2017 International Conference on
Unmanned Aircraft Systems (ICUAS)
June 13-16, 2017, Miami, FL, USA

SmartCaveDrone: 3D Cave Mapping Using UAVs as Robotic Co-Archaeologists

Guoxiang Zhang[1], *Student Member, IEEE,* Bo Shang[2], *Student Member, IEEE*
YangQuan Chen[3], *Senior Member, IEEE* and Holley Moyes[4]

Abstract— **This paper proposes the concept of drones capable of functioning as "Co-Archaeologists" that can map large caves and enter dangerous or hard-to-reach spaces. Using RGB-D data collected by drones, we will be able to produce accurate 3D models and semantic maps with proper lighting co-supervised by human archaeologists. This is going to be a major advance in archaeological practice, which can accelerate the speed of archaeological exploits by extending the archaeologists' sight and perception range. This will enable us to conduct 3D analyses so that we may answer new questions and create new insights into the archaeological record. The archaeologists will be able to visualize data collected by drones and instruct the drones' next step in real-time. These data will also be important in site management, data sharing and visualization. Human/drone interaction becomes important, not only for operating the equipment, but also for guiding drones to areas of interest to be mapped. Maps or real-time "fly-throughs" only make sense when they are organized by human interactions with the space. This human interaction is vital when visualizing and understanding a space and should be reflected in the imagery. We envision that this technology will be game changing in cave mapping and pertinent to anyone rendering interior spaces. It creates longer term impacts in archaeology and digital heritage and potentially creates a transformative way for further enhancing the performance of 3D mapping.**

I. Introduction

It has long been recognized that cave sites often contain the best-preserved material in the archaeological record. Cave archaeology has developed its own methodologies for mapping and recording sites, yet few sites are mapped to true 3D models, because it is a slow and tedious process for archaeologists to record and book-keep caves. They need to incrementally setup baseline along the cave and then measure distance from the baseline to the wall or objects of interest and mark wall or objects in a 2D map by hand. They may also use total stations to reduce the error introduced when moving baseline. This slow process has major negative impact on cultural relic preservation. One of our authors Prof. Holley Moyes is a leading specialist in cave archaeology and ancient religions. Over her career, she has worked to develop new mapping methodologies, but finds that caves in Belize are being looted more rapidly than they can be investigated. Typically, archaeological teams will visit a site and begin to record it in one year, but when they come back to finish data collection it has been looted, artifacts stolen, architecture destroyed and the archaeological record disturbed. Therefore, archaeologists need a faster, more efficient method of surveying and recording the sites. Drone 3D mapping offers a vast improvement over the current mapping and recording techniques such as hand-drawn maps, or even terrestrial LiDAR. This is a major advance in archaeological practice, enabling us to conduct 3D analyses and to visualize and share data with other researchers and the public. This will also be important for site management, such as tracking change of caves over time, because of the detail and accuracy of the data.

Our goal is to explore and record data using a drone platform to produce 3D and 2D semantic maps of caves, which contain three levels of information. In the coarse level, it shows the skeleton of a cave, including how large it is and where it goes; In the middle level, it can visualize the spatial distribution of objects of different categories, which can provide information for understanding the way ancient human use a cave for; In the detail level, accurate and colorized 3D models are used as a way to document artifacts and meaningful objects in the caves digitally, which can provide more information than photographs but still easy to share. Especially when augmented and virtual reality technology are becoming mature, which can make visualization of 3D model much easier.

Drones are beginning used in archaeology, but most of them are used in outdoor environments for the purposes of survey or documentation using visual camera, thermal/infrared/near-Infrared sensors or LiDAR to produce 3D models or images. When move to drone related cave mapping, there is still no well-working drone system that can be interactively work with archaeologists to explore, reconstruct and map caves, although there are many attempts from researchers in robotics, computer vision and computer graphics to attack 3D reconstruction problem. There are several challenges to solve before we can deliver a mature system, which we call it a *SmartCaveDrone* system. We will discuss these challenges later in details.

Our contribution is proposing the concept of drones that act as robotic co-archaeologists that can map ancient caves. We further identify challenges and opportunities of building this system. We have finished some basic provement of the effectiveness of the proposed method.

II. System Overview

As shown in Fig. 2, this system includes a quadcopter-based drone platform that has 3D sensors and sense-and-avoid subsystem on-board, a data processing and control center (DPCC) and most importantly, archaeologists as a supervisor to control this system at a high level.

