Computer Vision Project Report
M.Sc. in Data Science 2^{nd} Year, 2022
Chennai Mathematical Institute, Chennai

Structure From Motion and it's Application

by

Taher Azim	(MDS202043)
Sumeet Suley	(MDS202042)
Arnab Sen	(MDS202009)

Supervisor: Dr. Kavita Sutar [1]

[1] Lecturer, Chennai Mathematical Institute, ksutar@cmi.ac.in

Contents

Abstract

In the First Section

The main advantage of using Unmanned Aerial Vehicles (UAVs) is the relatively low cost of collecting data, especially when using photogrammetry on images of relatively small areas. Additionally, they have high operational flexibility and the results have a high spatial and temporal resolution. To further facilitate the use of UAVs in photogrammetry, we developed an algorithm to filter out points that indicate areas covered in low vegetation (grass, crops) from the generated point cloud. This paper presents a three-layer filtering algorithm based on convolutional neural networks (CNNs) created for this specific purpose. The modular structure of the algorithm makes it easy to expand on and improve. The proposed solution allows errors in the height of digital elevation model (DEM) points caused by the influence of vegetation to be reduced by as much as 60–70% in relation to height errors from the raw data of high grass. At the same time, the solution presented here is practical for low grass because it does not weaken the model. The algorithm significantly reduces the errors in the DEM, as well as the products derived from the DEM.

In the Second Section

Comprehensive data collection remains a challenge in the field of sediment research. The manual acquisition of fine-gridded data is almost infeasible even for a laboratory setup. Therefore, this paper demonstrates a simple and cost-effective SfM–MVS technique to acquire accurate morphological data. This data further can be used for assessing the scour development around a bridge pier. For this purpose, the experiments are conducted for clear-water scour around circular and hexagonal piers for three different discharges. The high-spatial resolution digital elevation models (DEMs) are generated using the SfM–MVS photogrammetry technique. A statistical analysis is performed between the checkpoints (observed data) and DEM predicted points, which revealed that the generated DEMs show high accuracy in all the cases. It is therefore concluded that the SfM–MVS technique can be applied to understand the morghological changes around any shape of the piers. Thus, the proposed image analysis method can be adopted for obtaining the high spatial resolution data for sediment transport research.

1 SFM Algorithm

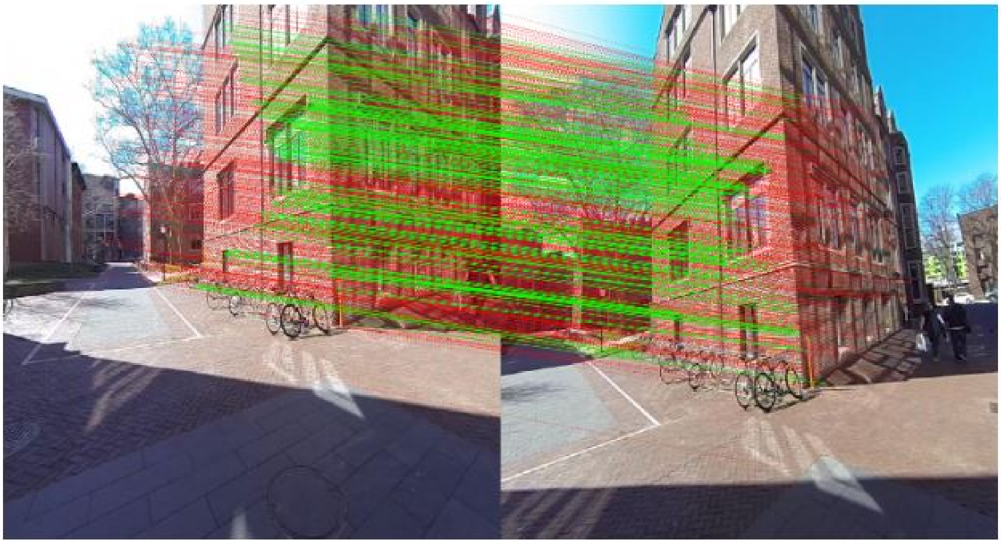
1.1 Structure From Motion Pipeline:

- Feature Matching and Outlier rejection using RANSAC - Estimating Fundamental Matrix - Estimating Essential Matrix from Fundamental Matrix - Estimate Camera Pose from Essential Matrix - Check for Cheirality Condition using Triangulation - Bundle Adjustment

- **Feature Matching and Outlier rejection using RANSAC:**

  Feature Matching are either done using Template matching or standard algorithm such as SIFT. The Template Matching method is mostly static and the process works only if the template/subset image is exactly contained in the full/target image. Slight deviations like the direction change, image intensity change, scale changes will not go well with Template Matching and results in very poor results. This makes the Template Matching method of object detection less usable and doesn't make to real-world applications.

  To overcome the above pitfalls of the Template Matching methods, SIFT (Scale Invariant Feature Transform) can be used. SIFT algorithm addresses the problems of feature matching with changing scale, intensity, and rotation. This makes this process more dynamic and the template image doesn't need to be exactly contained in the full/main image. This is considered one of the best approaches for feature matching and is widely used. Below example shows SIFT applied on image:

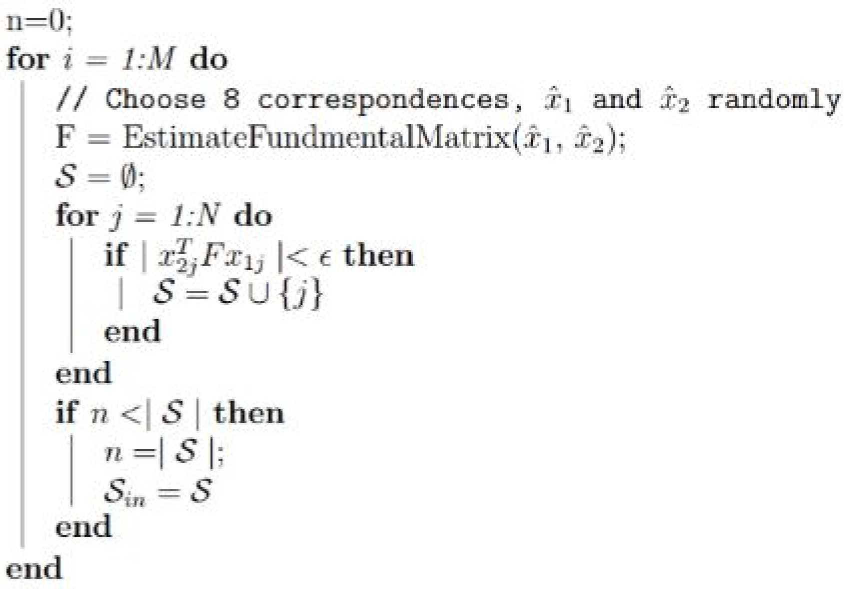


1.2 Outlier rejection using RANSAC:

Having done the feature matching we know that some features won't be accurate.To disregard such outliers we need to to accurately select correspondenses for estimating fundamental matrix such that maximum number of points satisfy the epipolar constraint within the prescrised error bound given below:

$$\begin{bmatrix} x'_i & y'_i & 1 \end{bmatrix} \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = 0$$

