3D Disaster Scene Reconstruction Using a Canine-Mounted RGB-D Sensor

Jimmy Tran
Ryerson University
Department of Computer Science
Toronto, Canada
q2tran@ryerson.ca

Alex Ufkes
Ryerson University
Department of Computer Science
Toronto, Canada
aufkes@ryerson.ca

Alex Ferworn, Mark Fiala
Ryerson University
Department of Computer Science
Toronto, Canada
<aferworn, mark.fiala>@ryerson.ca

***Abstract*— A 3D map of the interior of a disaster site that pinpoints the location of trapped victims would greatly aid search and rescue efforts. We propose using a canine-mounted RGB-D sensor; a trained rescue dog can carry an image sensor through the site to build a 3D model useful for rescuers. However, the registration of the data provides challenges beyond those typically faced in scene reconstruction due to the rapid motion and sudden pose changes. We provide a solution whereby a pre-processing step identifies good frames to combine from a stream of RGB-D image frames. These selected images are then combined into the larger model by calculating a relative pose using the 3D location of key points matched in the visible images. Results are presented of 3D models constructed using data collected from the canine platform.**

Keywords-3D Reconstruction, RGB-D Sensor, Canine Augmentation, Urban Search and Rescue

I. INTRODUCTION

Dogs have been used to search for people buried in rubble at least since WWII [1]. With training, their agility and sense of smell makes them the de facto standard for the traversal of rubble for the purposes of finding live trapped humans. Unfortunately, if a human handler cannot actually see where their dog is when the dog indicates a live human, it is difficult to know how to get to the location of the victim. The initial goal of Urban Search and Rescue (USAR) responders is to pinpoint the location of the victim and survey the structural condition of the interior environment.

There have been many attempts to replicate the mobility characteristics of dogs—on and in rubble—using response robots [2, 3]. The advantage of using a robot for this purpose is that it can be equipped with a wide variety of cameras and similar sensors which provide rescue teams with a lot of information about the route the robot took. This allows for the creation of rich visual models that can be used to plan a rescue [4]. Unfortunately, aerial robots cannot penetrate rubble and the terrains traversable by ground robots are limited compared to search dogs.

In this paper we explore the possibility of using dogs as a means of carrying sensors that are capable of producing feature-rich, visual models. Since a dog is not a stable sensor platform, the obvious problem with mounting an RGB-D sensor on a dog is that the data will suffer from severe motion blur. We wanted to investigate the failure and success conditions of the proposed system under these conditions to determine whether further research is warranted.

Figure 1: USAR dog wearing the Kinect harness.

We propose a method to filter image frames based on the statistics of matching performance on local image frames. Peaks in the matching performance are used to identify frames likely to provide a useful addition to the 3D map. We call the algorithm Intelligent Frame Selector (IFS). Our experiments and findings are presented in the rest of the paper.

Contributions

There are three contributions in this paper. First, we present a method of mounting an RGB-D sensor on the dog and collecting the data. To the best of our knowledge, this is the first set of RGB-D data collected using a dog as a platform. Second, we ran experiments and made observations that provided an analysis correlating the dog's gait with blurred images. Third, we developed an algorithm that looks ahead and skips over noisy frames to select appropriate frames for registration. The results presented in this paper shows that the algorithm is useful in creating more accurate models.

978-0-7695-4983-5/13 $26.00 © 2013 IEEE
DOI 10.1109/CRV.2013.15

II. Related Work

The use of sensors on dogs is discussed in [5, 6]. Canine Augmentation Technology (CAT) centers around a wearable harness designed to be as functional as possible without inhibiting the dog in any way. Various sensors and hardware have been attached to canines in this manner; including dual shoulder mounted cameras, onboard computers, wireless mesh routers, as well as a small, remote deployable robot [7].

