# Efficient 3D Visual Perception for Robotic Rock Breaking

Longchuan Niu[1], Ke Chen[2], Kui Jia[2], and Jouni Mattila[1]

***Abstract*— In recent years, underground mining automation (e.g., the heavy-duty robots carrying rock breaker tools for secondary breaking) has drawn substantial interest. This breaking process is needed only when over-sized rocks threaten to jam the mine material flow. In the worst case, a pile of overlapped rocks can get stuck on top of a crusher's grate plate. For a human operator, it is relatively easy to make the decisions about the rock locations in the pile and the order of rocks to be crushed. In an autonomous operation, a robust and fast visual perception system is needed for executing robot motion commands. In this paper, we propose a pipeline for fast detection and pose estimation of individual rocks in cluttered scenes. We employ the state-of-art YOLOv3 as a 2D detector to perform 3D reconstruction from point cloud for detected rocks in 2D regions using our proposed novel method, and finally estimating the rock centroid positions and normal-to-surface vectors based on the predicted point cloud. The detected centroids in the scene are ordered according to the depth of rock surface to the camera, which provides the breaking sequence of the rocks. During the system evaluation in the real rock breaking experiments, we have collected a new dataset with 4780 images having from 1 to 12 rocks on a grate plate. The proposed pipeline achieves 90.91% precision on overall detection with a real-time speed around 15Hz.**

## I. INTRODUCTION

Underground mining continues to progress to deeper levels for tackling the mineral supply crisis in the 21st century [1]. Human worker safety in mines deeper than a kilometer, along with time-consuming human shift worker logistics, is a massive mine operational cost challenge. This has increased demand for the level of autonomous robotics in mining. In deep mines, the extracted material is fed to crushers equipped with grate plates for stopping over-sized rocks (i.e., ore) from falling into the crusher jaws. The grate plate (e.g., a mesh size of 0.5 m x 0.5 m) prevents crusher jamming, but only if over-sized rocks remaining on the plate are immediately broken down into smaller pieces to ensure continuous mine mineral flow. Such rock breaking has been conventionally done by a human operator-driven heavy-duty hydraulic four-link anthropomorphic arm equipped with a hydraulic hammer tool, as shown in Fig. 1.

Recently, robotic rock breaking [2] has attracted wider attention owing to the controllable breaking procedure. Sensory rock perception plays an important role in robotic rock breaking as it provides the automatic over-sized rock detection and the motion target coordinates for the robotic rock

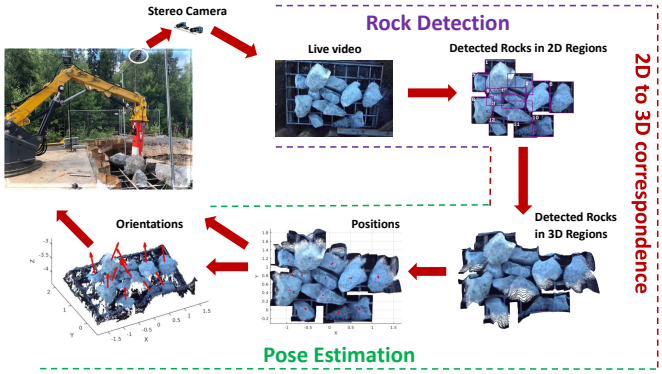



***Fig. 1*** *3D perception of rocks on a grate plate*

breaker arm. Some rock breaking systems with increased automation levels have been developed, such as the telerobotic rock breaker [3], vision-based mining automation controls [4], and 3D perception for mining robotics [5]. Some studies on rock breaking systems using force sensors [6] and stereo vision [7] have adopted algorithms for computing normals of rock surfaces. Nevertheless, none of the existing methods are capable of understanding the whole rock breaking scene in a complex environment.

