PERFORMANCE EVALUATION OF THE 1ST AND 2ND GENERATION KINECT FOR MULTIMEDIA APPLICATIONS

S. Zennaro, M. Munaro, S. Milani, P. Zanuttigh, A. Bernardi, S. Ghidoni, E. Menegatti

Department of Information Engineering, University of Padova, Italy.

{munaro, simone.milani, zanuttigh, ghidoni, emg}@dei.unipd.it, {simone.zennaro.90, andrea.bernardi85}@gmail.com

ABSTRACT

Microsoft Kinect had a key role in the development of consumer depth sensors being the device that brought depth acquisition to the mass market. Despite the success of this sensor, with the introduction of the second generation, Microsoft has completely changed the technology behind the sensor from structured light to Time-Of-Flight. This paper presents a comparison of the data provided by the first and second generation Kinect in order to explain the achievements that have been obtained with the switch of technology. After an accurate analysis of the accuracy of the two sensors under different conditions, two sample applications, i.e., 3D reconstruction and people tracking, are presented and used to compare the performance of the two sensors.

Index Terms— Kinect, depth estimation, 3D reconstruction, people tracking.

1. INTRODUCTION

Acquiring 3D information from real environments is an essential task for many applications in computer vision, robotics, and human-computer interfaces (HCI). All these applications require the availability of real-time depth acquisition systems. The recent introduction of matricial Time-of-Flight (ToF) cameras [1], structured light 3D scanners [2], and multi-camera systems has made real-time acquisition of 3D scenes with both static and dynamic elements available to the mass market.

Among these devices, the Xbox Kinect sensor v. 1.0 [2], which includes a standard RGB camera together with an infrared (IR) structured light sensor, has recently proved to be one of the most widely used sensors thanks to its versatility and the limited cost. Beyond the 3D acquisition and modeling of the scene [3, 4, 5], Kinect v1 has also been employed for tracking persons, recognizing their poses and gestures [6] and identifying their identity [7]. In addition to the HCI applications, the Kinect v1 device has been widely used in the control and navigation systems of robots[8]. Moreover, Maimone and Fuchs [9] presents a telepresence system employing multiple Kinect sensors. Unfortunately, despite the strong versatility and the wide range of new applications that these IR devices enable, the resulting depth signal is affected by a significant amount of noise, which degrades both the quality of the 3D reconstruction [10] and the performance of the algorithms that process depth information.

Starting from these premises, the second version of the Kinect device (Kinect v2) exploits a different technology to acquire depth maps from the scene. In this case, the structured-light IR camera has been replaced by a ToF sensor reducing the amount of noise and improving the accuracy of the measurements.

This paper aims at comparing these two devices. The performance is analyzed considering both the geometric modeling accuracy and the effectiveness in different applications exploiting depth data, e.g., 3D reconstruction and object tracking. Moreover, different acquiring environments have been considered in order to identify those cases in which the performance gap is more significant.

The paper is organized in the following way. Section 2 explains how the technology behind the two sensors works. The proposed measurement procedure and results are discussed in Section 3. The sensors have been tested also in two sample applications as discussed in Section 4. Finally Section 5 draws the conclusions.

2. OVERVIEW OF THE KINECT TECHNOLOGY

The Kinect sensor, introduced by Microsoft in 2010, has been the first example of depth camera targeted to the consumer market. The first generation sensor was a structured light camera but it has been replaced in 2013 by the second generation that instead works with a completely different principle, being a Time-Of-Flight (ToF) camera. In this section we briefly report the main characteristics of the two sensors, for a more detailed description of structured light and ToF technologies see [2].

2.1. Kinect v1

The first generation Kinect depth camera is a light-coded range camera. The two key components are a projector emit-

This work has been partially supported by the University of Padova projects Robotic3D.

a) b)

Fig. 1. First (a) and second (b) generation Kinect.

ting an IR light pattern at 830 nm and a video-camera working on the same wavelength that is able to see the pattern emitted by the projector. The pattern consists in a image of pseudo-random located dots. The basic principle is the active triangulation, i.e., the position of each 3D point is the intersection of the optical rays corresponding to a dot of the projector and the one of the considered pixel in the color camera.