Below is the RANSAC algorithm for the SFM:



```
n=0;
for $i$ = $1{:}M$ do
    // Choose 8 correspondences, $\hat{x}_1$ and $\hat{x}_2$ randomly
    F = EstimateFundmentalMatrix($\hat{x}_1$, $\hat{x}_2$);
    $\mathcal{S} = \emptyset$;
    for $j$ = $1{:}N$ do
        if $| x_{2j}^T F x_{1j} | < \epsilon$ then
            $\mathcal{S} = \mathcal{S} \cup \{j\}$
        end
    end
    if $n < | \mathcal{S} |$ then
        $n = | \mathcal{S} |$;
        $\mathcal{S}_{in} = \mathcal{S}$
    end
end
```

1.3 Estimating Fundamental Matrix:

The **F** matrix is only an algebraic representation of epipolar geometry and can both geometrically (contruct. the epipolar line) and arithematically. As a result, we obta $\mathbf{x}_i'\mathbf{TFx}_i = 0$ where $i = 1, 2, \ldots, m$. This is known as epipolar constraint or correspondance condition Longuet-Higgins equation). Since, **F** is a 3×3 matrix, we can set up a homogenrous linear system with unknowns:

$$\begin{bmatrix} x'_i & y'_i & 1 \end{bmatrix} \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = 0$$

$$x_i x'_i f_{11} + x_i y'_i f_{21} + x_i f_{31} + y_i x'_i f_{12} + y_i y'_i f_{22} + y_i f_{32} + x'_i f_{13} + y'_i f_{23} + f_{33} = 0$$

Simplifying for m correspondences,

$$\begin{bmatrix} x_1 x'_1 & x_1 y'_1 & x_1 & y_1 x'_1 & y_1 y'_1 & y_1 & x'_1 & y'_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_m x'_m & x_m y'_m & x_m & y_m x'_m & y_m y'_m & y_m & x'_m & y'_m & 1 \end{bmatrix} \begin{bmatrix} f_{11} \\ f_{21} \\ f_{31} \\ f_{12} \\ f_{22} \\ f_{32} \\ f_{13} \\ f_{23} \\ f_{33} \end{bmatrix} = 0$$

How many points do we need to solve the above equation? Think! Twice! Remember homography, where each point correspondence contributes two constraints? Unlike homography, in **F** matrix estimation, each point only contributes one constraints as the epipolar constraint is a scalar equation. Thus, we require at least 8 points to solve the above homogenous system. That is why it is known as Eight-point algorithm.

With $N \geq 8$ correspondences between two images, the fundamental matrix, F can be obtained as: By stacking the above equation in a matrix A, the equation $Ax = 0$ is obtained. This system of equation can be answered by solving the linear least squares using Singular Value Decomposition (SVD) as explained in the Math module. When applying SVD to matrix **A**, the decomposition USV $\mathbf{T^T}$ would be obtained with **U** and **V** orthonormal matrices and a diagonal matrix **S** that contains the singular values. The singular values σ_i where ^{2} $i \in [1, 9], i \in \mathbb{Z}$, are positive and are in decreasing order with $\sigma_9 = 0$ since we have 8 equations for 9 unknowns. Thus, the last column of **V** is the true solution given that $\sigma_i \neq 0 \forall i \in [1, 8], i \in \mathbb{Z}$. However, due to noise in the correspondences, the estimated **F** matrix can be of rank 3 i.e. $\sigma_9 \neq 0$. So, to enfore the rank 2 constraint, the last singular value of the estimated **F** must be set to zero. If F has a full rank then it will have an empty null-space i.e. it won't have any point that is on entire set of lines. Thus, there wouldn't be any epipoles.

1.4 Estimating Essential Matrix from Fundamental Matrix:

Since we have computed the **F** using epipolar constrains, we can find the relative camera poses between the two images. This can be computed using the Essential Matrix, **E**. Essential matrix is another 3×3 matrix, but with some additional properties that relates the corresponding points assuming that the cameras obeys the pinhole model (unlike **F**). More specifically, $\mathbf{E} = \mathbf{K^T F K}$ where **K** is the camera calibration matrix or camera intrinsic matrix. Clearly, the essential matrix can be extracted from **F** and **K**. As in the case of **F** matrix computation, the singular values of **E** are not necessarily $(1, 1, 0)$ due to the noise in **K**. This can be corrected by reconstructing it with $(1, 1, 0)$ singular values, i.e. $\mathbf{E} = U \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} V^T$

It is important to note that the **F** is defined in the original image space (i.e. pixel coordinates) whereas **E** is in the normalized image coordinates. Normalized image coordinates have the origin at the optical center of the image. Also, relative camera poses between two views can be computed using **E** matrix. Moreover, **F** has 7 degrees of freedom while **E** has 5 as it takes camera parameters in account.

1.5 Estimate Camera Pose from Essential Matrix:

The camera pose consists of 6 degrees-of-freedom (DOF) Rotation (Roll, Pitch, Yaw) and Translation (X, Y, Z) of the camera with respect to the world. Since the **E** matrix is identified, the four camera pose configurations: $(C_1, R_1), (C_2, R_2), (C_3, R_3)$ and $(C_4, R4)$ where $C \in \mathbb{R}^3$ is the camera center and $R \in SO(3)$ is the rotation matrix, can be computed. Thus, the camera pose can be written as: $P = KR \begin{bmatrix} I_{3\times3} & -C \end{bmatrix}$ These four pose configurations can be computed from **E** matrix.

Let $\mathbf{E} = UDV^T$ and $W = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

The four configurations can be written as:

1. $C_1 = U(:,3)$ and $R_1 = UWV^T$
2. $C_2 = -U(:,3)$ and $R_2 = UWV^T$
3. $C_3 = U(:,3)$ and $R_3 = UW^TV^T$
4. $C_4 = -U(:,3)$ and $R_4 = UW^TV^T$

It is important to note that the $\det(R) = 1$. If $\det(R) = -1$, the camera pose must be corrected i.e. $C = -C$ and $R = -R$.