For the automatic analysis of a scene, visual 3D perception requires fast and reliable initial detection with accurate object recognition and localization. However, this problem remains challenging due to piled rock scenes having arbitrary shapes, sizes, textures, and colours, as shown in Fig. 1. Pose estimation for objects with prior knowledge of shape was studied using 3D template matching technique in our earlier work [8]. For objects with unpredictable shapes, we have adopted a clustering algorithm for direct point cloud segmentation [9]. This method is used on secondary breaking in an unsupervised learning manner, but it suffers by missing texture-free visual cues for segmenting two rocks close to each other. A lack of contextual information in pure point analyses encourages us to conduct foreground highlighting in the RGB images to improve 3D rock detection. Moreover, such a setting has its significance in the practice of collision avoidance in robot on-line motions.

In this paper, we address the 3D visual perception of rocks via a pipeline visualized in Fig. 2, which consists of three stages: 2D rock detection, 2D-to-3D correspondence of regions, centroid position, and normal-to-surface vector estimation on object point cloud-based surfaces. At the first stage, the rocks displayed in the left image of Fig. 2 are detected as 2D regions (bounding boxes) by the state-of-the-art detector [10] (see Sec. IV-B). For the study, a stereo

[1]Automation Technology and Mechanical Engineering, Faculty of Engineering and Natural Sciences, Tampere University, FIN-33720, Tampere, Finland {longchuan.niu, jouni.mattila}@tuni.fi

[2]School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, P.R. China

camera system is used to reconstruct the geometry of a 3D scene based on stereo correspondences. Subsequently, the depth map is generated in the form of a gray-scale image describing its geometry. We utilize this property to recover a *3D point cloud* representation of a textured point cloud of rocks in 2D regions (produced by the 2D detector) to its corresponding point clouds in 3D regions. These are all performed at the 2D-to-3D correspondence stage (see Sec. IV-C), where scene background can be removed and we focus on analyzing rocks in the foreground. At the last stage, based on the predicted point sets for each rock, the centroid of the surface is discovered and its corresponding normal vector is estimated by searching for the best fit plane using the nearest points provided by random sample consensus (RANSAC) [11].

**Contributions** The novel contributions of this paper are fourthfold. Firstly, we developed an efficient 3D visual perception pipeline for the detection of visible rocks and individual rock 6D pose estimations in cluttered scenes. We achieved an average precision of 90.91% at a real-time speed around 15Hz. Secondly, instead of a conventional stereo-image rectification method, we proposed a plane-sweeping depth estimation method for establishing the 2D to 3D correspondence. Thirdly, on non-Euclidean structured points, we designed a method for estimating the centroid normal on detected rock surfaces. Finally, we collected and annotated 4780 different images for the rock detection in a real scale rock breaking robot set-up with the rocks weighing several hundreds kilos each. This dataset is the according to the authors' best knowledge of the first dataset of blasted overlapped rocks.

Experiment results on the new dataset verified the efficacy of the proposed method, which works even if a part of the object is occluded or truncated due to the presence of the robot arm or rocks in a pile. The dataset used for the training has been made available with this paper[1].

This paper is organized as follows: Section II introduces related research on object detection; Section III describes the research problem; Section IV details the methodologies used for the study; Section V explains the experiments that were carried out; and Section VI concludes the paper.

