Capturing Time-of-Flight Data with Confidence

Malcolm Reynolds, Jozef Doboš , Leto Peel†, Tim Weyrich, Gabriel J Brostow
†Advanced Technology Centre, BAE Systems, Bristol, UK
University College London

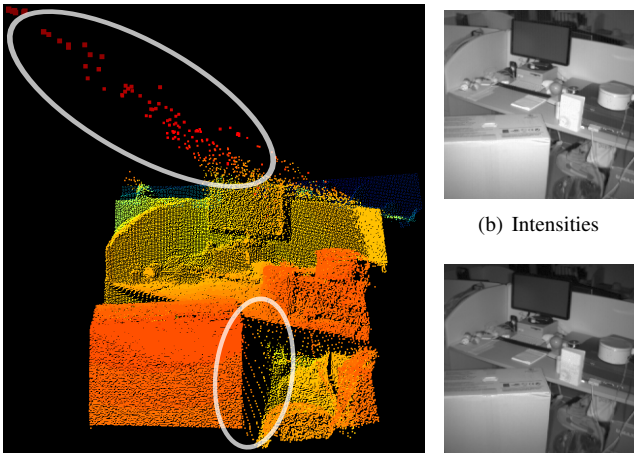
Abstract

Time-of-Flight cameras provide high-frame-rate depth measurements within a limited range of distances. These readings can be extremely noisy and display unique errors, for instance, where scenes contain depth discontinuities or materials with low infrared reflectivity. Previous works have treated the amplitude of each Time-of-Flight sample as a measure of confidence. In this paper, we demonstrate the shortcomings of this common lone heuristic, and propose an improved per-pixel confidence measure using a Random Forest regressor trained with real-world data. Using an industrial laser scanner for ground truth acquisition, we evaluate our technique on data from two different Time-of-Flight cameras[1]*. We argue that an improved confidence measure leads to superior reconstructions in subsequent steps of traditional scan processing pipelines. At the same time, data with confidence reduces the need for point cloud smoothing and median filtering.*

1. Introduction

Time-of-Flight (ToF) cameras have been successfully used for a variety of applications such as Simultaneous Localisation and Mapping (SLAM) [22, 25], 3D reconstruction of static scenes [10, 34], and object tracking and scene analysis [27, 33]. In contrast to stereo vision and triangulation-based scanners, the ToF camera operates from a single viewpoint and does not rely on matching of corresponding features, which greatly increases its robustness in the presence of traditionally difficult scene materials and internal occlusions. Compared to ToF *laser* scanners, ToF cameras deliver a lower-resolution 2D depth image at much higher framerates and at competitive costs of deployment. With costs expected to dramatically decrease in the near future, ToF cameras are likely to become a commodity item.

Despite the advantages of this new technology, it suffers from problems such as low signal-to-noise ratio, multiple light reflections and scattering. All these affect the recorded depth values and hence produce erroneous 3D points, whose reliability depends on various scene parameters. Previous works [21, 34] rely on the amplitude values (Fig. 1(c)) as an indicator of confidence for point removal and filtering. In agreement with Kolb *et al*. [16], our experiments reveal that simply thresholding low-amplitude values is insufficient to remove inaccurate pixels—valid points could be lost before *flying pixels* and other anomalies as depicted in Fig. 1(a) are removed.



(a) 3D Point Cloud (b) Intensities (c) Amplitudes

Figure 1. Flying pixels. (a) Color-coded ToF depth map (red indicates close geometry, blue indicates distant geometry). Erroneous depth readings are caused by materials with unsuitable reflectivity and large depth discontinuities (see bottom ellipse). These pixels collect modulated light from both the foreground and the background and give a resulting distance which is somewhere between the two. The top ellipse shows pixels flying towards the camera, but this is a different phenomenon, caused by highly specular reflections from the top of the monitor. (b) intensity and (c) amplitude images.

We observe that basing a confidence measure on the image amplitude alone disregards valuable additional information. Explicitly incorporating additional cues, such as depth and local orientation, however, may be challenging, due to the difficulty of developing a descriptive error model that covers their interplay. Instead, we propose to use *Random*

[1]For code, data, further results and other supplementary material please see `http://visual.cs.ucl.ac.uk/pubs/tofconfidence/`

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on February 17,2024 at 13:16:50 UTC from IEEE Xplore. Restrictions apply.

Forests [5] to train a regressor to infer the reliability of a depth sample from these input quantities. Random Forests have proven useful as a supervised learning technique that performs feature selection, and are even being used for real-time 3D classification tasks [32].

Analogously to Mac Aodha *et al*. [19] and Carreira and Sminchisescu [6], we obtain a per-point confidence assignment, which could be exploited in other data filtering and 3D reconstruction systems [23].

We qualitatively and quantitatively evaluate the robustness of our approach with two different ToF cameras, comparing it against baseline approaches of depth sample filtering. We show that our machine-learning-based approach outperforms previous methods, leading to a confidence measure of high discriminative power.

