A fast automatic extraction method for rock mass discontinuity orientation using fast k-means++ and fast silhouette based on 3D point cloud

Wei Wu¹,²,³, Keshen Zhang²*, Hehua Zhu¹,²,³
¹State Key Laboratory for Disaster Reduction in Civil Engineering, Tongji University, Shanghai 200092, China
²College of Civil Engineering, Tongji University, Shanghai 200092, China
³Key Laboratory of Geotechnical and Underground Engineering of Ministry of Education, Tongji University, Shanghai 200092, China

*Corresponding author: zhangkeshen@tongji.edu.cn

Abstract. In this paper, a fast automatic extraction method for rock discontinuity orientation using fast k-means++ and fast silhouette based on three-dimensional (3D) point cloud is proposed. The main contributions of the proposed method include: (1) a concept of base vector is proposed, (2) a fast k-means++ algorithm and a fast clustering validity index of silhouette are proposed and combined to automatically achieve a fast optimal grouping of discontinuities, and (3) the generation and optimization of independent discontinuity based on a double-clustering algorithm. The proposed method consists of three steps: (1) point cloud preprocessing, (2) optimal grouping of discontinuities, and (3) discontinuity generation and optimization. Several cases, including two polyhedral models under laboratory conditions and two rock slope models, are adopted to verify the accuracy and efficiency of the proposed method. A sensitivity analysis is performed to determine the optimal parameters. The results show that the proposed method can automatically achieve a fast and effective extraction of discontinuity orientations compared with the classical method, Slob (2010), Riquelme et al. (2014) and Li et al. (2016), and the efficiency is significantly improved compared with Rousseeuw (1987), Arthur and vassilvitskii (2007) and Li et al. (2016).

1. Introduction
Rock mass discontinuity is of vital importance to evaluate strength, permeability and stability of rock masses and largely determines the mechanical behavior of rock mass (ISRM, 1978; Hooke and Bray, 1981; Hudson, 2000). Orientation is one of the ten key parameters recommended by the International Society for Rock Mechanics to quantitatively describe rock discontinuities for orientation largely controls the possibility of unstable conditions or excessive deformations developing (ISRM, 1978). The traditional method of discontinuity information collection requires geo-engineers to manually measure the exposed rock surface using a geological compass or scanline. The main shortcomings of traditional method are: (1) lack of accuracy because of sampling difficulties, choice of sampling method, human bias and instrument error, (2) time-consuming and incomplete for the inaccessible areas, and (3) dangerous for geo-engineers because of the rock fall of unstable rock mass (Kemeny &
post, 2003; Abellán et al., 2014; Bolkas et al.2018). Alternatively, non-contact measuring methods, such as photogrammetry and Light Detection and Ranging (LIDAR), provide effective approaches to in situ measurement and allow discontinuities to be measured from images and 3D point clouds of rock mass exposures. (Roncella et al., 2004; Haneberg, 2008; Lato et al., 2009, 2010; Gigli and Casagli, 2011; Otoo et al., 2011; Fisher et al., 2014; Riquelme et al., 2014, 2015, 2016, 2018; Cacciari et al., 2016; Chen et al. 2016; Li et al.,2016; Zhu et al., 2016; Kong et al., 2020). Compared with the traditional methods, the main advantages includes: (1) high resolution, high precision and high speed, (2) ability to collect large-scale and multi-scale data, and (3) secure investigation for inaccessible or hazardous areas (Riquelme et al., 2014; Buyer and Schubert, 2017).

Recently, many studies are performed to extract discontinuity orientations based on 3D point cloud which generated from stereo photogrammetry or LIDAR. In some studies, discontinuities are selected manually from 3D point clouds and orientations are calculated based on the least square method (Abellán et al., 2006; Fernández, 2005; Sturzenegger and Stead, 2009). Currently, many studies use normal vector distribution of point cloud to group and extract discontinuity orientations. Specifically, some researchers use principal component analysis (PCA) to calculate normal vector of each point in the point cloud (Jaboyedoff et al. 2007; Riquelme et al., 2014). And some researchers use volumetric pixels (Gigli and Casagli, 2011; Guo rt al., 2017) or 3D TIN (Slob et al., 2010; Vöge et al., 2013; Chen et al., 2016) to calculate normal vector in a local region. In the analyzation of discontinuity grouping, Gigli and Casagli (2011) use manual operation on stereographic projection of normal vector to group discontinuity. Riquelme et al. (2014) use the method of density-based spatial clustering of applications with noise (DBSCAN) to semi-automatically group discontinuities, however this method requires to manually adjust the parameters for each model according to visual observation. Kong et al. (2020) use the method of fast search and find of density peaks (CFSFDP, Rodriguez and Laio, 2014) for discontinuity grouping, however this method also requires manual adjustment of parameters for each model to optimize the recognition results. Slob (2010) use the method of fuzzy c-means to group discontinuities, however this method requires visual observation to determine the optimal clustering number. Chen et al. (2016) used an improved k-means algorithm to automatically group discontinuities, however the efficiency of this method is low especially when deals with large scale of point clouds.

This paper is organized as follows: a fast automatic discontinuity orientation extraction method is introduced in Section 2, the proposed method is applied to polyhedron models and rock slope models in Section 3, features of the method is discussed in Section 4, and some conclusions are drawn in Section 5.

2. Methodology

The method for discontinuity orientation extraction consists of three steps: (1) point cloud preprocessing, (2) automatic optimal grouping of discontinuities based on a fast k-means++ algorithm and a fast clustering validity index, and (3) discontinuities generation and optimization.

2.1. Point cloud preprocessing

The raw point cloud of rock slope usually contain vegetation, unnecessary and sparse points which affect the precision of recognition and also increase processing time (Li et al., 2016). In addition, instrument errors, dust and dynamic disturbances in the open field can also generate noisy points (slob, 2010). As shown in Fig. 1a and Fig. 1b, there are many noisy points locate on the surfaces of the regular polyhedron models even if they are scanned under laboratory conditions (in Section 3.1). Therefore, raw point cloud in this paper are first resampled and denoised using moving least squares method (Alexa et al., 2003), and then a Delaunay triangulation algorithm is performed to generate 3D TIN model (Fig. 1c and Fig. 1d) of point cloud using Halcon software (MVTec, 2012).
Figure 1. Point cloud preprocessing. (a) Raw points of case A-1, (b) Raw points of case A-2, (c) Preprocessed points of case A-1, (d) Preprocessed points of case A-2.