Controlled by human operator and ground control computer, this collaborative drone system can fly into and explore unmapped caves. This system can take a qualitative map from archaeologists as initialization. Then it builds a map and localize the UAV using a state-of-the-art simultaneous localization and mapping (SLAM) system. A real-time 3D reconstruction result will be given to human operators during scanning process. Based on current scanning result,

Fig. 1. A photo of a cave [1]

[1] Mechatronics, Embedded Systems and Automation Lab, School of Engineering, University of California, Merced, Merced, CA, USA, gzhang8@ucmerced.edu

[2] College of Information Science and Engineering, Northeastern University, China, 110819, Email: cnpcshangbo@gmail.com

[3] Mechatronics, Embedded Systems and Automation Lab, School of Engineering, University of California, Merced, Merced, CA, USA, ychen53@ucmerced.edu

[4] School of Social Sciences, Humanities, and Arts, University of California, Merced, Merced, CA, USA, hmoyes@ucmerced.edu

978-1-5090-4494-8/17/$31.00 ©2017 IEEE

operators can naturally issue high level commands , such as which area should be covered more to get detailed information, where there are missed cave branches, and when to finish the scanning process, to improve scanning result. For the low-level commands, it relies on the on-board sensors and the DPCC. It has an active path planning subsystem that can issue control commands based on current map and location estimate, and navigate SmartCaveDrone through cave environment. In case of mapping or localization failure, the UAV has a sense-and-avoid subsystem on-board as a fail-safe. The sense-and-avoid subsystem continuously detects surrounding obstacles and makes maneuvers to avoid obstacles. The sense-and-avoid module can override any commands from the path planner and give feedback to the online 3D reconstruction system.

The DPCC is an off-board data processing and control center. We place the main computation power to an off-board computer, because then we can leverage massive parallel processing power of general purpose graphic processing unit (GPGPU). It enables running computational-heavy tasks to get real-time 3D modeling and efficient active path planning, which makes real-time human-drone interaction possible.

After data acquisition, an offline 3D refinement subsystem is employed to further improve the quality of 3D models. It jointly consider all the sensor reading for globally consistent 3D models. It can also take correction commands from archaeologists to remove mismatches or false loop closure from automatic process. Then a 3D mapping subsystem will take camera poses, RGB-D data and refined 3D models as input to segment objects in 3D and label their categories, whose final results are object level 3D maps or 2D projection maps. During this segmentation and mapping process, archaeologists can easily give coarse segmentation proposals when this system fails at detecting interesting objects. Computer algorithms will use these proposals to generate final results which can make fully use of geometric and photometric information.

We consider to use RGB-D sensors for 3D scanning, because some quadcoptors (*e.g.*, Intel Aero platform [2]) are beginning to have RGB-D sensors on-board. These sensors (*e.g.*, Microsoft Kinect and Intel Realsense) can provide reasonable good depth reading accuracy within a short range, even though they are much cheaper and lighter than laser scanners. Compared with laser scanners, RGB-D cameras can also provide RGB images, which makes it possible to colorize 3D models and leverage recent progress on image segmentation. In contrast, LiDARs are expensive. The commonly used LiDARs on small drones are from Hokuyo (a Japanese company) [3]. However, they can only scan in a 2D plane, which makes them only suitable for structured environment. Also, they usually costs thousands of dollars and are very easily get damaged during a crash landing. For these reasons, we will limit our discussion on RGB-D cameras related research.

III. Challenges and Opportunities

In order to make this system work, there are several challenges to overcome, which will create new research opportunities.

First, it is completely dark inside caves. A RGB camera is not going to capture anything in this environment. A normal head light or drone light can light the environment for drone operators to see the environments, however, the drone cameras are not able to get enough visual features for image processing. In this situation, the commonly used visual-based drone control is not going to work. The captured images do not have enough visual features for offline mapping. Therefore, it is necessary to have a dedicated lighting subsystem in order to have enough light to capture high quality images with colors. Since this is a UAV-based system with energy and weight limitations, it is important to use this energy in an optimal manner. So we propose to have optimal cooperative lightening, which is to light at the optimal location and angle which can provide the best lighting result at minimum amount of time with lowest power consumption.

Second, how to accurately reconstruct caves in 3D. This problem can be divided to several subproblems. For online 3D reconstruction, it is important to have a better loop closure detection method that can utilize more spatial and temporal information. Also, it is an interesting problem to decrease and bound global positioning drift of a SLAM system in cave environment, especially when we neither want to manually set up markers inside caves nor manually label data. In addition, how to add human decision to 3D reconstruction process. Human operators can understand the environment better while computer can do better in repeated easy tasks. A good way of combining strength from both will benefit this 3D reconstruction process. Meanwhile, this drone system should have some level of autonomy, which means it should be able to decide where to scan and how to act to reach its goal point with limited commands.

Third, how to semantically segment and label objects on the constructed 3D model of a cave. It is a crucial step. Since without semantic labels, a 3D model is just a huge surface or a cluster of points. Only with semantic segmentation and labels, archaeologists can perform spatial analysis of distribution of objects then understand better how ancient human use a cave.