In order to estimate the 3D geometry from a single frame, the projected pattern is uncorrelated along the rows: this permits recognizing each projected pixel by analyzing the samples within a window surrounding the pixel. The specific design of the light-coded pattern is one of the key core technologies of the sensor and has not been disclosed by Microsoft. Furthermore a reference pattern at a fixed distance has been acquired for calibration and the depth is computed from the disparity between the position of the sample in the reference pattern and in the acquired one (Fig. 2). The sensor is capable to acquire a 640×480 depth map at 30 fps, even if the real spatial resolution is smaller. The minimum measurable depth is 0.5 m while the sensor is able to provide reliable measures approximately up to 8 m.

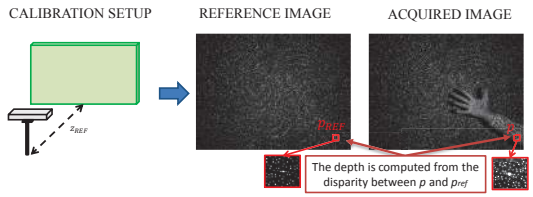


Fig. 2. The Kinect computes the depth from the disparity between the acquired pattern and a reference pattern at a known distance.

2.2. Kinect v2

The Kinect sensor has been replaced by a new device with the introduction of the Xbox One gaming device in November 2013. The second generation Kinect is a completely different device based on the ToF technology. The basic operating principle is the one of continuous wave ToF sensors [2], i.e., an array of emitters send out a modulated signal that travels to the measured point, gets reflected and is received by the CCD of the sensor (see Fig. 3). The sensor acquires a 512×424 depth map and a 1920×1080 color image at 15 to 30 fps depending on the lighting condition, since the sensor exploits an auto-exposure algorithm [11].

Some details of this technology have been disclosed in [11] revealing new innovative elements that overcome some critical limitations of the ToF sensors. First, the emitted light is modulated with a square wave instead of the sinusoidal modulation used by most ToF sensors and the receiver sensor exploits a differential pixels array, i.e., each pixel has two outputs and the incoming photons contribute to one or the other according to the state of a clock signal. This permits measuring the phase difference avoiding issues due to harmonic distortion. Another well-known critical issue is the wraparound error due to the periodic nature of the phase measure, which limits the maximum measurable distance. The Kinect deals with this issue using multiple modulation frequencies. This makes possible to extend the acquisition range since misleading measurements related to phase wrapping can be disambiguated by looking for consistent measures from the ones corresponding to different modulations. Finally, the device is also able to acquire at the same time two images with two different shutter times of 100 μs and 1000 μs. The best exposure time is selected on the fly for each pixel.

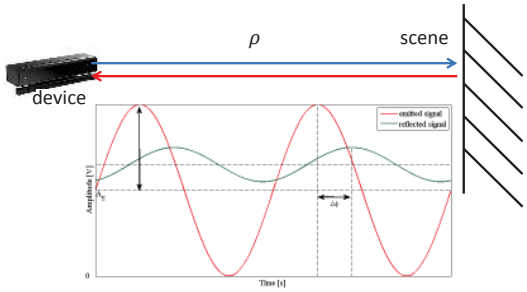


Fig. 3. Basic operation principle of a Time-Of-Flight camera.

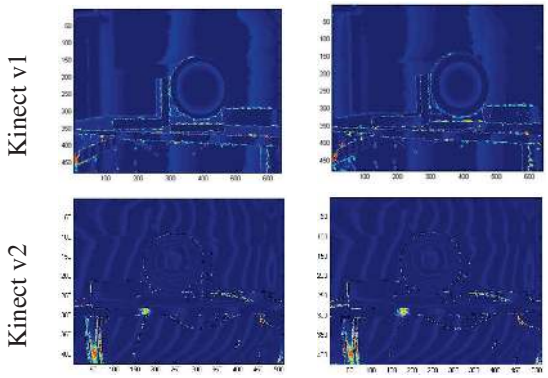
3. EVALUATION OF THE ACCURACY OF THE TWO SENSORS

3.1. Near Range Tests

In this Section, we describe the tests we used to evaluate the performance of Kinect v1 and v2 in the near range. We collected images of two small 3D scenes (2×2 meters) shown in Fig. 4 in four different lighting conditions: no light, low light, neon light and bright light (a incandescent lamp illuminating the scene with 2500W). For each scene, a ground truth point cloud was acquired via a NextEngine 2020i Desktop Laser Scanner (at the considered acquisition range the NextEngine scanner has an accuracy of 380 μm). In order to obtain an objective evaluation, the point clouds acquired by the Kinect sensors were compared with the ground truth according to the following procedure. First, we aligned and overlapped the test point cloud with the ground truth; then, we calculated the dis-

