Kinect vs. Low-cost Inertial Sensing For Gesture Recognition

Marc Gowing[1], Amin Ahmadi[1], Francois Destelle[1], David Monaghan[1], Noel O'Connor[1], Kieran Moran[2]

[1] INSIGHT: Centre for Data Analytics, Dublin City University, Ireland
[2] Applied Sports Performance Research, School of Health and Human Performance, Dublin City University, Ireland
{marc.gowing, anim.ahmadi, francois.destelle, david.monaghan, noel.oconnor, kieran.moran}@dcu.ie

Abstract. In this paper, we investigate efficient recognition of human gestures / movements from multimedia and multimodal data, including the Microsoft Kinect and translational and rotational acceleration and velocity from wearable inertial sensors. We firstly present a system that automatically classifies a large range of activities (17 different gestures) using a random forest decision tree. Our system can achieve near real time recognition by appropriately selecting the sensors that led to the greatest contributing factor for a particular task. Features extracted from multimodal sensor data were used to train and evaluate a customized classifier. This novel technique is capable of successfully classifying various gestures with up to 91 % overall accuracy on a publicly available data set. Secondly we investigate a wide range of different motion capture modalities and compare their results in terms of gesture recognition accuracy using our proposed approach. We conclude that gesture recognition can be effectively performed by considering an approach that overcomes many of the limitations associated with the Kinect and potentially paves the way for low-cost gesture recognition in unconstrained environments.

Keywords: Gesture recognition, Decision tree, Random forest, Inertial sensors, Kinect

1 Introduction

There is a growing trend towards moving away from the traditional keyboard and mouse as the primary computer interaction tools. In the past decade, a wealth of research in academia and industry [16, 7] has focused on finding new and more intuitive methods by which humans can interact with computers and computer-based content. Many such initiatives have been aimed at devising new algorithms and technologies for recognizing moving objects as well as human gestures and actions. Nonverbal behaviors such as hand movements, head gestures, body language, facial expression and eye contact play an import role within human communications. The recording and reconstruction of these human activities and gestures is one of the fundamental core building blocks in realizing

any advanced human/computer interaction system that is free from a keyboard and mouse.

Increasingly, this effort is driven by new application opportunities in 3D multimedia computing and modeling. The natural output of 3D multimedia capturing, processing, and scene reconstruction are novel virtual immersive environments that require more sophisticated control/interaction mechanisms than a simple point and click. In such scenarios, interaction or control based on human gestures seems a more intuitive and comfortable approach for end users [11, 8]. In certain applications, we will wish to visualize the real-time motion and actions of other users in the same immersive environment in order to experience a truly shared collaborative experience [4, 5]. In either case, approaches to real-time human gesture recognition in the real world are required so that the resulting gestures can be used to produce the required effect in the virtual world.

In recent years, the Microsoft Kinect has been a key driver for a new form of hands free interaction. As a low-cost and widely available approach to human motion sensing, the Kinect and the associated open source libraries have enabled researchers to target a range of next generation novel multimedia applications [3, 17]. However, the Kinect is not without its own limitations. Whilst it constitutes a practical approach to breaking the tether to mouse/keyboard, it is still rather limited in terms of its practical application, restricting movement sensing to indoor and to a limited spatial volume (typically 3m x 3m) [18]. In this paper, we consider other sensors that in theory allow movement sensing outside of these constraints, potentially opening up the possibility of human-computer interaction "*in the wild*" i.e. in unconstrained environments, that could then subsequently be mapped to novel multimedia experiences in immersive environments.

In general, the recording and reconstruction of human motion is referred to as Motion Capture (or MoCap for short). MoCap is a well-studied and broad research area that has been explored in multiple different research fields including computer vision, computer graphics and even body sensor networks [15, 12, 19, 14]. The various different approaches are numerous and include approaches based on mechanical, inertial, magnetic, acoustic and visual sensing etc. After a suitable MoCap system has been identified for a particular requirement and human motion has been captured, the next step is to perform gesture recognition. That is, it is required to infer some semantic meaning to the movements being performed. This can often be accomplished by manually annotating the captured motion followed by machine learning, i.e. a human observer will watch the action being performed and decide what movement or gesture has been recorded and then a suitable machine learning technique can be used to enable an automatic system to recognize similar actions for previously unseen data. It should be noted that the complexity and speed of such systems often increase exponentially when additional gestures are added to the system. The key tenets of the work that we present here are the need for real time applications that remain low-cost.

