MAGO APPROACH FOR SEMANTIC SEGMENTATION: THE CASE STUDY OF UAVID BENCHMARK DATASET

S. Gagliolo *, D. Sguerso

Geomatics Laboratory, Department of Civil, Chemical and Environmental Engineering (DICCA), University of Genoa, Italy –
sara.gagliolo@edu.unige.it, domenico.sguerso@unige.it

KEY WORDS: Semantic segmentation, Machine learning, UAV photogrammetry, Point cloud.

ABSTRACT:

The present work is focused on a semantic segmentation strategy implemented in the workflow of the tool MAGO (standing for "Adaptive Mesh for Orthophoto Generation"), considering the contribution of the 3D geometry and the colour information, both deriving from the point cloud of the scene. Moreover, the 2D source imagery, previously used to obtain the photogrammetric point cloud, is employed even to enhance the procedure with the recognition of moving objects, comparing the evolution of epochs.
The analysed context is an urban scene, deriving from the UAVid dataset proposed for the ISPRS benchmark. In particular, the so-called "seq18", a set of high-resolution oblique images taken by UAV (Unmanned Aerial Vehicle), has been used to test the semantic segmentation. The workflow includes the production of two Digital Surface Models (DSMs), containing the geometric and radiometric information, respectively, and their processing by means of the Harris corner detector, allowing the understanding of the image variability. Then, starting from the source geometry and colour information and combining them with their variability mapping, a preliminary classification is performed. Further criteria allow the segmentation of the humans and cars present in the scene. In particular, static objects are identified according to the content of the neighbour pixels in a certain kernel, while the evolution in time of moving elements is recognized by means of the comparison of the projected images belonging to the different epochs. The presented preliminary achievements show some criticalities that require further attention and improvement. In particular, the strategy could be enriched getting more information from the source 2D images, which at the moment are directly used only for the comparison of consecutive epochs.

1. INTRODUCTION

Semantic segmentation is a Computer Vision technique (Förstner and Wrobel, 2016) that aims to the recognition and the comprehension of the content of an image at the pixel level. This approach is widely used in remote sensing applications, especially in the analysis of urban scenarios (Ajmar et al., 2019; Huang et al., 2019, Schmitz et al., 2019, Zhou et al., 2019) or in the delineation of forest trees (Chen et al., 2021; Sothe et al., 2020; Kempf et al., 2019).
The segmentation approach could be based on imagery (Marmanis et al., 2018) or three-dimensional models (Ao et al., 2019), as well as on the combination of both 2D and 3D information (Ding et al., 2019). Typically, deep learning methods are applied to such procedure, including, to cite some examples, Conditional Random Fields (CRF; Pan et al., 2020; Lafferty et al., 2001), Markov Random Fields (MRF; Zoltan and Josiane, 2012), Spatial Pyramid Pooling (SPP; Zhengyu and Joohee, 2020), and Convolutional Neural Networks (Cresson, 2020; Martinez-Soltero et al., 2020; Ouyang and Li, 2021).
The present work is intended to describe the preliminary approach conceived by the authors, developed to obtain the segmentation. Both geometric and radiometric information are combined; moreover, the 3D point cloud of the object is used as source for the identification of static objects, while the contribution of 2D imagery allow evaluating the evolution in time of moving objects by comparing the projected images at different epochs. The analysed case study is represented by the UAVid dataset (Lyu et al., 2020), composed by high-resolution videos and imagery focusing on urban scenes, whose segmentation is based on eight object categories: buildings, roads, static cars, trees, low vegetation, humans, moving cars, and background clutter.
The proposed strategy consists in a machine-learning procedure, whose inputs are represented by the images and the photogrammetric point cloud obtained from their post-processing. Since the UAVid imagery is not exactly fitted for photogrammetric applications, the dataset employed as case study has been chosen paying attention that the image overlapping was sufficient to allow the 3D point cloud reconstruction by means of Structure From Motion (SFM; Ullman, 1979) technique.
The implemented functions have been introduced as a new module in the software MAGO (*Mesh Adattiva per la Generazione di Ortofoto*, literally Adaptive Mesh for Orthophoto Generation; Gagliolo, 2019; Gagliolo et al., 2019a and 2019b), implemented within the Geomatics Laboratory of the University of Genoa. This tool, written in C++ language, is originally born for the automatic reconstruction of high-resolution orthophotos of adjacent walls, automatically recognising their rotation, and it has been enriched with the function of semantic segmentation here presented. MAGO procedure already included a module for the automatic check of the geometry homogeneity, carried out by means of the evaluation of the Z coordinate trend, where Z is the direction normal to the representative surface. In the original purpose of the software, this module was useful to evaluate and apply the transformation required to put the point cloud in a service reference system with X and Y axes identifying the orthophoto plane and the Z direction along its normal vector.

* Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B2-2021-353-2021 | © Author(s) 2021. CC BY 4.0 License.

In the present work, taking cue from the described process, a new module for the detection of discontinuities is presented. The aim is to adopt this function for both geometric and radiometric segmentations, therefore joining the results of the two operations in a unique classification that takes into account both the aspects, assigning autonomously a category to each pixel.

The paper is organized as follows: in section 2, the case study is presented; in section 3, the strategy for the semantic segmentation is described; in section 4, the results of the application of the conceived approach on the testing dataset are shown; finally, conclusions and future perspectives of the work are reported.

2. THE UAVID DATASET

UAVid collection is a new high-resolution Unmanned Aerial Vehicle (UAV) semantic segmentation dataset focused on new challenges, including large-scale variation, moving object recognition and temporal consistency preservation. The proposed scenarios include urban and street scenes. The dataset consists of 42 video sequences (from “seq1” to “seq42”), which are captured with 4K high-resolution by the oblique point of view. The authors of the benchmark provided ten images extracted per each sequence, labelling the 420 resulting images with eight classes. Moreover, the sequences have been classified in three groups, i.e., training, test and validated sequences (Lyu et al., 2020).

Since the proposed approach requires the use of the 3D point cloud of the scene, the so-called “seq18” has been chosen as case study, given that the overlapping of provided frames was sufficient to obtain the photogrammetric reconstruction.

In Figure 1, the first epoch of the distributed “seq18” is shown.

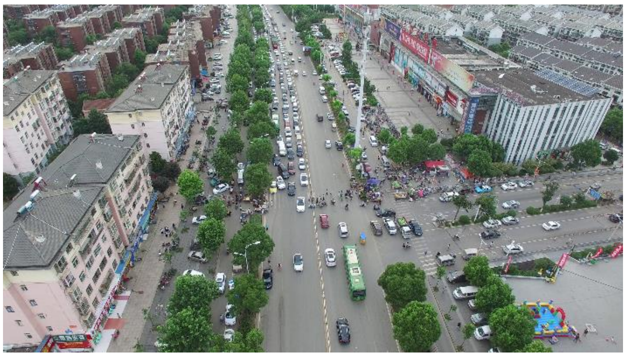

Figure 1. First epoch of “seq18”.

The 3D point cloud production has been achieved by means of the SFM software Agisoft Metashape© (Agisoft© LLC, 2019). In this regard, the choice of the input information for the photogrammetric post-processing represented the first issue. In facts, on the one hand, the extracted images provided for each sequence are a few number compared to the usual photogrammetric blocks, and their resulting point cloud is affected by the presence of outliers due to the scarcity of the correspondences, as testified by Figure 2. On the other hand, the use of a higher number of frames (91) directly extracted from the provided videos has been attempted, leading to the production of a visibly distorted point cloud (Figure 3). This behaviour could be due to the fact that the drone path followed almost a straight line, as in a single stripe dataset. Thus, the presence of a single stripe, without any Ground Control Point to stabilize, has been badly managed by the software in the set conditions, causing a fleeting reconstruction.