Also, it is a challenge on maintaining wireless communication between DPCC and drones, because caves can have irregular shape and they can be a few kilometers deep. But we do not try to solve communication problem in this paper, instead, we address it by making the drone be able to fly autonomously when needed, so active SLAM is studied in the paper as well. During autonomous flight, only flight critical components, such as cooperative lighting, sense-and-avoid and SLAM, are enabled. Time consuming processing will be done on a DPCC off-line.

We will review current art on these problems and discuss remaining challenges and new opportunities in details in following sections.

A. Cooperative Lighting

The lighting for 3D mapping is a challenging and complex task because 3D modeling needs the light distribution to be consistent, however, the lights have to move if we put lights on drones [4]. Quantitative analysis for the relationship between the lighting intensity distribution and the 3D mapping performance is important to optimize the power consumption of the drone, which has limited power storage on-board.

Lighting control is a hot topic for energy saving. In [5], a minimum energy point tracking algorithm is developed to achieve the minimum energy usage despite of environmental variations. Cooperative lighting control is an extension of the traditional fixed lighting control to a 3D-dimensional mobile lighting control. Cooperative lightening with drones means to use dedicated drones to help the drone with 3D scanners by lighting the area that is currently being mapped. Therefore, this scheme contains using two or more drones working together. Current technology includes the lead and following formation of drone swarms [6]. The lead and following formation contains positioning control by wireless communication or by using visual servoing. The challenges of this working scheme should be accurate relative positioning control without GPS and optimization for 3D mapping performance and power consumption.

Therefore, the cooperative lighting optimization problem should be a hot topic in the near future as the drone swarm technology is getting mature.

B. Online 3D Reconstruction

There are many research work [7], [4], [8], [9] in the literature to address the online 3D reconstruction problem, but it is still challenging to reliably reconstruct 3D models of indoor scenes in real time, especially when the indoor environment is a dark cave. First, RGB-D cameras have limited field of view and working range, which can bring in two problems: 1) more 3D model pieces need to be put together; 2) each view only cover a small portion of the

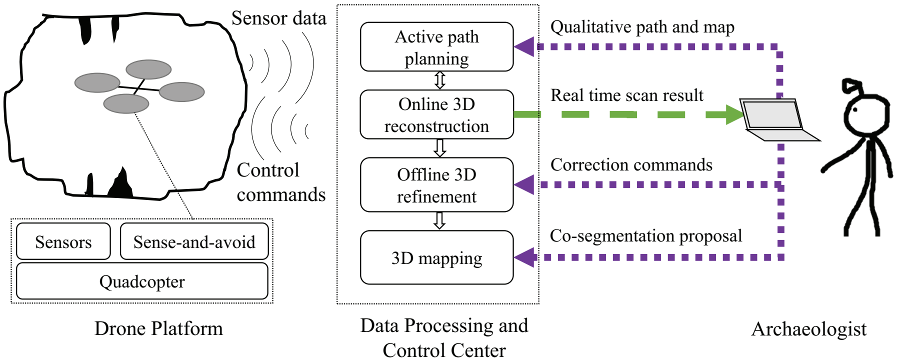

Fig. 2. System overview

scene with limited information, making tracking prone to failure. Especially when some areas of indoor environment do not have much shape and color variation. Second, cave environment usually contains sophisticated geometric structures and objects, and must be scan from complex camera trajectories for better coverage, which means one place can be observed multiple times from different view angle. In theory, this should give more opportunities to minimize reconstruction error by closing loops, but, in practice, it will cause problem due to there is no algorithm that can guarantee to fully detect and close all place revisits as loops . This can be problematic for generating good 3D models, since only a small mismatch that happens to a few frames of data can jeopardize reconstruction quality.

One important work in RGB-D base 3D reconstruction literature is KinectFusion [7], which first combines projective iterative closest point (ICP) and volumetric scene representation to build a real time dense 3D reconstruction system on a GPU. It first reveal the potential of real time dense 3D reconstruction system. Then Kintinuous [4] extends it so that it works on a scale larger than a single room. Later, Whelan *et al*. propose ElasticFusion [9], which based on deformation graphs and can jointly minimize geometric and photometric error.

