Shake'n'Sense: Reducing Interference for Overlapping Structured Light Depth Cameras

Alex Butler[1], Shahram Izadi[1], Otmar Hilliges[1], David Molyneaux[1,2], Steve Hodges[1], David Kim[1,3]

[1]Microsoft Research
7 J J Thomson Avenue
Cambridge, CB3 0FB, UK

[2]Computing and Communications,
Lancaster University,
Lancaster, LA1 4WA, UK

[3]Computing Sciences
Newcastle University,
Newcastle, NE1 7RU, UK

{dab,otmarh,shahrami,shodges,a-davmo,b-davidk}@microsoft.com

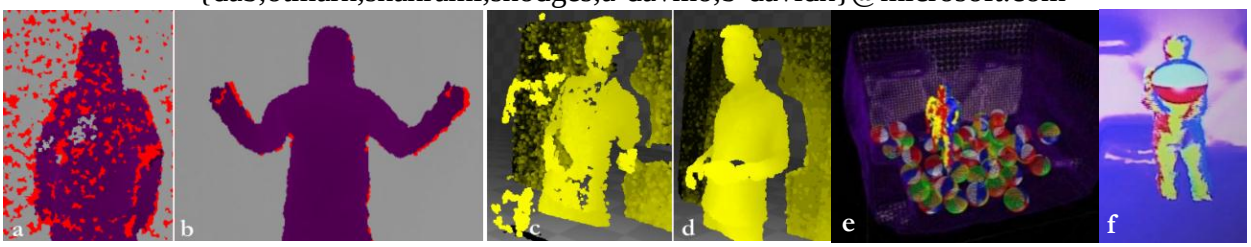

Figure 1: We present a novel yet simple mechanical technique for mitigating the interference when two or more Kinect cameras point at the same part of a physical scene. (a) Interference between overlapping structured light patterns from two regular Kinect cameras pointing at a person produces invalid and noisy depth pixels marked red. (b) Our method reduces noise and invalid pixels in the depth map. (c) The resulting point-cloud shows significant artifacts without our technique. (d) Point-cloud with our technique applied. (e) Our technique can be used to create an entire instrumented room with multiple overlapping Kinect cameras. (f) Meshed output accumulated from multiple Kinects shows reduced interference between cameras (color-coding indicates data from different cameras).

ABSTRACT

We present a novel yet simple technique that mitigates the interference caused when multiple structured light depth cameras point at the same part of a scene. The technique is particularly useful for Kinect, where the structured light source is not modulated. Our technique requires only mechanical augmentation of the Kinect, without any need to modify the internal electronics, firmware or associated host software. It is therefore simple to replicate. We show qualitative and quantitative results highlighting the improvements made to interfering Kinect depth signals. The camera frame rate is not compromised, which is a problem in approaches that modulate the structured light source. Our technique is non-destructive and does not impact depth values or geometry. We discuss uses for our technique, in particular within instrumented rooms that require simultaneous use of multiple overlapping fixed Kinect cameras to support whole room interactions.

ACM Classification: H5.2 [Information interfaces and presentation]: User Interfaces. - Graphical user interfaces.

General terms: Human Factors, Design, Algorithms.

Keywords: Kinect, depth camera, structured light, reducing interference, motion blur, instrumented rooms.

INTRODUCTION

In a relatively short time Kinect has had a huge impact in areas ranging from consumer gaming through to the maker and DIY communities and computer science research [4]. Whilst there has already been a great deal of research on depth sensing cameras, Kinect has now made such sensors cheap, commodity devices and dramatically broadened accessibility.

The depth sensing within Kinect is based on a fixed structured light source positioned at a known baseline from an infrared (IR) camera. Light from an IR laser diode passes through a diffractive optical element (DOE) to project a pseudo-random pattern of IR dots into the scene. The disparity between the known illumination pattern and the observed dots is used to calculate depth across the field of view of the IR camera. An on-board ASIC performs this calculation, generating a 16-bit 640x480 depth map at 30 frames per second.