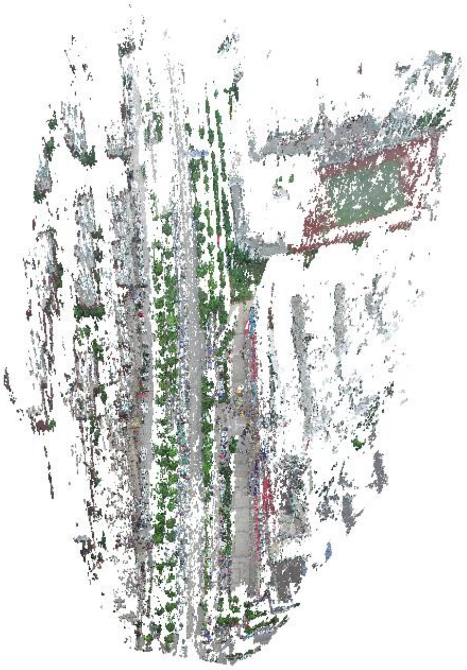

Figure 2. Resulting point cloud from the provided frames of “seq18”.

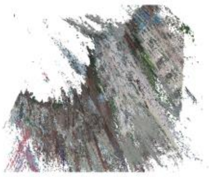

Figure 3. Lateral view of the point cloud obtained from 91 frames, with clearly visible distortions.

The following parameters have been set to carry out the workflow: the ten provided frames have been given as input, then the aerotriangulation has been performed at Medium quality, meaning that the image has been downscaled by factor of 4 (2 per each side). Finally, the dense cloud reconstruction has been launched at Ultra High quality, corresponding to process the images at their original resolution, and Mild depth filtering. Since no information about the Reference System was available, a pretended one has been attributed so that the building façades are vertical and the objects proportions are coherent with their standard measurements.

The resulting point cloud has been filtered using the Statistical Outlier Removal (SOR) and the noise filter available in the open source software CloudCompare (CloudCompare Development Team, 2021), applying the suggested default parameters; moreover, it has been subsampled with a minimum spacing between points of 0.1 m.

3. SEMANTIC SEGMENTATION APPROACH

As aforementioned, the proposed semantic segmentation approach combines the geometric and the radiometric criteria, in order to obtain a unique classification.

Starting from the 3D point cloud resulting from the 2D images provided by the ISPRS benchmark, two raster maps are produced: the former consists in the Digital Surface Model (DSM) of the scene, while the latter represents the corresponding nadiral greyscale map. Both these raster images contain Not a Number (NaN) values where the cell could not be filled with any source information from the 3D point cloud.

Both the raster maps are processed using the Harris Corner detector (Harris and Stephens, 1988), by means of the corresponding function implemented in the OpenCV (OpenCV Development Team, 2019) open-source library, available for C++ language. This technique allows to rate each image pixel with a mark *R*, according to the presence of a large variation in intensity with respect to the neighbours. In particular, the pixels are associated to the following groups basing on the obtained mark *R*:

- $R > 0$: corner, i.e., significant change in all directions;
- $R < 0$: edge, i.e., no change along the edge direction;
- $|R|$ small: flat region, i.e., no change in all directions.

The value *R* is obtained proceeding with the following steps. A greyscale 2D image, denoted as *I*, and a window *W(x, y)* of the image, shifted time by time of the quantity *(u, v)*, are assumed as input. The sum of squared differences (SSD) between these two patches, denoted as *E*, is given by:

$$E(u,v) = \sum_{x,y} W(x,y) \cdot [I(x+u, y+v) - I(x,y)]^2, \quad (1)$$

Approximating the quantity $I(x+u, y+v)$ by means of a first-order Taylor expansion, the function *E* could be written as:

$$E(u,v) \approx \sum_{x,y} W(x,y) \cdot [I(x,y) + I_x u + I_y v - I(x,y)]^2, \quad (2)$$

Thus:

$$E(u,v) \approx \sum_{x,y} W(x,y) \cdot [I_x^2 u^2 + 2 I_x I_y uv + I_y^2 v^2], \quad (3)$$