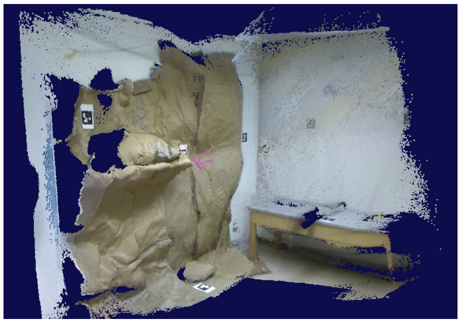

Fig. 3. Online 3D reconstruction result

These systems can provide reasonable good results, but they will fail at some situations, such as image blur during fast motion, especially fast rotation, and flat areas without color variations. In these cases, instead of trying to propose a solution can solve these problems within RGB-D data domain, we believe leveraging data from IMU sensors is a better choice. IMU should be able to provide a huge performance gain to current visual-SLAM approaches, but most of well known dense visual-SLAM systems and open source implementations do not support IMU sensor by default. Naively, IMU sensor can provide a good initialization, since ICP is sensitive to initialization and IMU is accurate in a short time period. Its potential has been revealed by [10] and [11]. Nießner *et al.* [11] use angular acceleration to bootstrap ICP of KinectFusion and show that the system is more robust to rapid motion. Usenko *et al.* [12] use a tightly coupled approach to get stereo camera base visual-inertial odometry. They show that the two sensors can complement each other: stereo vision allows the system to compensate for longterm IMU bias drift, while short-term IMU constraints help to improve vision frame-to-frame tracking, which makes the system more reliable even in area that does not have much visual information.

We argue that the biggest problem in a SLAM system is lacking the ability of accurately detecting all place revisiting, which is important for both tracking failure recovery and loop closure detection. Even with IMU, error will eventually propagate and increment to an unacceptable amount. Currently, whether a place revisit happens or not is detected by low-level computer vision algorithms, such as bag of word (BoW) based feature matching [13], which can not fully utilize all the information. There is a great need to improve current place recognition performance for SLAM systems.

C. *Offline 3D Reconstruction*

Other than different approaches to improve real time performance, another branch of 3D reconstruction is offline processing, which aims to achieve the best 3D reconstruction result by considering all the information in a jointly[14], [15], [16], rather than incremental manner. Structure from motion (SfM) [17], [18] and multi-view stereo (MVS) [19], [20], [21] have been actively explored and they can be used to recover 3D models from sets of images. After the emerging of RGB-D cameras, Xiao *et al.* [14] run 3D SfM on both depth and RGB images. They extract and match image features across images and then get their 3D coordinates from depth images, after that, SfM is conducted using 3D coordinates of keypoints. Choi *et al.* [15] produce state-of-the-art result among offline methods. They use RGB-D visual odometry from Kintinous [4] to merge several frames into a scene fragment, based on an assumption that RGB-D odometry is reliable in a short term. All fragment pairs are registered to each other, while line processes is used to filter false positives loops, which utilizes [22] as a back end.

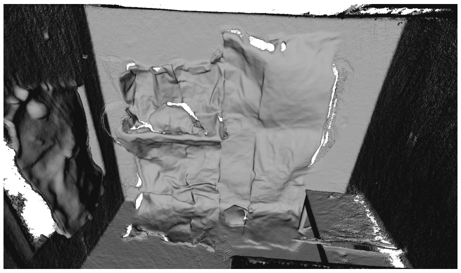

Fig. 4. Offline refined 3D model

Even through current best approaches can provide good results in their experiments, but they all work well under certain conditions and can fail at some cases. For example, in SUN3D, the appearance based place recognition can not recover all the global loops which can greatly reduce the reliability of the whole system, since one global loop detection failure can lead to many local loop detection failure, thus it can cause great 3D reconstruction quality degradation. Offline processing method [15] takes several hours to get a result on a room scale space. And time it consumes will grow quadratically when the size of the space becomes larger, because it will produce more fragments which means more pairs will need to be registered to each other. In order to get a workable solution for a much larger scale, efforts should still be made.

D. 3D Mapping

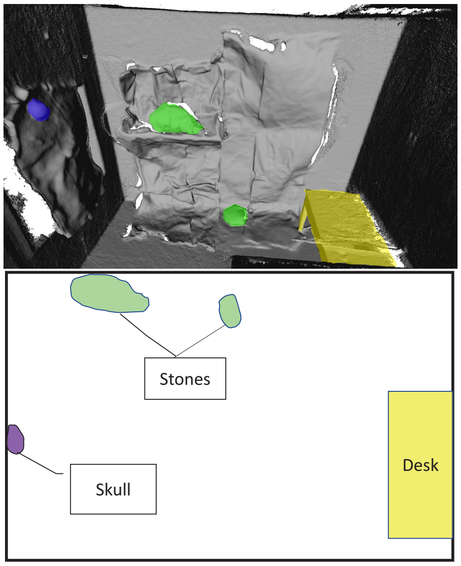

Fig. 5. Example 3D mapping result (upper) and corresponding 2D projection map (lower)

We define 3D mapping the processing of segmenting and labeling objects in generated high fidelity 3D models, as shown in Fig. 5. This step is crucial, because it can convert 3D models to high level abstracted 3D semantic maps, which enable more ways for archaeologists to analyze, such as spatial distribution of artifacts, 2D projection of semantic map which can be imported to a GIS software. This highly reduce the burden on archaeologists, because they spend most of their time on marking objects on their map during cave mapping.