There has been myriad work done with respect to 3D environment modeling both in general, and using the Kinect, an RGB-D sensor developed by Microsoft. Early work with the Kinect began with visual odometry and mapping on both ground and air platforms [8, 9], and evolved into full visual Simultaneous Localization and Mapping (SLAM) systems [10]. Some of the latest, most impressive results have been achieved by the KinectFusion project [11]. This system utilizes the Iterative Closest Point (ICP) algorithm to register the 3D depth data directly, without making use of the Kinect's color camera outside of texturing the finished model. One restriction of this system is that, due to memory constraints, it is spatially limited to a region approximately 7m3. This problem was solved by Kintinuous [12], which removes the spatial limitation by modifying the KinectFusion algorithm to allow environment sizes that can vary dynamically. However, preliminary tests with the KinectFusion algorithm implemented in the PCL library [13] showed that it was unable to model the data gathered by our canine mounted sensor.

III. Technical Approach

The problem we want to address is how to track the path of a search dog. Our proposal is to mount an RGB-D sensor on the dog's back, looking at an upward direction behind the dog. The sensor will record the journey of the dog. If the recorded data can be registered then the path can be recreated. The methods employed to test our hypothesis can be described in three phases: data collection is the recording of on-dog data; data analysis is the examination of how motion blur affects the registration process; and data processing is where different strategies are employed to perform registration on noisy data.

A. Data Collection

The sensor chosen is the Microsoft Kinect. Since its release, the Kinect has become ubiquitous among roboticists and researchers due to its low cost and ability to produce dense 3D point cloud data. For a complete description of the Kinect's capabilities, see [14]. The system records RGB-D data from the Kinect at a rate of 30Hz and at a resolution of 640x480.

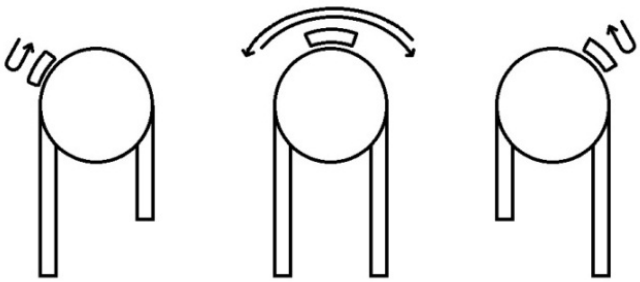

Figure 2: Depiction of sensor motion on a moving dog

The Kinect sensor is mounted on a rigid platform attached to a custom designed harness. The Kinect is angled upwards just below the dog's shoulders, ensuring that the dog obscures as little of the Kinect's view as possible. The recording unit consists of a battery pack and single board computer that interfaces with the Kinect sensor and stores the recorded data. It is attached to a bag that is strapped to the dog's chest. Figure 1 shows a picture of the system worn by a USAR dog.

B. Data Analysis

Even though efforts were made to ensure that the Kinect sensor stays as rigid on the dog's back as possible, swaying of the sensor from side to side still occurs. Through observations, Figure 2 was produced to demonstrate the motion of the sensor on the dog's back. As the dog's speed increases, the speed and severity of the sensor's swaying motion increase as well.

To identify usable frames within the motion blurred data, we began by analyzing the motion of the sensor with respect to the gait of the dog. It was observed that the sensor tends to oscillate very consistently with the dog's stride. When one of the dog's front legs reaches its highest point in the stride, the sensor is rolled the farthest away from that leg. This creates an inflection point where the sensor has rolled farthest to one side and begins rolling back in the other direction as the dog lowers that leg.

At these inflection points in the data, there are typically a handful of relatively useful frames with minimal motion blur (Figure 3). The downside is that these clear frames are also the most severely rotated and spatially separated. It also followed from this that the frames midway through the dog's gait were the most blurred and therefore unusable despite the fact that at these points the sensor is mostly level without any rotation.

Figure 3: Consecutive series of blurry frames (red) between clear frames (green) on the inflection points.

C. *Data Processing*

The method used to register frames is similar to the method presented in [15], which highlights the efficacy of the optic flow tracking algorithm [16] even under extreme conditions. Here, however, the pairs of usable image frames contain too much spatial separation for optic flow to be used reliably. Hence, descriptor matching is used to obtain image point correspondences. The proposed system utilizes a GPU implementation of the popular Speeded-Up Robust Features (SURF) [17] algorithm in the Open Computer Vision Library. From each pair of input images, SURF features are extracted and matched to produce a list of 2D image correspondences. These matching pairs of image points are projected into 3D using the depth data for each pixel provided by the Kinect, resulting in a new set of 3D correspondences.