The simultaneous use of multiple depth cameras can extend the coverage of the Kinect, overcome occlusions and create complete 360° 3D representations of environments and the objects they contain. However, for structured light sensors such as Kinect the depth signal severely degrades when multiple cameras are pointing at the same scene. This is due to the sensor projecting a structured light dot pattern onto the scene continuously, without modulation. There is *crosstalk* when dot patterns of devices interfere with one another. This issue, which also applies to other structured depth camera systems, is clearly demonstrated in Figure 1a and 1c.

The research community has begun to experiment with multiple Kinects. For example in [5] a 180° view of the user's head and surrounding environment is generated for telepresence using four Kinect cameras. However, considerable noise in the signal requires post-processing, including computationally expensive de-noising and hole-filling steps. Perhaps the most notable work in this area is LightSpace [7], which maps a 3D

CHI'12, May 5–10, 2012, Austin, Texas, USA.

space encompassing a horizontal and vertical surface to support a variety of on-surface and in-air interactions. Here the camera arrangement was carefully chosen to minimize overlapping areas due to the issue of crosstalk – trading off coverage and occlusion for a cleaner signal from each camera.

In this paper we describe a novel method that mitigates the issue of structured light crosstalk in a novel and easily reproducible way. Our approach needs no modification of the internal electronics, firmware or Kinect host software and does not degrade frame rate. Furthermore, we demonstrate quantitatively and qualitatively that the technique is effective in reducing noise without compromising sensed depth measurements. The technique can be generalized to other structured light depth sensing cameras and we believe its simplicity will enable other researchers to adopt it readily for their work. We discuss our mechanical setup in detail so that others can reproduce it.

SHAKE WELL BEFORE USE

The key behind Shake'n'Sense is to minimally vibrate a Kinect camera unit using an offset-weight vibration motor and thereby artificially introduce *motion blur*. This has similarities to vibration-based image super resolution displays [1], but uses motion for blurring the other camera signals. Both the structured light DOE illuminator and the IR camera of the Kinect will move in harmony, which means that depth sensing works as normal, albeit with a little induced blur. However, even minor almost imperceptible motion of the sensor in this way causes blurring of structured light patterns from other units which serves to eliminate most of the crosstalk.

To evaluate our Shake'n'Sense motion blur technique we minimally vibrate one or more of the Kinect cameras in a multi-camera setup. In our prototype we attach a custom offset-weight vibration motor to the casing of the Kinect using an acrylic mounting plate and rubber bands, see Figure 2. We have also designed a simple acrylic frame onto which the Kinect camera assembly can be mounted using additional rubber bands (Figure 2). These hold the camera in place but at the same time allow the whole unit to vibrate at the same frequency as the vibration motor. With this frame clamped firmly in place the number and the tension of rubber bands may be altered to adjust the amplitude of induced vibration.

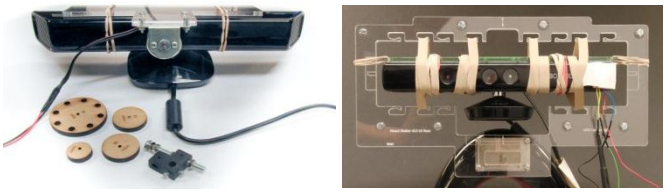

Figure 2: Left: our simple Shake'n'Sense setup for Kinect allows a custom rear-mounted offset-weight vibration motor to induce vibration in the entire unit. Right: the camera assembly is mounted in an acrylic frame using rubber bands allowing it to vibrate freely. An accelerometer is taped to the unit (at right) to enable the frequency of vibration to be measured.

The frequency of vibration can be controlled within the range of 15Hz to 120Hz. In order to accurately assess the optimal vibration frequency, we have attached an ADXL335 3-axis accelerometer to the camera's casing. Using this external sensing mechanism we have empirically derived an optimal frequency balancing power consumption, noise levels and sensing quality for our current design (see later).