2. Background

Optical range cameras [17, 30] measure the *time of flight* of near-infrared (NIR) light emitted by the camera's active illumination and reflected back from the scene. The illumination is *modulated* either by a sinusoidal or pseudo-noise signal, depending on the camera model. The phase shift of the received demodulated signals conveys the time between emission and detection, indicating how far the light has traveled and thus indirectly measuring the depth of the scene. Due to the periodicity in the modulation signal, these devices have a limited working range, usually only up to 7.5 m. Distances beyond this point are recorded modulo of the maximum depth, known as *wrap-around error* [24]. ToF cameras output images of distance and grayscale intensity (Fig. 1(b)). With knowledge of the camera intrinsics, the distance readings can be converted to a 3D point cloud (Fig. 1(a)). Some ToF sensors also record an amplitude value at each pixel, which indicates the amount of modulated light reaching the sensor and disregards external illumination and incoherent stray light (Fig. 1(c)). Similarly to intensity, this amplitude is influenced by scene depth, surface materials and orientation, and lens vignetting [15, 22]. Low amplitude suggests that a pixel has not received enough light to produce an accurate depth measurement, although saturation of high-intensity responses may also lead to invalid readings, despite a high amplitude [13]. If the amplitude values are not provided by the camera, intensity images have been used instead [10, 21].

2.1. Depth Corrections and Calibration

When dealing with measurement devices, it is important to properly understand the causes of potential errors and calibrate for them. Given the optical properties of ToF cameras, a standard intrinsic camera calibration sufficiently describes the focal length, principal focal point and lens distortion [14, 18]. To further improve the unreliable depth readings caused by systematic errors, Lindner and Kolb [18] estimate a higher-order polynomial function which takes into account global as well as per-pixel inaccuracies. Assuming such initial depth calibration, Beder *et al*. [2] further estimate the camera's 3D pose using a single image of a checkerboard. Unfortunately, as Boehm and Pattinson [3] recently showed, even a careful calibration is not sufficient for real-life scenarios. Despite calibration, they report errors in pose estimates that are an order of magnitude larger than the error modeled by the calibration. These errors are scene-dependent and cannot be overcome by calibration [3, 20]. Fuchs and May [12] and May *et al*. [21, 22], on the other hand, find known camera positions using a robotic arm to achieve high-precision results for both the initial calibration and final 3D reconstruction, using variations of the Iterative Closest Point (ICP) [26] algorithm. They demonstrate an overall error accumulation which has to be globally relaxed when closing a loop.

In contrast, Schiller *et al*. [28] use up to 80 images from each camera from their rigidly mounted multi-recording setup, improving on previous results of single ToF camera calibration methods. Similarly, Kim *et al*. [15] combine several video and ToF cameras, while taking into account the scene-dependent effects. They introduce ratios of normalized amplitude values, which are median and Gaussian filtered. Random noise, however, is assumed to be negligible.

2.2. 3D Reconstruction and Filtering

Feulner *et al*. [10] register consecutive ToF frames by detecting binary edge presence in the intensity images and aligning their corresponding 3D coordinates by maximizing the uncentered correlation coefficient. In contrast, Fuchs and May [12] filter points at depth discontinuities, as these exhibit the largest distance error. Swadzba *et al*. [34] present a full acquisition pipeline that relies on several preprocessing steps to improve depth accuracy: a varying *distance-adaptive median filter* is applied to the intensity, amplitude, and depth images, and points with low amplitude are thresholded. Subsequently, a custom-made neighborhood consistency filter detects and removes flying pixels at edge locations in the distance image. Nevertheless, all of the above produce only binary classification of erroneous depth readings.

Fusing high resolution color images with ToF data provides enhanced range maps, as shown by Yang *et al*. [36]. In their work, a bilateral filter aggregates probabilities of depth estimates based on color affinity, iteratively reducing the level of up-sampled blur that occurs due to interpolation in discontinuous areas. Our main inspiration, however, comes from the work of Schuon *et al*. [29]. Their LidarBoost method combines several noisy ToF frames into a single, high-resolution depth image. By assigning zero confi-

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on February 17,2024 at 13:16:50 UTC from IEEE Xplore. Restrictions apply.

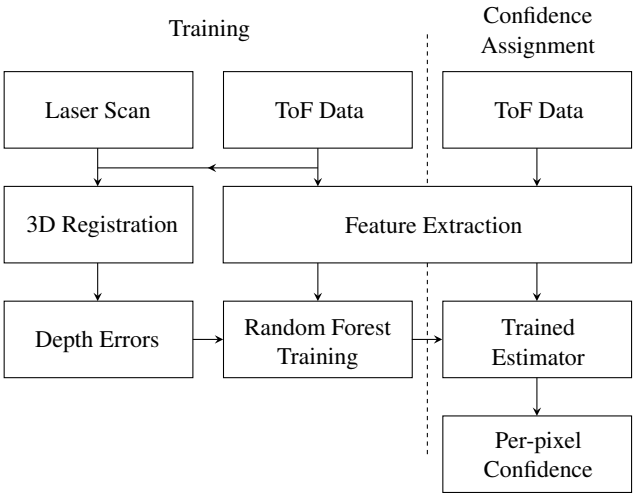


Figure 2. Pipeline for computing a confidence value for each pixel in a ToF depth-scan. The offline supervised training stage builds a Random Forest for regression by correlating extracted features (see §3.4) with depth errors. These errors were computed by comparing the raw depths from the ToF camera against the "ideal" geometry of the training scene.