2.2 Fast optimal grouping of discontinuity

In many recent studies of rock discontinuity recognition, normal vectors, which representatively generated using volumetric pixels (Gigli and Casagli, 2011; Guo et al., 2017), neighboring points (Jaboyedoff et al., 2007; Riquelme et al., 2014) and 3D TIN (Slob et al., 2010; Vöge et al., 2013; Chen et al., 2016), are applied to discontinuity grouping. In this paper, normal vector are generated from 3D TIN obtained in Section 2.1. In order to group discontinuity and determine the optimal group number, manual extraction methods (Gigli and Casagli, 2011) and semi-automatic extraction algorithms (Riquelme et al., 2014; Kong et al., 2020) are often adopted with manual operation or manual adjustment of parameters. In addition, Slob (2010) use the method of fuzzy c-means to group discontinuities; however this method requires to determine the optimal clustering number by visual observation. Chen et al. (2016) used an improved k-means algorithm to automatically group discontinuities, however the efficiency of this method is low especially when deals with large scale of point clouds.
In this part, a new concept of base vector is proposed in Section 2.2.1, a fast k-means++ clustering algorithm is proposed in Section 2.2.2 and a fast clustering validity index of silhouette is proposed in Section 2.2.3.

### 2.2.1. Base vector

In this paper, given the base vector set $\text{VEC}_{\text{base}} = \{\text{vec}_1, \text{vec}_2, ..., \text{vec}_{N_b}\}$ ($N_b$ the number of base vector), vectors in $\text{VEC}_{\text{base}}$ satisfy the following conditions: (1) the angle between any two adjacent vectors is $\\text{Deg}$ (the optimal value of $\\text{Deg}$ is determined in Section 3.2), (2) unit vectors, and (3) coordinate $z \geq 0$. In other words, the base vector can be regarded as points uniformly distributed on the unit hemispherical plane with $z \geq 0$. Because according to geometry, it proves to be difficult generating uniform points on the unit hemispherical surface, we use the method of regular icosahedron subdivision to generate uniform points approximately. Therefore, the $\\text{Deg}$ value actually corresponds to the number of subdivision $N_{\text{sub}}$. Base vectors are generated as follows:

**Algorithm 1. Base vector generation.**

**Input:** $V_{20}$, coordinate set of regular icosahedron; $F_{20}$, triangular mesh set of regular icosahedron; $N_{\text{sub}}$, subdivision number

**Output:** $\text{VEC}_{\text{base}}$, base vector set

**Begin**

$F_{\text{tmp}} \leftarrow F_{20}$ // temporary triangular mesh set

$V_{\text{tmp}} \leftarrow V_{20}$ // temporary coordinate set

for $i \leftarrow 1$ to $N_{\text{sub}}$

for each triangular mesh in $F_{\text{tmp}}$

given the vertex $V_{\text{tmp0}}$, calculate the midpoint coordinates of each edge of the triangular mesh and save in $V_{\text{tmp1}}$, set $V_{\text{tmp1}} \leftarrow \|V_{\text{tmp1}}\|$, $V_{\text{tmp}} \leftarrow V_{\text{tmp}} \cup V_{\text{tmp1}}$, calculate the triangular mesh generate by $V_{\text{tmp0}}$ and $V_{\text{tmp1}}$ and save in $F_{\text{tmp}}$, $F_{\text{tmp}} \leftarrow F_{\text{tmp}} \cup F_{\text{tmp1}}$

end for

end for

$\text{VEC}_{\text{base}} \leftarrow \{\text{vec}|\text{vec} \in V_{\text{tmp}}, \text{z coordinates of vec } \geq 0\}$

**End**

Fig. 2 shows the results of base vector generation with different subdivision number, and Table. 1 shows the relationship among subdivision number, mean angle of any two adjacent base vectors and base vector number.

![Figure 2. Base vector. (a) 5-time subdivision, (b) 6-time subdivision, (c) 7-time subdivision.](image)

### 2.2.2. Fast k-means++

K-means algorithm has become one of the most commonly used clustering algorithm due to its simplicity (Xhafa et al., 2016). However one of its main disadvantages is the sensitivity to the selection of initial clustering centroids (Peña et al., 1999; Likas et al., 2003). Traditional k-means selects initial clustering centroids by experience or random way, which makes the accuracy of the algorithm low. Arthur and Vassilvitskii (2007) proposed an algorithm of k-means++, which optimized the selection method of initial clustering centroids and effectively improved the convergence speed and accuracy of k-means algorithm. However, the algorithm still randomly selects the first initial
clustering centroid, which affects the quality of the initial clustering centroid selection and the accuracy of clustering, and the convergence (the time complexity is $O(N)$) is still slow when deals with large number of points. Based on k-means++, Chen et al. (2016) proposed an improved k-means algorithm which selected initial clustering centroids based on normal vector density. This method improved the quality of initial clustering centroids and final clustering, which effectively achieved discontinuity grouping. However the calculation of initial clustering centroid requires a large number of iterations (time complexity is $O(N^2)$) which significantly slow when deals with large number of points.

In this section, a fast k-means++ is proposed which improves the initial clustering centroid selection and speed of clustering convergence. Given the normal vector set of point cloud generated by 2.1 VEC and base vector generated by 2.2.1 VEC$_{\text{base}}$. The proposed fast k-means++ is described as follows.