## II. RELATED WORK

Object detection is widely studied, and a number of methods based on deep learning has been proposed [12]–[17]. Most existing methods operate using 2D Euclidean convolution on images, which can be categorized into two main groups. The first group is object proposals and image classification, such as region-based convolutional neural networks (RCNN) [18], fast RCNN [19], and faster RCNN [15]. These methods begin by generating thousands of region proposals within the images, and then apply a convolutional classifier to filter the proposals by classification score thresholds. This two-stage setting increases networked training difficulties due to independent training on each individual component in the pipeline. The second group is single shot-based detection, such as SSD [14] and YOLO. Recently, the YOLO detector [10], [12], [13] has become a viable alternative to RCNN variants by achieving superior detection efficacy. Not many 2D-driven 3D object detection studies [20], [21] have been based on both RGB-D images and point clouds. Specifically, utilizing a mature 2D object detector's output to generate 3D object proposals, this reduces the search in entire 3D dense point cloud. Even though light detection and ranging (LiDAR) generated point clouds can be used for outdoor applications, compared to RGB images, LiDAR point clouds are unordered and too sparse for distinguishing the severe inter-occlusion between the rocks. This makes the direct application of these methods challenging in a rock breaking scenario. In light of this, our method maps 2D pixels within predicted bounding boxes into rock point cloud surfaces, which generate a visible rock surface as 3D proposals. Our proposed method works effectively in the robotic rock breaking scenario, which is verified in Sec. V.

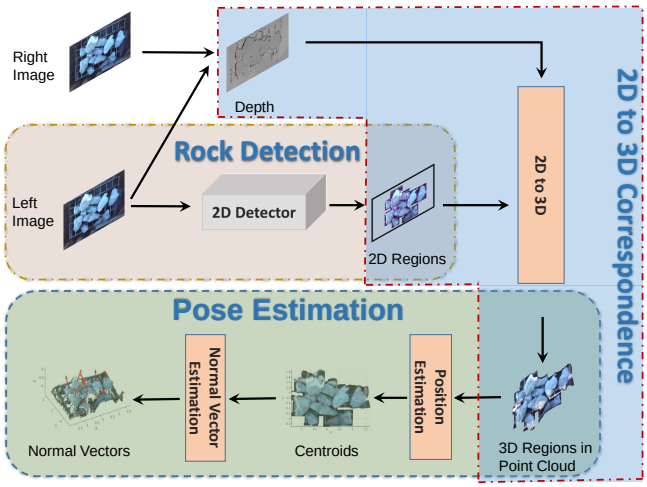



***Fig. 2*** *Pipeline of the proposed visual perception system*

## III. PROBLEM STATEMENT

As mentioned, automatic rock breaking requires fast and reliable detection and localization of every rock in a given scene. Oversized rocks on the grate plate can range from one rock or few rocks scattered around to many rocks in a complex pile overlapping each other. In our real-world robotic rock breaking set-up, we utilize a top-mounted stereo camera to provide video and images for automatic rock recognition and analysis. Given live video or still stereo images as input, the goal is to achieve real-time and sophisticated rock detection in a reference camera (left camera) coordinate, since individual 6D poses have to be shown to the operator and sent to the robot controller.

## IV. METHODOLOGY

For obtaining required rock poses for the controller, the rock centroid positions $[x, y, z]$ and orientations (i.e. normal-to-surface vectors at their centroids), we conduct three phases

[1]https://doi.org/10.5281/zenodo.2581287

in our visual perception system. The first phase is detection, where we employ a 2D object detector [10] for rock detection (see Sec. IV-B). The second phase is 2D to 3D correspondence, where 3D rock surfaces in a point cloud are generated via projection from 2D regions (see Sec. IV-C). In the final pose estimation phase, estimation methods for the centroid position and normal-to-surface vectors are applied. Fig. 2 illustrates the whole pipeline of the proposed system.

### A. A NEW DATASET FOR ROCK BREAKING

The procedure of data collection and annotation for the new dataset generation in the rock breaking application was organized as follows. The videos were recorded with various amounts of rocks on a grate plate under different outdoor illumination conditions by using a top-mounted stereo camera. However, due to the complex image gathering process in an outdoor environment, the position of camera was not entirely fixed. Therefore, slight camera movements during the video recordings can occur, which leads to background subtraction process failure. In view of this, object detection is considered the best possible approach to cope with the diverse background. In the gathered dataset, 4780 videos were recorded using a pre-calibrated stereo camera compressed in a lossless format in 720p at 15fps. They were further processed offline to extract selected frames into left and right images, which were used to generate depth maps as well as point clouds with color information (in ply format).