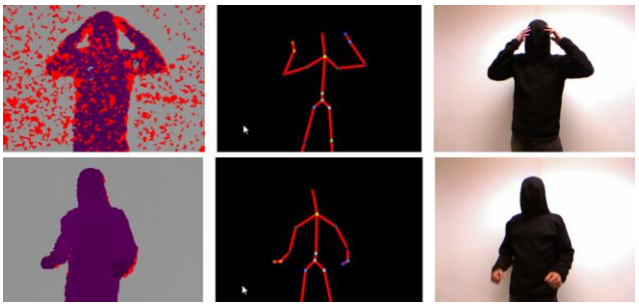

Figure 3: Top row: depth noise caused by multiple overlapping Kinect patterns (left); the Kinect SDK skeletal tracker loses accuracy (middle); corresponding RGB image (right). Bottom row: our technique significantly reduces noise (left); skeletal tracker shows correct pose (middle); Motion blur in RGB image is imperceptible (right).

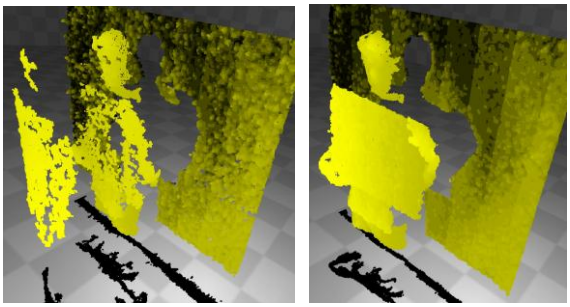

Figure 4: Point-cloud renderings of a person standing in front of a wall holding a sheet of card. Left: significant error in depth due to cross-talk including depth values being hallucinated. Right: our method applied.

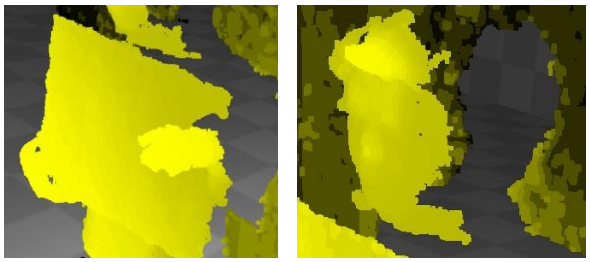

Figure 5: Detail-preserving properties of our method. Left: Scene with user holding card during camera shake. No edge fattening or depth flattening effects observable on cardboard, hands and user's face (shown in detail right).

QUALITATIVE RESULTS

By vibrating each Kinect independently, interference is dramatically reduced. Figures 1, 3 and 4 demonstrate the effect on recovered depth values; notice how the number of holes and "hallucinated" readings are dramatically improved. Figure 3 illustrates how extreme noise can impair the Microsoft Kinect SDK skeletal tracker and how camera shake can help stabilize it. Note that the RGB image does not exhibit perceptible motion blur despite the vibration. Figure 4 shows a 3D rendering of such a scene; without noise removal the the user is barely recognisable and many depth values are incorrect.

Figure 5 sheds further light on the accuracy of depth measurements and image detail using our technique. Here we show close-ups of the scene, with no noticeable fattening occurring

around the edges of the user, which one might expect as side-effect of motion blur. Also, small details such as the user's hands, fingers, hair, eyebrows and nose are preserved.

QUANTITATIVE EXPERIMENTS

To objectively quantify the Shake'n'Sense technique, we conducted a number of quantitative experiments to measure the impact on noise in the depth-map *(Q1)*. To ensure others can reproduce and build on top of our technique we also conducted experiments to pin-point the optimal vibration frequency *(Q2)*. While the latter only directly applies to our specific hardware setup we believe that the method can inform other configurations as well.