(a) (b)

Fig. 4. Scenes used for the near-range tests. a) `ball_and_book`, b) `tea_and_bear`.



a) no light b) bright light

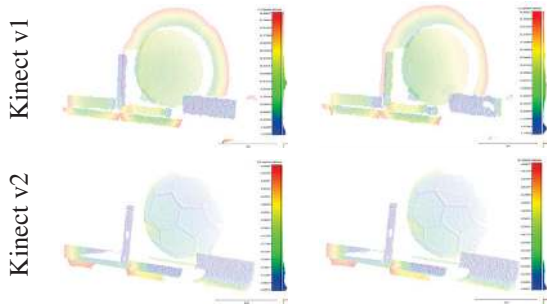
Fig. 5. Near field results with Kinect v1 (*first row*) and Kinect v2 (*second row*): a) standard deviation with no light; b) standard deviation with bright light.

tance between the points in the two clouds; finally, the points with a distance higher than 10 cm are filtered out to exclude the non-overlapping parts and the distance of the test point cloud from the ground truth is recomputed excluding these points.

3.1.1. Scene `ball_and_book`

For the `ball_and_book` scene, in Fig. 5, we report the standard deviation of every pixel for the depth images acquired with the two sensors. Only the results obtained with no light and bright light are shown since the results in the other cases are very similar. For both sensors, we can see that the standard deviation is higher at the object borders, where it is more difficult to correctly estimate depth. However, we did not notice a considerable change in presence of bright light with respect to the case of no light, thus showing a high invariance of these sensors to illumination changes.

Finally, Fig. 6 shows the results of the comparison of Kinect v1 point clouds with the ground truth. The color coding allows to identify those parts where the distance from the ground truth is low (blue), medium (green) and high (red). It can be seen how depth estimation is worse at the object borders, in accordance with the standard deviation results. It can also be noticed how the high distance (red) zones are slightly more extended for the case of bright light, thus proving that



a) no light b) bright light

Fig. 6. Comparison with the ground truth: a) No light; b) bright light.

artificial light do affect depth estimation for Kinect v1, even if to a small extent. On average, as reported in Table 1, Kinect v1 obtained a distance from the ground truth of 45 mm and 47 mm for the cases of no light and bright light, respectively. The corresponding test performed with Kinect v2 is shown in the second row of Fig. 6. The point clouds look visually the same in the two lighting conditions and the mean distance to the ground truth confirms that Kinect v2 is invariant to incandescent artificial light. Moreover, Kinect v2 obtained a mean distance from the ground truth of 23.6 mm, which is 48% lower than what obtained with Kinect v1, while the standard deviation for this near range test is the same for both Kinects.

	No light		Bright light	
	mean dist.	std. dev.	mean dist.	std. dev.
Kinect 1	45.29	25.09	46.98	25.4
Kinect 2	23.62	25.5	23.59	25.45

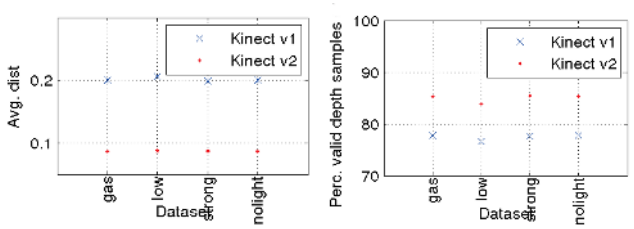
Table 1. Accuracy in $[mm]$ on the `ball_and_book` scene.

3.1.2. Scene `tea_and_bear`

For the `tea_and_bear` scene, we obtained the accuracy results illustrated in Fig. 7a, which are coherent with those obtained on the `ball_and_book` dataset. For this test, we also report the percentage of valid points in the acquired point clouds in Fig. 7b. It can be noticed how Kinect v2 obtains about 10% more valid points than Kinect v1. We can then state that the Kinect v2 is more precise for near range depth estimation and it is more invariant to artificial lighting.