- Initial clustering centroid selection

In this paper, the distance between any two normal vectors $\text{vec}_i$ and $\text{vec}_j$ is defined as:

$$\text{dist}(i, j) = \arccos(|\text{vec}_i \cdot \text{vec}_j|)$$  \(1\)

Firstly, calculate the density $D$ of the base vector (as shown in Fig. 3b), take the base vector with the largest value in $D$ as the initial clustering centroid and put it into $U$, then remove it from $D$. Secondly, select the base vector, whose minimum distance from the existing initial clustering centroids of $D$ is the largest, as another initial cluster centroid and merge it into $U$, then delete the index from $D$. In this way, the remaining initial clustering centroids are selected. The specific algorithm is as follows:

**Algorithm 2. Initial clustering centroids**

Input: VEC, normal vector set of point cloud; VEC$_{\text{base}}$, base vector set; $k$, clustering number

Output: U, set of initial clustering centroids

Begin

D={}; // set of initial base vector density
U={}; // set of initial clustering centroids

for each vector in VEC
    calculate the distance between the vector and all vectors in VEC$_{\text{base}}$ according to Eq. (1) and save in $\text{dist}_1$, find $\text{dist}_1[]$ which equals to min{$\text{dist}_1$}, then $D[] ← D[] + 1$
end for

find $D[]$ which equals to max{$D$}, $U ← U ∪ \{\}$, VEC$_{\text{base}} ←$ VEC$_{\text{base}} − \{\}$;

for the 2$^{\text{nd}}$ to $k^{\text{th}}$ initial clustering centroids

for each vector in VEC$_{\text{base}}$
    calculate the minimum distance the based vector to all vectors in $U$ according to Eq. (1) and save in $\text{dist}_2$
end for

find $\text{dist}_2[]$ which equals to max{$\text{dist}_2$}, then $U ← U ∪ \{\}$, VEC$_{\text{base}} ←$ VEC$_{\text{base}} − \{\}$
end for
Figure 3. The process of fast k-means++ based on case B-1. (a) Normal vectors, (b) Base vector with density, (c) Clustering results of base vector, (d) Clustering results of normal vector. Point size in (b) represents the relative density of base vector, however the size is not linear to density with considering of display. In (c)-(d), each color represents a cluster of normal vectors and a group of discontinuities.

2.2.3. Fast silhouette.

A main shortcomings of k-means++ algorithm is that it requires to determine the clustering number in advance (PENIA, 1999). Chen et al. (2016) used a clustering validity index silhouette (Rousseeuw, 1987) to evaluate clustering quality and determine the optimal clustering number, which is the grouping number of discontinuities. However the traditional silhouette (Rousseeuw, 1987) requires to calculate the distance between any two different samples in the data set, which is time-consuming especially when deals with large scale of data.

In this section, a fast silhouette algorithm is proposed to evaluate the clustering quality efficiently. The main steps are as follows: given base vector density $D$, clustering results of base vector $R_{\text{base}}$ and clustering number $k$. For the cluster $k_{\text{tmp}} \in \{1, \ldots, k\}$, the silhouette value of $i^{\text{th}}$ base vector of cluster $k_{\text{tmp}}$ is defined as:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$  \hspace{1cm} (3)

In Eq. (3), $a(i)$ is defined as the distance between the $i^{\text{th}}$ base vector and all others base vectors in cluster $k_{\text{tmp}}$ as:

$$a(i) = \frac{\sum_{j \in R_{\text{base}, k_{\text{tmp}}}} |d_j| \cdot \text{dist}(\text{vec}_i, \text{vec}_j) \cdot d_j}{|R_{k_{\text{tmp}}}| - 1}$$ \hspace{1cm} (4)

In Eq. (4), $\text{vec}_i$ denotes the $i^{\text{th}}$ base vector in cluster $k_{\text{tmp}}$, $d_j$ denotes the $j^{\text{th}}$ base vector density in $D$ and $|R_{k_{\text{tmp}}}|$ denotes the number of
base vector in cluster $k_{tmp}$. In Eq. (3), $b(i)$ is defined as the smallest value of the mean distance between the $i^{th}$ base vector and all base vectors in other cluster except $k_{tmp}$ as

$$b(i) = \min_{m \in \{1, 2, \ldots, k\}, m \neq k_{tmp}} \left\{ \frac{\sum_{j \in R_{base, m}} \text{dist}(\text{vec}_i, \text{vec}_j) \cdot d_j}{|R_j|} \right\}$$

(5)

Silhouette value is ranged from -1 to 1, a larger silhouette value indicates a better clustering quality. In this paper, the mean silhouette value of all base vectors with density is used to evaluate the quality of clustering, and the clustering number with local maximum silhouette is defined as the optimal. Specifically, the algorithm is organized as follows:

**Algorithm 3. Determination of the optimal clustering number**

| Input: VEC, normal vector set of point cloud |
| Output: $k_{opt}$, the optimal clustering number |

**Begin**

generate base vector $\text{VEC}_{base}$ according to section 2.2.1

$\text{Sil} = \{}$ // silhouette set of each clustering number

set $k_{tmp} \leftarrow 2$

while 1

fast k-means++ is performed to obtain clustering result $R_{base}$ and base vector density $D$

according to Section 2.2.2, calculate mean silhouette value of cluster $k_{tmp}$ according to Eq. (3) and save in $\text{Sil}[k_{tmp}]$

if $\text{Sil}[k_{tmp} - 1] < \text{Sil}[k_{tmp}] > \text{Sil}[k_{tmp} + 1]$ then

$k_{opt} \leftarrow k_{tmp}$

done while

else then

$k_{tmp} \leftarrow k_{tmp} + 1$

done if

end while

**End**

2.3. Discontinuity generation and optimization

In the above sections, discontinuities are optimally grouped based on normal vectors of triangular meshes. In this step, triangular meshes in the same cluster are segmented into independent discontinuities. Chen et al. (2016) used the common edge of triangular meshes as a criterion segmentation. However, this method is easy to falsely connect discontinuities, which belongs to different groups, and identify them as the same discontinuity. For example, as shown Fig. 4c, the region in the black circle is identified as one independent discontinuity according to Chen et al. (2016), which fails to distinguish discontinuities No. 22 and No. 52 (as shown in Fig. 4f).
In this section, a double-clustering algorithm is performed to generate independent discontinuities accurately. Firstly, discontinuities are preliminarily generated according to Section 2.2. Secondly, each discontinuity is re-clustered according to Section 2.2 to optimize the discontinuity recognition quality.