The visual representation of 4 different possible camera poses have been shown below:

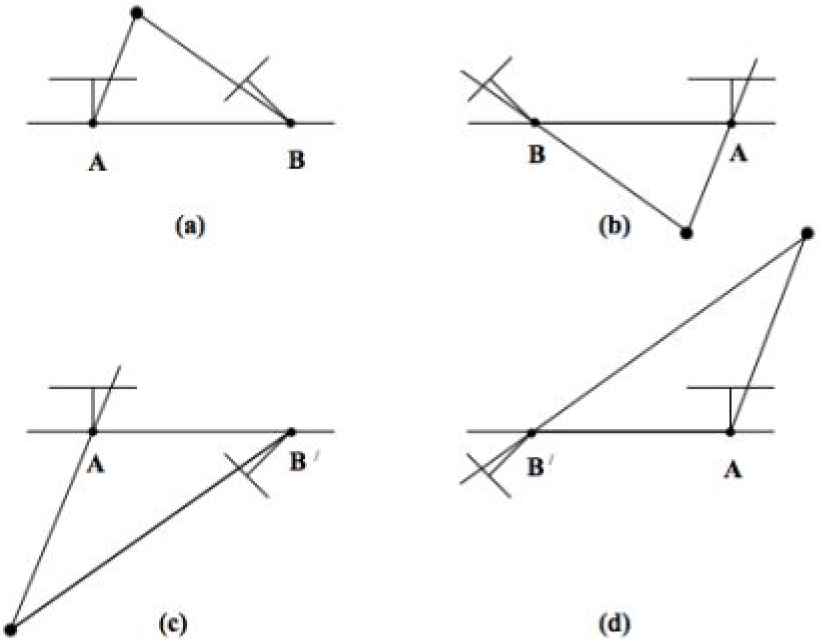


Fig a is the correct camera poses since point lies in front of camera.

1.6 Check for Cheirality Condition using Triangulation:

In the previous section, we computed four different possible camera poses for a pair of images using essential matrix. Though, in order to find the correct unique camera pose, we need to remove the disambiguity. This can be accomplish by checking the cheirality condition i.e. the reconstructed points must be in front of the cameras. To check the cheirality condition, triangulate the 3D points (given two camera poses) using linear least squares to check the sign of the depth Z in the camera coordinate system w.r.t. camera center.

1.7 Linear Least Square to estimate the 3D coordinate given poses:

Since we have the camera poses estimated we can construct $\mathbf{P}$ and $\mathbf{P}'$ for the two cameras.

$$\mathbf{x} = \mathbf{P}\mathbf{X}$$

$$\mathbf{x}' = \mathbf{P}'\mathbf{X}$$

$$\mathbf{A}\mathbf{X} = \mathbf{0} \quad \mathbf{A} = \begin{bmatrix} u_3^T - \mathbf{p}_1^T \\ \mathbf{v}\mathbf{p}_3^T - \mathbf{p}_2^T \\ u'\mathbf{p}_3'^T - \mathbf{p}_1'^T \\ v'\mathbf{p}_3'^T - \mathbf{p}_2'^T \end{bmatrix}$$

A 3D point X is estimated by solving the above equation by SVD via finding Least Square solution to above equation.

A 3D point X is in front of the camera iff: $r_3(\mathbf{X} - \mathbf{C}) > 0$ where r_3 is the third row of the rotation matrix (z-axis of the camera). Not all triangulated points satisfy this conidition due of the presence of correspondence noise. The best camera configuration, (C, R, X) is the one that produces the maximum number of points satisfying the cheirality condition.

1.8 Bundle Adjustment:

Once you have computed all the camera poses and 3D points, we need to refine the poses and 3D points together, initialized by previous reconstruction by minimizing reprojection error.

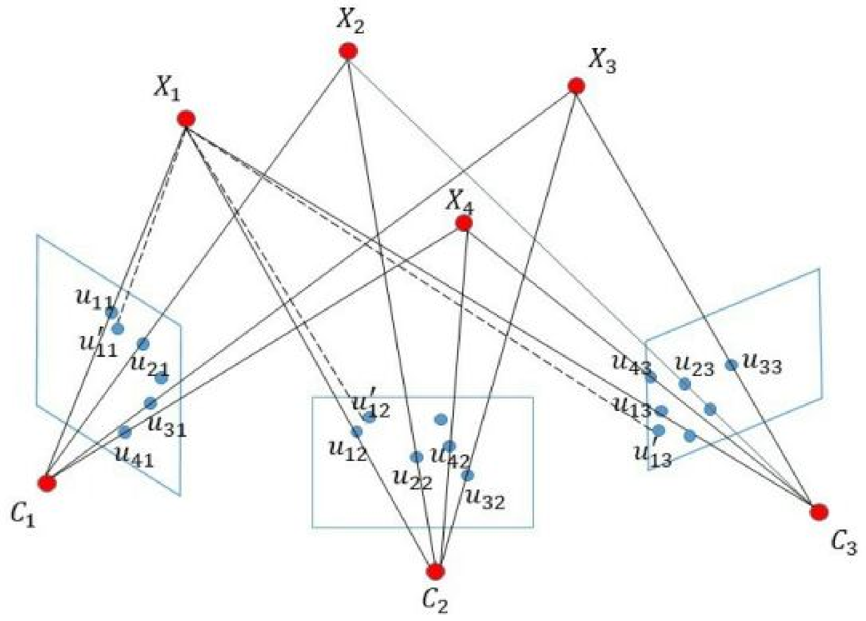


Figure 5. Bundle Adjustment: u_{ij} is the observations, X_i is the 3D points in world frame, C_j is the camera center, u'_{ij} is the re-projected 2D points. The solid line represents the projection procedure, the dotted line represents the re-projection procedure.

It tries to minimize the sum of errors between 2D observations and the predicted 2D points, where the predicted points are re-projected from 3D structures by camera parameters. It measures the accuracy of the computed 3D structures and camera parameters. As shown in Figure, bundle adjustment is actually a nonlinear least square problem, which described by the following equation:

$$\min \sum_{i=1}^{n} \sum_{j=1}^{m} \left(u_{ij} - \pi\left(C_j, X_i\right)\right)^2$$

where u_{ij} is the observed point coordinate in pixel level, which represents the i th 3D point X_i is observed by the j th camera $C_j \cdot \pi\left(C_j, X_i\right)$ is the nonlinear operation. The following non-linear optimization can be solved using Gauss-Newton Merthod.