The Yolo Mark tool[2] was used for left image annotation. To alleviate manual labelling, an automatic labelling tool was implemented. This required the manual labelling of 1000 images, which then were used to train a coarse 2D detector to label the remaining 3780 images. After automatic labelling, the labelled images were still checked one by one. The quality of automatic labelling is known to be highly dependent on the quality of previously labelled data as well as the coverage of the data set. Therefore, a random data selection mechanism was implemented for this purpose.

### B. OBJECT DETECTION

As aforementioned, YOLO [10] was adopted for rock detection in 2D, making it an essential step for further processing. This kind of 2D detector formulates object detection into a regression problem, which addresses localization and recognition in a unified framework via simultaneous prediction of bounding box confidence and class probabilities. To this end, the whole image is divided into regular grids before the network predicts the object's centroids from the given set of candidates for various bounding boxes and object classes. Owing to its efficient detection, we are utilizing the latest network structure [10]. More specifically, the detection network (a variant of darknet-53) consists of 106 convolutional layers, where the prediction is performed at three different scales by predicting 10 times the numbers of boxes, producing more accurate results when detecting small objects.

[2]https://github.com/AlexeyAB/Yolo_mark

### C. 2D-3D CORRESPONDENCE

The estimation of scene geometry from a stereo camera setup is usually called a *depth-from-stereo* problem. Conventional stereo-matching methods based on stereo-image rectification [22] might underperform due to the introduction of artificial camera transforms and excessive image interpolation steps. In addition, a deviation from a geometrically parallel camera configuration is possible, thus introducing substantial image deformation in rectification-based methods [22].

A *plane-sweeping depth estimation* method allows for direct processing of captured imagery [23] by means of calibrated camera parameters. Fig. 3 illustrates the plane-sweeping principle of the depth-estimation method. The method assumes that the entire scene can be divided into a number of front-to-parallel planes where stereo correspondences could be found. The depth hypotheses can be selected according to the possible depth range ($z_{min} \leq z \leq z_{max}$) and a finite number of layers, to achieve a balance between fidelity and computational complexity.

Another advance of this method is its suitablity for parallel computing, and therefore, a dense 3D reconstruction of a complex scene can be realized in real time through GPU acceleration.

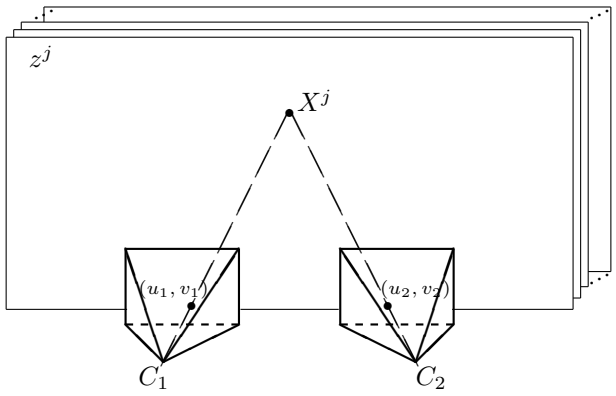



***Fig. 3*** *An illustration of the plane-sweeping principle of the depth-from-stereo estimation methods*

For every hypothetical depth $z^j$, one can project a pixel $(u_1, v_1)$ from a reference camera to a 3D space, using pre-calibrated camera matrix $C_1$:

$$\mathbf{X}^j = C_1^{-1}(u_1 \cdot z^j, v_1 \cdot z^j, z^j, 1)^T = C_1^{-1}\dot{\mathbf{x}}_1, \qquad (1)$$

where $\dot{\mathbf{x}}_1 = (u_1 \cdot z^j, v_1 \cdot z^j, z^j, 1)^T$ is the homogeneous projective coordinate of a current pixel, $\mathbf{X}^j$ is the resulting point coordinate in a 3D space, and $j = 1, .., N$ where $N$ is the selected number of layers.