This problem can be solved in 3D space directly [23], [24], [25], either on point cloud [23], [25] or 3D mesh [24]. Valentin *et al.* [24] create a 3D mesh and compute geometric features on it directly. Visual features are computed on the images and combined with geometric features to obtain a coarse per mesh face label result. Conditional random field (CRFs) are defined over the 3D mesh to get a geometric consistent segmentation. Tateno *et al.* [26] segment each input depth map by employing normal edge analysis, then they run 3D descriptor directly on each 3D segment derived from the incremental segmentation stage, and match it with 3D descriptor features computed on the full 3D object models. Finally they use a confidence to update multiple observations. When confidence drops to zero, this means that the segment has changed.

Another way is to address this problem in 2D image space and combine results from different frames together. This approach can leverage the recent progress in image segmentation, scene parsing, object detection and recognition [27], [28]. There also exist many object detection and segmentation methods for RGB-D, among which most treat depth as a fourth channel [29], [30], and use either hand-generated features and classifiers [31], or a convolutional neural network (CNN) [30]. Hermans *et al.* [32] get 2D semantic segmentation result by a soft classification for each pixel which corresponds to a 3D point in the point cloud, then the result of multiple observations are merged in 3D using a Bayesian update, which takes the current belief for a 3D point and updates it with the new predictions. The spatial consistency is achieved by applying dense pairwise CRFs over the 3D point cloud. Vineet *et al.* [33] extracts 2D features and evaluates unary potentials based on random forest classifier predictions to get semantic segmentation. It transfers these into a 3D volume, then volumetric CRFs are used to enforce temporal consistency.

Currently, all these methods are focus on either general indoor or outdoor scene segmentation. In order to use them in a cave environment and detect artifacts, modifications should be made. For methods work in 3D space, they usually have a database of model descriptors [26] or train classifiers on 3D models [25], so for these methods, a database of 3D artifacts models should be built. For 2D image based methods, [34] has shown that transfer learning and fine-tuning can adapt models trained on general computer vision tasks to a specific task. What we need is an image database with human annotation.

There are remaining challenges. First, all the prior methods can be considered as data driven approach, which means it only works well when objects during testing look similar to the ones in training, but artifacts may have many variations. A big problem is how to define a model that can learn knowledge from archaeologists. Second, for cases which is too hard for computer to solve, it is necessary to have an effective yet easy to use approach to help bootstrap or improve results with human-in-the-loop. Third, most approaches, especially the ones with CRFs, can not run in real time, but a real time system can benefit SLAM with object level constraints.

E. Autonomous Cave Exploration

Autonomous cave exploration is an import ability for a SmartCaveDrone, because most caves have hash or even dangerous environment. With SLAM approach on-board, SmartCaveDrone can incrementally build a map and localize itself at the same time, but SLAM only passively process input data and does not give any control on how to move sensors to scan a complete map. Active-SLAM is the technique that can make decisions and control on where to scan next without a predefined path. Also, sense-and-avoid should be added to a SmartCaveDrone. It acts as a fail-safe, whenever there is a sensor failure, false localization or a wrong decision.

Active-SLAM is solved in three major steps [35], [36]: 1) propose possible next way points based on current map and trajectory estimate; 2) evaluate all vantage points proposals based on a defined utility function which contains uncertainty metric and cost for reaching this point, and select the one that can reduce uncertainty most as the next goal point; 3) execute actions to reach next goal point generated by a path planning solver.

While most of prior research developed the fundamental theory and run simulations and tests on 2D ground robots, recent research tries to adapt this framework to a UAV platform. Heng *et al.* [37] propose an efficient RGB-D vision based approach to perform 3D exploration and coverage in unknown environments. They assume the pose of MAV is known, so there is no trajectory uncertainty, and map uncertainty is reflected by number of explored and unknown grids in their octmap representation. Thus they define the information gain to be the number of expected observable unexplored voxels that are enclosed in the corresponding view frustum. Exploration is achieved through maximizing information gain in a 3D occupancy map and consider execution cost. Most of their efforts are focused on pre-computating expect sensor reading due to limited on-board resources. Next-best-view planning algorithms [38] iteratively determine the best viewing configuration determined by the amount of unmapped space that can be explored. a RRT or RRT* based planner is used to generate potential vantage points. Then, the destination of first edge of the best branch is determined the next-best-view configuration then executed. It does no assume pose is known, but resort to visual-inertial odometry to provide accurate pose estimate, so they do not consider pose uncertainty and make no effort of improving localization accuracy.