The transformation between these sets of corresponding 3D points can be determined using Singular Value Decomposition (SVD). To remove outliers such as poor or false feature matches, a Random Sample and Consensus (RANSAC) loop [18] was used. This loop randomly selects a set of four matching pairs of points and computes a hypothesis transformation. This hypothesis is then applied to all the matched points from one frame. If the result of applying the transformation to a given point is within a certain threshold of the matched point, the matching pair is considered to be an inlier. After a predetermined number of iterations, the transformation that yields the highest number of inliers is selected. A least-squares SVD is then performed on the set of inlier matches which yields the final, accepted 3D transformation between the two frames.

We first tried to input the collected data into our visual odometry system without any selective rejection of blurry frames. As expected this produced poor, inaccurate models.

Next the system was modified to use a human operator to manually choose the set of frames to match with each other. This experiment was a proof of concept to see if human intuition and recognition of clear frames would help the system produce better models.

Finally an algorithm was developed to automate the manual process of using the aid of a human operator. We call the algorithm Intelligent Frame Selector (IFS). The first step in the algorithm is to produce a large matrix of matches between every single frame and the next N number of frames where N is the window size. This creates a histogram of the number of match inliers found in each frame. The second step is to generate a list of local maxima. Figure 4 shows a sample histogram of matches and the red bars highlights the local maxima.

It was observed that these local maxima each lie across a series of non-blurred frames from the inflection point of the sensor motion described in section 2.2. The algorithm selects the frame lying on the furthest local maximum that has a greater number of inliers than an acceptable threshold. The acceptable threshold is dynamically calculated. It is the average inlier count of every frame in the window with an inlier count higher than four.

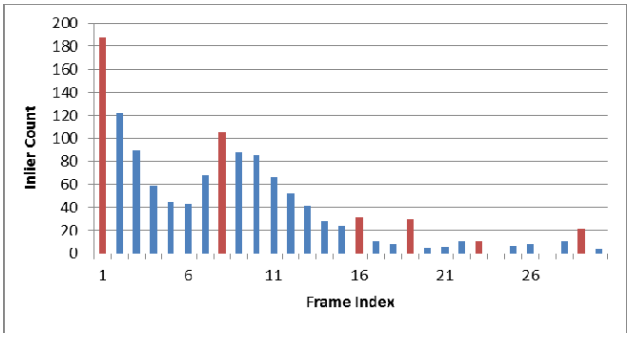

Figure 4: Sample histogram of a window of matches.

Registration is performed between the chosen frame and the origin frame. The process is then repeated from this new frame. Essentially the algorithm steps from local maximum to another local maximum, avoiding blurry frames as much as possible.

One of the parameters that can be adjusted is the size of the search window. This would depend on the data and the environment. From experimentation, we found that the window size should be adjusted based on the speed of the dog. For all but the slowest tests, within a span of 30 frames the dog had moved a distance greater than the ideal range of the Kinect's depth camera, making it impossible to match beyond 30 frames.

Searching for matches across an entire window for each new frame may seem expensive but in reality there is only a small amount of redundant computation. Features are extracted once per frame and then stored until the window passes over that frame. Matching is the only process performed multiple times per frame, but it typically takes under 2ms with GPU acceleration.

IV. EXPERIMENTS

Our test subject for this paper was Dare, a Federal Emergency Management Agency (FEMA) certified USAR dog. Under the guidance of Dare's handler, we ran several trials.

At each trial, Dare walked/ran down a long, straight, interior hallway (Figure 5) at three different speeds, slow, medium and fast. In the slow and medium speed trials, Dare was led by his handler to control the pace. In the fast speed trial, Dare started at one end of the hallway and was called by his handler at the other end. Dare`s handler advised us that the fast speed trials show the typical pace for Dare when performing a search in a real scenario.

