consider are: drinking, paying attention, using laptop (we are going to refer to it as laptop only), checking watch, checking

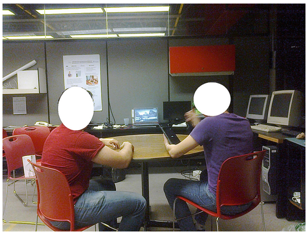

Fig. 2. Two kinects

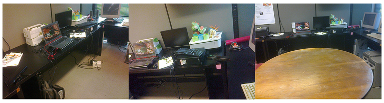

Fig. 3. Configuration of the meeting room and kinects

cellphone, writing, using tablet, attending a call, pointing, raising hand, and participating. Each joint coordinate is captured on a frame, a group of frames have a group of joint's coordinates. The skeleton consists on 8 joints, corresponding to the head, shoulders, elbows and hands. The Kinect sensor captures 30 frames each second, so the raw data comprise for each 30th of a second the 3D position of each of the 8 considered joints, that is, 24 numbers for each Kinect, the double for a set of two Kinect sensors.

When we use Two Kinect sensors, each one of them has a different reference for determining the position of body joints, so we need to apply a coordinate transformation to one of them in order to make its measures compatible with the other Kinect. We are going to use one Kinect sensor as a reference and transform the measurements made by the second Kinect using standard methods [20]

Once we have the measurements of two Kinect sensors aligned to the same coordinate system, we can observe that there are small but significant differences in the corresponding position estimations, so we need to apply a *consolidation* operation, in which we take the middle point between the positions reported by each of the two Kinect sensors. This is the position we consider for the two-Kinect experiments.

3.2 Feature Extraction

Once the capture and consolidation process is finished, we have a data matrix and a corresponding synchronized video for establishing the ground truth. The next step in our method is to complement the raw data with derived features that could help the classification process. While most authors use only static (pose, posture) features for activity recognition, we decided to use also dynamic features, such as the body joints speed and acceleration. Also, we included statistical features like Andersson [21] from a 3 s window (that is, 90 frames). Of course, whether or not a feature is useful can only be established once the classification task is performed and precision is established, though there are feature selection techniques such as Principal Component Analysis (we used PCA against our feature set, giving that position is the most important one).

So, the features we considered are:

- Static: Joint Position (x,y,z)
- Dynamic: Absolute Linear Speed
- Dynamic: Absolute Acceleration
- Statistical: Mean position over a 3 s window
- Statistical: Median of position
- Statistical: Standard deviation of position

3.3 Classifiers

In this study we used and compared three different classification algorithms, namely Naive Bayes, Support Vector Machines and k-nearest neighbor.

4 Experiments and Results

The purpose of the validation experiments we are going to present is to compare the performance of activity recognition using one Kinect and two ones, so we test the hypothesis that two Kinect sensors will be able to overcome to a certain degree the occlusion problems. We also want to test the precision of each classifier at recognizing the activities of users. Besides precision, also the specificity and sensitivity were calculated. We use a time window of 3 s to classify the activity performed, 1 s equals 30 frames, so we analyze 90 frames. We use this time window to extract features, the total of extracted features are 11.

For collecting data we need to setup and configure a computer with two Kinect sensors. We used open source tools, in particular OpenNI/NiTE, SimpleOpenNI, and Processing, installed on a Linux machine. OpenNI/NiTE is an open source framework that provides a set of application programming interface (API) that provide support for body motion tracking, hand gestures, and voice command recognition. SimpleOpenNi is a wrapper for processing. It supports all the functions of OpenNI, it provides a simple access to the functionality of OpenNI. Processing is a programming language.

For the experiments with two Kinect sensors these ones were arranged with a 60 degree angle as in Fig. 2, so that users' body parts that were occluded to one Kinect could be visible for the other one. This arrangement has been used in other works, but optimization of the exact arrangement remains to be done (see Sect. 5), and indeed depends on the exact room shape and furniture arrangement.

