Gait & Posture 48 (2016) 83–88

Contents lists available at ScienceDirect

Gait & Posture

journal homepage: www.elsevier.com/locate/gaitpost

An evaluation of 3D head pose estimation using the Microsoft Kinect v2

John Darby [a,*], María B. Sánchez [b], Penelope B. Butler [c], Ian D. Loram [b]

[a] School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University, Manchester, UK
[b] School of Healthcare Science, Manchester Metropolitan University, Manchester, UK
[c] The Movement Centre, Oswestry, Shropshire SY10 7AG, UK

ARTICLE INFO

Article history:
Received 8 July 2015
Received in revised form 18 April 2016
Accepted 28 April 2016

2010 MSC:
00-01
99-00

Keywords:
Head posture
Non-invasive
Real-time
Kinect v2
Assessment

ABSTRACT

The Kinect v2 sensor supports real-time non-invasive 3D head pose estimation. Because the sensor is small, widely available and relatively cheap it has great potential as a tool for groups interested in measuring head posture. In this paper we compare the Kinect's head pose estimates with a marker-based record of ground truth in order to establish its accuracy. During movement of the head and neck alone (with static torso), we find average errors in absolute yaw, pitch and roll angles of $2.0 \pm 1.2^\circ$, $7.3 \pm 3.2^\circ$ and $2.6 \pm 0.7^\circ$, and in rotations relative to the rest pose of $1.4 \pm 0.5^\circ$, $2.1 \pm 0.4^\circ$ and $2.0 \pm 0.8^\circ$. Larger head rotations where it becomes difficult to see facial features can cause estimation to fail ($10.2 \pm 6.1\%$ of all poses in our static torso range of motion tests) but we found no significant changes in performance with the participant standing further away from Kinect – additionally enabling full-body pose estimation – or without performing face shape calibration, something which is not always possible for younger or disabled participants. Where facial features remain visible, the sensor has applications in the non-invasive assessment of postural control, e.g. during a programme of physical therapy. In particular, a multi-Kinect setup covering the full range of head (and body) movement would appear to be a promising way forward.

1. Introduction

Sensor systems able to accurately estimate the position of a person's head have underpinned research in a wide range of areas including: automatic identity recognition (e.g., by explaining variations in the 2D appearance of a person's face [1]); interpretation of non-verbal gestures (e.g., by tracking small head movements and establishing the focus of visual attention [2]); human–computer interaction (e.g., through gesture- and gaze-driven interfaces for users with disabilities [3]); and physical therapy and rehabilitation (e.g., through real-time biofeedback systems for the learning of head control [4]).

In each of these applications, the potential utility of the sensor is increased when the impact of its operation on participants is minimised. Specifically, sensors will ideally: (i) be non-invasive (not requiring any hardware to be in contact with the participant's head); and (ii) not require the participant to engage in any training or calibration procedures. For example, in the context of physical therapy – our own application area of interest – it would be desirable if a sensor could be used to monitor head control in younger children and/or those with learning difficulties, who are not necessarily willing to wear hardware sensors, or able to respond to verbal requests for calibration movements [5].

The topic of *non-invasive* head pose estimation has therefore received considerable attention from computer vision researchers, aiming to produce accurate estimates of head pose from one or more conventional colour (RGB) cameras [6]. In more recent years the arrival of cheap, colour+depth sensing (RGB-D) cameras has significantly progressed the state of the art in both full-body [7] and head pose estimation [8]. For example, the system in [9] is able to produce pose estimates from single depth images with rotational errors of less than 6°, and translation errors of less than 15mm. And where it is possible for users to actively participate in an interactive calibration phase, even lower errors are possible [10].