Nominal Distance (m)	Mean (bad pixels)	Std. Dev.	Min (bad pixels)	Max (bad pixels)	Test Condition
1.0	0.00	0.00	0.00	0.00	Static
1.0	0.00	0.00	0.00	0.00	Shaking
1.0	107	12.7	78.0	150	Static+CamB
1.0	0.00	0.00	0.00	0.00	Shaking+CamB
1.5	0.00	0.00	0.00	0.00	Static
1.5	0.00	0.00	0.00	0.00	Shaking
1.5	211	20.8	162	273	Static+CamB
1.5	0.00	0.00	0.00	0.00	Shaking+CamB
2.0	0.00	0.00	0.00	0.00	Static
2.0	0.00	0.00	0.00	0.00	Shaking
2.0	311	25.0	257	392	Static+CamB
2.0	0.00	0.00	0.00	0.00	Shaking+CamB
2.5	0.00	0.00	0.00	0.00	Static
2.5	0.00	0.00	0.00	0.00	Shaking
2.5	989	111	661	1140	Static+CamB
2.5	0.00	0.00	0.00	0.00	Shaking+CamB

Table 1a: Invalid pixel count with standard deviation and min-max measurements of corrupted pixels under four different test conditions. These are run across 150 frames and at four different distances from the planar wall surface. Results shown to three significant figures.

Nom. Dist. (m)	Mean Depth (mm)	Std. Dev.	Mean Min (mm)	Mean Max (mm)	Abs. Min (mm)	Abs. Max (mm)	Test Condition
1.00	1003	0.04	1002	1003	984	1037	Static
1.00	1002	0.05	1001	1002	984	1034	Shake
1.00	1002	0.04	1002	1002	973	1044	Static+CamB
1.00	1002	0.04	1002	1002	984	1034	Shake +CaB
1.50	1503	0.04	1503	1503	1464	1556	Static
1.50	1503	0.03	1503	1503	1470	1563	Shake
1.50	1504	0.04	1504	1504	1163	1622	Static+CamB
1.50	1503	0.05	1503	1503	1464	1571	Shake +CB
2.00	2028	0.06	2028	2028	1952	2120	Static
2.00	2029	0.07	2029	2029	1941	2120	Shake
2.00	2031	0.12	2031	2032	1677	3528	Static+CamB
2.00	2029	0.04	2029	2030	1941	2120	Shake +CB
2.50	2501	0.24	2500	2502	2400	2639	Static
2.50	2501	0.19	2501	2502	2417	2639	Shake
2.50	2506	0.27	2506	2507	2274	2765	Static+CamB
2.50	2504	0.28	2503	2505	2400	2659	Shake +CB

Table 1b: Mean depth, standard deviation, min/max of the average frame depth and absolute min/max depth statistics (four significant figures). Data was averaged across 150 frames for each of the four test conditions.

We used the following setup for our experiments: A Kinect camera (A) mounted using the Shake'n'Sense system was clamped to a horizontal surface (125cm high) and initially 1m away from a perpendicular wall. A second Kinect camera (B) was placed 50cm away from the first Kinect such that it observed (and projected interference onto) the same field of view as camera A.

The depth map from A was captured for 150 frames without the camera B being active (condition 1). The Shake'n'Sense technique was activated for camera A and depth data captured for basic comparison without IR interference (condition 2). Shake'n'Sense was turned off and camera B was then turned on and another 150 frames were captured from A (condition 3). Finally, Shake'n'Sense was re-activated and another 150 frames from camera A were captured (condition 4).