2.3.1. Discontinuity preliminary generation.

In this section, triangular meshes of the same cluster are segmented into independent discontinuities according to their spatial adjacency. The segmentation criterion is that if two triangular meshes share at least one common vertex, then they are segmented into one discontinuity. Given the triangular mesh set of 3D TIN (generated in Section 2.1) F, the specific steps are as follows:

**Algorithm 4. Discontinuity preliminary generation**

Input: F, triangular mesh set; R_vec, normal vector clustering result; k_opt, optimal clustering number

Output: GP, independent discontinuity set

Begin

for i ← 1 to k_opt // for each cluster

    gp = {} // temporary results in this cluster
    Facet ← {F[i] | i ∈ R_vec[i]}
    while 1
        gp ← Facet{1} and Facet{1} ← {}
        while 1
            tmp ← gp ∩ Facet
            if tmp = {} then
                GP ← GP ∪ gp
            end while
            else then
                gp ← gp ∪ tmp and Facet ← Facet − tmp
            end if
        if Facet={} then
            end while
        end if
    end for

2.3.2. Discontinuity optimization.
In order to reduce the false segmentation (as described in Section 2.3) and improve the accuracy of discontinuity recognition, this section perform a re-clustering algorithm to further recognize the distinct orientation area in preliminary discontinuities. Specifically, the normal vectors of each discontinuity generated by Section 2.3.1 is treated as the input of Section 2.2 and Section 2.3.1, and the maximum silhouette value Sil_opt, which corresponding to the optimal clustering number, is calculated and serves as a re-clustering index. An optimization threshold T (the optimal value is determined in Section 3.2) is defined. If Sil_opt is larger than T, then the corresponding preliminary discontinuity is replaced by the re-clustering results, and if Sil_opt is equal to or smaller than T, the corresponding discontinuity is preserved. The specific steps are as follows:

**Algorithm 5. Discontinuity optimization**

| Input: | VEC, normal vector set of point cloud; GP, preliminary discontinuity set |
| Output: | GP<sub>opt</sub>, optimized discontinuity set |

**Begin**

```
GP<sub>opt</sub> = {};
for i ← 1 to n // n denotes the number of discontinuities in GP
vec ← {VEC[j] | j ∈ GP[i]}
input vec to Section 2.2, calculate clustering result R<sub>vec</sub>, optimal clustering number k<sub>opt</sub> and optimal silhouette value Sil<sub>opt</sub>
If Sil<sub>opt</sub> > T then
G<sub>P</sub> = GP<sub>opt</sub>, R<sub>vec</sub> and k<sub>opt</sub> to Section 2.3.1, calculate the generated discontinuity set g<sub>P</sub> and GP<sub>opt</sub> ← GP<sub>opt</sub> ∪ g<sub>P</sub>
else then
GP<sub>opt</sub> ← GP<sub>opt</sub> ∪ GP[i]
end if
end for
```

2.3.4. Plane fitting and orientation calculation.

In the above sections, independent discontinuities are generated and the triangular meshes are assigned to the belonging discontinuities. Because of curvature and roughness of rock mass surfaces, point cloud is usually undulated and exists noise (Chen et al., 2016). Therefore, discontinuity orientation is often calculated based on fitting planes. Specially, the least square method is often applied to plane fitting (Fernández, 2005; Sturzenegger and Stead, 2009a; Gigli and Casagli, 2011; Riquelme et al. 2014), however this method requires manual selection of parameters to denoise. Alternatively, the random sample consensus method (RANSAC) is widely applied to plane fitting (Ferrero et al., 2009; slob, 2010; Chen et al., 2016) for the robustness to noise.

In this section, RANSAC algorithm is adopted to plane fitting and is performed as introduced by Fischler and Bolles (1981). Given the normal vector of a fitting plane (l, m, n), the orientation is calculated as follows (Priest, 1993):

\[
\theta = \tan^{-1}\left(\frac{m}{l}\right) + D \\
\delta = \tan^{-1}\left(\frac{n}{\sqrt{l^2 + m^2}}\right)
\]

where D denotes the angle coefficient for dip direction calculation and is determined as:

\[
D = \begin{cases} 
0^\circ & l \geq 0 \text{ and } m \geq 0 \\
180^\circ & l \leq 0 \text{ and } m < 0 \\
360^\circ & \text{others}
\end{cases}
\]

3. Application
3.1. Data description
The raw point cloud of rock slope

3.1.1. Case A. Case A includes two polyhedron models (Fig. 1a).
Case A-1 is a cube consisting of 40,414 raw points, and case A-2 is a polyhedron composed of six upper faces of a dodecahedron (Fig. 1b), consisting of 60,488 raw points, both the two models are scanned with laboratory conditions by Riquelme et al. (2014). After the data processing described in 2.1, case A-1 generates 28,701 points and 57,319 triangular meshes (Fig. 1c), and case A-2 generates 40,414 points and 80,581 meshes (Fig. 1d). Obviously, there are five independent interfaces in case A-1 that can be divided into three groups according to the parallel relationship, and six structural surfaces in case A-2 model are divided into six groups. Therefore, these two models are mainly used as benchmark data to verify the conceptual accuracy of the proposed method.

3.1.2. Case B. Case B includes two real rock slope models (Fig. 5).
Case B-1 was a road slope located in Torroja, Spain and was scanned by Slob (2010) collecting a raw point cloud of 1,715,256 points. The analyzed region was selected as shown in Fig. 5a and generated 619,035 points and 1,223,012 triangular meshes according to Section 2.1. Because discontinuities in case B-1 are relative regular, this model was selected to verify the accuracy and determine the optimal parameters of the proposed method. Case B-2 was another road slope located in Colorado, USA and the raw data, which consisted of 1,515,722 points, was publicly available at Rockbench repository (Lato et al., 2013). The analyzed region was selected as shown in Fig. 5b and generated 1,228,056 points and 2,444,078 triangular meshes according to Section 2.1. Case B-2 was selected to compare the results of the proposed method and other methods.

Figure 5. Case B. (a) Case B-1 (b) Case B-2. Analyzed region is in the black rectangle.

3.2. Sensitivity analysis
The proposed method consists of two main parameters, including base vector density (Section 2.2.1) and discontinuity optimization threshold T (Section 2.3.2), which affect the accuracy and efficiency. In this section, the main steps of the proposed algorithm is validated, and the optimal values of the parameters are determined. Three models, including case A-1, case A-2 and case B-1 are analyzed in this section.