A problem for those interested in making non-invasive estimates of head pose is that access to state of the art methods is not straightforward; the reimplementation of relevant algorithms

* Corresponding author. Tel.: +44 1612471542.
E-mail address: j.darby@mmu.ac.uk (J. Darby).

http://dx.doi.org/10.1016/j.gaitpost.2016.04.030
0966-6362/

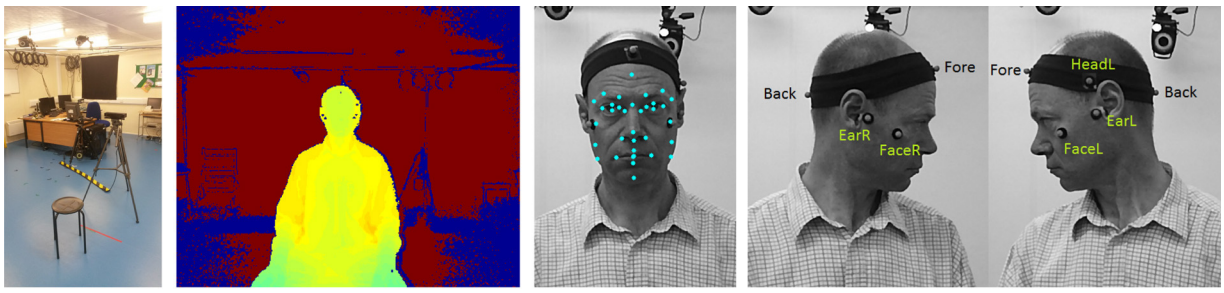

Fig. 1. Head pose estimation: (a) Kinect placement; (b) Kinect's view of a participant (full depth image); (c) facial features estimated by Kinect's high definition face tracking; (d) the Vicon marker set used for assessment.

requiring considerable time and expertise. Notable exceptions are the system described in [11] for which source code is made available on the authors' web pages, but which only returns yaw/pitch rotation estimates, and the subscription-based commercial system[1] developed from [10], which requires participation in an interactive training phase. This situation has recently changed with the release of the Microsoft Kinect v2 sensor and its software development kit (SDK).

The high definition face tracking (HDFT) component of the Kinect v2 SDK allows for real-time non-invasive head pose estimation without the need for calibration or training. The SDK code examples can be easily and quickly extended to record estimates, and apart from the cost of the sensor itself, use of the system is completely free. Here we present an evaluation of the sensor's accuracy in estimating the rotational and translational components of 3D head pose, and consider a number of experimental conditions likely to be relevant to those interested in using the sensor in a clinical setting; e.g., to evaluate head control in the context of a programme of physical therapy.

2. Methods

2.1. Data collection

Eight participants (age: 21–64; 1M; 7F) were asked to perform a series of pre-defined head movements (see Section 3 and the supplementary materials for full details). Their movements were simultaneously recorded using: (i) Kinect v2's HDFT[2]; and (ii) a marker-based Vicon motion capture system.[3] The work was approved by the Ethics Committee for the Faculty of Science and Engineering at Manchester Metropolitan University and complied with the principles laid down by the Declaration of Helsinki. All participants gave informed consent to the work.

2.2. Kinect head pose estimation

Kinect was placed on a tripod (height of 1.15 m) and angled to frame the upper body of participants when they were seated in the centre of the Vicon capture volume (a distance of approximately 1.1 m), see Fig. 1a and b. Head rotations relative to the Kinect coordinate system were recorded by subscribing to the HDFT stream [12] and writing the timestamped `FaceOrientation` quaternion to file each time a tracking event was generated. The `HighDetailFacePoints` enumeration was also used to extract the locations of the main facial features from the resulting 3D face mesh, see Fig. 1c and the supplementary materials for full details and code listing.

[1] www.faceshift.com.
[2] Kinect for Windows SDK v2.0.1410.19000, Microsoft, USA.
[3] 10-camera MX system, Nexus 1.8.5, Vicon Motion Capture, Oxford, UK.

2.3. Vicon head pose estimation

The marker set shown in Fig. 1d was used to estimate head rotations relative to the Vicon coordinate system (see also Section 2.4). A combination of clips and a head band were used to keep long hair from obstructing the markers, or the participant's face. Participants were otherwise dressed normally.