When goes to 3D, it becomes more challenging, since there is much more grids in 3D than in 2D, and there are more possible actions since a UAV can move in 3D free space with 6 degree of freedom while in 2D there are only 3 degree of freedom. Also, when evaluate candidate actions, it is more computational expensive to run 3D ray-casting on-board, so a DPCC with massive data processing power is very necessary. Also, a UAV has limited power resources, so an highly efficient exploration policy is a desired property as well. It is better to be capable of leveraging prior knowledge of operators, such as the topology shape of the cave, area of interest, and area that is not necessary to map.

1) sense-and-avoid: The sense-and-avoid technology is necessary for the 3D mapping mission because the long-term mapping and position estimation have errors that can lead to collisions. It can be classified by cooperative sense-and-avoid methods and non-cooperative sense-and-avoid methods. The cooperative ones requires known position. This scheme will not work in the 3D mapping mission since there is no GPS location and the obstacle will not send any signal by themselves. Therefore, only non-cooperative sense-and-avoid methods will be discussed.

Current technology of non-cooperative sense-and-avoid can be classified by the sensors they use. Popular sensors for sense-and-avoid are ultrasound sensors, laser range finders, web cameras, RGB-D sensors and combo sensors. The ultrasound sensor is a traditional one. The typical detect angle is often about 30 degrees. The detection range is not the same in the 30 degrees, so the ultrasound sensor often has large blind areas. The ultrasound sensor returns the nearest distance in the detection range, which means the detection is very rough. In real applications, ultrasound sensor often lead to collisions. Therefore, the ultrasound sensor method is a very basic one and not practical for real applications. The laser range finder can detect a 2D surface, so the applications are limited to structured environments. Olivares-Mendez *et al.* successfully use computer vision to detect an obstacle in the way of the flight path [39], however, the obstacle is something with special features like a cone. Therefore, it does not work in caves where obstacles are not marked with special features. RGB-D sensors have been used to detect obstacles [40], [41]. An RGB-D camera often has the same detection angle as the paired RGB camera, so they can only provide detection for the front part of a drone. Therefore, multi-RGB-D cameras or 360 degree RGB-D camera can be a more effective scheme for sense-and-avoid purpose.

Currently, the non-cooperative sense-and-avoid is still not mature. The sense-and-avoid failure is one of the neck-bottle problems preventing the drones from acquiring more applications. Therefore, quantitative evaluations for sensor effectiveness in different lighting environments are going to be hot research topics.

F. Other Opportunities

The current technologies still have a very big gap to enable drones to be co-archaeologists. The hardware infrastructure still greatly limits drones' applications. Great opportunities are with longer battery life, longer detection range, less power consumption and less weight for sensors, faster on-board computers and better data processing algorithms. The positioning in caves is a big challenge since there is no GPS signal in caves. An idea to solve this problem is use beacons to help localization. The beacons can be more helpful by collecting data and storing information for later drones. Later drones can communicate with beacons on the ground to know this place is already exploited. Algorithms should be developed to find out the best positions to put beacons. The beacons are also good relay devices to help communications in caves.

Another interesting topic is the cooperation between the archaeologists and drones. We want to take advantages of the archaeologists' field knowledge to help drones accelerate the path planning process. Our scheme is to let the drones scan the cave to collect information of the surroundings first. Then archaeologists give a qualitative path planning base on their experience. Last, the drone finishes the flight path based on the qualitative path planning and avoiding obstacles at the same time. This scheme is shown in Fig. 6.

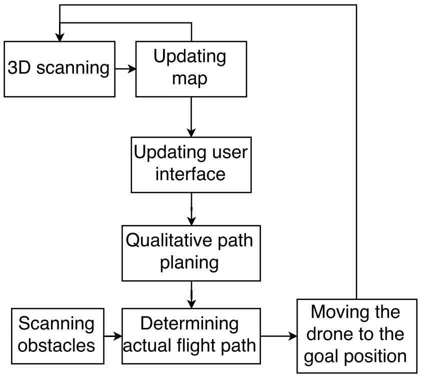

Fig. 6. The scheme to use archaeologists' knowledge to help path planning

IV. Conclusions

We propose the idea of SmartCaveDrone system, which can closely work with archaeologists. It sure will change the way archaeologists work. They will be able to explore caves faster and safer and gather richer information. After exploring related research fields, we conclude this SmartCaveDrone system has broader impacts. It can create new opportunities in the field of UAV, computer vision, robotics and human-robot interaction, because this system will see, understand, remember and think caves in 3D as we do and naturally interact with operators. It will give UAV higher level of autonomy, which can promote new possibility of leveraging UAVs in new area.