Demonstration of structure-from-motion (SfM) and multi-view stereo (MVS) close range photogrammetry technique for scour hole analysis

1 Introduction

The flow of water in a river erodes the materials from the riverbed, banks, around the structures and transports them to a certain distance until it gets deposited. As a consequence, foundation of a structure is caved in by the erosive force of flowing water resulting in the formation of scour. During high flood events, if the dynamics of exchange between water and structure is not evaluated accurately while designing the foundation, the excessive scour can give rise to the destruction of the structure. The local scour, which is a decrease in bed level in the vicinity of an obstruction due to the flow-structure interaction, is found to be one of the prominent components that can cause extensive destruction. Therefore, numerous studies have focused on local scour estimation around vertical piles or bridge piers in steady and unsteady currents. The pier geometry is also found to be one of the prominent causes for aggravating the scour evolution phenomenon. This study carried out using manual data collection methods that employed traditional instruments such as point gauges or sensors. In case of scour analysis, the main concern is the identification of the maximum scour depth and three-dimensional (3D) description of scour hole due to limited measured data by manual methods. The data is commonly collected at some selected points. A set of experiments are conducted in a laboratory flume to demonstrate the applicability of automated Close Range Photogrammetry (CRP) for scour hole analysis. Two different piers, namely, circular and hexagonal are chosen. 3 circular and 3 hexagonal piers.

2 Methods

In this section we talk about how to conduct the procedure, work flow and cautions.

2.1 Setup

- Two different pier geometries are chosen, namely, circular (usual) and hexagonal (streamlined pier).
- The circular and the hexagonal piers are firmly secured below the sand bed level by 25 cm and 15 cm respectively preceding to the experimental runs.
- The bed materials for the study are selected with a median particle size (d_50) of 0.68 mm. The geometric standard deviation (σ_g) of sand is 1.15, which implies that the sand is uniform.
- It is observed that the fully developed flow is attained at approximately 4.7 m away from the inlet of the flume as the vertical velocity distribution is logarithmic.
- During the experiments, the average approaching flow velocity (U) is taken as approximately 90% of the critical velocity (U_C) of the uniform sediment bed to fulfil the nearly limiting clear-water condition. The semi-logarithmic empirical equation proposed by Dey and Raikar is used to compute the critical velocity (U_C). This equation dependents on the approaching flow depth (h) and is given by

$$(\frac{U_C}{U_{*C}}) = [5.75 * log_{10}(\frac{h}{2d_{50}})] + 6$$

- The pier is then located at approximately 5.75 m from the inlet. The centre of the pier is taken as (0, 0) as the local co-ordinate for the measurement of scour hole.

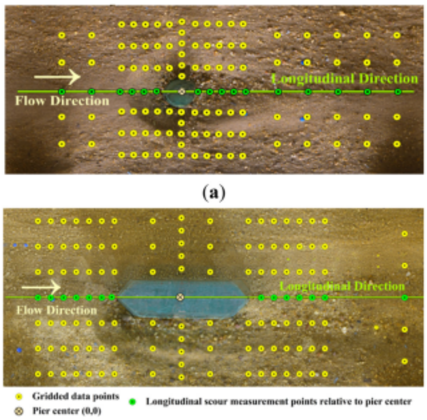
2.2 Experiment Procedure

1. The respective pier model is embedded in the flume and the sediment bed surface in the flume is levelled to 100 cm from the bottom reference with the help of a point gauge.
2. The thickness of laid sediment bed is 25 cm and uniform throughout the flume.
3. Three different discharges are used to study erosion around each geometrical pier. The water is slowly released into the flume.
4. The required water depth and surface slope are achieved by controlling the tail-gate.
5. Each experiment is run for three hour duration for scour hole development.
6. The process was observed to be close to the equilibrium condition. It take upto 3 hours.
7. The flow depth in the flume is maintained to 10 cm in all the cases by adjusting the tailgate.
8. The water in the flume is allowed to drain off by carefully operating the tail gate after complete experimental run.
9. The flume is then allowed to dry up so that the bed is clearly visible for data acquisition.

Point to be noted that, The runs were shortened because the purpose of this study is only to demonstrate the capability of the SfM–MVS technique to map scour hole in 3D due to sediment transport and not the scour phenomenon.

2.3 Data acquisition and ground truth data

- The distribution of grid points in the figure are established based on the dimension of scour hole.
- The profile of maximum scour depth is observed at the discrete points along the longitudinal direction marked by the green colour line and points.

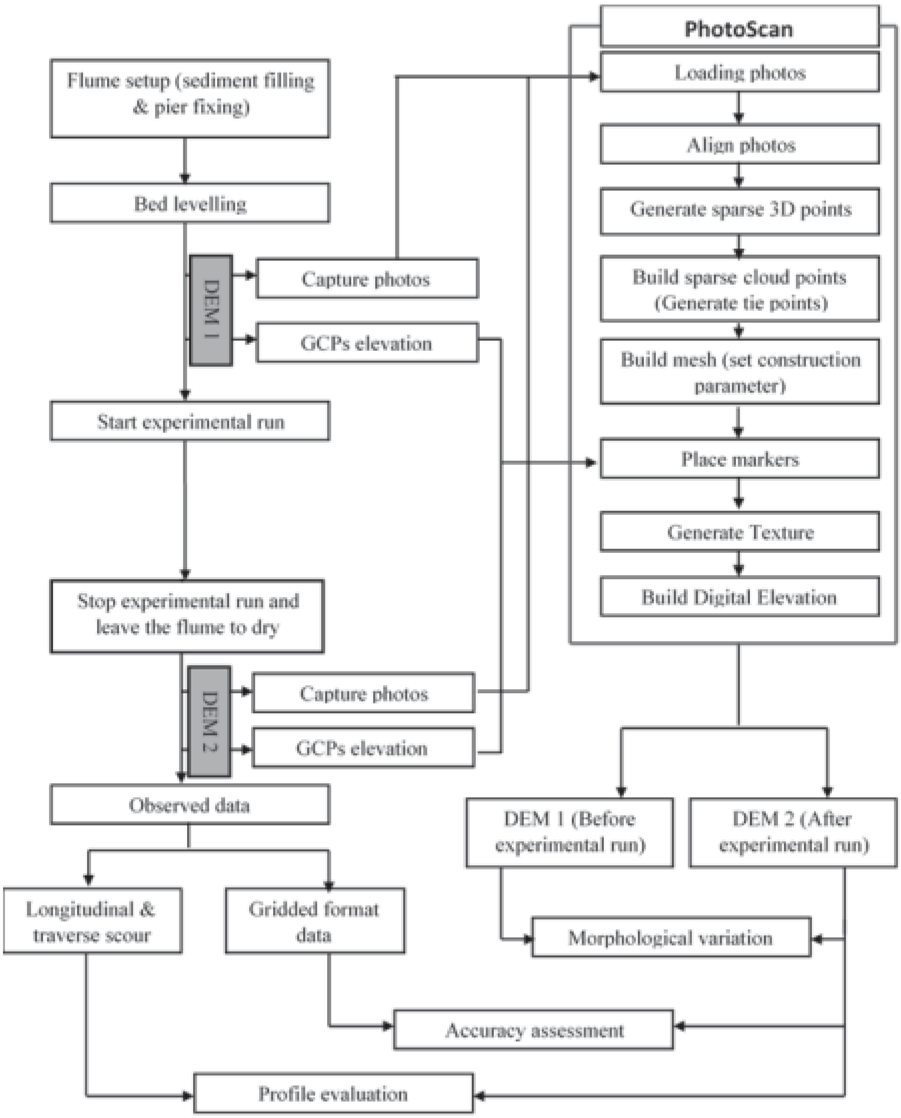


(a)