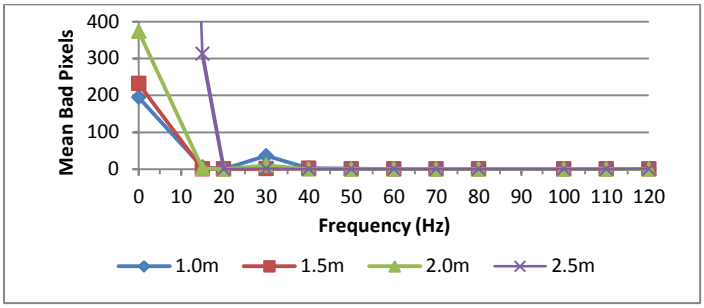

Figure 6: Mean bad pixel counts versus vibrational frequency of camera at different distances from a plane wall. Note: data is omitted for 90Hz because vibrations at that frequency could not be mechanically induced by the simple motor drive. The 2.5m trace for 0Hz was 1650 and not shown here for graph-scaling clarity.

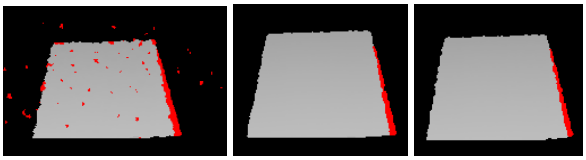

Figure 7: A single, angled plane segmented from background by depth-thresholding. Left: Noisy depth-map with two overlapping patterns. Centre: Ground-truth from single Kinect. Right: Camera shake almost identical to ground-truth.

Mean Bad Pixel Count	Std. Dev.	Min (bad pixels)	Max (bad pixels)	Test Condition
296.79	9.82	272.00	322.00	Static
458.39	16.55	407.00	499.00	Static+CamB
269.55	9.24	243.00	292.00	Shaking+CamB

Table 2a: Mean invalid pixel counts with standard deviation and min/max values averaged over 150 frames for the three test conditions.

Mean Depth (mm)	Std. Dev.	Mean Min (mm)	Mean Max (mm)	Abs. Min (mm)	Abs. Max (mm)	Test Condition
1243.94	0.09	1243.69	1244.16	1066.00	1600.00	Static
1243.29	0.20	1242.75	1243.76	1063.00	1542.00	Static+CamB
1241.62	0.29	1241.10	1242.27	1066.00	1596.00	Shaking+CamB

Table 2b: Mean depth, standard deviation, min/max of the average frame depth and absolute min/max depth statistics. Data was averaged across 150 frames for the three test conditions indicated.

Image statistics were run on data captured in all the four conditions, and shown in Table 1a and 1b. The results indicate that the Shake'n'Sense method dramatically reduces interference, indicated by the number of holes and depth measurement noise (shown as the standard deviation across the depth values). In fact there is hardly any measurable difference in terms of depth mean, variance and standard deviation when compared to the ground truth of a single Kinect.

Our second experiment evaluates the effect of vibration frequency on crosstalk noise elimination. The frequency of the

shake was varied from 15Hz to 120Hz in 10Hz increments. 150 frames were captured for each frequency and averaged. Figure 6 shows little variation above 40Hz, although the raw data indicates that the optimal frequency for the shake with our particular mechanical setup is between 60 Hz and 80Hz.

When looking at more complex scenes however, one might imagine small negative side-effects caused by the vibration of the Kinect sensor. In Figure 5, we show qualitatively that edge-fattening and depth-flattening are minimal, if existent at all. However, to quantifiably assess these effects, we placed a 57cm x 46 cm planar target in front of the wall, 125cm away from Kinect A. It was tilted at a 45-degree angle and segmented out via thresholding as shown in Figure 7. Statistics were measured on this foreground plane for each of the three conditions, see Table 2. As assumed from our qualitative assessment, there is practically no deviation from the ground truth when we use Shake'n'Sense. We have verified this with two, three and four overlapping Kinect illumination patterns.