dence to low amplitude pixels (*i.e*. thresholding), the results are further improved. This technique was later built upon by Cui *et al*. [8] to achieve state-of-the-art reconstruction results. In their work, one point cloud is randomly selected as a reference model, and remaining frames from the sequence are aligned with it. At each point, a multi-variate Gaussian with fixed uniform covariance is centered, and the maximum likelihood estimate is found through Expectation Maximization (EM). Unfortunately, all sources of systematic error (such as surface orientation and reflectance, camera's integration time, or depth) are neglected, and the only assumption is an increasing bias growing from the image center.

In contrast, our approach derives confidence measures across both systematic and random errors. As a single-frame method operating directly on the camera output, our method is further amenable to combination with many of the approaches mentioned above, providing an improved prior for depth sample confidences.

3. Learning ToF Confidence

Each model of ToF camera has its own set of characteristics and inaccuracies. To make meaningful predictions about the confidence of a camera's output, we propose that these characteristics can be learned by comparing data from the camera with ground truth distance. The objective is to assign a confidence value in the range $[0, 1]$ to each point produced by a ToF camera, where 1 signifies a distance reading believed to be completely accurate while 0 signifies a reading that should be ignored. This assignment of confidence is performed with a trained Random Forest operating in regression mode. Regression is preferable here to classification because confidence is continuous, and different applications will stipulate their own expectations of accuracy. Other supervised learning systems such as SVMs [7] could also be used. Random Forests were chosen because they automatically perform feature selection and they require no cross-validation due to the use of out-of-bag estimates during training [5]. We acquire training data of the form $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$. Each feature vector $\boldsymbol{x}_i$ is encoded from a single point recorded by the ToF camera, and the corresponding y_i is the target value, or confidence. These confidences are computed from the difference between the ToF's reported distance and the ground truth distance for each reading. When the disparity between these two values is low, the correct label is close to 1, whereas for pixels where the camera's reported distance is significantly in error, the label value is close to 0. Once the Random Forest has been trained, we can compute a prediction y^* for an unseen $\boldsymbol{x}^*$, allowing a confidence map to be created for each new image (see Fig. 2).

(a) `room` scene (b) `stairs` scene

Figure 3. Intensity images of training scenes from the laser scanner.

3.1. Data Capture

To compute ground truth label values for each pixel, a true distance image $\hat{\boldsymbol{D}}$ must be generated. This image is the same resolution as the output from the ToF camera, and each pixel contains the true distance, *i.e*. the reading that the ToF would have produced were it completely accurate. The viewpoint of this perfect depth image must match the actual location of the ToF camera to be meaningful. In practice, we obtained ground truth by fixing a ToF camera close to a high-end, Time-of-Flight laser scanner, so that the view frustra overlap as much as possible (see Fig. 4). In principle, its measurements may be subject to similar errors as the ToF camera. The scanner's precision and resolution are significantly higher than the camera's [3, 8, 11], which makes it sufficient for ground-truth in our case.

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on February 17,2024 at 13:16:50 UTC from IEEE Xplore. Restrictions apply.

3.2. Calibration

The ToF camera intrinsics were computed using images of a checkerboard and the Bouguet toolbox [4]. The intensity images $\boldsymbol{I}$ have properties similar to a standard grayscale camera, so no modifications to the calibration code are required.

The intrinsic parameters included focal lengths in x and y, the principal point, and a 5-parameter lens distortion model. The extrinsics of the scanner relative to the ToF camera were computed by placing targets in the scene. Paper photogrammetry targets (see Fig. 3) with printed circles were attached to available flat surfaces throughout the scene, spanning the range of depths and spread across the joint field of view of both systems. The 2D location of ellipse centers was determined with sub-pixel accuracy in $\boldsymbol{I}$, and in an equivalent intensity image from the laser scanner. The 2D target centers in $\boldsymbol{I}$ and the 3D locations of the target centers determined by the laser scanner were used to compute an estimate of the extrinsics which minimized the squared reprojection error.

Due to the low resolution of current ToF cameras, the extrinsics calibration based on target detection may only be accurate enough to initialise registration. Experiments with targets at the back and sides of the capture volume showed that background geometry reprojected accurately but foreground objects were offset sideways in $\hat{\boldsymbol{D}}$ compared to the ToF recorded $\boldsymbol{D}$. Using unweighted rigid ICP, the initial registration was refined by aligning the laser point cloud with the ToF point cloud. A streamlined one-shot registration technique could be useful if large quantities of training data were acquired in the future, possibly leading to even better confidence estimates.