3.2.1. Fast k-means++ and fast silhouette.
In this paper, fast k-means++ and fast silhouette is essentially achieved by transforming the calculation of N normal vectors to the calculation of n base vectors (n<<N and n is constant). Therefore, the accuracy and efficiency of the proposed algorithm depend on the density of the base vector. Large density of basis vectors will improve accuracy but reduce efficiency, and small density of basis vectors will improve efficiency but reduce accuracy. Considering the relationship between base vector density and subdivision number (Table 1), firstly, the base vectors of 4 to 7 subdivision numbers are performed to evaluate the accuracy of fast silhouette and select the optimal base vector density. Secondly, the accuracy of fast k-means++ is analyzed based on the optimal base vector density. Fig. 6 is the comparison between fast silhouette and traditional silhouette (Rousseeuw, 1987), and it can be summarized as: (1) the optimal clustering number \( k_{opt} \) of fast silhouette, which corresponding to the largest silhouette value, is the same as traditional silhouette, (2) the difference between fast silhouette and traditional silhouette towards silhouette value corresponding to the optimal clustering number is significantly small except the result of 4-time subdivided base vector (as shown in Fig. 6c), which means the accuracy of fast silhouette is guaranteed, and (3) the silhouette difference of other clustering numbers except \( k_{opt} \) between the fast and traditional algorithm is not as small as (2), especially for the 4-times subdivided base vectors (as shown in Fig. 6a). Because the fast silhouette in this paper aims to determine the optimal clustering number \( k_{opt} \), therefore the base vector subdivision number larger than 4 is proven to be effective. Considering the calculation efficiency, the optimal base vector \( b_{opt} \) is determined as 5-time subdivided base vector.

**Figure 6.** Comparison between the fast silhouette and the traditional silhouette. (a) Case A-1, (b) Case A-2, (c) Case B-1. The traditional silhouette value is plotted in red. The fast silhouette value calculated using base vectors generated by 4-time subdivision is plotted in yellow, 5-time subdivision is plotted in green, 6-time subdivision is plotted in blue, and 7-time subdivision is plotted in pink.

Fig. 7 shows the comparison between the fast k-means++ and the traditional k-means++ (Arthur and vassilvitskii, 2007) based on the optimal base vector \( b_{opt} \). It can be summarized as: (1) the optimal clustering number \( k_{opt} \) of fast k-means++ is the same as the traditional k-means++, and both algorithms accurately reflect the grouping characteristics of surface or discontinuity of the three analyzed models (as shown in Fig. 8), and (2) the difference of fast k-means++ and traditional k-means++ towards silhouette value corresponding to \( k_{opt} \) is significantly small, which means the clustering quality of fast k-means++ is guaranteed.

**Figure 7.** Comparison between the fast k-means++ and the traditional k-means++. (a) Case A-1, (b) Case A-2, (c) Case B-1.
Figure 8. The optimal clustering results base on fast k-means++ and fast silhouette. (a) Case A-1, (b) Case A-2, (c)In each figure, each color represents a cluster of normal vectors and a group of discontinuities. Although the silhouette values of the fast k-means++ may fluctuate in the non-optimal clustering number (as k=4 in Fig. 7b and k=7 in Fig. 7c), the accuracy of the fast k-means++ will not be affected because the main aim of fast k-means++ is to determine the optimal clustering number.

3.2.2. Discontinuity optimization.

As described in Section 2.3.2, discontinuity optimization is performed to reduce the false segmentation and improve the accuracy of discontinuity recognition. A threshold T is defined to control the accuracy of discontinuity optimization. In this section, the proposed method is applied to case B-1 based on different values of T. As shown in Fig. 9a, when T is set to 1, which means no discontinuity optimizations, some independent discontinuities falsely contain areas with different orientations (especially for regions in the black circles compared with Fig. 9b, in which T=0.6). Therefore, a small T is required to make more discontinuity optimized. However, over small settings of T (as shown in Fig. 9c, in which T=0.5) make the discontinuities trivial and irregular in shape, which is because the small undulation and curvature of the surface are extracted as new discontinuities. And when T is set as 0.6 (as shown in Fig. 9b), the optimized discontinuities are comparatively accurate. Therefore, the optimal threshold T_opt is determined as 0.6.

Table. 2 shows part of the recognized independent discontinuity orientations (Fig. 9d) of case B-2 based on the optimal base vector base_opt and optimal threshold T_opt. And the results of the proposed method and manual measurement (Slob, 2010) is quite close, which means the accuracy of the proposed method under the optimal parameters is guaranteed.
3.3. Application and comparison to a rock slope
In this section, the proposed method is applied to case B-2 and the discontinuity recognition results are compared with other methods. The parameters are set as the optimal values determined in Section Fig. 10 shows the change of silhouette values under different clustering numbers. Because the silhouette value corresponding to clustering number 3 is the largest, the discontinuities in case B-2 are divided into 3 groups (Fig. 4a), which is the same as the results of Chen et al. (2016). And Fig. 4d-f
show part of the recognized discontinuities belonging to each group, and the discontinuities are numbered which accords with Riquelme et al. (2014) for comparison.

Figure 10. The values of fast silhouette based on Case B-2.

Fig. 4c shows the discontinuity recognition results of Chen et al. (2016) when the clustering number was 3. Compared with Fig. 4c and Fig. 4d-f, it indicates that the proposed method is more accurate in independent discontinuity extraction than Chen et al. (2016). For example, the region in the black circle of Fig. 4c is recognized as one discontinuity, which should be two different discontinuities No. 22 and No. 52 in Fig. 4f. And as shown in Table. 3, compared with other methods, the results of the proposed method is closer to the classical method for the maximum deviation and mean deviation are the smallest.

In addition for case B-2, the optimal clustering number of the proposed method is different from Riquelme et al. (2014). And actually when clustering number is 5 (Fig. 4b) instead of 3 (Fig. 4a), the discontinuity grouping results of the proposed method are quite similar to Riquelme et al. (2014). That is because the method of Riquelme et al. (2014) is a semi-automatic method which requires manually adjust parameters according to stereographic projection of normal vectors to make the results better accord with subjective observation, however the proposed method is an automatic method which does not require manually interfere during processing. And making the results of discontinuity grouping more accord with subjective observation is the next aim of the proposed method. Additionally, both the two method can accurately recognize discontinuities according to Table. 3.