Recent advancements in microelectronics and other technologies mean that inertial sensors are gaining popularity to monitor human movements in a number of

sporting [2], rehabilitation [6] and everyday activities [9]. MEMS inertial sensing technology is already integrated by default into many consumer devices - virtually every smart phone and many computer games controllers (e.g. the Nintendo Wii). MEMS inertial sensors are being widely used in MoCap research due to the following reasons:

- They are miniaturized and lightweight so they can be placed on any part or segment of a human body without hindering performance.
- The cost of such sensors is falling dramatically as they start to persuade mass market consumer devices.
- They can be utilized to capture human movement/actions in real unconstrained environments (e.g. outdoor environments with variable lighting conditions) to obtain accurate results.
- They can be used to provide real time or near real time feedback.

On the other hand, Microsoft Kinect comes with an RGB camera and a depth sensor, which in combination provide full-body 3D motion capture capabilities and gesture recognition. This inexpensive technology is also widely being used for gesture recognition mainly due to the following reasons:

- Using Kinect allows users to avoid wearing body sensors when performing the movements.
- We can extract skeleton data using off-the-shelf software such as Kinect SDK and OpenNI.
- Kinect sensors can be used to obtain real time feedback.

In order to investigate the relative benefits of both approaches in terms of gesture recognition, in this paper we have investigated the use of wearable inertial sensors and a Microsoft Kinect depth sensor to classify a wide range of activities performed by five different subjects. We compare the gesture recognition results obtained from inertial sensors with that from the Kinect sensor using a customized Random Forest decision trees. In addition, we simulated an ultra low-cost system where only a small number of low rate inertial sensors are available by using down sampled data (from 256 Hz to 32Hz) from only three sensors (three out of eight worn sensors).
The paper is organized as follows. In Section 2 we explain the dataset and also describe the sensor modalities used in this paper. In Section 3 we fully explain the methodology. We then provide our results and discussion section and finally conclude and highlight our contributions.

2 Dataset

We use a gesture recognition dataset that includes recordings of human subjects performing various gestures and activities (17 in total) such as simple actions, training exercises and diverse sports activities. The dataset encompasses recordings of five subjects whose actions were captured using eight wearable inertial sensors and one Microsoft Kinect. The inertial system was mounted on eight

different places on the body and then s/he was asked to carry out random movements to ensure that s/he felt comfortable with the system and that the system was not limiting their movements. Next, the subject was asked to perform a series of actions. The performed actions can be divided into the following categories:

1. Simple actions (hand waving, knocking on the door, clapping, throwing, punching, push away by both hands).
2. Training exercises (jumping jacks, lunges, squats, punching and kicking).
3. Sports activities (golf drive, golf putt, golf chip, tennis forehand, tennis backhand, weight lifting, walking).

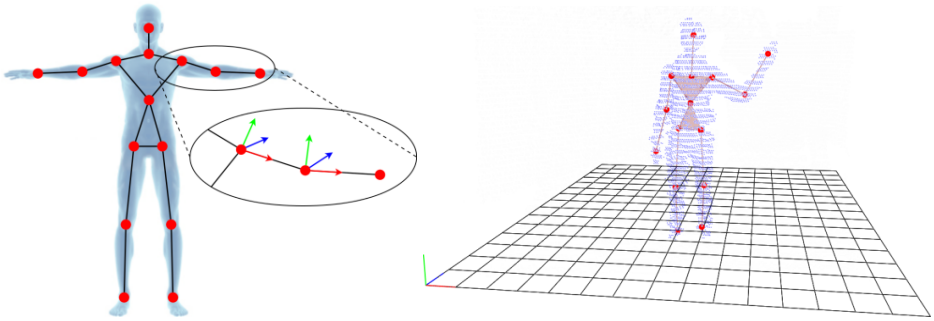
Once the sensors were switched on and worn by a subject, they were tapped separately to use the acceleration spike to synchronize all the inertial sensors. After the synchronization process, all participants performed each action/movement five times before starting the next action.
One Kinect camera was also setup about two meters away in front of the subject to capture the front body part movement. The inertial sensors were synchronized with the recorded images by clapping three times before each recording so that a specific event could be identified. The data set along with all annotations is available for download from: http://mmv.eecs.qmul.ac.uk/mmgc2013.