3.3. Computing Ground Truth Depth

After calibration, a projection of the laser geometry into the ToF camera is computed for comparison against the ToF's own depth image. The laser scan has much higher resolution than the ToF, (a typical scan was 1251×1055 compared to 200×200 for the ToF, while covering roughly the same field of view), but still consists of points not surfaces. As a result, many laser scanned points at various depths fall within the point spread function of a ToF pixel, as shown in Fig. 4. One could choose the closest, *i.e.* frontmost of the points for $\hat{\boldsymbol{D}}_{i,j}$, but flying pixels occasionally occur in laser data as well and could provide erroneous depths. Neither the minimum nor the mean of these points' depths would be appropriate (Fig. 4). All laser points which project into the sensor region of pixel (i, j) are used to compute $\hat{\boldsymbol{D}}_{i,j}$. We fit a Mixture of 3 Gaussians using EM to the depth values of these points. $\hat{\boldsymbol{D}}_{i,j}$ is set to the minimum of the means of the fitted Gaussians. Finally, due to parallax and occlusion, the ToF camera may be sampling parts of the surface geometry that were unseen by the laser scanner. By placing the ToF camera very close to the laser scanner, we minimize the areas of the ToF image for which no ground truth exists.

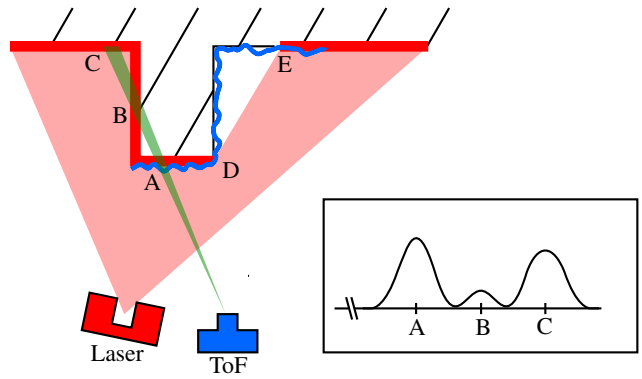


Figure 4. Overhead view of a scene acquired by the laser scanner (red) and ToF camera (blue). We project the laser geometry into the ToF camera (green) without considering occlusions. The inset depth histogram shows how laser points from distinct locations can project into a single ToF pixel, but only points from A are valid. The area visible to the ToF between D and E is not covered by the laser scanner and so no training data is acquired here.

Each depth value in $\hat{\boldsymbol{D}}$ must be converted to a training label $y \in [0, 1]$. Intuitively, the difference between $\hat{\boldsymbol{D}}$ and the ToF's $\boldsymbol{D}$ is normalized by $\hat{\boldsymbol{D}}$ to find the relative depth error. Relative depth error is used because we assume that it is unrealistic to expect ToF cameras to have uniform absolute accuracy at all depths [3, 11, 15]. The relative depth error is passed through the $\arctan$ sigmoidal function, reflected, and rescaled so that a relative distance error of 0 becomes 1, and a hypothetical infinite distance error becomes 0. The final confidence label y is

$$y_{ij} = 1 - \frac{2}{\pi} \arctan\left(\alpha \frac{abs(\hat{\boldsymbol{D}}_{i,j} - \boldsymbol{D}_{i,j})}{\hat{\boldsymbol{D}}_{i,j}}\right), \quad (1)$$

where the parameter α controls how quickly y tends to 0 as relative distance error tends to infinity.

3.4. Feature Extraction

A feature vector $\boldsymbol{x}_{i,j}$ is computed for each pixel (i, j) in the ToF image. To allow the Random Forest to generalize the properties of pixels with high and low confidence, three types of data are used to construct the feature vector. Note that Random Forests calculate the importance of each entry in the feature vector as part of the training process. This feature selection property makes it possible to use a large number of features that *may* be correlated with confidence, and have the training process automatically determine which ones were in fact useful. The categories of features are: local features from a single point (i, j), spatial features calculated from a neighbourhood surrounding (i, j), and global features calculated from the entire frame.

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on February 17,2024 at 13:16:50 UTC from IEEE Xplore. Restrictions apply.

Local Features. The primary feature values at each ToF pixel (i, j) consist of the intensity $\boldsymbol{I}_{i,j}$, signal amplitude $\boldsymbol{A}_{i,j}$ if available, and distance $\boldsymbol{D}_{i,j}$. These elements are included to allow learning of different physical phenomena, *e.g.* pixels not receiving enough light to compute an accurate distance where the scene has low reflectivity in NIR light spectrum, or receiving too much light so that pixels saturate. Distance is included because, in accordance with the inverse square law, surfaces a larger distance away are likely to have a higher error magnitude. The radial distance of each pixel is also included in the feature vector to allow the possibility of learning different error characteristics near to the edge of an image.

Spatial Features. Filters are used to incorporate local spatial information about the scene. Neglecting perspective projection of the camera, we find approximations to the normal angle in x and y. A Laplacian filter is convolved with both the distance and intensity images. This filter is commonly used for edge detection, so it is included in the feature vector to allow the forest to learn the relationship between error and depth discontinuities. Our initial observations indicated that flying pixels at depth discontinuities were among the most noticeable artifacts in ToF data, but that not all sharp depth changes were incorrect. As well as the 3×3 Laplacian kernel, 5×5 and 7×7 versions are used to incorporate information at different scales.