3.2. Far Range Tests

In this section, we describe the experiments we performed to evaluate Kinect v1 and v2 depth accuracy along the whole depth range that these sensors can measure. In particular, we placed the sensors at increasing distances from a flat wall and we captured 100 point clouds at each distance. The distance to the wall given by the Kinect sensors has been measured as the distance of the plane fitted to the acquired point cloud



(a) (b)

Fig. 7. Accuracy of point clouds from 20 Kinect v1 and Kinect v2 acquisitions with respect to laser scan model. Datasets belong to the `tea_and_bear` 3D scene. a) average distance of point clouds over the 20 different acquisition; b) average number of valid depth samples.

a) b)

Fig. 8. a) Platform used for the far range tests and b) robotic platform used for the people tracking and following application.

by means of the RANSAC algorithm. The result has then been averaged over all the point clouds acquired at the same distance and reported in Fig. 9a. Fig. 9b, instead, shows the standard deviation of the inliers to the plane estimated with RANSAC.

To get the ground truth for these measurements, we used the setup shown in Figure 8a, which consists of three laser meters with a precision of $\pm$ 2 mm placed in a triangle configuration. This provided a reliable ground truth distance to evaluate the performances of the sensors and permitted verifying the orthogonality of the sensor locations with respect to the wall. As it can be inferred from the graphs, the accuracy of Kinect v2 is almost constant at all distances, while that of Kinect v1 increases quadratically with the distance due to quantization and depth computation errors. The same trend can be observed for the standard deviation, which is very low and mostly constant for Kinect v2. This validates once again how the ToF technology of Kinect v2 allows to drastically reduce the errors in depth estimation.

In Fig. 10, we also report the resolution of Kinect v1 and v2 as measured at the various distances. The resolution of Kinect v2 is constantly below 1 mm, since this sensor is able to estimate continuous depth variations, while that of Kinect v1 quadratically increases with the distance coherently with the theoretic resolution.

At 7 m, the Kinect v2 accuracy resulted to be below 6 cm and its standard deviation below 10 cm. Since the Kinect v1 obtained values ten times bigger at that far range, the improved accuracy of the second generation Kinect is evident.

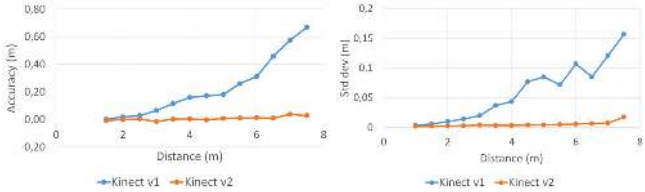


Fig. 9. a) Depth accuracy and b) depth standard deviation of Kinect v1 and v2 at all ranges.

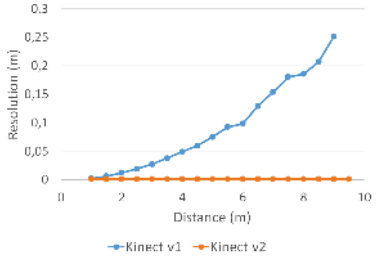


Fig. 10. Depth resolution of Kinect v1 and v2 at all ranges.

3.3. Outdoor Tests

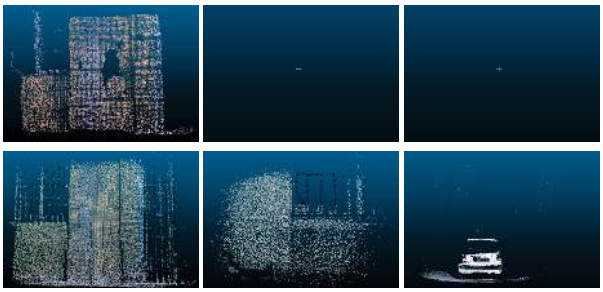
One of the main limitations of the first generation Kinect was the total blindness in presence of sunlight. This problem was due to the fact that Kinect v1 estimates depth by triangulating the position of an infrared pattern of points it projects on the scene. Since sunlight also contains the infrared wavelength used for the Kinect v1 pattern, the Kinect v1 cannot recognize the pattern if some sunlight is present because the scene is filled with that infrared wavelength. For this reason, we performed some tests outdoors in order to see if the ToF technology used in Kinect v2 allows it to estimate depth also in outdoor scenarios. The first column of Figure 11 shows the point clouds obtained with Kinect v1 and v2 at 2.5 meters from an outdoor wall with indirect sunlight. Both Kinects managed to obtain good point clouds, even though those generated by Kinect v1 were not as dense as those of Kinect v2 (it is possible to notice a hole in the middle due to intensity saturation). Anyway, under direct sunlight, Kinect v1 was not able to estimate any valid depth sample while Kinect v2 generated a partial point cloud in the central part of the image (second column of Figure 11). Finally, a similar result was obtained when acquiring a car directly hit by the sunlight (the point clouds are in the last column).