2.1 Sensors

We chose wearable inertial sensors and the Microsoft Kinect since they are low-cost and are each gaining in popularity in the area of human movement monitoring and gesture recognition due to their accuracy and potential for real time applications.

Kinect Since very recently, computer game users can enjoy a novel gaming experience with the Xbox, thanks to the introduction of the Microsoft Kinect sensor, where *your body is the controller* [1]. Like the Nintendo Wii sensor bar, the Kinect device is placed either above or below the video screen. However, the Kinect adds the capabilities of a depth sensor to those of a RGB camera, recording the distance from all objects that lie in front of it. The depth information is then processed by a software engine that extracts, in real time, the human body features of players, thus enabling the interaction between the physical world and the virtual one. The Kinect dataset recordings of subjects' activities were captured using the OpenNI drivers/SDK. We employed the widely known NiTE framework to track the 3D skeleton for each subject from the Kinect sensor which in turn can be used to extract subjects' joint positions and angular velocities. Estimation of 3D joints position and orientation is illustrated in Fig 1.

[1] http://www.xbox.com/en-US/KINECT



(a) (b)

Fig. 1. Estimation of 3D joint positions from the Kinect sensor. (a) The Kinect skeleton and the local coordinate system at each joint. (b) A real scene point cloud and the visualization of its skeleton computation in real time.

Table 1. Technical specifications of the inertial sensor units.

Features	**Values**
Resolution (Acc, Gyr, Mag)	12 bit, 16 bits, 12 bits respectively
Sampling rate	Scalable up to 512 Hz
Sensor range	Acc: scalable up to 8G Gyro: scalable up to 2000/s Mag: scalable up to 8.1G
Connectivity	Bluetooth-Class 1(100m range), Micro SD card
Dimension	$57 \times 38 \times 21mm$
Weight	49g including housing and battery

WIMU In general, a Wireless/Wearable Inertial Measurement Unit, or WIMU, is an electronic device consisting of a microprocessor board, on-board accelerometers, gyroscopes and magnetometers and a wireless connection to transfer the captured data to a receiving client. WIMUs are capable of measuring linear acceleration, angular velocity, and gravitational forces and are often used in MoCap systems. Technical specifications of the WIMUs we have utilized are summarized in Table 1. In the dataset eight WIMUs were attached to different parts of the subjects to capture their activities. In particular, sensors were attached on the left/right wrist, left/right ankle, left/right foot, waist and chest of all participants. Placement of inertial sensors on a subject's body is depicted in Fig 2.

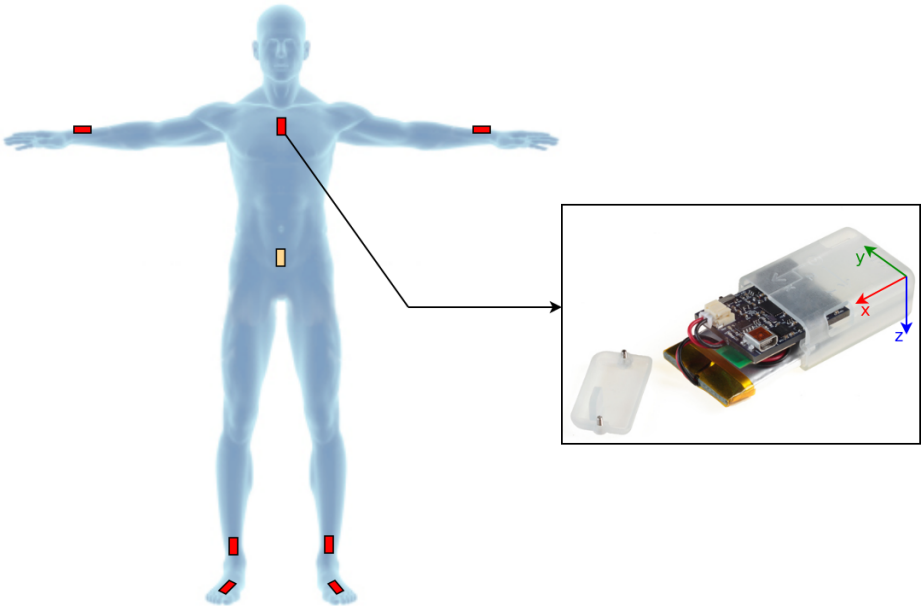


Fig. 2. Placement of inertial sensors on subjects body is shown. The lighter colored sensor indicates that the sensor is attached to the subject's lower back.