(b)

- These scour depth data are collected by point gauge at the gridded points. A set of photos are captured before and after each experimental run using a DSLR camera for the purpose of the automated CRP analysis.
- The camera is mounted vertically downwards from a hanging system at a distance of 1.2 m above the flume surface. The camera is operated in manual mode.
- The blurred images are discarded from the original set of photographs. A total of 25-70 (depending on the size of scour hole) images are required for each run.
- It is to be noted that each and every ground feature that needs to be reconstructed must be covered by at least three images taken from different positions and orientations, and preferably more in number to enhance the reconstruction accuracy.
- The SfM technique is less sensitive to colour and resolution of the images, and hence position and orientation of the camera is not a matter of concern.

2.4 Flow Chart


PhotoScan
Flume setup (sediment filling & pier fixing)
Loading photos
Align photos
Bed levelling
Generate sparse 3D points
DEM 1
Capture photos
Build sparse cloud points (Generate tie points)
GCPs elevation
Build mesh (set construction parameter)
Start experimental run
Place markers
Generate Texture
Stop experimental run and leave the flume to dry
Build Digital Elevation
DEM 2
Capture photos
GCPs elevation
Observed data
DEM 1 (Before experimental run)
DEM 2 (After experimental run)
Longitudinal & traverse scour
Gridded format data
Morphological variation
Accuracy assessment
Profile evaluation

2.5 Cautions

- Blurred images can be a cause of mismatching the conjugated points, so they need to be removed from the set of images before the processing. Significant overlaps among the images are needed to be maintained.

- The degree of orientation from which images are taken should be maintained up to 10° for better accuracy.

- Convergent images can be used along with parallel images in order to mitigate the non-linear dome effect in the DEM.

2.6 SFM and DEM

The automated CRP available in the commercial Agisoft PhotoScan software is used to generate high spatial resolution bed surface topography.

1. The PhotoScan uses SfM principle and also takes the advantages of the latest multi-view 3D technology.

2. SfM component detects feature matching common points in all the photographs, which are stable under viewpoints and match them. Then it generates a descriptor for each point from its local neighbourhood to detect more correspondences across the photos based on the SIFT algorithm.

3. Then SfM computes the locations of the feature points in a relative coordinate frame by creating sparse 3D points.

4. The custom algorithm is used to identify the conjugated features between the images and create the 3D sparse point clouds.

5. A greedy algorithm is then used to find approximate camera position, orientation and altitude with respect to 3D sparse point clouds.

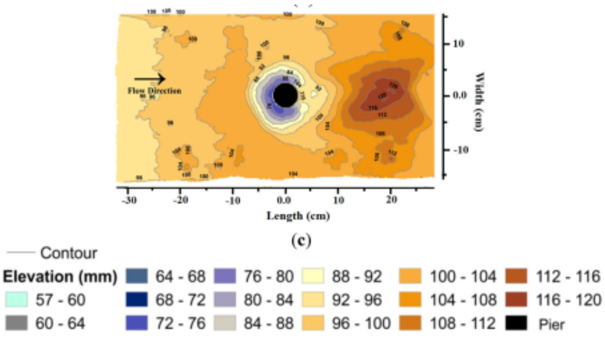
6. 3D Sparse Point Clouds $\hookrightarrow$3D Dense Point Clouds $\hookrightarrow$ 3D Surface

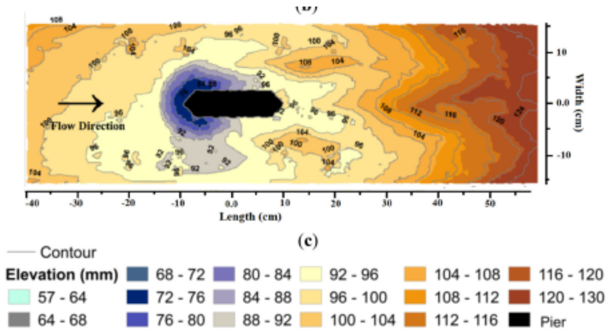
7. The texture map is generated after the reconstruction of the 3D surface. As a result, the DEM is generated which can be exported by specifying the required spatial resolution parameters.

NB: Since the Photoscan software is expensive, in order to minimize the cost, a set of freely downloadable packages can be used systematically to execute different steps of the SfM–MVS technique for generating high resolution DEM.

3 Results Accuracy Assessment :

3.1 Result

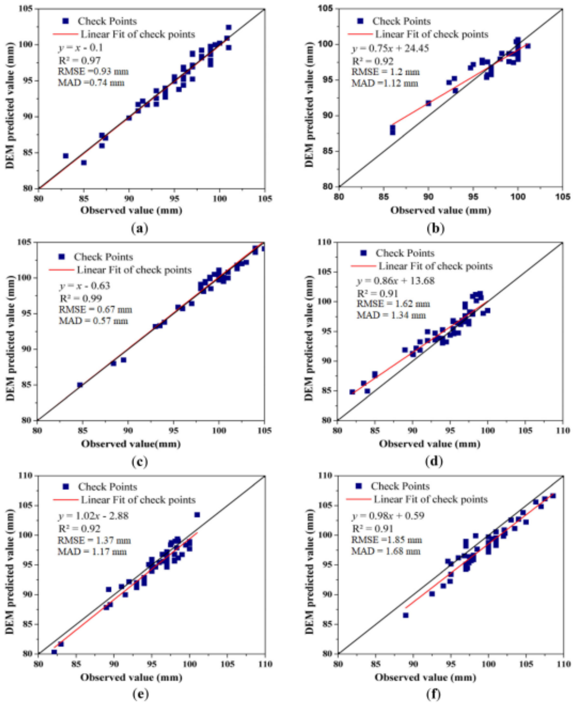




As the figures are showing these are two types of piers circular (left) and hexagonal (right one), the scours are blacks. The more we move from the center the the place became more elevated.

3.2 Accuracy Assessment

- Using RMSE(Root Mean Square Error) and MAD(Mean Absolute Deviation) for quantitative accuracy between predicted elevation in DEM and observed data.
- The error analysis proves that the predicted DEM elevation for each run is adequate to represent complex bed surface morphology.
- The accuracy could be better if the ultrasonic sensors had high precision.
- The accuracy obtained in the DEMs is still to be highly accurate and reliable for morphological analysis.