These experimental results allow us to conclude that Kinect v2 showed promising results in outdoor scenarios. Even though the accuracy decreases, it is still able to produce point clouds of outdoor scenes up to 3.5 m of distance.

4. COMPARISON OF THE PERFORMANCE ON SAMPLE APPLICATIONS

4.1. 3D Reconstruction

Until a few years ago, 3D reconstruction required very expensive structured light and laser scanners and a lot of manual



(a) (b) (c)

Fig. 11. Point clouds for Kinect v1 (*first row*) and Kinect v2 (*second row*) in different light conditions. a) a wall in shadow; b) a wall under sunlight; c) a car under sunlight.

work for the subsequent data processing. The introduction of consumer depth cameras and of novel 3D reconstruction algorithms have opened the way to novel reconstruction systems that are both cheaper and simpler to use.

The employment of the Kinect sensor for 3D reconstruction is very attractive and different approaches for this task have been proposed, based on the ICP algorithm or on volumetric approaches. As expected the accuracy of the Kinect is smaller than the one of a laser scanner, but it allows to obtain reasonable results. We employed a 3D reconstruction pipeline derived from the one of [4]. A set of depth maps and color views has been acquired by moving the sensor around the object or scene that is going to be acquired. Then outlier samples are removed and a modified version of the bilateral filter is applied to the depth map [12]. Salient points are then extracted considering both color and depth data and fed to a modified version of the ICP algorithm that uses both color and depth information for optimal 3D views registration. Finally a global optimization is applied and color data is added to the reconstructed 3D scenes.

In order to compare the performances of the two devices for 3D reconstruction we acquired the complete geometry of the 3D scene of Fig. 4b with the two versions of the Kinect sensor and we reconstructed the corresponding point clouds. Two snapshots of the resulting point clouds are shown in Fig.12. In this test the performances of the two sensors are more similar since the data has been acquired in the close range where the performances of the two sensors are more comparable and the reconstruction algorithm is able to remove some of the artifacts of the Kinect v1. However in some regions, e.g., the tea box, it is possible to see that the Kinect v2 has better performances (an enlarged version of the images is available at `http://www.dei.unipd.it/~sim1mil/r3d/kin1v2`). When acquiring larger scenes, e.g., a whole room as in the laboratory example of Fig. 13, that have been reconstructed from more than 400 views, the second generation Kinect allows to obtain a bigger improvement in the quality of the reconstruction. Notice that for both sensors the largest artifacts are concentrated in proximity of edges, where both the first and second generation Kinect have a more limited accuracy with respect to flat regions.

Fig. 12. Snapshot of the 3D reconstruction of `Tea_and_bear` scene obtained from: a) Kinect v1 data; b) Kinect v2 data.

a) b)

Fig. 13. Snapshot of the 3D reconstruction of a research laboratory obtained from: a) Kinect v1 data; b) Kinect v2 data.

4.2. People Tracking

We also compared Kinect v1 and v2 performance for people tracking applications. For this purpose, we collected a video of a group of people moving in front of the two Kinects at distances up to 10 meters and we applied the state-of-the-art people tracking algorithm described in [13]. In Figure 14 and at `http://youtu.be/DIS4en2rznc`, we show a qualitative comparison of the tracking performance of Kinect v1 and v2 at far range, while Table 2 reports a quantitative evaluation of the tracking results obtained with the CLEAR MOT [14] metrics against a manual ground truth.

a) Kinect v1 b) Kinect v2

Fig. 14. Qualitative tracking results obtained with the two generations of Kinect. For every person, the ID number and the distance from the sensor is reported.

Kinect v2 obtained 20% more tracking accuracy (MOTA) than Kinect v1, thanks to the higher precision of its point

Table 2. Tracking evaluation.