V. Future Work

After all these aforementioned modules available, we will further explore human-in-the-loop and story-telling. Since we want our SmartCaveDrone to act as co-archaeologists, it will work tightly with archaeologists to help its exploring and mapping process. At least for now, human still have better understanding of surrounding environment, and better decision making ability. Archaeologists may have work in ancient cave environment for years, they can quickly identify which is the area that contains interesting artifacts. Then our SmartCaveDrone should understand not only voice commands but also body language, such as gesture of pointing to somewhere. During post processing, operators will play an important role of improving final result of 3D reconstruction and mapping. This means this system should understand intention of operators and have a heavy human computer interaction, which we will explore later. In addition, we want to explore story-telling as well. Give all the information, taking advantage of virtual reality, we want to interactively tell a story of how ancient people use a cave, which should take knowledge of archaeologists into account and recent progress in machine learning and natural language processing.

Acknowledgments: We thank Christopher Currier for his help in building the drone platform. This work is supported by CITRIS Seed Grant 2016-2018 entitled "Drones for Cave Archeology and 3D Mapping."

References

[1] U.S. National Park, "Mammoth Cave," http://kentuckylakes.com/sites/default/files/imagecache/galleryformatter_slide/slides/969754-mammoth-cave_2117_600x450.jpg, 1990, [Online; accessed Feb 24, 2017].

[2] "Intel® Aero Ready to Fly Drone," http://click.intel.com/intel-aero-ready-to-fly-drone.html, accessed: 2017-02-23.

[3] S. range finder (SOKUIKI sensor), "UTM-30LX," https://www.hokuyo-aut.jp/02sensor/07scanner/utm_30lx.html, [Online; accessed Feb 26, 2017].

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

[5] C. Yin, B. Stark, Y. Chen, and S.-m. Zhong, "Adaptive minimum energy cognitive lighting control: Integer order vs fractional order strategies in sliding mode based extremum seeking," *Mechatronics*, vol. 23, no. 7, pp. 863–872, Oct. 2013, 00031. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957415813001621

[6] N. Gallardo, K. Pai, B. A. Erol, P. Benavidez, and M. Jamshidi, "Formation control implementation using kobuki turtlebots and parrot bebop drone," in *Proc. of the 2016 World Automation Congress (WAC)*, July 2016, pp. 1–6.

[7] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "KinectFusion: Real-time dense surface mapping and tracking," in *Proc. of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality*, Oct 2011, pp. 127–136.

[8] C. Kerl, J. Sturm, and D. Cremers, "Dense visual SLAM for RGB-D cameras," in *Proc. of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nov 2013, pp. 2100–2106.

[9] T. Whelan, S. Leutenegger, R. S. Moreno, B. Glocker, and A. Davison, "ElasticFusion: Dense SLAM Without A Pose Graph," in *Proc. of Robotics: Science and Systems*, Rome, Italy, July 2015.

[10] O. Kähler, V. A. Prisacariu, C. Y. Ren, X. Sun, P. Torr, and D. Murray, "Very High Frame Rate Volumetric Integration of Depth Images on Mobile Devices," *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, no. 11, pp. 1241–1250, Nov 2015.

[11] M. Nießner, A. Dai, and M. Fisher, "Combining inertial navigation and ICP for real-time 3D surface reconstruction." in *Proc. of the Eurographics (Short Papers)*. Citeseer, 2014, pp. 13–16.

[12] V. Usenko, J. Engel, J. Stückler, and D. Cremers, "Direct visual-inertial odometry with stereo cameras," in *Proc. of the 2016 IEEE International Conference on Robotics and Automation (ICRA)*, May 2016, pp. 1885–1892.

[13] D. Gálvez-López and J. D. Tardós, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, October 2012.

[14] J. Xiao, A. Owens, and A. Torralba, "SUN3D: A database of big spaces reconstructed using SfM and object labels," in *Proc. of the 2013 IEEE International Conference on Computer Vision*, Dec 2013, pp. 1625–1632.

[15] S. Choi, Q. Y. Zhou, and V. Koltun, "Robust reconstruction of indoor scenes," in *Proc. of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 5556–5565.

[16] Q.-Y. Zhou, J. Park, and V. Koltun, "Fast global registration," in *Proc. of the European Conference on Computer Vision*. Springer, 2016, pp. 766–782.

[17] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, "Building rome in a day," in *Proc. of the 2009 IEEE 12th International Conference on Computer Vision*, Sept 2009, pp. 72–79.

[18] D. Crandall, A. Owens, N. Snavely, and D. Huttenlocher, "Discrete-continuous optimization for large-scale structure from motion," in *Proc. of the CVPR 2011*, June 2011, pp. 3001–3008.

[19] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *Proc. of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1, June 2006, pp. 519–528.