- **R2:** The slope of the best fit regression for H2 shows slightly greater than 1 and the y-intercept is negative as shown in figure. The elevation values are over predicted in the case of H1 while underpredicted in H2 and H3 cases. R2 estimations show a highly accurate positive linear relationship between the computed DEM elevation and the measured elevation values.

4 Conclusion

This paper demonstrates the ability of SfM–MVS CRP technique to reconstruct high-resolution bed morphology and its changes during the scour process due to the introduction of the geometrical piers. Following conclusions can be derived from the study.

- A set of experiments are conducted in a laboratory flume to demonstrate the applicability of automated CRP for scour hole analysis. Two different piers, namely, circular and hexagonal are chosen. Therefore it is clearly evident that the automated CRP can be applied to any shape of piers.
- The accuracy achieved in the DEMs is in the scale of mm to sub-mm. Therefore, it can be concluded that the SfM technique can be used to collect high resolution data for studying morphological changes.
- The SfM technique is easy to implement, cost effective and efficient than the classical photogrammetric technique.

Therefore, it can be concluded that the automated SfM CRP technique is an effective alternative for collecting simultaneous and high spatial resolution topographic data for morphological analysis, even for a laboratory setup.

Application of convolutional neural networks for low vegetation filtering from data acquired by UAVs

1 Introduction

In recent years, the use of unmanned aerial vehicles (UAVs) to acquire high resolution spatial data has been the subject of many studies. A DEM represents the actual topography of the earth surface. Point cloud filtration to remove non-ground points is a critical and difficult step when generating the DEM, especially for areas with a very diverse topography. There have been many algorithms developed to filter point clouds. These algorithms can be divided into several categories: interpolation-based, slope-based, segmentation-based, and morphological methods. Over the last year, there were individual attempts to use CNNs to filter and classify the data from UAVs, including DEM extraction. For example, Gevaert et al. (2018) proposed a two-stage criterion to obtain training examples. The first stage uses simple morphological filters to pre-classify the points as ground and non-ground points. The second stage selects information on the geometry of the objects from the first stage and uses radiometric data from the photographs. The combination of these steps allows the selection of appropriate training examples for the CNN.

2 Materials and Methodology

The filtration method proposed here uses a CNN to analyse images generated based on the height of points as determined using photogrammetry from a UAV. In the following subsections, the methodology used to collect measurement data (UAV, reference) as well as the data processing used to filter out the points that reflect vegetation are described.

2.1 Datasets

The study used datasets acquired at two different locations in the south of Poland. Data acquired at the village of Łaziska were used to train and validate the network, while data from the village of Jerzmanowice were used for testing (filtering and evaluation using the trained neural networks). In both cases, the measurements included a UAV dataset and reference measurements for the height of points located on the ground mainly using a total station. At the same time, during the measurements, objects were divided into regions with the same type and condition of plant coverage. The initial data processing focused on creating dense point clouds and determining coordinates for the reference points in a homogeneous coordinate system.

2.1.1 Details of the Łaziska dataset

The total area of the Łaziska dataset was approximately 1.4 ha, 40of which was used as the training set, with the rest used as the validation set. The entire research area in Łaziska was measured using two methods: a tacheometric method as the reference and a photogrammetric method. All observations were connected to four points in the control network, whose coordinates were determined to within an accuracy of 2–3 mm.

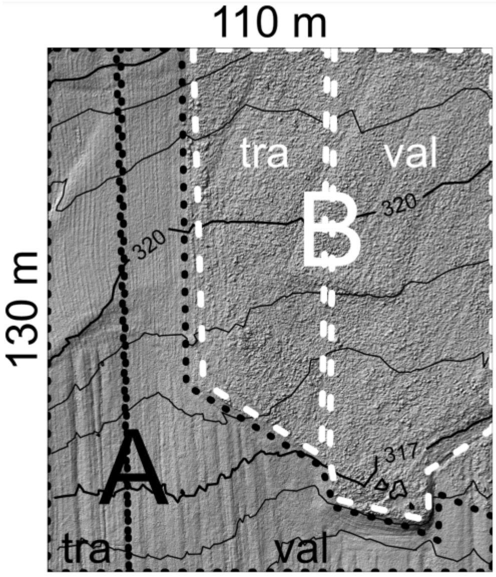


Figure 1: Division of the Łaziska dataset: A - low grass (mowed), B - high grass (about 60 cm tall), tra - data used for network training, val - data used for training validation.

2.1.2 Details of the Jerzmanowice dataset

The total area of the Jerzmanowice dataset used to assess the filtering accuracy was approximately 2.7 ha. The corresponding figure presents the division of the area into regions of groundcover with a uniform type and height. The analysis excluded all trees in the measurement area Similar to the Łaziska dataset, the research area in Jerzmanowice was characterized using reference measurements and photogrammetry with a UAV. The vast majority of the reference data regarding the terrain were collected using the tacheometric method, and only small parts of the areas covered by high grass were measured with the Global Navigation Satellite Systems real time network (GNSS RTN) technique.

2.2 Data Processing

The main part of the processing was performed using the CNN. The data processing was divided into the following three stages:

1. Classification of the point cloud into the high grass (HG) or low grass (LG) classes. Fragments of cloud considered to belong to the LG class were also considered as belonging to the terrain (Ground) class.

2. Classification of point cloud fragments that represented the HG class for points reflecting the Ground class and the remaining points (NonGround class). Points found to belong to the Ground class were retained, while those belonging to the NonGround class were discarded from the point cloud.

3. The determination (regression) and corrections of heights of the points in the cloud for the cloud fragments that were classified simultaneously as the HG class and Ground class.

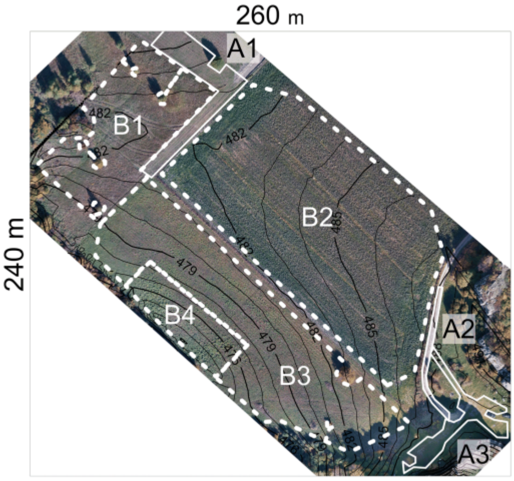


Figure 2: Division of the Jerzmanowice dataset into regions of groundcover with a uniform type and height: A1 - mowed grass, A2 - paved road, A3 - mowed and trodden grass, B1 - wild meadow (grass height of approximately 40–80 cm), B2 - field covered with young broad bean (height of approximately 60 cm), B3 - meadow (grass height of approximately 10–30 cm), B4 - meadow (grass height of approximately 30–50 cm).