	MOTP	MOTA	FP	FN	ID Sw.
Kinect v1	75.22%	41.22%	2.5%	55.6%	27
Kinect v2	80.45	61.56%	7.2%	30.7%	22

cloud at far range, which allows to better detect the shape of people.

At last, we tested the people tracking application in the outdoor scenario reported in Figure 15. Kinect v1 was not able to estimate depth, thus tracking was impossible, while Kinect v2 was able to produce point clouds of people up to 4 meters, thus enabling people tracking up to that range.

Fig. 15. Qualitative result of tracking people outdoors with Kinect v2.

5. CONCLUSIONS

In this paper, we compared the depth data that can be obtained with the first and second generation of Microsoft Kinect sensors. Kinect v2 proved to be two times more accurate in the near range and even ten times more accurate after 6 meters of distance. Moreover, the new sensor presents an increased robustness to artificial illumination and sunlight. We also verified the performance of Kinect v1 and Kinect v2 in 3D reconstruction and people tracking: the accuracy of both applications significantly improves in different environments with Kinect v2. Further research will be devoted to the evaluation of the sensors performances in proximity of the edges and in dynamic environments. Moreover, we will investigate in which situations Kinect v2 does not provide a reliable depth estimation.

6. REFERENCES

[1] S.B. Gokturk, H. Yalcin, and C. Bamji, "A Time-Of-Flight Depth Sensor - System Description, Issues and Solutions," in *Proc. of CVPRW 2004*, June 27 – July 2, 2004, vol. 3, p. 35.

[2] C. Dal Mutto, P. Zanuttigh, and Guido M. Cortelazzo, *Time-of-Flight Cameras and Microsoft Kinect*, SpringerBriefs in Electrical and Computer Engineering, Mar. 2012.

[3] S. Milani and G. Calvagno, "Joint denoising and interpolation of depth maps for MS Kinect sensors," in *Proc. of ICASSP 2012*, 2012, pp. 797–800.

[4] E. Cappelletto, P. Zanuttigh, and G.M. Cortelazzo, "Handheld scanning with 3D cameras," in *Proc. of MMSP*. IEEE, 2013, pp. 367–372.

[5] D.S. Alexiadis, D. Zarpalas, and P. Daras, "Real-time, full 3-d reconstruction of moving foreground objects from multiple consumer depth cameras," *IEEE Trans. on Multimedia*, vol. 15, no. 2, pp. 339–358, 2013.

[6] G. Pozzato, S. Michieletto, E. Menegatti, F. Dominio, G. Marin, L. Minto, S. Milani, and P. Zanuttigh, "Human-robot interaction with depth-based gesture recognition," in *Proc. of IAS-13 Workshops*, 2014.

[7] T. Leyvand, C. Meekhof, Yi-Chen Wei, Jian Sun, and Baining Guo, "Kinect identity: Technology and experience," *Computer*, vol. 44, no. 4, pp. 94 –96, Apr. 2011.

[8] Hyeun J.M., D. Fehr, and N. Papanikolopoulos, "A solution with multiple robots and Kinect systems to implement the parallel coverage problem," in *Proc. of MED 2012*, July 2012, pp. 555–560.

[9] A Maimone and H. Fuchs, "Encumbrance-free telepresence system with real-time 3D capture and display using commodity depth cameras," in *Proc. of ISMAR 2011*, Oct. 2011, pp. 137–146.

[10] S. Milani, E. Frigerio, M. Marcon, and S. Tubaro, "Denoising Infrared Structured Light DIBR Signals Using 3D Morphological Operators," in *Proc. of 3DTV Conference 2012*, Zurich, Switzerland, Oct. 2012.

[11] J. Sell and P. O'Connor, "The Xbox One System on a Chip and Kinect Sensor," *Micro, IEEE*, vol. 34, no. 2, pp. 44–53, Mar 2014.

[12] E. Cappelletto, P. Zanuttigh, and G.M. Cortelazzo, "3D scanning of cultural heritage with consumer depth cameras," *Multimedia Tools and Applications*, 2014.

[13] M. Munaro and E. Menegatti, "Fast RGB-D people tracking for service robots," *Autonomous Robots Journal*, vol. 37, pp. 227–242, 2014.

[14] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the CLEAR MOT metrics," *Journal of Image Video Processing*, pp. 1:1–1:10, January 2008.