Every obtained 3D point $\mathbf{X}^j$ can be further projected onto the sensor plate of a second camera using a similar equation:

$$\dot{\mathbf{x}}_2 = C_2\mathbf{X}^j = (u_2 \cdot z^j, v_2 \cdot z^j, z^j, 1)^T, \qquad (2)$$

where $\dot{\mathbf{x}}_2$ is a projective pixel position in a second camera image plane, and the actual pixel coordinates can be

recovered as:

$$u_2 = \frac{\dot{\mathbf{x}}_2.x}{\dot{\mathbf{x}}_2.z}, \quad v_2 = \frac{\dot{\mathbf{x}}_2.y}{\dot{\mathbf{x}}_2.z}.$$

We can construct a 3D cost volume in which pixel dissimilarities are calculated between the original pixel in the reference camera and the corresponding pixel in the second one:

$$C(u, v, j) = \|I_1(u_2, v_2) - I_2(u_1, v_1)\|, \tag{3}$$

where $I_1$ and $I_2$ denote the first (reference) and second images, respectively.

Through appropriate cost aggregation [23], the depth map can be recovered as such:

$$Z_1(u, v) = z^{\hat{j}}, \hat{j} = arg\min_j \tilde{C}(u, v, j), \tag{4}$$

where $\tilde{C}(\cdot)$ denotes the aggregated cost volume. The coordinates of the point cloud in the reference camera can now be reconstructed using equation (1), replacing $z^{\hat{j}}$ with the estimated value.

### D. POSE ESTIMATION

*1) Position:* The position of a rock is characterized in camera coordinates, indicating it is the geometric center of the bounding box in $x-y$ plane, as it is projected from image coordinates. This position estimation approach is sufficient, as those oversized rocks are with a dimension of at least 500 mm x 500 mm in $x-y$ plane, which allows some millimeter-level deviation.

*2) Orientation, Normal-to-surface vectors:* Given the location of the centroid of each rock, we estimate its normal vector for the best fitting plane of a nearby point cloud surface. For this goal, the principle of a RANSAC algorithm [11] searches for the best plane among a 3D point cloud surface.

A general plane equation is given as:

$$ax + by + cz + d = \mathbf{n}^T \hat{\mathbf{x}} = 0, \tag{5}$$

where $\mathbf{n} = [a, b, c]^T$ is the normal vector of plane parameters to estimate and $\hat{\mathbf{x}} = [x, y, z, 1]^T$ is the homogeneous point coordinate of the cloud.

The algorithm starts by randomly selecting three points from the cloud, fitting the plane parameters, and detecting all points of the point cloud that belong to the same plane by a given threshold. The process is repeated multiple times, until the plane equation containing the largest number of inliers is determined, the plane is considered as the best fitting plane.

As the point cloud estimated with the stereo-camera setup usually does not capture highly slanted or parallel-to-the-optical axis planes, inliers can be selected using a predefined threshold value $\theta$, as points whose distance to plane is lower than a threshold

$$(x, y, z) \in Z^3 : 0 \leq |ax_i + by_i - z_i + c| \leq \theta. \tag{6}$$

The threshold $\theta$ can also control the expected proximity of an object surface to a plane model. For object surfaces containing many bumps or cavities, larger values of $\theta$ can be beneficial.

## V. EXPERIMENTS

### A. Settings

The whole data for rock detection was split into training, validation, and testing sets for fair comparison. Specifically, 70% of the images (in 1280 x 720 resolution) were selected for training, 20% for validation, and the remaining 10% for testing. During parameter tuning, we used training data to fit network parameters by evaluating the performance on the validation set.

***(a)*** *Left image of stereo camera taken at the secondary breaking site*

***(b)*** *An example of point cloud generated from left and depth image*

***Fig. 4*** *Input images for visual perception*

### B. Implementation Details

We set our visual perception system on Ubuntu with the following environment settings:

- OpenCV 3.4.0
- PCL 1.7.1
- CUDA 10.0
- CuDNN 7.4

We implemented all schemes in C++ with OpenCV library and Point Cloud Library (PCL).