3 Methodology

To facilitate real time gesture recognition, we chose to extract features based on a sliding window approach. A fixed window size of one second was chosen based on an experimentally derived average duration of all gestures, with a 50% overlap. At each step, we compute temporal features from each modality.
For the final recognition system, different features were calculated from the one-second windows to compress the information of interest in the data. We initially calculate a large set of features for each sensor signal, and apply a standard sequential feature selection technique [1] to identify the most discriminative features prior to training our classification model. The following sections outline the procedure in more detail.

3.1 Feature Selection

In the dataset, each subject performs the full list of actions as one long continuous sequence of motion. Therefore, each sequence contains many instances of unlabeled data, where the subject is deemed to not be performing one of the predefined actions. Rather than segmenting the motion sequence into examples of predefined actions only, we instead chose to include the full sequence in our training phase. We use any instances of unlabeled data as negative examples

in our training set. The rationale for taking this approach is that, during real time operation there will be many instances where a subject is not performing any action at all, and it is more desirable to classify the gesture as unknown rather than misclassifying it as one of the known actions. A sliding window of 1 second, with a 0.5 second step size was chosen as the duration of each gesture is relatively short. The annotation label was applied to the data if the start time of the sliding window was within 0.5 seconds of the action start time or 0.5 seconds of the action end time. These features were computed for each Kinect joint orientation (u, v and w axis), accelerometer (x, y and z axis), magnetometer (x, y and z axis) and gyroscope sensor (x, y and z axis). To extract features from the Kinect, we relied on the orientation estimates provided by the NiTE skeleton tracker. We opted to use joint orientation estimates instead of positions due to the fact they are robust to changes in the user's global root orientation, and they do not require retargeting/normalization to an average skeleton of all users. From two subsequent local coordinate systems $P : \{u, v, w \in \mathbb{R}^{3\times3}\}$ and $P' : \{u', v',\ w' \in \mathbb{R}^{3\times3}\}$ linked to a skeleton joint, we define a unique rotation matrix M as:

$$P' = M.P. \tag{1}$$

Let the quaternion q be the modelization of the 3D rotation from P to P'. We divide this 3D rotation in two separate 2D rotations modeled by quaternions:

$$\begin{cases} q_u : \{A_u^x, A_u^y, A_u^z, \varphi_u\} \\ \\ q_v : \{A_v^x, A_v^y, A_v^z, \varphi_v\} \end{cases} \tag{2}$$

where $A_u, A_v \in \mathbb{R}^3$ are the axis of rotation from u to u' and from v to v', while $\varphi_v, \varphi_v \in \mathbb{R}$ are the deviation angles:

$$\begin{cases} A_u \equiv u \times u' \\ \varphi_u = u.u' \end{cases} \begin{cases} A_v \equiv v \times v' \\ \varphi_v = v.v' \end{cases} \tag{3}$$

We now compose these quaternion, $q = q_u.q_v$ and create the unique rotation matrix M from P to P'. From there, we directly obtain the Euler angles prior to feature extraction, to provide the quantity of rotation about the local x, y and z-axis.

The feature selection process involved extracting a large number of features for every signal in the motion sequence, and then reducing this using the well-known standard sequential forward selection technique to reduce the number of features and thereby to improve computational cost and to obtain near real time performance. Heuristic features including SMA [10] and Inter axis correlation [20, 13] were derived from a fundamental understanding of how a specific movement would produce a distinguishable sensor signal. For instance, there are obvious correlations, using Pearson correlation test, between left and right wrist movements during all golf swings, walking, pushing with two hands, weight lifting and clapping. Correlation in x and y signals between the left wrist and the right wrist during clapping action (5 times) is shown in Figure 3.