Table 3

| Discontinuity index | Classical method | Riquelme et al. (2014) | Chen et al. (2016) | The proposed method | Riquelme et al. (2014) | Chen et al. (2016) | The proposed method |
|---------------------|------------------|------------------------|-------------------|---------------------|------------------------|-------------------|---------------------|
|                     |                  |                        |                   |                     |                        |                   |                     |
| 11                  | 249.2±0.2        | 246.2±0.2              | 244.6±0.8         | 244.9±0.8          | 3                      | 1.2                | 4.6                 | 4.8                 | 7.1                 | 3.8                 |
| 12                  | 264.2±0.7        | 256.9±0.2              | 256.2±0.2         | 257.1±0.2          | 7.3                    | 4.7                | 8                   | 4.8                 | 7.1                 | 3.8                 |
| 13                  | 264.2±1.9        | 70.3±3.5               | 251.36±2          | 252.6±3.7          | 13.7                   | 6.1                | 13                  | 5.7                 | 12.0                | 4.8                 |
| 14                  | 252.6±6.35      | 252.7±3.5              | 251.4±3.9         | 251.6±3.7          | 0.1                    | 1.2                | 2.6                 | 1.2                 | 1.0                 | 2.8                 |
| 15                  | 248.7±3.7        | 249.7±3.5              | 258.3±6.8         | 249.4±3.6          | 1                      | 1.1                | 2.1                 | 0.2                 | 0.7                 | 0.1                 |
| 16                  | 254.9±2.9        | 70.3±3.5               | 258.3±5.9         | 249.1±3.4          | 4.3                    | 6.1                | 4.3                 | 6.1                 | 5.7                 | 4.7                 |
| 17                  | 249.3±5.9        | 251.3±3.7              | 253.2±3.5         | 253.6±3.3          | 5.2                    | 3.2                | 3.3                 | 2.4                 | 3.7                 | 2.6                 |
| 21                  | 338.7±8.4        | 339.5±3.8              | 157.6±8.8         | 157.2±8.8          | 0.8                    | 0.9                | 1.1                 | 1.4                 | 1.5                 | 0.4                 |
| 22                  | 347.5±7.9        | 166.3±7.6              | 166.3±7.6         | 168.1±7.4          | 1.2                    | 2.4                | 1.2                 | 0.3                 | 0.6                 | 4.2                 |
| 23                  | 341.9±9.5        | 160.2±9.9              | 157.5±8.9         | 159.2±9.7          | 0.8                    | 0.4                | 3.5                 | 2.6                 | 1.8                 | 2.3                 |
| 24                  | 353.7±6.4        | 173.6±6.9              | 353.7±7.8         | 352.7±7.4          | 0.1                    | 0.5                | 0.4                 | 1.4                 | 0.8                 | 1.7                 |
| 31                  | 314.7±7.2        | 136.6±8.2              | 314.7±8           | 310.2±7.8          | 2.5                    | 5.4                | 0.6                 | 2.8                 | 3.9                 | 0.6                 |
| 32                  | 302.4±5.9        | 131.2±8.2              | 306.9±9.9         | 125.7±7.2          | 8.8                    | 6.8                | 14.1                | 14                  | 2.8                 | 1.3                 |
| 33                  | 330.2±8.3        | 143.9±9.7              | 145.6±8.9         | 147.6±8.6          | 6.3                    | 6.7                | 4.6                 | 6.9                 | 2.6                 | 3.9                 |
| 41                  | 286.1±5.8        | 97.6±6.3               | 286.5±9.8         | 285.1±6.7          | 8.5                    | 4.3                | 0.1                 | 0.9                 | 1.0                 | 1.8                 |
| 42                  | 274.2±5.1        | 91.1±5.0               | 272.6±4.7         | 271.8±4.9          | 3.1                    | 0.9                | 1.6                 | 3.5                 | 2.4                 | 1.2                 |
| 43                  | 272.3±4.6        | 96.6±4.8               | 277.3±4.9         | 278.1±4.8          | 0.6                    | 1.6                | 0.1                 | 2.9                 | 0.9                 | 2.0                 |
| 51                  | 305.7±7.6        | 123.4±7.6              | 305.7±7.6         | 304.1±7.5          | 1.6                    | 1.4                | 1.6                 | 4.4                 | 0.7                 | 2.4                 |
| 52                  | 290.2±6.7        | 105.8±9.9              | 109.3±7.6         | 112.3±8.8          | 4.4                    | 2.9                | 0.9                 | 9.6                 | 2.1                 | 1.8                 |

Maximum deviation

13.7 ± 6.8 ± 16.3 ± 14 ± 12 ± 4.8

Mean deviation

3.7 ± 2.85 ± 2.75 ± 5.7 ± 3.2 ± 1.6
4. Discussion

4.1. Efficiency of the proposed method

In this section, the running time of the proposed method is compared with traditional k-means++ (Arthur & Vassilvitskii, 2007), Chen et al. (2016) and traditional silhouette (Rousseeuw, 1987). All algorithms are programmed based on Matlab (2018a) and performed on an i7-8750H CPU and a 16GB RAM. Parameters are set as the optimal values determined in Section 3.2. And the running time is selected as the mean time of clustering numbers from 2 to 10. Advantages and disadvantages

The raw point cloud of rock slope

As shown in Table 4, the efficiency of fast k-means++ is obviously higher than the traditional k-means++ and Chen et al. (2016). Specifically when the point number is 1,228,056, the efficiency of the proposed method is 41 times of traditional k-means++ and 59,544 times of and Chen et al. (2016), and the efficiency of fast silhouette is 609,101 times of traditional silhouette. In addition, the efficiency multiple of the proposed method compared with the other three methods is significantly rising with the number of point increases (Fig. 11). That is because, given the point number N, the time complexity of the traditional k-means++ is $O(N)$, Chen et al. (2016) is $O(N^2)$ and the traditional silhouette is $O(N^2)$, which are linear growth (Fig. 11a) or square growth (Fig. 11b-c) with the point number increases. However the proposed method transforms the calculation of N vectors to n base vectors and n is a constant which far smaller than N, which effectively reduces the time complexity of the algorithm and improves the efficiency.

![Figure 11](attachment:image1)

**Figure 11.** Comparison of the proposed method to other methods. (a) Fast k-means++, (b) Fast k-means++, (c) Fast silhouette.