The experiments consist of three types of meetings. One is an exposition of a class, another type is a discussion meeting, and the third type is a work meeting. The meetings are acted, but the subjects do not follow a script. In each meeting there are two participants seated in front of a table, eventually with a laptop or a block for notes or beverages such as coffee or a snack for eating. Of course, the laptops are going to cause visual occlusion, which is one of the challenges in the activities recognition task.

On average, meetings are 51 min long, and we collected 4 meeting rounds for each type, thus 12 experiments, both for one and for two Kinect sensors, giving a total of 24 individual meetings, each one together with a corresponding video for establishing the ground truth (see below).

Each individual meeting data collection, taken by the Kinect at 30 frames per second, generates a raw data file of around one hundred thousand lines. Overall, the total number of data text lines collected for experiments in the 24 meetings was 2,200,086 text lines.

Then we enrich the tables with the derived features that were presented in Sect. 3.2, including the Mean, Median, Standard Deviation, Variance of the skeleton joint position in a time window of 3 s, as well as the velocity and acceleration of the skeleton's joints.

For all experiments, we are going to use the 80 % of the data collected for training the classifiers, the rest, 20 % are going to be used for testing them.

For the testing data, we use the ground truth as recognized by a human experimenter. Activities are tagged in the frames matrix directly from the video associated to the meeting, which has been synchronized with the Kinect data capture. Needless to say this was an extremely meticulous task. So, the system succeeds in predicting an activity when that prediction matches the tag given by the human experimenter.

The first experiment was an exposition of a class. Users attended the class. Relevant activities performed by the users were *Point*, *Paying Attention*, *Drinking Water* and *Raise Hand*.

In the 4 rounds of this experiment we collected 58,235, 110,204, 114,013 and 58,237 lines of skeleton position data from the Kinect sensor (one Kinect) and 114,071, 120,803, 86,990 and 72,113 lines for two Kinect sensors.

In Table 1 we present the precision comparison for experiment 1 using one and two Kinect sensors. Later on we will present the global analysis of experiments.

Table 1. Precision comparison between classifiers, with one Kinect and two Kinect sensors for experiment 1

	One Kinect			Two Kinects		
	Classifiers			Classifiers		
	NB	SVM	KNN	NB	SVM	KNN
Raise hand	1.70 %	74.26 %	81.21 %	23.37 %	77.19 %	82.96 %
Point	8.19 %	82.34 %	92.01 %	3.85 %	14.47 %	75.00 %
Paying attention	95.44 %	90.78 %	98.17 %	98.56	81.95 %	98.56 %
Drinking water	6.06 %	6.37 %	90.76 %	36.17 %	100 %	100 %
Average	27.84 %	63.43 %	90.53 %	40.48 %	68.40 %	89.13 %

In experiment 2, a discussion meeting is being held. Relevant activities are: *Participate*, *Paying attention*, *Drinking Water*, *Using Tablet*, *Eat*, *Raise Hand*.

This experiment also has 4 rounds, both for one and for two Kinect sensors, giving 8 experiment sessions. In this experiment we collected 34,781, 74,996, 102,328 and 114,070 lines of raw data for one Kinect, and 67,123, 110,183, 88,872 and 59,250 lines for two Kinect.

In Table 2 we present the comparison between classifiers and one and two Kinect sensors.

Experiment 3 is about a simulated work meeting, and the experimental settings are very similar to the preceding ones. The relevant activities are *Paying attention*, *Drinking Water*, *Check cellphone*, *Using laptop* and *Participate*; we think they are self-explanatory.

As the preceding ones, this experiment has 4 rounds for one Kinect and 4 for two Kinect. The data collected was composed of 51,251, 94,574, 115,023 and 92,187 lines of data for one Kinect and 98,368, 116,413, 93,345 and 93,406 line for two Kinect sensors.