2.4. Post processing

A spatial transformation between the coordinate systems of the Kinect and Vicon was estimated (see the supplementary materials for full details) and the Vicon marker data rotated and translated to lie in the Kinect coordinate system. Vicon head rotations were then extracted by using Visual 3D (C-Motion, US) to create a virtual head segment between the ear and face markers, see Fig. 1d and the supplementary materials for full details. Finally, a time synchronisation between the two sensors was estimated (see the supplementary materials for full details) and the Vicon data downsampled using a simple nearest neighbour interpolation to provide a record of ground truth corresponding with the timestamp of every HDFT event, e.g. see Fig. 2.

2.5. Data analysis

Following other approaches, yaw, pitch and roll errors were computed and are presented in units of degrees [9,13,14]. Errors were calculated as the modulus of the difference in the pose estimates generated by the Kinect sensor and the Vicon system. A millimetre error between facial features was also computed, but rather than using the nose (e.g., [9,11]) the average error between the Vicon cheekbone markers and the cheekbone features found by the HDFT (see also Fig. 1c) was computed, so as to additionally account for roll errors. The HDFT can also return tracking failure events (e.g. see Fig. 2), and these were computed as a "missed" percentage of the total number of HDFT events recorded during each movement.

For all experimental conditions described in Section 3, averages for each of the measures above (yaw, pitch and roll rotation errors, cheekbone translation errors, and the fraction of missed frames) were computed across the movements of each individual participant and one-way repeated measure ANOVA tests used to determine significant differences between conditions.

3. Conditions tested

We evaluated the performance of Kinect against Vicon under each of the following experimental conditions. In addition to providing a quantification of its accuracy under optimal conditions,