The quadratic approximation could be written in matrix form as

$$E(u,v) \approx [u \quad v] \cdot M \cdot \begin{bmatrix} u \\ v \end{bmatrix}, \quad (4)$$

where *M* is a second moment matrix computed from image derivatives:

$$M = \begin{bmatrix} \sum_{x,y} W(x,y) \cdot I_x^2(x,y) & \sum_{x,y} W(x,y) \cdot I_x(x,y) I_y(x,y) \\ \sum_{x,y} W(x,y) \cdot I_x(x,y) I_y(x,y) & \sum_{x,y} W(x,y) \cdot I_y^2(x,y) \end{bmatrix}, \quad (5)$$

$$M = \begin{bmatrix} \sum I_x I_x & \sum I_x I_y \\ \sum I_x I_y & \sum I_y I_y \end{bmatrix} = \sum \begin{bmatrix} I_x \\ I_y \end{bmatrix} \cdot [I_x \quad I_y] = \sum \nabla I \cdot (\nabla I)^T, \quad (6)$$

Each horizontal section of the function *E(u, v)* is the equation of an ellipse. The diagonalisation of the *M* matrix allows to obtain the lengths of the ellipse axes and their orientation, by means of the eigenvalues λ_1 and λ_2 and the corner response measure *R*, respectively.

$$M = R^{-1} \cdot \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \cdot R, \quad (7)$$

As previously described, the method implemented in OpenCV takes into account the corner response measure, using the value *R*, which is calculated as:

$$R = \lambda_1 \lambda_2 - \alpha(\lambda_1 + \lambda_2)^2 = \det(M) - \alpha\, trace(M)^2, \quad (8)$$

where α is an empirically determined constant ranging within 0.04 and 0.06. In the present work, the value 0.04 has been adopted.

The input parameters for the *cv::cornerHarris* function are the source image, the destination image, the kernel size, the aperture parameter for the Sobel operator (Duda and Hart, 1973), and the constant α.

Once both the geometric and the radiometric input maps have been processed with this technique, each cell has been classified according to the obtained *R* value. In particular, the service images containing the *R* marks obtained from the radiometric and the geometric contribution are called R_{colour} and R_{geom}, respectively. Moreover, a synthesis of the two contributions is stored in the matrix R_{class}, giving a label based on the following criterion, such that the obtainable R_{class} values are resumed in Table 1.

if			then
$R_{colour} > 0$	&	$R_{geom} > 0$	$R_{class} = +4$
$R_{colour} < 0$	&	$R_{geom} < 0$	$R_{class} = -4$
$R_{colour} < 0$	&	$R_{geom} > 0$	$R_{class} = +3$
$R_{colour} > 0$	&	$R_{geom} < 0$	$R_{class} = -3$
$\|R_{colour}\|$ small	&	$R_{geom} > 0$	$R_{class} = +2$
$R_{colour} > 0$	&	$\|R_{geom}\|$ small	$R_{class} = -2$
$\|R_{colour}\|$ small	&	$R_{geom} < 0$	$R_{class} = +1$
$R_{colour} < 0$	&	$\|R_{geom}\|$ small	$R_{class} = -1$
$\|R_{colour}\|$ small	&	$\|R_{geom}\|$ small	$R_{class} = 0$

Table 1. R_{class} attribution conditions.

Resuming, the labelling operation allows identifying the level of variability associated to each pixel, considering the changes in X and Y directions and in geometric and radiometric information.

The following step is the processing of the classification map by grouping in homogeneous regions the neighbour pixels with the same assigned value (Nikhil and Sankar, 1993). The set of regions that derives from this segmentation process is called partition.

The needed phases for the achievement of the partition are detailed in the following. First, the labelling results from the *R*-based classification determine for each pixel the so-called event in which it is involved, tagging if pixels are pertaining or not to an area. Then, the operation of grouping allows joining the pixels with the same label in a cluster. This operation generates regions, i.e., a set of neighbouring pixels, called connected components. The label assigned to each pixel is an integer that identifies the belonging region of the pixel.