From each video frame, we extracted a left image (an example is shown in Fig. 4a) together with a right image, computing its depth map (by means of the proposed plane sweeping method) to generate a point cloud (an example is shown in Fig. 4b). In parallel, the left images with labelled bounding boxes were provided to train the rock detector.

### C. Evaluation of Rock Detection

We adopted the off-the-shelf YOLO detector using a variant of darknet-53 [24] in view of its solid detection performance as well as its efficiency during inference. We trained the darknet using our data by setting a learning rate of 0.001, which converges at an average loss of 0.12. We achieved good detection results during testing. Threshold was 0.25, true positive (TP) was 7262, false positive (FP) was only 16, false negative (FN) was 23, and the average intersection-over-union (IoU) was 89.69%. Moreover, an

average precision of 90.91% was reached. We validated the stability of the model using images at different scales and rotations to retain result robustness. Fig. 5 illustrates the detection result from an offline video, and the detected objects provided by YOLOv3 are highlighted with 2D bounding box.

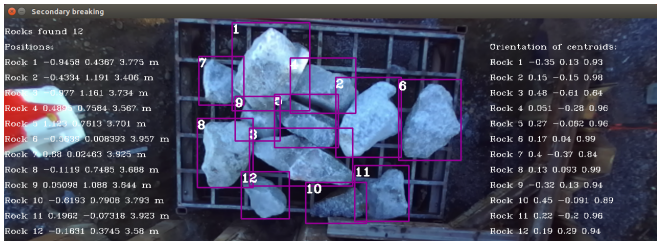
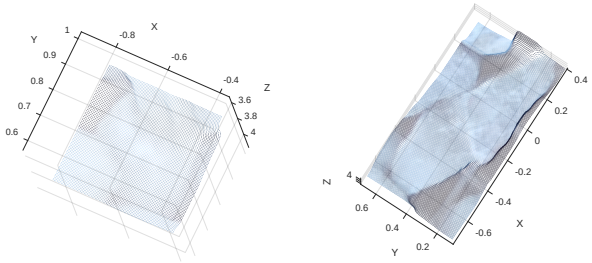



***Fig. 5*** *Detection and Localization of Rocks at approx. 15Hz*

In addition, this single shot-based rock detector can efficiently localize all rocks at a video frame rate around 15 Hz. In Fig. 5, it can be seen that the rock 3 has a sharp edge in the middle, which is hard to segment properly using unsupervised learning methods [9], while rock 9 is occluded and truncated by rocks 1 and 8, which is harder to recognize using the aforementioned method.

### *D. Evaluation on Pose Estimation*



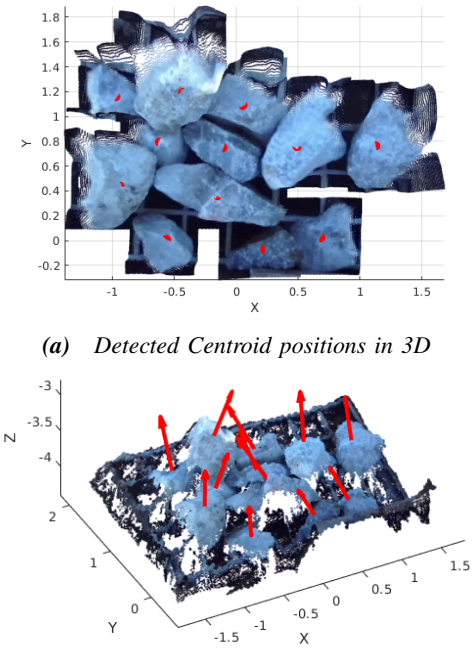
***(a)*** *Rock 9* ***(b)*** *Rock 3*

***Fig. 6*** *Examples of the point cloud for rock 3 and rock 9 segmented by the projected 3D bounding boxes*

Here we conduct experiments to evaluate the results of estimating the position and orientation of individual rocks within 3D regions. For each detected 2D region, every pixel within has its 3D corresponding point in 3D point cloud with X,Y,Z and RGBA color. After 2D to 3D correspondence mapping, we obtained their 3D regions in a point cloud. Fig. 6 indicates rocks 3 and 9 in point clouds, through which 6D pose estimation can be performed.