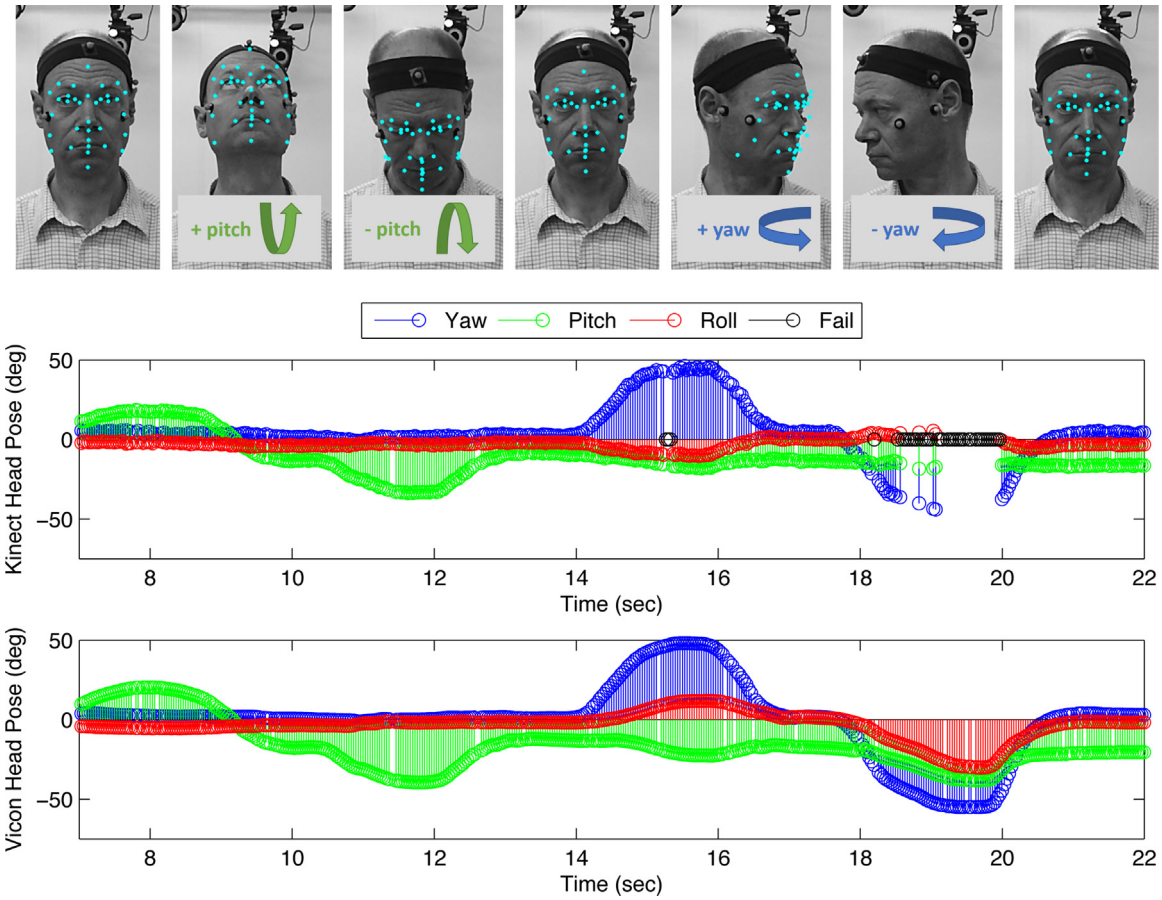

Fig. 2. Typical HDFT results for a pitch (up-and-down) followed by yaw (left-to-right) head rotation: (top) Example images with HDFT results overlaid in cyan; (middle) Corresponding yaw (blue circles), pitch (green circles) and roll (red circles) estimates from the HDFT, with tracking failures, e.g. as the participant looks to their right during 18–20 s, shown as black circles; (bottom) Corresponding Vicon estimates for the same sequence. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

we have also tried to anticipate relevant scenarios for those wishing to use the Kinect in a clinical setting.

3.1. Baseline: range of motion

This condition is clinically relevant because it facilitates a basic evaluation of head control, in isolation. Whilst sitting comfortably with their torso and shoulders still, participants were asked to complete a set of movements that demonstrated their full range of head motion. For example, testing comfortable limits in each of yaw (left-to-right), pitch (up-and-down), and roll (side-to-side) rotations, see also Fig. 2 and the supplementary materials for full descriptions. This condition was used as a baseline, against which performance in subsequent conditions was compared.

3.2. Relative rotations

Each movement in the baseline tests began and ended in the participant's resting pose: the head pose they naturally take up when seated comfortably. In order to remove any impact from our particular choice of Vicon parameterisation on results, the rotations measured in Section 3.1 were recomputed relative to each participant's resting pose, and rotational accuracy recalculated in relative (rather than absolute) terms.

3.3. Extremes of motion

Visually, missed HDFT events appear to correspond with larger rotations of the head, see for example the period between 18 and 20 s in Fig. 2. However, such rotations can easily occur when participants are free (and able) to move their upper body (relative to a stationary sensor). Participants were asked to complete a further set of range of motion movements, but with their torso and shoulders free to move, allowing them to rotate their head further relative to the Kinect, see for example the images in Fig. 3, and the supplementary materials for full details.

3.4. Occlusion of facial features

The HDFT tracks facial features in order to estimate head pose, but there are many reasons the face may become occluded in a clinical setting, e.g., children raising their hands to their head. Participants were asked to repeat the yaw and pitch movements from Section 3.1 with their hands over their mouth, see supplementary materials for full details and images.

3.5. Face shape calibration

The HDFT allows the shape of a person's face to be learned by having them participate in an interactive calibration where they perform small head rotations in response to requests by the Kinect sensor. This extra pre-processing step can improve facial feature tracking performance [12] and, although such a calibration is not always possible (e.g., for young or disabled participants), it may, where feasible, have an impact on the accuracy of head pose estimation. Participants were guided through the face shape calibration procedure (see the supplementary materials for full details) before being asked to repeat the range of motion tests in Section 3.1.