FURTHER SIMPLIFICATION

Whilst our prototype Shake'n'sense design was intentionally chosen to illustrate the simplicity of our approach, we are eager that our approach is as accessible as possible. To that end, we have experimented with a number of variations and have concluded that an even simpler physical setup is viable. Figure 8 shows an offset-weight vibration motor hot-melt glued to the top of the casing of a Kinect camera which is itself mounted to a rigid object (in this case a table) using self-adhesive Velcro tape. In a qualitative evaluation of this setup, the compliance in the neck of the motorized Kinect stand coupled with that of the Velcro results in a similar elimination of interference between two Kinect cameras as reported above. We believe that this remarkably simple setup may be sufficient for many multi-Kinect scenarios and encourage others to adopt it.

Figure 8: Our simplified setup with a motor hot-melt glued to the Kinect which is mounted on a table with Velcro. A small wooden disc acts as the offset weight.

DISCUSSION

We have shown both qualitatively and quantitatively that our Shake'n'Sense technique reduces noise in a number of multi-Kinect configurations and environments. The motivation to overcome this cross-talk stems from our original goal of creating complete 360° occlusion-free 3D representations of larger environments.

Our prototypes are clearly proof-of-concept designs and we can imagine a number of improvements. It is conceivable that a tiny vibration mechanism, perhaps based on piezo actuators for example, could be used to cause vibration of the structured light source and the image sensor only, rather than the entire camera body. This would use less power and reduce the audible noise associated with our prototype. It may also be useful to combine physical movement with other techniques, such as time-division or wavelength multiplexing.

To further motivate our Shake'n'Sense method, we end by reporting on an instrumented interactive space with 4 ceiling mounted Kinect cameras, each at the midpoint of a wall in a rectangular room (4m x 3m x 2.75m high) and angled down at 45°. We calibrated the intrinsic parameters of each camera and predicted their extrinsic pose relative to each other [0]. A background mesh was captured without users in the room, using an offline Poisson surface reconstruction [3]. The foreground aligned point clouds representing users and other non-static parts of the scene are meshed in real-time. Figure 8 shows a snapshot of the live mesh representing 'foreground' objects which is rendered in four colors, each color corresponding to the Kinect generated that part of the mesh. The mesh representation of the scene may be used to interact with virtual objects in a physics simulation. Examples of this are shown in Figure 9. In practice, this leads to a variety of playful interactions as shown in the accompanying video.

Figure 9: Left: a photo of our setup shows 3 of the ceiling-mounted Kinect cameras (red circles); the extrinsic parameter calibration pattern can also be seen (blue circle). Right: renderings of reconstructed background mesh (lilac) and live foreground mesh of occupant (multi-color).

CONCLUSIONS

We have shown how the data from multiple Kinect cameras may be integrated into a single mesh model of a scene. We have demonstrated that crosstalk between the multiple structured illumination patterns may be mitigated simply by vibrating or gently shaking each device. We hope that other practitioners will build on top of this technique to build 360 capture systems such instrumented rooms without the issue of cross-talk. Our aims are now to explore specific interaction techniques based on these types of instrumented spaces.

ACKNOWLEDGEMENTS

Note related research that has been conducted independently and in parallel to our work and will be appearing in [6].

REFERENCES

1. Berthouzoz, F., Fattal, R., Resolution Enhancement by Vibrating Displays, To Appear ACM TOG 2012.
2. Hartley, R. and Zisserman, A. Multiple View Geometry in Computer Vision. Cambridge University Press, 2003.
3. Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. 2006. Poisson surface reconstruction. In Proc. Geometry processing (SGP '06). Eurographics, 61-70.
4. Kinect Hacks, http://www.kinecthacks.net/
5. Maimone, A. and Fuchs, H. Encumbrance-free Telepresence System with Real-time 3D Capture and Display using Commodity Depth Cameras, Forthcoming.
6. A. Maimone and H. Fuchs. Reducing Interference Between Multiple Structured Light Depth Sensors Using Motion. To appear in IEEE Virtual Reality 2012.
7. Wilson, A., and Benko, H. Combining Multiple Depth Cameras and Projectors for Interactions On, Above, and Between Surfaces. In *Proceedings UIST 2010*.