There are several reasons for using the hallway. The straight path makes it easy to calculate the speed of the dog and also to visually judge the quality of a 3D model. Additionally, the hallway is visually plain and the little textures and features that it has are repeated consistently. The scarcity of visual features and geometric 3D features makes it one of the most challenging environments for visual odometry systems. Our motivation is that the interior of a disaster site is unpredictable. The presence of large amounts of dust and debris tend to obscure defining visual characteristics, which puts severe strain on feature detection and matching methods. Real disaster environments may be poorly lit with presence of smoke. Currently our system does not account for those situations.

It is difficult to obtain ground truth data with our experimental setup without an expensive motion capture system. Instead visual inspection is used to evaluate the model produced by our system. To create a baseline comparison, we recorded a dataset of the hallway with an RGB-D sensor mounted on a wheeled-cart. The model created from this dataset represents what the data should look like if it was on an ideal platform (smooth and slow moving). Table 1 shows the details of the six dog trials and one cart trial.

Figure 5: Plain interior hallway

	Distance (m)	Speed (m/s)	Accompanied
Slow1	44	1.00	Yes
Slow2	44	1.08	Yes
Medium1	44	1.60	Yes
Medium2	44	1.65	Yes
Fast1	44	2.35	No
Fast2	44	2.37	No
Cart	44	0.65	N/A

Table 1: Hallway trials

V. RESULTS

In the slow and medium speed trials, our system was able to generate four complete models from one end of the hallway to the other. In the high speed trials, our system produced a model that represented approximately two thirds of the hallway. There was a section of the hallway where there was a long stretch of plain wall and at the fastest speed of the dog, the system failed to produce a suitable registration.

From top to bottom, Figure 6 shows a comparison between the model created from Cart trial data with models created from the Slow1, Medium1 and Fast1 trials. These models were created using the IFS algorithm.

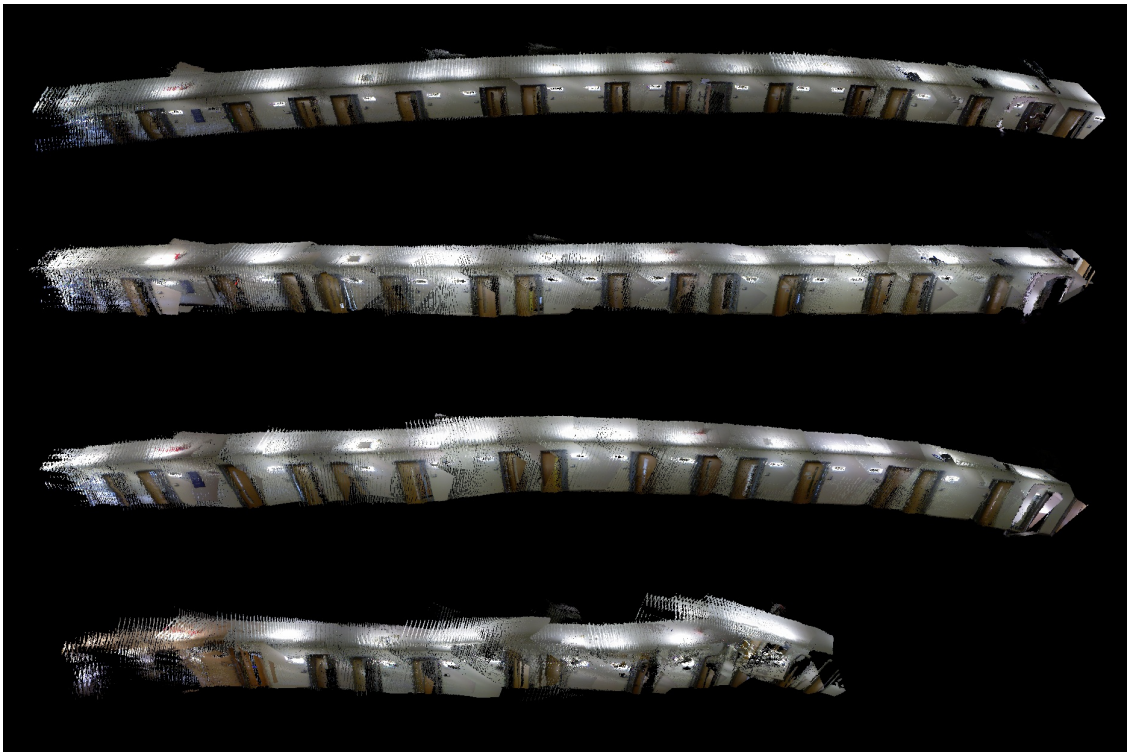

Figure 6: From to top bottom, model generated from Cart, Slow1, Medium1 and Fast1 datasets.