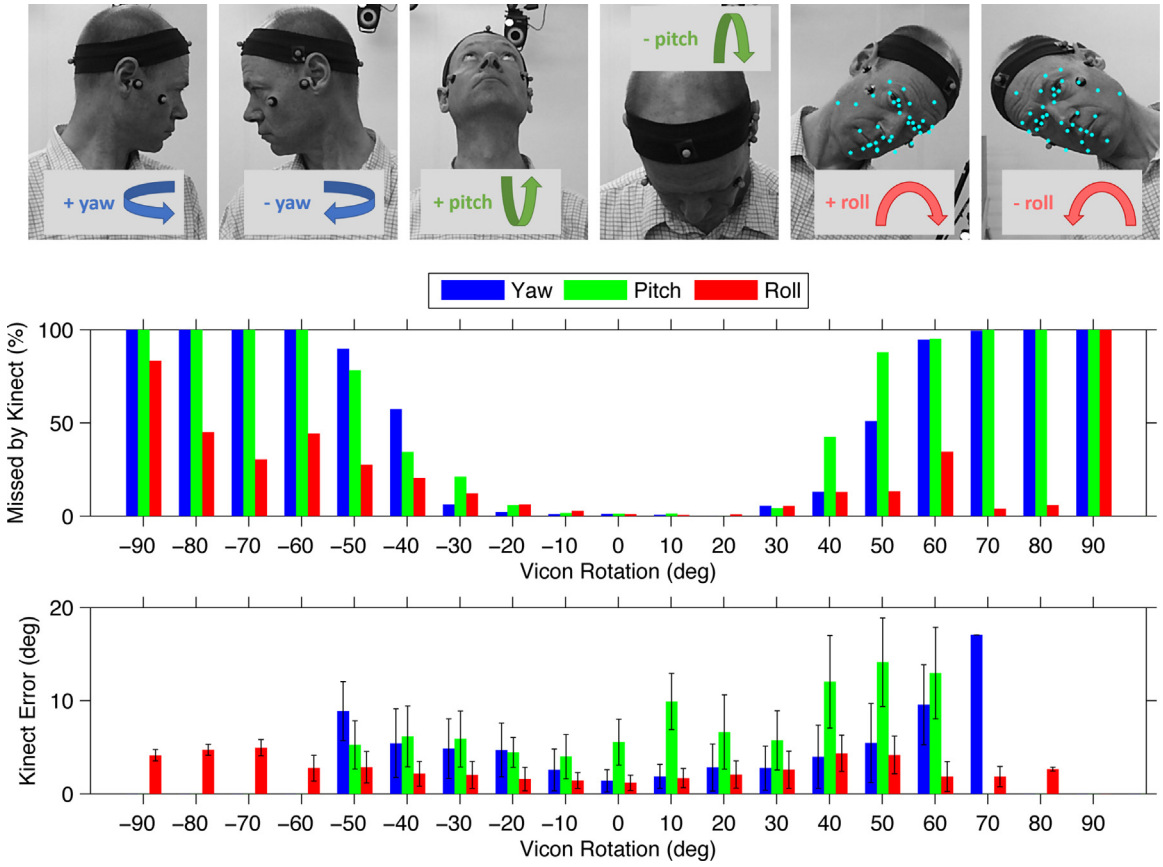

Fig. 3. Free torso rotations: (top) Example images from yaw, pitch and roll rotations with any HDFT results overlaid in cyan; (middle) The percentage of missed frames across all participants computed across 10° bins (centred at −90°, −80°, …, +90°) during yaw (blue), pitch (green), and roll (red) rotations; (bottom) The average error in Kinect's rotational estimates across each bin. Kinect's estimation of roll rotations is more robust at large rotations, probably because facial features tend to remain fully visible even at the ends of range (see images). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.6. Sensor range

When a participant's body is fully visible in the depth image Kinect can additionally estimate their body pose [7]. This is an additional measure likely to be of interest to those studying postural control [15,16], but it is not possible in the close-up experimental setup used here and elsewhere (Fig. 1b, [9,11,14]). Participants were asked to stand and move back from the sensor until full-body pose estimation became possible (a distance of approximately 1.7 m from Kinect, varying slightly with participant height), before then repeating the range of motion tests (Section 3.1), see supplementary materials for full details and images.

3.7. Rotated viewpoint

Guaranteeing an anterior view of participants, as shown in Fig. 1b, may not always be possible or practical. For example, if a physical therapist must work in front of a patient during an evaluation [17]. Participants were asked to repeat the range of motion tests (Section 3.1) while seated at a 45° angle to the Kinect sensor (see supplementary materials for full details and images).