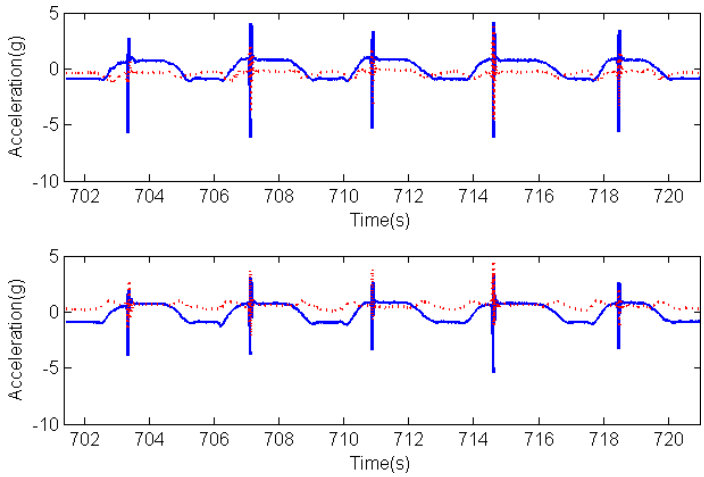


Fig. 3. Correlation in x (solid line) and y (dashed-line) acceleration signals between the right wrist (top) and left wrist (bottom) during clapping action is shown.

3.2 Classification

We investigated a number of different fast decision trees in order to choose one technique to classify all the activities in the dataset using both modalities. We examined Random Forest, Random Tree and C4.5 algorithms and compared the results. Results are shown in Table 2. We have chosen the random forest method, as it provided the best accuracy among the algorithms investigated. In general, random forest does not over fit and it is fast, which makes it suitable for our near real time application. In addition, in this work, we extracted features for all subjects from triple axis inertial sensors and a Microsoft Kinect which results in a large amount of data. Therefore, we needed to apply a method such as random forest, which runs efficiently on large datasets. This method is also capable of providing feedback on what inputs/features are more important so we could enhance our model by removing low-priority features to speed up and implement towards real time application. We found that a random forest consisting of 30 trees provided the best results on our dataset. The results to classify different activities using wearable inertial sensors and Kinect are shown in Table 3.

3.3 Reducing the number of sensors

By observing all the gestures and activities in the dataset, it can be seen that all the activities have upper body movement component so they can be distinguished by upper body movements. Therefore, we only extracted features from three upper body (right wrist, left wrist and chest) sensors to classify the gestures. The advantages of employing a smaller number of sensors are as follows:

Table 2. Comparison of three decision tree classification algorithms for gesture classification. CV stands for Cross Validation.

Modality	**Random Forest CV (%)**	**C4.5 CV (%)**	**Random Tree CV (%)**
Kinect	80.3157	73.0883	68.9196
Acc	89.0725	80.5624	78.8357
Gyr	86.4085	78.9097	74.0257
Mag	88.3325	81.5491	77.7750
Acc + Gyr	89.9359	81.6971	79.6251
Acc+Mag	90.6759	83.4238	79.8717
Gyr+Mag	88.1845	81.5491	77.1830
Acc+Gyr+Mag	90.6512	83.9418	80.7351

Table 3. Results of activity recognition using a range of multimodal sensors.

Modality	**Sub. 1 CV (%)**	**Sub. 2 CV (%)**	**Sub. 3 CV (%)**	**Sub. 4 CV (%)**	**Sub. 5 CV (%)**	**All Subjs. CV (%)**
Kinect	85.1544	82.797	86.5340	87.6494	87.7039	80.3157
Acc	90.0238	94.5545	90.5152	90.4382	87.9548	89.0725
Gyr	87.6485	92.9455	88.2904	86.8526	83.1870	86.4085
Mag	89.3112	93.1931	92.1546	88.1806	87.9548	88.3325
Gyr+ Mag	89.4299	93.0693	92.3888	88.9774	86.7001	88.1845
Gyr + Acc	91.4489	93.8119	91.3349	89.7742	89.5859	89.9359
Acc+ Mag	90.3800	94.5545	92.0375	91.8991	90.2133	90.6759
Acc+Gyr+Mag	90.9739	94.8020	92.2717	91.8991	90.3388	90.6512

- The entire gesture recognition system would be cheaper and less prone to set up/synchronization error.
- Reducing the number of sensor nodes allows us to reduce the amount of features to be extracted during the training phase. This can lead to less computational cost which in turn can result in providing near real time feedback.
- This can address a typical scenario for many applications - where end-users may only have a small number of sensors at their disposal.