Based on the point cloud of case A1-A2 and case B1-B2, the total running time of the proposed method is 3.3s for 20,781 points, 4.91s for 40,414 points, 94.6s for 619,035 points and 159.22s for 1,228,056 point, which significantly faster than the traditional k-means++, the traditional silhouette and the discontinuity extraction method of Chen et al. (2016).

4.2. Advantages and disadvantages

According to the above sections, the advantages of the proposed method mainly include: (1) high efficiency, compared with Chen et al. (2016), the efficiency enhancement is more obvious with the number of point increases, (2) simplicity, only two main parameters are required and the optimal values are determined, and the optimal values do not need to change for different models until the results are inaccurate or there are specific requirements, and (3) automation, the proposed method does not need manual interference during processing. And the main disadvantage of the proposed method is

| Case of point | Number of point | a. Traditional k-means++ Arthur (2007) | b. Chen et al. (2016) | Clustering time (s) | Running time | Efficiency to a. 100% | Efficiency to b. 100% | c. Traditional silhouette Rousseeuw (1987) | Fast silhouette | Running time | Efficiency to c. 100% |
|---------------|----------------|-------------------------------------|-----------------------|---------------------|--------------|-----------------------|-----------------------|-----------------------------------------------|----------------|--------------|---------------------|
| A-1           | 28701          | 1.68                                | 84.15                 | 0.18                | 9            | 467                   | 80.16                 | 0.23                                                          | 349             |              |
| A-2           | 40414          | 3.39                                | 145.39                | 0.28                | 12           | 519                   | 165.89                | 0.27                                                          | 614             |              |
| B-1           | 619035         | 101.25                              | 76971.05              | 2.86                | 35           | 26913                 | 56122.95              | 0.33                                                          | 170070          |              |
| B-2           | 1228056        | 211.07                              | 300628.37             | 5.20                | 41           | 59344                 | 225367.3              | 0.37                                                          | 609101          |              |
the result of discontinuity grouping is not as accurate as some semi-automatic methods such as Riquelme et al. (2014), which requires to be improved in further studies.

5. Conclusion
This paper proposes a fast automatic extraction method for rock discontinuity orientation based on 3D point cloud. The main contributions include: (1) a concept of base vector is proposed, (2) a fast k-means++ algorithm and a fast clustering validity index of silhouette are proposed and combined to achieve a fast optimal grouping of discontinuities, and (3) the generation and optimization of independent discontinuities based on a double-clustering.

A sensitivity analysis is performed to select the optimal parameters of the proposed method according to the application to both laboratory models and rock models. The optimal base vector is selected to be generated by a 5-time subdivision of icosahedron, and the optimal discontinuity optimization threshold T is 0.6.

The accuracy of the proposed method is validated based on the comparison with traditional methods and other methods. The main advantages include high efficiency, simplicity and automation. And the main disadvantage is the discontinuity grouping is not as accurate as some semi-automatic methods, which requires to be studied in further researches.

In conclusion, the proposed method can automatic achieve a fast and effective discontinuity extraction, which can serves as a supplement to the traditional geological surveys.

References

[1] Abellán A, Oppikofer T, Jaboyedoff M, Rosser J, Lim M, Lato J 2014 J. Terrestrial laser scanning of rock slope instabilities. Earth Surf. Process. Landf 39 80-97
[2] Abellán A, Vilaplana J, Martínez J 2006 J. Application of a long-range Terrestrial Laser Scanner to a detailed rockfall study at Vall de Núria (Eastern Pyrenees, Spain). Engineering Geology 88(3-4) 136-148
[3] Alexa M, Behr J, Cohen-Or D, Fleishman S, Levin D, Silva T, 2003 J. Computing and rendering point set surfaces. Visualization and Computer Graphics, IEEE Transactions 9 (1) 13-151
[4] Arthur D, Vassilvitskii S, 2007 J. K-Means++: The Advantages of Careful Seeding// Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, ACM, 1
[5] Bolshakova N, Azuaje F, 2003 J. Cluster validation techniques for genome expression data. Signal Processing 83 (4) 825-833
[6] Cacciari P, Futain M, 2016 J. Mapping and characterization of rock discontinuities in a tunnel using 3D terrestrial laser scanning. Bulletin of Engineering Geology and the Environment 75 (1) 223-237
[7] Chen J, Zhu H, Li X, 2016 J. Automatic extraction of discontinuity orientation from rock mass surface 3D point cloud. Comput. Geosci 95 18-31
[8] Fernández O, 2005 J. Obtaining a best fitting plane through 3D georeferenced data. Struct. Geol 27 855-858
[9] Ferrero A, Forlani G, Roncella R, Voyat H, 2009 J. Advanced Geostructural Survey Methods Applied to Rock Mass Characterization. rock mechanics & rock engineering 42 (4) 631-665
[10] Ferrero A, Migliazza M, Umili G, 2014 J. Rock mass characterization by means of advanced survey methods// Eurock -the Isrm European Symposium
[11] Fisher E, Shakoor A, Watts F, 2014 J. Comparing discontinuity orientation data collected by terrestrial LiDAR and transit compass methods. Eng. Geol 181 78-92
[12] Gaich A, Pötsch M, Schubert W, 2006 J. Acquisition and assessment of geometric rock mass features by true 3D images, in: American Rock Mechanics Association, editor 50 Years of Rock Mechanics: Landmarks and Future: American Rock Mechanics Association.
[13] Gigli G, Casagli N, 2011 J. Semi-automatic extraction of rock mass structural data from high
resolution LIDAR point clouds. International Journal of Rock Mechanics and Mining Sciences 48 (2) 187-198

[14] Guo J, Liu S, Zhang P, Wu L, Zhou W, Yu Y, 2017 J. Towards semi-automatic rock mass discontinuity orientation and set analysis from 3D point clouds. Computers & Geosciences 103 164-172

[15] Haneberg W, 2008 J. Using close range terrestrial digital photogrammetry for 3-D rock slope modeling and discontinuity mapping in the United States, Bull Eng Geol Environ 4 457-69

[16] Haneberg W, 2008 J. Using close range terrestrial digital photogrammetry for 3-D rock slope modeling and discontinuity mapping in the United States. Bull. Eng. Geol. Environ 67 (4) 457-469

[17] Hartigan J and Wong M, 1979 J. Algorithm as 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. SeriesC (Applied Statistics) 28 (1) 100-108.