4. Results

Table 1 shows the results for each of the eight participant's baseline range of motion tests (Section 3.1), with the average performance across all participants highlighted in grey. Missed frames occurred predominantly during maximal yaw rotations, which were larger on average (59°) than pitch (39°) or roll (33°) rotations. Table 2 compares this baseline performance against all other conditions tested (see Section 3), with any significant differences detected by the ANOVA tests identified using Bonferroni post hoc tests and highlighted in bold.

When the rotations in the baseline condition are recomputed relative to the participants' resting poses (second row of Table 2) there is a significant reduction in average pitch errors ($p < 0.001$).

The free torso tests (third row of Table 2) produced no significant change in pose estimation accuracy, but a significant increase in the number of missed frames ($p < 0.001$).

Fig. 3 presents an analysis of missed frames and estimation accuracy during the free torso yaw, pitch and roll rotations (average maximum extents of 79°, 87° and 66°, respectively). Both quantities are presented as a function of the rotational components measured by the Vicon system. For yaw and pitch rotations, the fraction of missed frames grows very rapidly outside the range 35°, reaching 100% beyond 65°. Estimation during roll rotations was more robust, only reaching 100% beyond 85°. The error in the estimates that were returned (not missed) by Kinect grows with the magnitude of the true rotation, but again this effect is less marked for roll rotations.

When facial features are occluded by hands (fourth row of Table 2), the number of missing frames is significantly higher ($p < 0.001$), and their distribution is shifted right across the participants' range of motion (see supplementary materials for a comparison with Fig. 3).

No aspect of performance was statistically significantly different after face shape calibration (fifth row of Table 2), or

Table 1
Head pose estimation accuracy during the static torso range of motion (baseline) tests for all 8 participants. The highlighted grey row shows the mean and standard deviation of each quantity across all participants.

	Cheekbone (mm)	Yaw (°)	Pitch (°)	Roll (°)	Missed (%)
P1	6.6	2.7	4.7	3.6	5.1
P2	7.9	1.2	8.0	2.0	19.5
P3	5.4	1.4	5.9	2.8	8.7
P4	11.5	4.5	12.1	3.4	18.0
P5	13.3	1.2	12.2	2.5	13.0
P6	14.9	1.1	6.5	2.6	3.7
P7	11.6	2.4	5.0	2.0	4.5
P8	12.7	1.8	3.8	1.7	9.1
Average	10.5 ± 3.4	2.0 ± 1.2	7.3 ± 3.2	2.6 ± 0.7	10.2 ± 6.1

with participants standing back to allow simultaneous full-body pose estimation (sixth row of Table 2).

Estimation from a rotated viewpoint (seventh row of Table 2) lead to a significant rise in: missed frames ($p < 0.001$), which occurred during even moderate positive yaws; cheekbone errors ($p = 0.045$); and all rotational components ($p = 0.026$, $p < 0.001$, $p < 0.001$ for yaw pitch and roll, respectively). However this condition did allow for reliable pose estimation at the participants' extremes of range (large negative yaws, see supplementary materials for example images).

5. Discussion

Visibility of facial features appears to be the determining factor in whether Kinect's HDFT is able to provide an estimate of head rotation; it likely relies on an algorithm for describing facial structure/appearance (e.g. [18]), extended to depth data. Large yaw rotations made with a static torso (where one side of the face becomes self-occluded) consistently produced missing frames, and participants raising their hands to their mouths caused a significant rise in missing frames, distributed right across the entire range of motion (see also supplementary material). Similarly, combined head and torso rotations resulting in greater overall rotation led to almost 100% missing frames beyond 55° for both yaw and pitch rotations, but the Kinect was much more robust to combined roll rotations where facial features tend to remain visible. When estimates are returned by the Kinect they are reliably low in error and comparable with other state of the art approaches [9,14] (a summary of RGB-D methods from the literature is included in the supplementary materials). If the extremes of motion are of particular interest, then Kinect can be placed at an angle relative to the participant, but with significant increases in average error and missed frames across the full range of motion. For complete coverage, a multi-Kinect setup which favours the sensor returning minimum relative rotation would appear to be a promising way forward.