Table 2. Precision comparison between classifiers, with one Kinect and two Kinect sensors for experiment 2

	One Kinect			Two Kinects		
	Classifiers			Classifiers		
	NB	SVM	KNN	NB	SVM	KNN
Participate	17.06	98.56	96.48	43.00	89.01	100.00
Paying attention	61.90	79.17	80.30	89.22	87.30	98.11
Drinking water	9.02	85.57	79.21	13.12	93.11	93.12
Using tablet	95.25	82.23	97.40	84.22	92.11	90.94
Eat	24.55	88.89	96.00	34.22	89.22	94.41
Raise hand	6.23	87.06	88.24	6.15	93.76	95.81
Average	35.67	86.91	89.61	44.99	90.75	95.40

Results of this experiment are presented in Table 3.

Analyzing the data in the tables, we can see that, concerning the classifiers, Naive Bayes fall behind the other two, which alternate being the best depending on the activity (for instance, in experiment 1, raising hand, SVM is better, but in writing KNN is better). In the average of one kinect, we can see that k-NN performs slightly better than two kinects, because maybe an occlusion happened, and it lowered the classifier performance. Overall, the NB precision was under 40 %, which is clearly very low, and on average SVM was below 80 %, while k-NN precision was well above 90 %, establishing a clear advantage above SMV and obviously NB. While the explanation for the low NB performance could be the data independence assumption, which we think is not respected in this task, the advantage of k-NN over SVM could be explained in terms of the high non-linearity of our data; the SVM hyperplanes for data separation imply some degree of linearity that could not be respected in our data.

Table 3. Precision comparison between classifiers, with one Kinect and two Kinect sensors for experiment 3

	One Kinect			Two Kinects		
	Classifiers			Classifiers		
	NB	SVM	KNN	NB	SVM	KNN
Paying attention	20.96	94.65	80.00	45.98	81.79	100.00
Drinking water	16.13	70.53	85.73	30.76	77.94	96.88
Check cellphone	16.43	33.44	96.91	26.12	71.22	97.22
Using laptop	89.78	94.43	93.98	95.95	92.96	99.34
Participate	21.01	87.16	90.42	38.56	98.31	100.00
Average	32.86	76.04	89.41	47.47	84.44	98.69

We also see that having two Kinect sensors raises precision noticeably, namely above 7 percent –which was of course something to expect due to occlusion resolution in the case of the two Kinect sensors– but not dramatically. The reasons why having two Kinect sensors does not raise precision 10 points or more is difficult to grasp, as the skeleton detection itself is done inside the Kinect system and could not be analyzed by us.

5 Conclusions

In this paper we have presented a data-driven method for recognizing the activities of users in meeting settings using two Kinect sensors. Indeed, what meeting participants do in organizations get-together is important for assessing their utility and eventual improvements.

We used as input the position of body joints, as given by two Kinect sensors, together with derived features such as velocity and acceleration, as well as statistical features over time windows. We established a reasonably good precision in activity recognition for common activities in meetings such as listening, drinking, taking notes, using a laptop, raising hands for participation, etc.

We compared the performance of three different classifiers for the activity recognition task, namely the Naive Bayes, Support Vector Machine, and the k-Nearest Neighbour, finding that k-NN is the best classifier for this task, followed by SVM and far behind Naive Bayes.

Further, we compared in a rigorous way the performance of the activity recognition using one against two and one sensors, and established exactly how much is the improvements of the two Kinect arrangement. This had never been done to our knowledge.

As future work, we intend to optimize the Kinect arrangement, that is, the angles, the distances and so on, which of course depend of specific meeting rooms, and even tables. We also want to complement the skeleton detection of Kinect with the analysis of sound, as taken from the Kinect microphones, because sound is an additional clue for activities such as participation in a meeting; a two-Kinect arrangements could be particularly suited for detecting the sound source, which could give information about which participant in the meeting was talking.