Two pixels p and q with the same label f belongs to the same connected component C if there is a sequence of points $(p_0, p_1, \dots, p_n)$ of value f belonging to C where $p_0 = p$ and $p_n = q$ and p_i is neighbour to p_{i-1} for $i = 1, \dots, n$.

The value of each pixel is replaced by the smallest label value of its neighbours belonging to same connected component; the pixel above and the one on the left are considered, obtaining a 4-connected grouping. This operation is carried out recursively from the left to the right and from the top to the bottom. Thus, bottom-up row scan follows, considering the 4-connected neighbourhood made by the lower pixel and the one on the right. The replacement of the values is iterated until no more label changes are applicable.

Until now, the greyscale colour space has been chosen to apply the geometric segmentation using the Harris corner detector for the gathering of the discontinuities. Nevertheless, this colour space does not easily allow recognizing the hue of the analysed pixel, as well as the well-known RGB (Red Green Blue). Thus, the authors decided to convert in the HSV (Hue Saturation Value) range the original map of colours, obtained from the coloured point cloud. Starting from the HSV associated to each cell, several masks are arranged in order to identify the pixel membership. In particular, the threshold criteria are listed in the following; they have been chosen according to the authors interpretation and not using a superimposed classification. The OpenCV interpretation of the input values is due to the bytes coverage and requires that the H is within 0° and 180° instead of 360°, and S and V are within 0 and 255 instead of between 0 and 1. Table 2 resumes the input parameters for colour masking, according to OpenCV convention.

Colour	H	S	V
White	0 – 180	0 – 24	230 – 255
Red	0 – 14	25 – 255	100 – 255
	165 – 180	25 – 255	100 – 255
	0 – 14	100 – 255	25 – 255
	165 – 180	100 – 255	25 – 255
Brown	0 – 14	25 – 99	25 – 99
Purple	135 – 180	25 – 99	25 – 99
Green	45 – 74	25 – 255	25 – 255
	15 – 44	25 – 99	25 – 99
Blue	105 – 134	25 – 255	25 – 255
Cyan	75 – 104	25 – 255	25 – 255
Yellow	15 – 44	25 – 255	100 – 255
	15 – 44	100 – 255	25 – 255
Magenta	135 – 164	25 – 255	100 – 255
	135 – 164	100 – 255	25 – 255
Black	0 – 180	0 – 255	0 – 24
Grey	0 – 180	0 – 24	25 – 230

Table 2. HSV categories.

In such regions having a resulting value R_{class} equal to 0 or 1, i.e. with no significant radiometric and null or low geometric variation respectively, the HSV masking is applied homogeneously, on the basis of the most recurrent value in the area.

In the first step of the segmentation, the criteria coming from R_{class} or directly the height information resulting from the DSM are combined with the colour inferred from the HSV masking, as listed in Table 3.

Class	Label	Conditions		
0	no data	NaN		
1	buildings	HSV ∉ green	&	$Z > 6$ m
2	roads	grey	&	$R_{class} < 2$
4	trees	green	&	$R_{class} \geq 2$
5	low vegetation	green	&	$R_{class} < 2$
8	background clutter	remaining		

Table 3. Applied criteria for the preliminary labelling.

In this phase, the static scene is distinguished in five macro-areas, including buildings, roads, trees and low vegetation, as well as the remaining background.

Further criteria need to be implemented in order to point out also the three remaining categories, i.e., static cars, humans and moving cars, isolating them from the generic background.

In this regard, the actual potentialities of the algorithm are not suitable to discern humans from cars. Thus, the label static cars (class 3) and moving cars (class 7) are changed to static and moving objects respectively, while the category humans (class 6) is suppressed.

Regarding the segmentation of static objects, they are extracted from the generic background checking the presence of at least a certain number of cells labelled as road (class 2) or static object (class 3) in the neighbourhood of the analysed pixel by using a kernel. In particular, the road and the static object cells need to be more than the half of cells filled with categories different from the background and the empty ones (classes 8 and 0, respectively).