Results obtained from utilizing three upper body sensors are illustrated in Table 4.

Table 4. Activity classification using reduced number of inertial sensors.

Modality	**Sub. 1 CV (%)**	**Sub. 2 CV (%)**	**Sub. 3 CV (%)**	**Sub. 4 CV (%)**	**Sub. 5 CV (%)**	**All Subs. CV (%)**
Selected ACC+Gyr+Mag	90.7363	93.2693	89.8126	90.4382	88.0803	88.7025

Table 5. Activity classification using down-sampled data.

Modality	**Sampling Frequency**	**Sub. 1 CV (%)**	**Sub. 2 CV (%)**	**Sub. 3 CV (%)**	**Sub. 4 CV (%)**	**Sub. 5 CV (%)**	**All Subs. CV (%)**
Selected ACC+Gyr+Mag	32 Hz	89.5487	93.1931	88.8759	89.1102	87.8294	87.8999

3.4 Using down sampled data

In a further attempt to reduce the cost of the system amd the amount of computation at each window, we down sampled the inertial sensors data from 256Hz to as low as 32Hz, effectively reducing the computation time by a factor of 8. The results are shown in Table 5. Down sampling can simulate the output of inertial sensors systems manufactured to lower specification.

4 Results

We have compared three different fast decision tree techniques including random forest, random tree and C4.5 to classify activities in our dataset. We trained all the classifiers on each modality and performed 10 fold cross validation to test the performance. The comparisons of these techniques are summarized in Table 2. It is clearly shown that the highest accuracy can be obtained by using random forest followed by the C4.5 technique.

As can be seen in Table 3, using accelerometers, gyroscopes or magnetometers can provide more overall accuracy to classify activities than using the Kinect sensor. Table 3 also shows that by combining the data from accelerometers, gyroscopes and magnetometer, maximum accuracy can be achieved. However, even using any one of these modalities on its own its still outperforms the Kinect. The output from the random forest classifier for the selected sensors (left/right wrist sensor and chest sensor) is illustrated in Table 4. As shown, reducing the number of sensor nodes does not degrade the overall accuracy since all the activities studied in this work contain upper body movement components. In addition, we have investigated the effect of down-sampled data to simulate utilizing low-cost inertial sensors for gesture recognition. Not only does it simulate the low-cost scenario, it also decreases the computational cost to achieve near real time application. We have down sampled the data from 256 Hz to 32 Hz for this experiment.

Table 5 summarizes the results obtained from down-sampled inertial sensors. As can be seen, this does not significantly affect the accuracy as normal human movements are not as fast as 32 repetitions per second (32Hz) and thus cheap inertial sensors with 32Hz sampling rate can be considered to be used to capture human activities.
Finally, decision trees are relatively fast to classify activities and therefore are suitable techniques for near real time or real time applications. Once the features were selected, it took between 0.5 seconds to 0.76 seconds to produce a random forest of 30 trees model for each subject's dataset on a MacBook Pro 2.33GHz framework. In the future, further optimization techniques are required to enhance the performance of the devised system to achieve real time feedback.

5 Conclusions

In this paper we described a novel inertial-based system that automatically classifies a large range of activities (17 different gestures) using a customized random forest decision tree. Our system achieved near real time gesture recognition by appropriately selecting the sensors that led to the greatest contributing factor for a particular task. Our technique is capable of classifying various gestures successfully using inertial sensors with up to 91% overall accuracy, making it extremely competitive with the MS Kinect. We have fully analyzed our system for a wide range of MoCap solutions thus providing a look up table to enable potential researchers to choose an appropriate MoCap solution based on their specific accuracy requirements. We managed to achieve a high level of accuracy for a low-cost system which is capable of providing feedback in near real time.
Our results point to the fact that the Kinect is clearly not the only option to be considered for applications requiring MoCap. Its attraction is rooted in its low-cost and lack of instrumentation but it is inherently limited in terms of the scenarios in which it can be implemented. Low-cost inertial sensors, on the other hand, do not suffer from many of the limitations associated with the Kinect and can operate in outdoor unconstrained environments. We have shown in this paper that very similar or even higher accuracy to Microsoft Kinect can be achieved with a very small amount of human instrumentation. This potentially paves the way for novel future multimedia applications whereby human motion and interaction can be captured in a range of challenging environments, not just indoors in front of a computer.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programmes (FP7/2007-2013) under grant agreement no. ICT-2011-7-287723 (REVERIE project) and (FP7/2013-2016) under grant agreement no. ICT-2011-8.2.601170 (REPLAY project).