Gabor filters [9] describe a family of kernels, which differ from the Laplacian in that they can be set to a particular orientation. Whereas a large response from the Laplacian indicates the presence of an edge, Gabor filters at different orientations θ produce a response which can determine the direction of the edge. Filters at $0°$, $45°$, $90°$ and $135°$ were computed over the image. Due to the more complex form of the Gabor wavelet, a 13×13 filter was used over the whole image.

Global Features. Inspired by Kim *et al.* [15] we allow the forest to learn the relationship between some feature at a particular point and the distribution of feature values across the whole image. For every feature listed above except radial distance, the minimum, maximum, mean, and median values across the whole frame were included. Because of this, each feature already mentioned contributes 5 elements to the feature vector, giving a final dimensionality of 91 (18 features $\times$ 5, plus radial distortion). Our final feature vector takes advantage of many of the cues used deterministically in previous works to probabilistically assess the quality of ToF data.

4. Experiments

To evaluate the confidence assignment method, scenes were captured with a PMD CamCube 2.0 ToF camera. This camera has a resolution of 200×200 px, and provides both amplitude and intensity images. One scene of a staircase (`stairs` Fig. 3(b)) and one scene of an office (`room` Fig. 3(a)) were captured, both with a number of clutter items placed at different depths in the scene. The objects were chosen to include a wide variety of surface reflectance properties, including plastic, wood, textiles, polystyrene, and brushed metal. The object placement was intended to create multiple depth discontinuities of varying magnitudes, and to span a variety of different errors on which to train the model. For these scenes, a laser scanner was used to acquire ground truth, so the scenes could be used for training and quantitative experiments. Additional scenes were captured without ground truth (see supplementary material for full details) to be used for qualitative evaluation, so that our algorithm was tested on a total of 4 different scenes.

4.1. Quantitative Results

To generate quantitative results, a forest was trained on the `stairs` scene and tested on the `room` scene. 125 trees were generated with a maximum depth of 25, forest accuracy of 0.0001 and the labels were generated with $\alpha = 7$.

It is standard practice when dealing with noisy point cloud data to delete points which are deemed to have the highest error, so that further processing must cope only with more reliable data. Previous work such as Schuon *et al.* [29] used amplitude as a per-point quality indicator, deleting points with low values. The probabilistic confidence output from our algorithm can be thresholded in the same way to produce an alternative binary classifier for points.

During the review process a "geometric" baseline algorithm was suggested, which classified each point by computing the distance to the closest other point in each scan, so a small distance would indicate a reliable point. We include comparisons of our method to both this method and the amplitude baseline.

To quantify the relative improvement of our method over simply using the amplitude or examining geometric properties of the data, we present ROC curves on `room` in Fig. 5. After discarding points for which no laser geometry or filter output was available, 31k points were left for classification. The ground truth in this *binary* classification task was computed by thresholding relative distance error from (1). Fig. 5(a) shows the classifiers' accuracy at correctly identifying which points are within 4% of the true depth, and Fig. 5(b) shows the same for 25%. ROC curves are generated by sweeping each metric from minimum to maximum value, where the metrics are amplitude, confidence posterior probability, and inverse distance to sample point.

The average ROC Area Under Curve (AUC) calculated across the error tolerance range $\{1 \ldots 25\}\%$ (see Table 1) shows our technique outperforming both the amplitude and geometric baselines on average, and in fact in each of these

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on February 17,2024 at 13:16:50 UTC from IEEE Xplore. Restrictions apply.

Threshold (relative error, %)	Amp AUC	Geom AUC	Conf AUC
4	0.639	0.599	**0.729**
25	0.685	0.722	**0.757**
$\{1 \ldots 25\}$ (average)	0.654	0.661	**0.736**

Table 1. AUC of ROC curves for Amplitude (Amp), Geometric (Geom) and Confidence (Conf) algorithms.

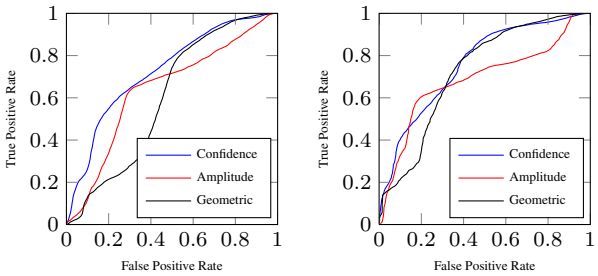


Figure 5. ROC curves. Comparison of our method against amplitude and geometric filtering on: (a) less than 4% error with $21,290$ actual positives and $9,931$ actual negatives, (b) less than 25% error with $29,102$ actual positives and $2,119$ actual positives. 4% error corresponds to an error margin of 16cm on a surface 4 meters away.

25 cases, the AUC for confidence thresholding is largest. Further graphs for these settings can be found in the supplementary material. Our technique no longer has a clear advantage over the baselines for tolerances above 25%, but very few bad points remain. Over 93% of the data already falls within this tolerance.