Fig. 7 illustrates detected 3D regions overall, along with estimated centroids and normal vectors for each region. Estimated centroids for each rock are drawn as red spots (as shown in Fig. 7a where they geometrically reside at the center of each rock, even for all occluded rocks).

The estimation of normal vectors was performed using kd-tree to search for the neighbors around each centroid point, and applying the RANSAC method for finding the best fitting plane. As no ground-truth normal vectors were available, we visualized the normal vectors together with the rocks for



***(a)*** *Detected Centroid positions in 3D*

***(b)*** *Estimated Normal Vectors in 3D*

***Fig. 7*** *Estimation of Centroids' Positions and Normal Vectors for each 3D region*

quality evaluation. Fig. 7b presents the result of estimating the normal vectors shown with red arrows for each rock. As a result, those normal vectors were perpendicular to the estimated main surface plane of each rock. More qualitative results are shown in Fig. 8.

## VI. CONCLUSIONS

We have proposed a novel fast method for 3D object detection and target pose estimation for complex scenes containing irregularly shaped and sized blasted rocks that can be in an overlapping pile. Even though object detection using bounding boxes has been widely studied, its extension to 3D in such complicated scenes remains a challenge, especially in a real outdoors environment. On one hand, in real-world outdoor applications, the 3D bounding boxes detector with LiDARs is not an efficient method for solving complex scenes with many sharp changes in the depth and overlying edges that are only visible on the images. On the other hand, 3D detection methods operating solely on dense point clouds can be computationally expensive, rendering the required real-time operation hardly feasible. This paper has presented an efficient online method by taking advantage of fast 2D object detection combined with the 2D to 3D plane-sweeping stereo matching method for 3D object detection. Given secondary rock breaking as an application, the proposed robotic visual perception method can meet the requirements for autonomous breaking required for the mining industry with its reliable object detection, real-time performance, and substantial accuracy on object pose estimation. The experiment results veried the efficiency of the proposed method with 90.91% detection accuracy at