There are three potential sources of error in our evaluation that are independent of Kinect's performance. First, the automatic spatial calibration between the two sensor systems (Section 2.4). For example, Table 1 shows higher cheekbone errors for P5–P8, versus P1–P4, and these groups were recorded on two separate days following two separate calibration procedures. Second, the automatic time synchronisation between the two sensor systems. Although this is mitigated by their high sampling rates (which bound the error if tracking is accurate) and the fact that all movements studied were slow. Third, our parameterisation of head pose through Vicon marker placement. Pitch errors are consistently higher than yaw and roll, but an analysis of rotations measured relative to the rest pose showed a significant reduction in pitch errors. This suggests the Vicon parameterisation used here (see Section 2.4) may overestimate pitch rotations versus the parameterisation used by Kinect (the precise details of which are unknown). Similarly, our comparison between cheekbone translations involves points on the surface of the skin estimated by Kinect, versus locations of Vicon markers sitting just above the surface of the skin.

It is possible that placing Vicon markers on the surface of participants' faces may have interfered with the HDFT's performance. But it was notable that the face shape calibration procedure (Section 3.5) still operated quickly and normally with all but one of the eight participants (who was calibrated after repeating the interactive procedure). Face calibration did not, however, impact

Table 2
Comparison between the baseline (grey) and all other conditions. Significant differences, as identified by the ANOVA tests, are highlighted in bold.

	Cheekbone (mm)	Yaw (°)	Pitch (°)	Roll (°)	Missed (%)
Range of motion	10.5 ± 3.4	2.0 ± 1.2	7.3 ± 3.2	2.6 ± 0.7	10.2 ± 6.1
Relative	-	1.4 ± 0.5	$\mathbf{2.1 \pm 0.4}$	2.0 ± 0.8	-
Free torso	11.7 ± 3.1	2.7 ± 1.2	7.6 ± 3.6	2.4 ± 0.6	$\mathbf{38.2 \pm 7.7}$
Occlusion	10.3 ± 3.7	1.8 ± 1.0	6.6 ± 2.6	1.3 ± 0.5	$\mathbf{50.9 \pm 17.5}$
Calibrated	9.1 ± 2.5	2.0 ± 1.0	6.6 ± 3.2	2.6 ± 0.5	10.5 ± 5.0
Standing	15.0 ± 3.2	2.5 ± 0.9	6.7 ± 3.0	3.3 ± 1.0	12.5 ± 4.3
Rotated	$\mathbf{16.3 \pm 9.3}$	$\mathbf{5.1 \pm 3.2}$	$\mathbf{13.9 \pm 5.3}$	$\mathbf{6.8 \pm 4.0}$	$\mathbf{49.0 \pm 21.6}$

on the quality of head pose estimation. The fact that Kinect offers accurate head pose estimation without an interactive calibration step is an important result for those applications where participant cooperation is not always possible (e.g., due to age and/or disability [3,4]). Having participants move far enough from the sensor that their full bodies were visible also had no significant impact on performance. This result will be of interest to those wishing to make simultaneous non-invasive measurements of movement in other parts of the body using Kinect's skeletal tracking functionality [7,15,16].

The software used in this study was a combination of the standard v2 SDK examples, extended to record the various Kinect datastreams (e.g., RGB, depth, skeletal estimates). HDFT events were recorded, on average across the whole dataset at a rate of 22 Hz, but higher rates are possible when not attempting to record other signals simultaneously. Any group can immediately benefit from Kinect head pose estimation simply by extending the v2 SDK's "HDFaceBasics" example to write the `FaceAlignment.FaceOrientation` quaternion to file (see the supplementary materials for full details).

6. Conclusion

Where facial features remain visible (e.g., anterior view with a static torso), the Kinect v2 sensor is able to offer state of the art head pose estimation accuracy in real time and without the need for calibration. Occlusion of the facial features (e.g., by the hands or through large rotations of the head involving the torso) can cause tracking to fail and no pose estimate to be returned. However, this is in useful contrast to returning unreliable, high-error pose estimates which must be reviewed and excluded during post processing. The sensor's low cost, its easy to use SDK, and its ability to simultaneously estimate body poses, mean it has considerable potential as a tool for those interested in making non-invasive measurements of posture. In particular, a multi-Kinect setup covering the full range of head (and body) movement would appear to be a promising way forward.