4.2. Qualitative Results

It is informative to examine which points are deleted by each ToF assessment algorithm. We compare amplitude thresholding, the thresholding technique of Swadzba *et al*. [34], and confidence thresholding, referred to as methods A, B, and C (Fig. 6). Method B performs amplitude thresholding after smoothing the amplitude image, and additionally discards points based on edge detection. It is unclear how to vary both the amplitude and edge parameters to draw a fair ROC. As our technique is only being used to threshold points and not actually change their positions, we omit their distance smoothing step.

Using the recommended parameters, method B removes 27% of the $35,344$ points in `room`. Thresholds for methods A and C were adjusted to remove the same number of points for direct comparison, although we would use a lower threshold on method C to improve results. As the Fig. 6 insets show, methods A and B are not as comprehensive in their removal of flying pixels at depth discontinuities. Please see the supplementary material to better inspect the 3D data.

4.3. Feature Importance

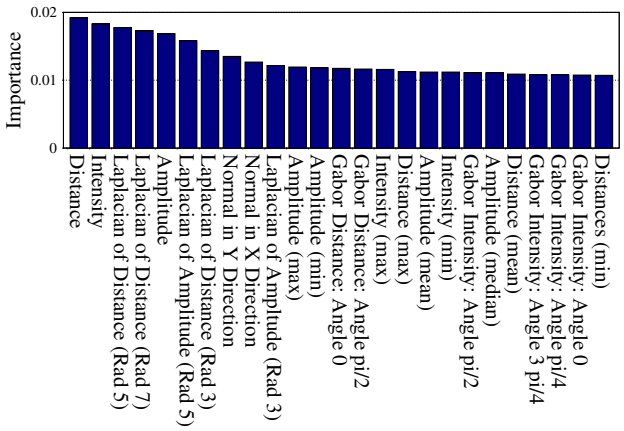


Figure 7. The 25 most important Random Forest features after training on `stairs` and `room` data.

The Random Forest training process calculates the relevance of each entry of the feature vector, which allows us to analyze which features carry the most information as shown in Fig. 7. A full graph is included in the supplementary material. Distance is the most important factor, followed by intensity, the Laplacian of the distance image at two different scales, and finally the amplitude. For this analysis, the forest has been trained on both `stairs` and `room`, which allows the variation in the global features to contribute to the confidence model. It is shown that the global features based on amplitude contain the most discriminative content. Note that when training on only one scene, all the global features are constant across the whole training set and so have the same level of importance.

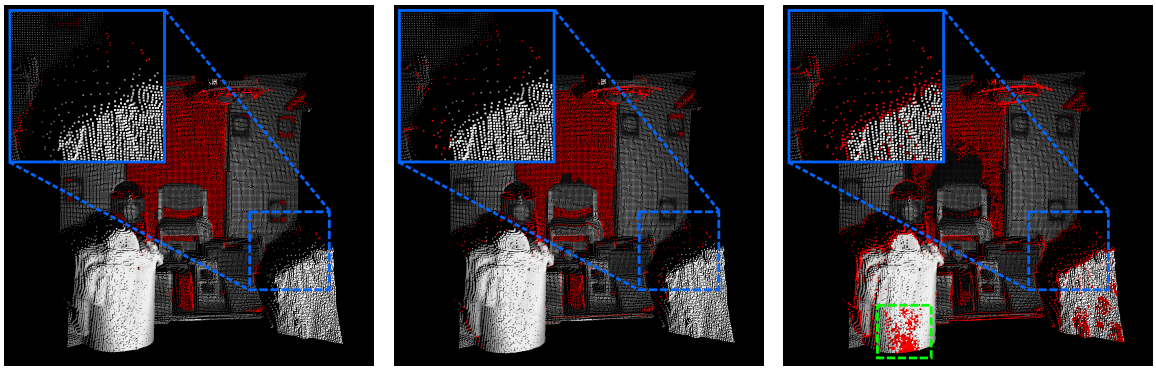
4.4. Application to Different Model ToF

To evaluate the effectiveness of our technique on a different model of ToF camera, we also collected training data for the Mesa SwissRanger SR-3100. This camera does not provide an amplitude image, only intensity and distance, so the feature vector was correspondingly smaller. Fig. 8 shows a photo of a test scene which included large depth discontinuities and many flying pixels. The thresholded point cloud once more shows successful elimination of flying pixels and other forms of outliers.