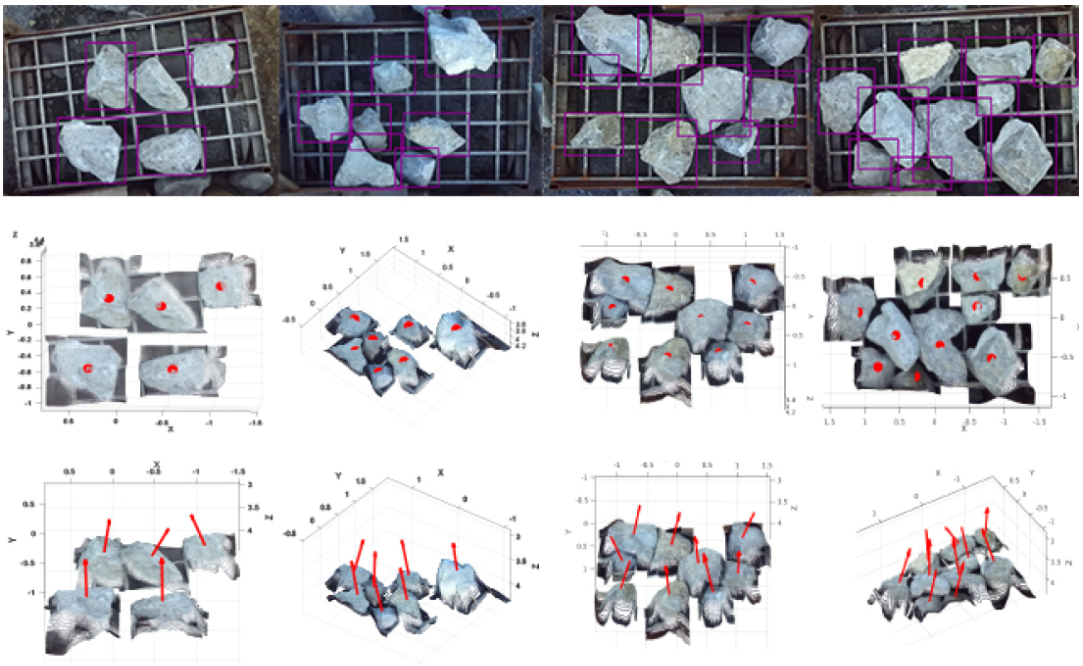


***Fig. 8*** *More visualization results of detection (top), position (middle), and normal estimation (bottom)*

15Hz in real outdoors worksite conditions. Our next research objective is to experimentally verify the success rate of real rock breaking with the machine vision estimated rock surface position as "a sweet spot" for the productive robotized operation.

## REFERENCES


[1] P. G. Ranjith, J. Zhao, M. Ju, R. V. De Silva, T. D. Rathnaweera, and A. K. Bandara, "Opportunities and challenges in deep mining: A brief review," *Engineering*, vol. 3, no. 4, pp. 546–551, 2017.

[2] J. J. Green and D. Vogt, "Robot miner for low grade narrow tabular ore bodies: the potential and the challenge," 2009.

[3] E. Duff, C. Caris, A. Bonchis, K. Taylor, C. Gunn, and M. Adcock, "The development of a telerobotic rock breaker," in *Field and Service Robotics*. Springer, 2010, pp. 411–420.

[4] P. Corke, J. Roberts, and G. Winstanley, "Vision-based control for mining automation," *IEEE Robotics & Automation Magazine*, vol. 5, no. 4, pp. 44–49, 1998.

[5] ——, "3d perception for mining robotics," in *Field and Service Robotics*. Springer, 1998, pp. 46–52.

[6] H. Takahashi and T. Monden, "Automatic breaking system of large rocks by use of force sensors," in *INTERNATIONAL SYMPOSIUM ON ROBOTICS*, vol. 30. Citeseer, 1999, pp. 705–710.

[7] A. Iamrurksiri, T. Tsubouchi, and S. Sarata, "Rock recognition using stereo vision for large rock breaking operation," in *Field and Service Robotics*. Springer, 2014, pp. 383–397.

[8] L. Niu, S. Smirnov, J. Mattila, A. Gotchev, and E. Ruiz, "Robust pose estimation with a stereoscopic camera in harsh environments," *Electronic Imaging*, vol. 2018, no. 9, pp. 1–6, 2018.

[9] L. Niu, M. M. Aref, and J. Mattila, "Clustering analysis for secondary breaking using a low-cost time-of-flight camera," in *2018 Ninth International Conference on Intelligent Control and Information Processing (ICICIP)*. IEEE, 2018, pp. 318–324.

[10] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[11] 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, no. 6, pp. 381–395, 1981.

[13] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[16] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *European conference on computer vision*. Springer, 2016, pp. 354–370.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[19] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[20] J. Lahoud and B. Ghanem, "2d-driven 3d object detection in rgb-d images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4622–4630.

[21] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 918–927.

[22] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[23] S. Smirnov, A. Gotchev, and M. Georgiev, "Comparison of cost aggregation techniques for free-viewpoint image interpolation based on plane sweeping," in *Ninth International Workshop on Video Processing and Quality Metrics for Consumer Electronics - VPQM*, 2015.

[24] J. Redmon, "Darknet: Open source neural networks in c," http://pjreddie.com/darknet/, 2013–2016.