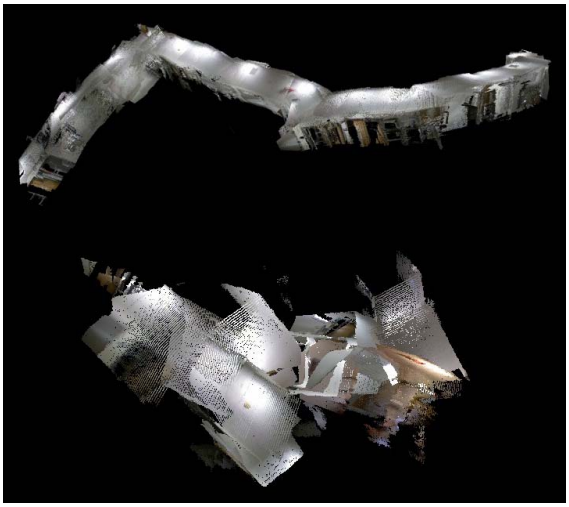

Figure 7: Models generated without IFS (top-Slow1, bottom-Fast1 dataset)

Figure 8: Models generated by human assisted process (top-Slow1, bottom-Fast1)

As a comparison, we tested our system without the IFS algorithm. The models produced from the slow trials have several badly broken parts and the models from the fast trials are not even recognizable. This is shown in Figure 7.

We also made a comparison between IFS and the human assisted process. As shown in Figure 8 the models created using human assistant in the Slow1 actually suffers from more drift than the IFS generated model. In the Fast1 dataset, the IFS and human assisted models look similar.

Table 2 shows the process time of each dataset using our RGB-D system without IFS, with IFS and with human assistance.

	Video Length (s)	Frame Count	Process Time (mm:ss)		
			Without IFS	IFS	Human Assisted
Slow1	44	1320	0:56	4:25	~30:00
Slow2	44	1320	0:54	4:32	~30:00
Medium1	27.5	825	0:37	2:45	~25:00
Medium2	26.7	801	0:30	2:46	~25:00
Fast1	18.7	561	0:22	1:30	~20:00
Fast2	18.6	558	0.22	2:28	~20:00

Table 2: Process time comparison

To further validate IFS we wished to compare it against another system that use ICP which does not rely on a crisp RGB images. We ran all of the datasets on the PCL implementation of KinectFusion but it failed to produce a model. This is likely due to the fact the the ICP algorithm

used by KinectFusion requires a good initialization. With high speed data, subsequent frames are too spatially separated.

VI. Conclusion and Future Work

This paper presents work on recreating the path of a search dog in a GPS denied environment while building a 3D map. We highlighted the challenges of using visual odometry on a canine mounted platform. Our results demonstrated that the IFS algorithm can help a visual odometry system produce accurate models in these extreme conditions. While IFS does increase process time slightly, this is acceptable in a USAR scenario. Due to the logistics of using a canine-mounted system, all processing must be done offline. Thus the increase in processing time is negligible when compared to the timeline of the rescue operation. The IFS algorithm is also significantly faster than the tedious human assisted process.

The positive results from our experiments warrant further study in the canine-mounted visual odometry problem. Our next challenge would be to determine limitations of this system in confined space. The obvious limitation is the minimum workable distance of the Kinect. We also want to test our system on winding paths and under various lighting conditions.

Furthermore, we want to test other map making systems. Although we were unable to obtain any useful results using KinectFusion, we plan to experiment with combining IFS with KinectFusion to see if better results can be achieved on this type of data.

We also plan to explore different ways of mounting the sensor on the dog. We suspect that mounting the Kinect further down the dog's back may significantly reduce the rolling back and forth of the sensor. An additional mounting technique being considered is to use two sensors, one on each shoulder looking sideways rather than backwards. It was observed that during the fast trials, the dog would often move very close to the wall causing large portions of the image to be inside the minimum range of the depth camera. Mounting cameras on each side ensures that at least one of them has a clear view.