5. Discussion and Future Work

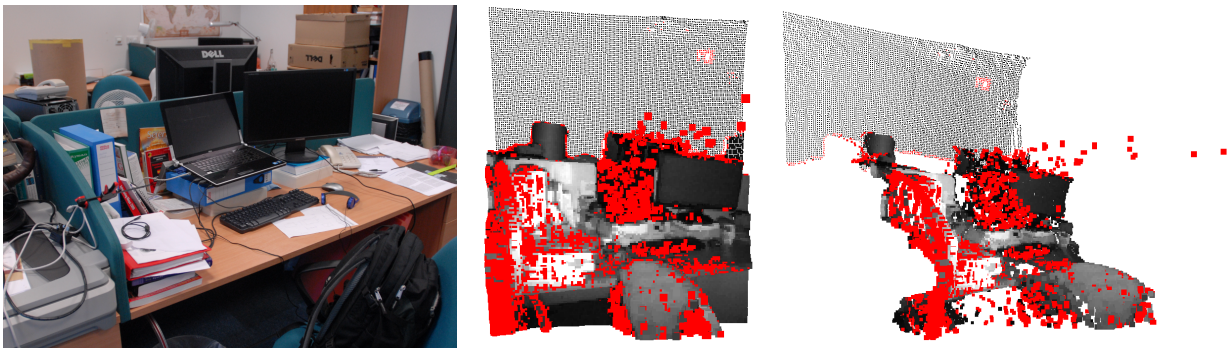
We have presented a method to assign per-point confidences to ToF scans using a supervised learning approach. Both qualitative and quantitative results show marked improvement on contemporary methods for the removal of inaccurate points. The algorithm is particularly good at de-

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on February 17,2024 at 13:16:50 UTC from IEEE Xplore. Restrictions apply.



(a) Amplitude (method A) (b) Swadzba *et al*. [34] (method B) (c) Confidence (method C)

Figure 6. Points removed by a method are shown in red. The green box in (c) shows an artifact of our method, see §5.



(a) (b) (c)

Figure 8. Cluttered desk environment captured by Mesa SwissRanger 3100. (a) photo of the scene, (b) and (c) are respective frontal and side views of the obtained point cloud where points with confidence below 0.6 are shown in red.

tecting flying pixels. Demonstrations using different camera models suggest that our confidence assignment method is hardware agnostic. We anticipate methods such as Schuon *et al*. [29] could be improved with our confidences. ICP is a common algorithm for merging point clouds, and could be augmented by applying confidence weighting to each matching pair of points. Our confidence measure could potentially be adapated to support the semantic cues used for SfM in Bao and Savarese [1]. Similarly, the initial mesh models reconstructed using multi-view stereo and then refined based on shading in Wu *et al*. [35] could instead start from our ToF data, and be more malleable where the confidence was low.

As with all supervised learning methods, for a high level of performance, a representative training set is required. Despite the limited size and variety in our training set, the technique has proven successful on unseen test data. Some artifacts remain, such as the incorrectly deleted points in the green square in Fig. 6(c). We surmise that adding data from a wider range of scenes, including bright nearby objects, would further improve generalisation ability and performance. The registration problems detailed in §3.2 could be solved either with higher-resolution ToF cameras or a better distribution of targets within the scene. Applying a non-linear calibration for the systematic depth [3, 8, 11] could improve the raw distance readings which would then be used to compute feature vectors.

An area we have not yet explored is the possibility of using a variant of the standard Random Forest or indeed a different Machine Learning algorithm altogether. Recent developments such as Adaptive Random Forests [31] have the potential to increase the speed of the confidence assignment, which is currently around 5 seconds for a 200×200 frame.

Acknowledgements

We are grateful to Jan Boehm and Stuart Robson, for sharing their laser scanner and expertise, and to Marc Pollefeys for lending us the SR-3100. Thanks to Maciej Gryka, Oisin Mac Aodha, and Frédéric Besse who helped during data collection, and to Jim Rehg for valuable discussions. We would also like to thank the reviewers for their feedback and suggestions. The student authors were supported by the UK EPSRC-funded Eng. Doctorate Centre in Virtual Environments, Imaging and Visualisation (EP/G037159/1).

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on February 17,2024 at 13:16:50 UTC from IEEE Xplore. Restrictions apply.

References

[1] S. Y. Bao and S. Savarese. Semantic structure from motion. *CVPR*, 2011.

[2] C. Beder and R. Koch. Calibration of focal length and 3D pose based on the reflectance and depth image of a planar object. *Int. J. Intell. Syst. Technol. Appl.*, 5:285–294, 2008.

[3] J. Boehm and T. Pattinson. Accuracy of exterior orientation for a range camera. In *ISPRS Commission V Mid-Term Symposium 'Close Range Image Measurement Techniques'*, volume 38, pages 103–108, 2010.

[4] J.-Y. Bouguet. *Visual methods for three-dimensional modeling*. PhD thesis, California Institute of Technology, Pasadena, CA, USA, 1999. AAI9941097.

[5] L. Breiman. Random forests. *Machine Learning*, 45(45):5–32, 2001.

[6] J. Carreira and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *CVPR*, pages 3241–3248, 2010.

[7] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, first edition, 2000.

[8] Y. Cui, S. Schuon, C. Derek, S. Thrun, and C. Theobalt. 3D shape scanning with a time-of-flight camera. *CVPR*, pages 1173–1180, 2010.

[9] J. G. Daugman. Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research*, 20(10):847–856, 1980.

[10] J. Feulner, J. Penne, E. Kollorz, and J. Hornegger. Robust real-time 3D modeling of static scenes using solely a time-of-flight sensor. In *6th IEEE Workshop on Object Tracking and Classification Beyond and in the Visible Spectrum*, pages 74–81, 2009.