The use of neural networks for data processing requires training to determine the network weights. The CNN training requires both input data and target responses in a process known as supervised learning. The input data for all networks were prepared in the same way as described in the followig sections describe the preparation of the training data (target responses) for the networks used in the individual processing stages as well as the structure of the neural networks.

2.2.1 Preparation Of Input Data

Neural network input data should have a clearly defined structure. All input variables, both in the training stage and the subsequent use of a trained network, should have values from a specific closed set or interval. These intervals should be defined at the training stage and should be the same when using trained networks. The input data were presented as fragments of a greyscale image, with an 8-bit depth. The images were created as follows:

- The entire point cloud was divided into adjacent cells (squares), with dimensions of 5 × 5 cm. In each cell, the point with the lowest height was selected and attributed as the height of the cell, while the remaining points were discarded. The point cloud processed in this way is almost uniformly dense, except for the single cells that contain no points. Cell values were transcribed into the matrix labelled R.
- The missing entries in the R matrix were filled using a nearest neighbour interpolation method.
- A G matrix was created by applying a Gaussian filter to the R matrix. The square, Gaussian filter had a size of 105 × 105 cm (21 × 21 elements), with a standard deviation of 25 cm (5 elements).
- An M matrix was created as the difference between the R and G matrices: M = R - G

The network input images were 65 × 65 pixels. For this reason, the actual images recorded on the disk had to be a larger size, i.e. (65 × 65) * 20.5. As a result, during training when rotating the image by some angle, each network input had a specific value. After rotation, the images were

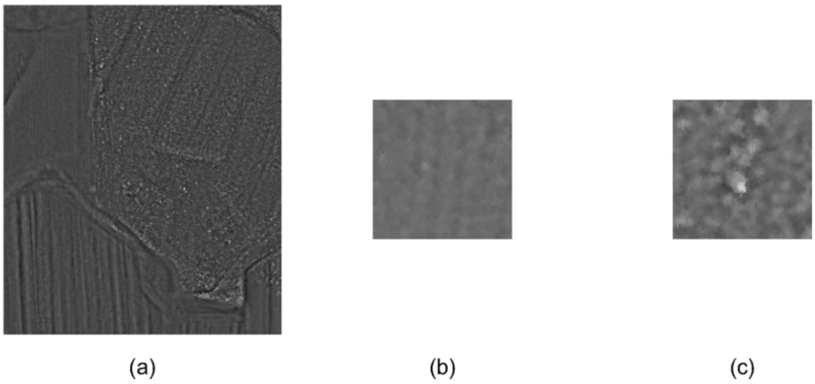


Figure 3: . Full image I for the Łaziska dataset (a) and its exemplary fragments used as the inputs to neural networks from the regions covered by low (b) and high (c) grasses.

cropped at their centre to fit in a window of 65 × 65 pixels. For all the CNNs, the pixel considered was located in the middle of a given example. This pixel corresponded to the values of the network outputs. Data from the Łaziska site were used for training and validation in all cases, and the data from the Jerzmanowice site were used to evaluate the trained network.

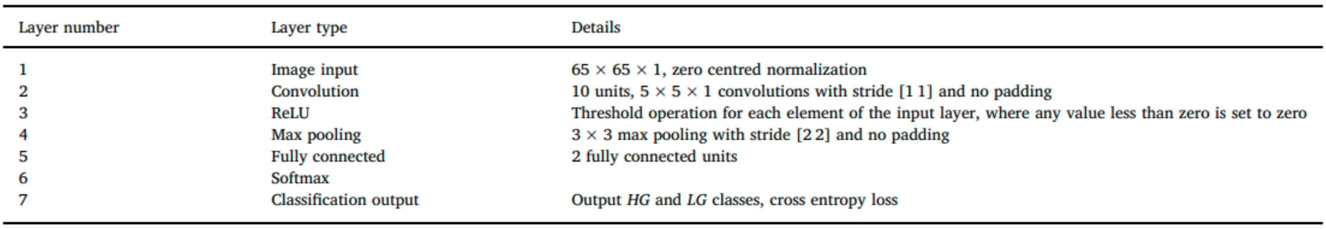
2.2.2 Classification into points with high and low grass

The targets for the HG and LG classes were assigned based on field observations. The structure of the CNN used to implement this stage of processing is summarized in the corresponding table. A standard network structure was used. It was chosen to maintain simplicity. Because the processing speed was not a key criterion, and the effectiveness of the classification turned out to be completely satisfactory, the selected network architecture was considered correct for the given task. The training examples were cropped from the full image by shifting the window vertically and horizontally every 3 pixels (the examples were created for approximately 1/9 pixels). The total number of training and validation examples was around 540,000.

2.2.3 Classification into points located on the ground and others

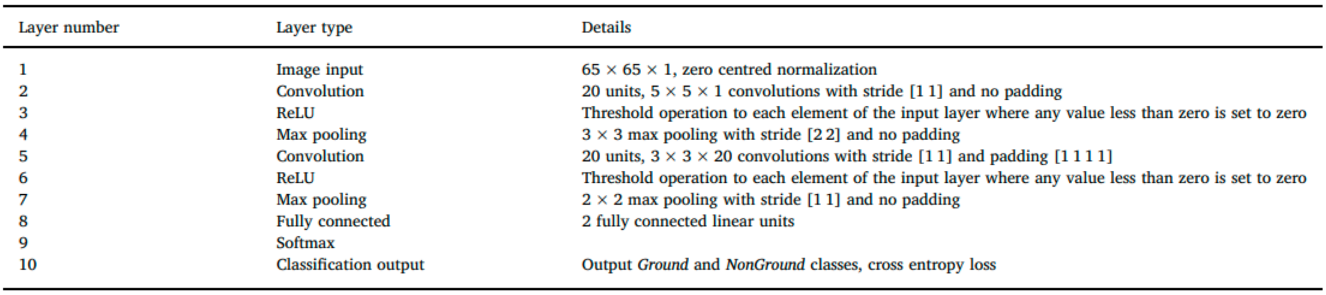
Only data from the area covered with high grass were used to train the CNN classifying points for the Ground and NonGround classes. The assignment of target classes was based on the reference measurements. The reference surface was created using a linear interpolation between the reference points. If the point of the cloud had a height that was less than 5 cm from the reference model, it was considered to be a point on the ground (Ground class). Otherwise, the point was assigned to the NonGround class.

Only around 10 percent of the points from the Łaziska dataset used for training were assigned to the Ground class. This caused an even selection of examples from a full picture, resulting in a very uneven number of examples for the classes in the training set. This could lead to difficulties in effective training, because the training set will have a relatively small number of examples that belong to one of the classes.