References

1. Tapia, E.M., Intille, S.S., Larson, K.: Activity recognition in the home using simple and ubiquitous sensors. In: Ferscha, A., Mattern, F. (eds.) PERVASIVE 2004. LNCS, vol. 3001, pp. 158–175. Springer, Heidelberg (2004)
2. Aarts, E., Wichert, R.: Ambient intelligence. In: Bullinger, H.J. (ed.) Technology Guide. Principles - Application - Trends, pp. 244–249. Springer, Heidelberg (2009)
3. Demiris, G., Hensel, B.K., Skubic, M., Rantz, M.: Senior residents' perceived need of and preferences for "smart home" sensor technologies. Int. J. Technol. Assess. Health Care **24**(01), 120–124 (2008)
4. Garcia-Ceja, E., Brena, R.: Long-term activity recognition from accelerometer data. Procedia Technol. **7**, 248–256 (2013)

5. Niu, W., Long, J., Han, D., Wang, Y.-F.: Human activity detection and recognition for video surveillance. In: 2004 IEEE International Conference on Multimedia and Expo, ICME 2004, vol. 1, pp. 719–722. IEEE (2004)
6. Zhang, Z.: Microsoft kinect sensor and its effect. MultiMedia IEEE **19**(2), 4–10 (2012)
7. Mazurek, P., Morawski, R.Z.: Application of naïve bayes classifier in a fall detection system based on infrared depth sensors. In: Proceedings of 8th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (2015)
8. Stohne, V.: Real-time filtering for human pose estimationusing multiple kinects (2014)
9. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques: Concepts and Techniques. Elsevier, New York (2011)
10. Mazurek, P., Wagner, J., Morawski, R.Z.: Acquisition and preprocessing of data from infrared depth sensors to be applied for patients monitoring. In: The 8th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (2015)
11. Ravi, N., Dandekar, N., Mysore, P., Littman, M.L.: Activity recognition from accelerometer data. In: AAAI, vol. 5, pp. 1541–1546 (2005)
12. Song, Y., Demirdjian, D., Davis, R.: Multi-signal gesture recognition using temporal smoothing hidden conditional random fields. In: 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011), pp. 388–393. IEEE (2011)
13. Cottone, P., Re, G.L., Maida, G., Morana, M.: Motion sensors for activity recognition in an ambient-intelligence scenario. In: 2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), pp. 646–651. IEEE (2013)
14. Megavannan, V., Bhuvnesh Agarwal, R., Babu, V.: Human action recognition using depth maps. In: 2012 International Conference on Signal Processing and Communications (SPCOM), pp. 1–5. IEEE (2012)
15. Cohen, I., Li, H.: Inference of human postures by classification of 3d human body shape. In: IEEE International Workshop on Analysis and Modeling of Faces and Gestures, AMFG 2003, pp. 74–81. IEEE (2003)
16. Zhang, C., Tian, Y.: Rgb-d camera-based daily living activity recognition. J. Comput. Vis. Image Process. **2**(4), 12 (2012)
17. Gordon, D., Hanne, J.-H., Berchtold, M., Miyaki, T., Beigl, M.: Recognizing group activities using wearable sensors. In: Puiatti, A., Gu, T. (eds.) MobiQuitous 2011. LNICST, vol. 104, pp. 350–361. Springer, Heidelberg (2012)
18. Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., Bajcsy, R.: Berkeley MHAD: A comprehensive multimodal human action database. In: 2013 IEEE Workshop on Applications of Computer Vision (WACV), pp. 53–60. IEEE (2013)
19. bin Abdullah, M.F.A., Negara, A.F.P., Sayeed, M.S., Choi, D.J., Muthu, K.S.: Classification algorithms in human activity recognition using smartphones. Int. J. Comput. Inf. Eng. **6**, 77–84 (2012)
20. Kaenchan, S., Mongkolnam, P., Watanapa, B., Sathienpong, S.: Automatic multiple kinect cameras setting for simple walking posture analysis. In: 2013 International Computer Science and Engineering Conference (ICSEC), pp. 245–249. IEEE (2013)
21. Andersson, V.O., de Araújo, R.M.: Person identification using anthropometric and gait data from kinect sensor. In: AAAI, pp. 425–431 (2015)