[11] S. Fuchs and G. Hirzinger. Extrinsic and depth calibration of ToF-cameras. In *CVPR*, pages 1–6, 2008.

[12] S. Fuchs and S. May. Calibration and registration for precise surface reconstruction with time-of-flight cameras. *Int. J. Intell. Syst. Technol. Appl.*, 5:274–284, 2008.

[13] S. B. Gokturk, H. Yalcin, and C. Bamji. A time-of-flight depth sensor - system description, issues and solutions. In *CVPR Workshop*, volume 3, page 35, 2004.

[14] T. Kahlmann, F. Remondino, and H. Ingensand. Calibration for increased accuracy of the range imaging camera swissranger. In *ISPRS*, volume XXXVI, pages 136–141, 2006.

[15] Y. M. Kim, D. Chan, C. Theobalt, and S. Thrun. Design and calibration of a multi-view TOF sensor fusion system. *CVPR Workshop*, pages 1–7, 2008.

[16] A. Kolb, E. Barth, R. Koch, and R. Larsen. Time-of-flight sensors in computer graphics. *Eurographics State of the Art Reports*, pages 119–134, 2009.

[17] R. Lange and P. Seitz. Solid-state time-of-flight range camera. *IEEE Journal of Quantum Electronics*, 37 (3):390–397, 2001.

[18] M. Lindner and A. Kolb. Lateral and depth calibration of PMD-distance sensors. In *ISVC 2*, pages 524–533, 2006.

[19] O. Mac Aodha, G. J. Brostow, and M. Pollefeys. Segmenting video into classes of algorithm-suitability. In *CVPR*, 2010.

[20] S. May, D. Droeschel, D. Holz, S. Fuchs, E. Malis, A. Nüchter, and J. Hertzberg. Three-dimensional mapping with time-of-flight cameras. *J. Field Robot.*, 26(11–12):934–965, 2009.

[21] S. May, D. Droeschel, D. Holz, C. Wiesen, and S. Fuchs. 3D Pose Estimation and Mapping with Time-of-Flight Cameras. In *IROS Workshop on 3D Mapping*, 2008.

[22] S. May, S. Fuchs, D. Droeschel, D. Holz, and A. Nüchter. Robust 3D-Mapping with Time-of-Flight Cameras. In *IROS*, pages 1673–1678, 2009.

[23] M. Pauly, N. J. Mitra, J. Giesen, M. Gross, and L. J. Guibas. Example-based 3D scan completion. In *SGP*, page 23, 2005.

[24] J. Poppinga and A. Birk. A novel approach to efficient error correction for the SwissRanger time-of-flight 3D camera. *RoboCup 2008*, pages 247–258, 2009.

[25] A. Prusak, O. Melnychuk, H. Roth, I. Schiller, and R. Koch. Pose estimation and map building with a time-of-flight-camera for robot navigation. *Int. J. Intell. Syst. Technol. Appl.*, 5(3/4):355–364, 2008.

[26] S. Rusinkiewicz and M. Levoy. Efficient variants of the ICP algorithm. In *3DIM*, pages 145–152, 2001.

[27] T. Schamm, M. J. Zollner, S. Vacek, J. Schroder, and R. Dillmann. Obstacle detection with a photonic mixing device-camera in autonomous vehicles. *Int. J. Intell. Syst. Technol. Appl.*, 5(3/4):315–324, 2008.

[28] I. Schiller, C. Beder, and R. Koch. Calibration of a pmd camera using a planar calibration object together with a multi-camera setup. In *ISPRS*, volume Vol. XXXVII. Part B3a, pages 297–302, 2008.

[29] S. Schuon, C. Theobalt, J. Davis, and S. Thrun. Lidarboost: Depth superresolution for ToF 3D shape scanning. *CVPR*, pages 343–350, 2009.

[30] R. Schwarte, Z. Xu, H.-G. Heinol, J. Olk, R. Klein, B. Buxbaum, H. Fischer, and J. Schulte. New electro-optical mixing and correlating sensor: facilities and applications of the photonic mixer device (PMD). In *SPIE*, 1997.

[31] A. Schwing, C. Zach, Y. Zheng, and M. Pollefeys. Adaptive random forest - how many "experts" to ask before making a decision? *CVPR*, 2011.

[32] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. *CVPR*, 2011.

[33] A. Swadzba, N. Beuter, J. Schmidt, and G. Sagerer. Tracking objects in 6D for reconstructing static scenes. In *CVPR Workshop: Time of Flight Camera based Computer Vision*, 2008.

[34] A. Swadzba, B. Liu, J. Penne, O. Jesorsky, and R. Kompe. A comprehensive system for 3D modeling from range images acquired from a 3D ToF sensor. In *ICVS*, 2007.

[35] C. Wu, B. Wilburn, Y. Matsushita, and C. Theobalt. High-quality shape from multi-view stereo and shading under general illumination. *CVPR*, 2011.

[36] Q. Yang, R. Yang, J. Davis, and D. Nistér. Spatial-depth super resolution for range images. In *CVPR*, pages 1–8, 2007.

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on February 17,2024 at 13:16:50 UTC from IEEE Xplore. Restrictions apply.