References

1. Aha, D.W., Bankert, R.L.: A comparative evaluation of sequential feature selection algorithms. In: Learning from Data, pp. 199–206. Springer (1996)
2. Ahmadi, A., Rowlands, D.D., James, D.A.: Development of inertial and novel marker-based techniques and analysis for upper arm rotational velocity measurements in tennis. Sports Engineering 12(4), 179–188 (2010)
3. Alexiadis, D.S., Kelly, P., Daras, P., O'Connor, N.E., Boubekeur, T., Moussa, M.B.: Evaluating a dancer's performance using kinect-based skeleton tracking. In: Proceedings of the 19th ACM international conference on Multimedia. pp. 659–662. ACM (2011)
4. Bowman, D.A., Hodges, L.F.: User interface constraints for immersive virtual environment applications. Tech. rep., Atlanta: Graphics, Visualization and Usability (1995)
5. Bowman, D.A., Hodges, L.F.: Formalizing the design, evaluation, and application of interaction techniques for immersive virtual environments. Journal of Visual Languages & Computing 10(1), 37–53 (1999)
6. Culhane, K., O'Connor, M., Lyons, D., Lyons, G.: Accelerometers in rehabilitation medicine for older adults. Age and ageing 34(6), 556–560 (2005)
7. Dix, A.: Human computer interaction. Pearson Education (2004)
8. Jaimes, A., Sebe, N.: Multimodal human–computer interaction: A survey. Computer vision and image understanding 108(1), 116–134 (2007)
9. Junker, H., Amft, O., Lukowicz, P., Tröster, G.: Gesture spotting with body-worn inertial sensors to detect user activities. Pattern Recognition 41(6), 2010–2024 (2008)
10. Karantonis, D.M., Narayanan, M.R., Mathie, M., Lovell, N.H., Celler, B.G.: Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. Information Technology in Biomedicine, IEEE Transactions on 10(1), 156–167 (2006)
11. Lyons, D.M.: System and method for permitting three-dimensional navigation through a virtual reality environment using camera-based gesture inputs (Feb 27 2001), uS Patent 6,195,104
12. Mannini, A., Sabatini, A.M.: Machine learning methods for classifying human physical activity from on-body accelerometers. Sensors 10(2), 1154–1175 (2010)
13. Mathie, M.: Monitoring and interpreting human movement patterns using a triaxial accelerometer. Ph.D. thesis, The University of New South Wales (2003)
14. Menache, A.: Understanding motion capture for computer animation and video games. Morgan Kaufmann (2000)
15. Moeslund, T.B., Granum, E.: A survey of computer vision-based human motion capture. Computer Vision and Image Understanding 81(3), 231–268 (2001)
16. Myers, B.A.: A brief history of human-computer interaction technology. interactions 5(2), 44–54 (1998)
17. Ren, Z., Meng, J., Yuan, J., Zhang, Z.: Robust hand gesture recognition with kinect sensor. In: Proceedings of the 19th ACM international conference on Multimedia. pp. 759–760. ACM (2011)
18. Roth, H., Vona, M.: Moving volume kinectfusion. In: BMVC. pp. 1–11 (2012)
19. Sturman, D.J.: A brief history of motion capture for computer character animation. SIGGRAPH 94, Character Motion Systems, Course notes 1 (1994)
20. Yang, J.Y., Wang, J.S., Chen, Y.P.: Using acceleration measurements for activity recognition: An effective learning algorithm for constructing neural classifiers. Pattern recognition letters 29(16), 2213–2220 (2008)