Layer number	Layer type	Details
1	Image input	65 × 65 × 1, zero centred normalization
2	Convolution	10 units, 5 × 5 × 1 convolutions with stride [1 1] and no padding
3	ReLU	Threshold operation for each element of the input layer, where any value less than zero is set to zero
4	Max pooling	3 × 3 max pooling with stride [2 2] and no padding
5	Fully connected	2 fully connected units
6	Softmax	
7	Classification output	Output *HG* and *LG* classes, cross entropy loss

Figure 4: Structure of the network used for the classification of vegetation into the high grass (HG) and low grass (LG) classes.



Layer number	Layer type	Details
1	Image input	65 × 65 × 1, zero centred normalization
2	Convolution	20 units, 5 × 5 × 1 convolutions with stride [1 1] and no padding
3	ReLU	Threshold operation to each element of the input layer where any value less than zero is set to zero
4	Max pooling	3 × 3 max pooling with stride [2 2] and no padding
5	Convolution	20 units, 3 × 3 × 20 convolutions with stride [1 1] and padding [1 1 1 1]
6	ReLU	Threshold operation to each element of the input layer where any value less than zero is set to zero
7	Max pooling	2 × 2 max pooling with stride [1 1] and no padding
8	Fully connected	2 fully connected linear units
9	Softmax	
10	Classification output	Output *Ground* and *NonGround* classes, cross entropy loss

Figure 5: Structure of the network used to classify the points into the Ground and the NonGround classes

2.2.4 Setting the correction to the point heights classified as Ground

Training was performed on image fragments corresponding to pixels classified as high grass for both Ground and NonGround classes. The CNN was later used to process (for the Jerzmanowice dataset) only examples classified by the network as Ground. The structure of the network was selected following the second stage of processing (Ground/NonGround classification). There were also attempts to use slightly larger CNN structures, but this did not result in any improvement

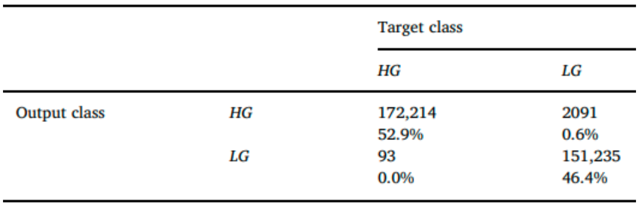
Training examples were cropped from the full image by shifting the window every 3 pixels both vertically and horizontally. As a result, the total number of training and validation examples was around 300,000.

In Figures 4 and 5, we can see the Strucutres of the CNNs used to classify the image points.

3 Results

The Łaziska dataset was used to train and validate the neural networks, while the Jerzmanowice dataset was used to test the trained networks.

The results for this network are best summarized by the confusion matrix presented in Figure 6. The effectiveness of the neural network was considered to be very high, as 99.3 percent of the cases were correctly classified.



		Target class	
		HG	*LG*
Output class	*HG*	172,214 52.9%	2091 0.6%
	LG	93 0.0%	151,235 46.4%

Figure 6: Confusion Matrix

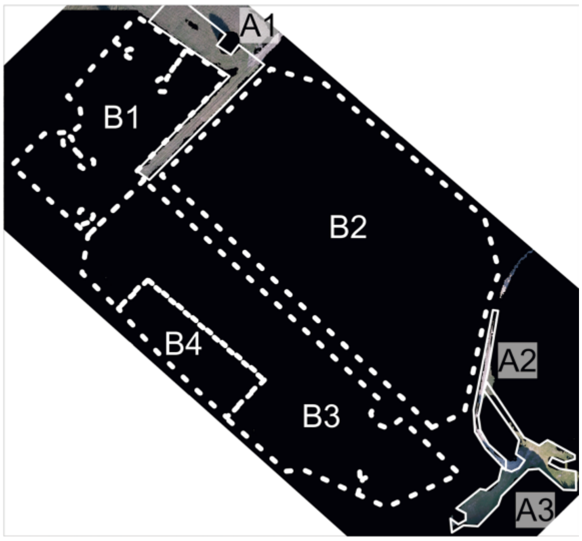


Figure 7: Points classified by the CNN as high grass (HG class - black pixels)

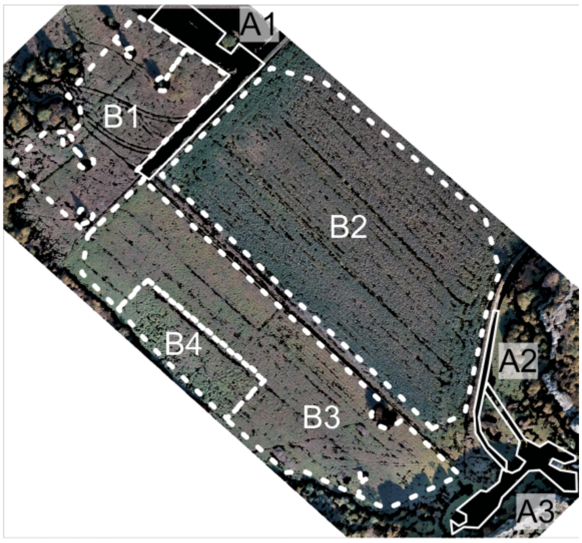


Figure 8: Points of the Jerzmanowice dataset classified as terrain (Ground class - black pixels)

4 Conclusion

The use of CNNs with the proposed input data formulation enabled the correct classification of areas as low or high grass in nearly all cases, and therefore there is no need to improve the filtering stage. In contrast the classification of the cloud points as ground or non-ground points using the CNN gave promising results, but requires further processing or training using a more extensive dataset. In this study, the CNN used to classify points as ground or non-ground gave similar DEM accuracies to those obtained using an algorithm based on local minima. The introduction of a correction set by the CNN to the height of the cloud points located in areas overgrown with high grass resulted in significant reductions in the systematic deviations of the height of the cloud points from the true values

References

1. Structure from Motion Theory https://cmsc426.github.io/sfm/

2. Demonstration of structure-from-motion (SfM) and multi-view stereo (MVS) close range photogrammetry technique for scour hole analysis https://www.ias.ac.in/article/fulltext/sadh/046/0227

3. Application of convolutional neural networks for low vegetation filtering from data acquired by UAVs https://www.sciencedirect.com/science/article/abs/pii/S0924271619302254

4. Oliveto G and Hager W H 2005 Further results to time- dependent local scour at bridge elements. J. Hydraul. Eng. 131: 97-105

5. Chiew Y M and Melville B W 1987 Local scour around bridge piers. J. Hydraul. Res. 25: 15–26

6. Wikipedia page of Structure from motion here.