[20] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, "Towards Internet-scale multi-view stereo," in *Proc. of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 1434–1441.

[21] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1362–1376, Aug 2010.

[22] P. Agarwal, G. D. Tipaldi, L. Spinello, C. Stachniss, and W. Burgard, "Robust map optimization using dynamic covariance scaling," in *Proc. of the 2013 IEEE International Conference on Robotics and Automation*, May 2013, pp. 62–69.

[23] J. Papon, A. Abramov, M. Schoeler, and F. Wörgötter, "Voxel cloud connectivity segmentation - supervoxels for point clouds," in *Proc. of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 2027–2034.

[24] J. P. C. Valentin, S. Sengupta, J. Warrell, A. Shahrokni, and P. H. S. Torr, "Mesh based semantic modelling for indoor and outdoor scenes," in *Proc. of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 2067–2074.

[25] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese, "3D semantic parsing of large-scale indoor spaces," in *Proc. of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 1534–1543.

[26] K. Tateno, F. Tombari, and N. Navab, "When 2.5D is not enough: Simultaneous reconstruction, segmentation and recognition on dense SLAM," in *Proc. of the 2016 IEEE International Conference on Robotics and Automation (ICRA)*, May 2016, pp. 2295–2302.

[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2016.

[28] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *Proc. of the 2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 1529–1537.

[29] X. Ren, L. Bo, and D. Fox, "RGB-(D) scene labeling: Features and algorithms," in *Proc. of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 2759–2766.

[30] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Proc. of the European Conference on Computer Vision*. Springer, 2014, pp. 345–360.

[31] S. Gupta, P. Arbeláez, and J. Malik, "Perceptual organization and recognition of indoor scenes from RGB-D images," in *Proc. of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 564–571.

[32] A. Hermans, G. Floros, and B. Leibe, "Dense 3D semantic mapping of indoor scenes from RGB-D images," in *Proc. of the 2014 IEEE International Conference on Robotics and Automation (ICRA)*, May 2014, pp. 2631–2638.

[33] V. Vineet, O. Miksik, M. Lidegaard, M. Nießner, S. Golodetz, V. A. Prisacariu, O. Kähler, D. W. Murray, S. Izadi, P. Pérez, and P. H. S. Torr, "Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction," in *Proc. of the 2015 IEEE International Conference on Robotics and Automation (ICRA)*, May 2015, pp. 75–82.

[34] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. of the 27th International Conference on Neural Information Processing Systems*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, pp. 3320–3328. [Online]. Available: http://dl.acm.org/citation.cfm?id=2969033.2969197

[35] J. L. Blanco, J. A. Fernández-Madrigal, and J. Gonzalez, "A novel measure of uncertainty for mobile robot SLAM with Rao-Blackwellized particle filters," *International Journal of Robotics Research*, vol. 27, no. 1, pp. 73–89, Jan. 2008. [Online]. Available: http://dx.doi.org/10.1177/0278364907082610

[36] H. Carrillo, P. Dames, V. Kumar, and J. A. Castellanos, "Autonomous robotic exploration using occupancy grid maps and graph SLAM based on Shannon and Rényi entropy," in *Proc. of the 2015 IEEE International Conference on Robotics and Automation (ICRA)*, May 2015, pp. 487–494.

[37] L. Heng, A. Gotovos, A. Krause, and M. Pollefeys, "Efficient visual exploration and coverage with a micro aerial vehicle in unknown environments," in *Proc. of the 2015 IEEE International Conference on Robotics and Automation (ICRA)*, May 2015, pp. 1071–1078.

[38] A. Bircher, M. Kamel, K. Alexis, H. Oleynikova, and R. Siegwart, "Receding horizon "Next-Best-View" planner for 3D exploration," in *Proc. of the 2016 IEEE International Conference on Robotics and Automation (ICRA)*, May 2016, pp. 1462–1468.

[39] M. A. Olivares-Mendez, P. Campoy, I. Mellado-Bataller, and L. Mejias, "See-and-avoid quadcopter using fuzzy control optimized by cross-entropy," in *Proc. of the 2012 IEEE International Conference on Fuzzy Systems*, June 2012, pp. 1–7.

[40] Z. Y. Dang and J. D. Lee, "Fuzzy-based obstacle avoidance for a mobile robot navigation in indoor environment," in *Proc. of the 2015 IEEE International Conference on Consumer Electronics - Taiwan*, June 2015, pp. 55–56.

[41] J. Stowers, M. Hayes, and A. Bainbridge-Smith, "Biologically inspired uav obstacle avoidance and control using monocular optical flow amp; divergence templates," in *Proc. of the 5th International Conference on Automation, Robotics and Applications*, Dec 2011, pp. 378–383.