[18] Hoek E, Bray J, 1981 J. Rock Slope Engineering. CRC Press.

[19] Huang H, Wu S, Gong M, Cohen-Or D, Ascher U, Zhang H, 2013 J. Edge-aware point set resampling. ACM Trans. Graph.

[20] Hudson A, Harrison P, 2000 J. Engineering Rock Mechanics-An Introduction to the Principles. Elsevier, Burlington.

[21] ISRM, 1978 J. Suggested methods for the quantitative description of discontinuities in rock masses. International Journal of Rock Mechanics & Mining Sciences & Geomechanics Abstracts 15 (6) 319-368

[22] Jaboyedoff M, Oppikofer T, Abellán A, Derron M.-H, Loye A, Metzger R, Pedrazzini A, 2012 J. Use of LIDAR in landslide investigations: a review. Nat. Hazards 61 5-28.

[23] Kemeny J, Post R, 2003 J. Estimating three-dimensional rock discontinuity orientation from digital images of fracture traces. Computers & Geosciences. 29 (1) 65-77.

[24] Kong D, Wu F, 2019 J. Automatic identification and characterization of discontinuities in rock masses from 3D point clouds. Engineering Geology 265

[25] Lato M, Diederichs S, Hutchinson J, Harrap R, 2009 J. Optimization of LiDAR scanning and processing for automated structural evaluation of discontinuities in rock masses. Int. J. Rock. Mech. Min. 46 (1) 194-199.

[26] Lato J, Diederichs S, Hutchinson J, 2010 J. Bias correction for view-limited Lidar scanning of rock outcrops for structural characterization. Rock. Mech. Rock. Eng 43 (5) 615-628

[27] Li X, Chen J, Zhu H, 2016 J. A new method for automated discontinuity trace mapping on 3D rock mass surface model. Computers & Geosciences 89 118-131

[28] Likas A, Vlassis N, Verbeek J, 2003 J. The global k-means clustering algorithm. Pattern Recognition: The Journal of the Pattern Recognition Society 36 (2) 451-461

[29] MVTEc Software GmbH, 2012. Halcon Version 11 [HDevelop Reference Manual]. Retrieved from (http://www.mvtec.com/halcon/)

[30] Otoo J, Maerz N, Duan Y, Liu X, 2011 J. LiDAR and optical imaging for 3-D fracture orientations. In: Proceedings of 2011 NSF Engineering Research and Innovation Conference, Atlanta, Georgia.

[31] Peña J, Lozano J, Larrañaga P, 1999 J. An empirical comparison of four initialization methods for the k-means algorithm. Elsevier Science Inc.

[32] Priest S, 1993 J. Discontinuity Analysis for Rock Engineering. Chapman & Hall

[33] Qiu C, Xiao J, Han L, Muhammad N, 2014 J. Enhanced interval type-2 fuzzy c-means algorithm with improved initial centroid. Pattern Recognition Letters 38 86-92

[34] Rao A, Srinivas V, 2006 J. Regionalization of watersheds by hybrid-cluster analysis. Journal of Hydrology (Amsterdam) 318 (1-4) 0-56

[35] Riquelme A, Abellán A, Tomás R, 2015 J. Discontinuity spacing analysis in rock masses using 3D point clouds. Eng. Geol. 195 185-195

[36] Riquelme A, Abellán A, Tomás R, Jaboyedoff M, 2014 J. A new approach for semi-automatic rock mass joints recognition from 3D point clouds. Comput. Geosci. 68 38-52
[37] Riquelme A, Cano M, Tomáš R, Abellan A, 2017 J. Identification of Rock Slope Discontinuity Sets from Laser Scanner and Photogrammetric Point Clouds: A Comparative Analysis. Procedia Engineering 191 838-845

[38] Rodriguez A, Laio A, 2014 J. Clustering by fast search and find of density peaks. Science 344 (6191) 1492-1496

[39] Rousseeuw P, 1987 J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math 20 53-65

[40] Slob S, 2010 J. Automated Rock Mass Characterisation Using 3-D Terrestrial Laser Scanning (Ph.D. dissertation). Delft University of Technology, Delft, Netherlands.

[41] Slob S, Knapen B, Hack R, Turner K, Kemeny J, 2005 J. Method for automated discontinuity analysis of rock slopes with three-dimensional laserscanning. Transp. Res. Rec.: J. Transport. Res. Board 187-194.

[42] Sturzenegger M, Stead D 2013 J. Close-range terrestrial digital photogrammetry and terrestrial laser scanning for discontinuity characterization on rock cuts. Engineering Geology 106(3-4) 163-182

[43] Vöge M, Lato J, Diederichs S, 2013 J. Automated rock mass discontinuity mapping from 3-dimensional surface data. Eng. Geol 164 155-162

[44] Xhafa F, Bogza A, Santi Caballé, Leonard B, 2016 J. Apache Mahout’s k−Means vs. Fuzzy k−Means Performance Evaluation// International Conference on Intelligent Networking & Collaborative Systems. IEEE.

[45] Zhu H, Wu W, Chen J, 2016 J. Integration of three dimensional discontinuous deformation analysis (DDA) with binocular photogrammetry for stability analysis of tunnels in blocky rock mass. Tunnelling & Underground Space Technology Incorporating Trenchless Technology Research 51 30-40

Acknowledgement
This work was supported by the Special Fund for Basic Research on Scientific Instruments of the National Natural Science Foundation of China [4182780021]; National Natural Science Foundation of China [41902275]; National Natural Science Foundation of China [41130751]; Science and Technology Projects for Western Transportation Construction [2011ZB04]; Fundamental Research Funds for the Central Universities [22120180006]; Shanghai Sailing Program [18YF1424400]; Emeishan-hanyuan Highway Program by Sichuan Lehan Expressway Co., Ltd.[LH-HT-45]; Taihang Mountain Highway Program by China Power Construction Hebei Transportation Expressway investment Development Co., Ltd. [kh0023020184285]; Guizhou Reverence High-speed Double-track Program by Guizhou Highway Engineering Group Co., LTD. [[2018]3011].