Finally, we would like implement a hardware solution using Inertia Measurement Units and mechanical stabilizer to reduce the motion blur.

Acknowledgment

The authors would like to thank Constable Dan Bailey of the Ontario Provincial Police.

References

[1] P. Harris, "The Blitz hero with a nose for survivors, Rip the dog's bravery medal could fetch £10,000," in *Daily Mail*, ed. UK, 2009.

[2] A. Jacoff, A. Downs, A. Virts, and E. Messina, "Stepfield pallets: Repeatable terrain for evaluating robot mobility," in *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, 2008, pp. 29-34.

[3] A. Ferworn, A. Sadeghian, K. Barnum, D. Ostrom, H. Rahnama, and I. Woungang, "Canine as Robot in Directed Search," in *IEEE International Conference of Systems of Systems (SoSE'07)*, San Antonio, TX, USA, 2007, pp. 1-5.

[4] A. Ferworn, J. Tran, A. Ufkes, and A. D'Souza, "Initial Experiments on 3D Modeling of Complex Disaster Environments Using Unmanned Arial Vehicles " in *9th IEEE International Symposium on Safety, Security, and Rescue Robotics SSRR 2011*, Kyoty, Japan, 2011.

[5] A. Ferworn, A. Sadeghian, K. Barnum, H. Rahnama, H. Pham, C. Erickson, D. Ostrom, and L. Dell'Agnese, "Urban search and rescue with canine augmentation technology," presented at the IEEE International Conference of Systems of Systems (SoSE'06), Los Angeles, CA, USA, 2006.

[6] A. Ferworn, A. Sadeghian, K. Barnum, D. Ostrom, H. Rahnama, and I. Woungang, "Rubble Search with Canine Augmentation Technology," in *IEEE International Conference of Systems of Systems (SoSE'07)*, San Antonio, TX, USA, 2007, pp. 1-6.

[7] M. Gerdzhev, J. Tran, A. Ferworn, and D. Ostrom, "DEX - A design for Canine-Delivered Marsupial Robot," in *Safety Security and Rescue Robotics (SSRR), 2010 IEEE International Workshop on*, 2010, pp. 1-6.

[8] M. Fiala and A. Ufkes, "Visual Odometry Using 3-Dimensional Video Input," in *Computer and Robot Vision (CRV), 2011 Canadian Conference on*, 2011, pp. 86-93.

[9] A. S. Huang, A. Bachrach, P. Henry, M. Krainin, D. Maturana, D. Fox, and N. Roy, "Visual odometry and mapping for autonomous flight using an RGB-D camera," in *International Symposium on Robotics Research (ISRR)*, 2011.

[10] H. A. Wurdemann, E. Georgiou, L. Cui, and J. S. Dai, "SLAM Using 3D Reconstruction via a Visual RGB and RGB-D Sensory Input," 2011.

[11] R. A. Newcombe, A. J. Davison, S. Izadi, P. Kohli, O. Hilliges, J. Shotton, D. Molyneaux, S. Hodges, D. Kim, and A. Fitzgibbon, "KinectFusion: Real-time dense surface mapping and tracking," in *Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on*, 2011, pp. 127-136.

[12] T. Whelan, M. Kaess, M. Fallon, H. Johannsson, J. Leonard, and J. McDonald, "Kintinuous: Spatially Extended KinectFusion," in *RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras*, Sydney, Australia, 2012.

[13] Willow Garage, "Point Cloud Library (PCL)," 1.6.0, Available: www.pointclouds.org., 2012

[14] Wikipedia. (2011). *Kinect*. Available: http://en.wikipedia.org/wiki/kinect.

[15] J. Tran, A. Ufkes, M. Fiala, and A. Ferworn, "Low-cost 3D scene reconstruction for response robots in real-time," in *2011 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR),* , 2011, pp. 161-166.

[16] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th international joint conference on Artificial intelligence*, 1981.

[17] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Computer Vision–ECCV 2006,* vol. 3951, pp. 404-417, 2006.

[18] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM,* vol. 24, pp. 381-395, 1981.