Source of funding

No external funding was received specifically for this project. The second author's studentship is jointly funded by Manchester Metropolitan University and The Movement Centre.

Acknowledgements

We are grateful to Pauline Holbrook, Sarah Bew, Lynne Ford and Richard Major for help with study conception and acquisition of data.

Conflict of interest statement

There is no conflict of interest.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.gaitpost.2016.04.030.

References

[1] Zhao W, Chellappa R, Phillips PJ, Rosenfeld A. Face recognition: a literature survey. ACM Comput. Surv. (CSUR) 2003;35(4):399–458.
[2] Stiefelhagen R. Tracking focus of attention in meetings. In: Proceedings of the 4th IEEE International Conference on Multimodal Interfaces; 2002.p. 273.
[3] Manresa-Yee C, Varona J, Perales FJ, Salinas I. Design Recommendations for Camera-Based Head-Controlled Interfaces that Replace the Mouse for Motion-Impaired Users. Universal Access in the Information Society; 2013. p. 1–12.
[4] James R. Biofeedback treatment for cerebral palsy in children and adolescents: a review. Pediatr. Exerc. Sci. 1992;4(3).
[5] Fehlings D, Switzer L, Findlay B, Knights S. Interactive computer play as "motor therapy" for individuals with cerebral palsy. Semin. Pediatr. Neurol. 2013;20(2):127–38.
[6] Murphy-Chutorian E, Trivedi MM. Head pose estimation in computer vision: a survey. IEEE Trans. Pattern Anal. Mach. Intell. 2009;31(4):607–26.
[7] Shotton J, Sharp T, Kipman A, Fitzgibbon A, Finocchio M, Blake A, Cook M, Moore R. Real-time human pose recognition in parts from single depth images. Commun. ACM 2013;56(1):116–24.
[8] Fanelli G, Gall J, Van Gool L. Real time 3D head pose estimation: recent achievements and future challenges. In: International Symposium on Communications Control and Signal Processing. 2012. p. 1–4.
[9] Fanelli G, Dantone M, Gall J, Fossati A, Van Gool L. Random forests for real time 3D face analysis. Int. J. Comput. Vis. 2013;101(3):437–58.
[10] Weise T, Bouaziz S, Li H, Pauly M. Realtime performance-based facial animation. ACM Trans. Graph. (TOG) 2011;30(4):77.
[11] Fanelli G, Gall J, Van Gool L. Real time head pose estimation with random regression forests. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2011.p. 617–24.
[12] Kinect for Windows SDK 2.0, https://msdn.microsoft.com/en-us/library/dn785525.aspx, accessed (05.07.15).
[13] Breitenstein MD, Kuettel D, Weise T, Van Gool L, Pfister H. Real-time face pose estimation from single range images. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008; 2008.p. 1–8.
[14] Baltrušaitis T, Robinson P, Morency L-P. 3D constrained local model for rigid and non-rigid facial tracking. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2012.p. 2610–7.
[15] Clark RA, Pua Y-H, Fortin K, Ritchie C, Webster KE, Denehy L, Bryant AL. Validity of the Microsoft Kinect for assessment of postural control. Gait Posture 2012;36(3):372–7.
[16] Clark RA, Pua Y-H, Oliveira CC, Bower KJ, Thilarajah S, McGaw R, Hasanki K, Mentiplay BF. Reliability and concurrent validity of the Microsoft Xbox One Kinect for assessment of standing balance and postural control. Gait Posture 2015;42(2):210–3.
[17] Butler P, Saavedra S, Sofranac M, Jarvis S, Woollacott M. Refinement, reliability and validity of the segmental assessment of trunk control. Pediatr. Phys. Ther. 2010;22(3):246–57.
[18] Cootes TF, Edwards GJ, Taylor CJ. Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. 2001;(6):681–5.