If at least one of the neighbour cells is road, a further check on the difference between the analysed pixel and the mean of the heights in the surrounding road cells is performed, i.e., if the Z coordinate of the analysed pixel is lower than three meters over the road average height, the matching with class 3 is confirmed.

The last step is the recognition of the moving objects, which is achieved thanks to the comparison of the 10 epochs provided in the UAVid source imagery.

The frames, obtained from the acquisition of the camera as central projections, are orthogonally projected using the tool MAGO firstly on a plane with a similar orientation to the original image attitude, then on the XY plane. The intermediate phase, which takes into account a service plane approximately at the same inclination of the original image, allows MAGO to optimize the research of the matching points that compose the adaptive mesh (Gagliolo et al., 2019b).

Once the projections are performed, the resulting greyscale maps are subtracted, in order to highlight the difference from the previous to the following epoch. The pixels that are not visible in at least one of the single analysed views are excluded from the comparison. Moreover, a threshold of 30 in the range of greyscale tones is applied to exclude changes barely perceivable by the human eye.

4. RESULTS AND DISCUSSION

The described procedure has been applied on the case study of the “seq18” belonging to the UAVid dataset proposed for the ISPRS benchmark.

First, the 3D point cloud built starting from the provided frames sequence has been processed to obtain two DSMs, the former containing the geometric information in terms of Z coordinate median (Figure 4), while the latter containing the radiometric information converted to greyscale values (Figure 5). The following Figure 6 and Figure 7 are focused on a portion of interest, depicted in red in Figure 4 and Figure 5. This box, located where the obtained point cloud is sufficiently satisfying, has been chosen to carry out the test. The poor quality of the obtained point cloud for the analysed dataset is due to the fact that the shooting of the source images was not planned to obtain a 3D survey.

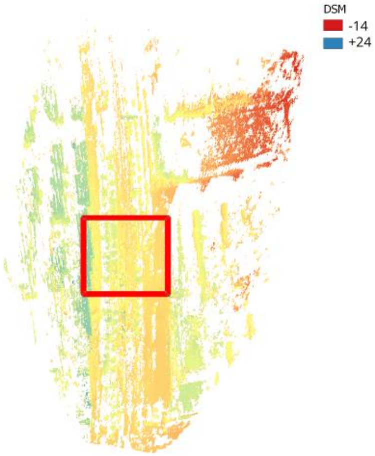

Figure 4. DSM with the highlighting of the focusing box.

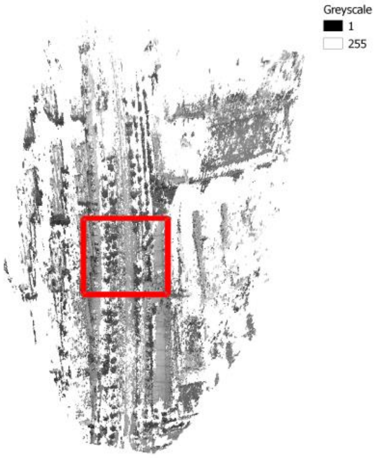

Figure 5. Corresponding greyscale map with the highlighting of the focusing box.

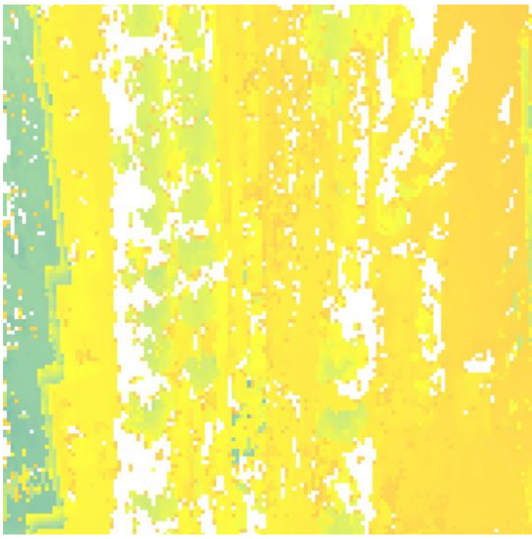

Figure 6. Detail of DSM focused on the indicated box.

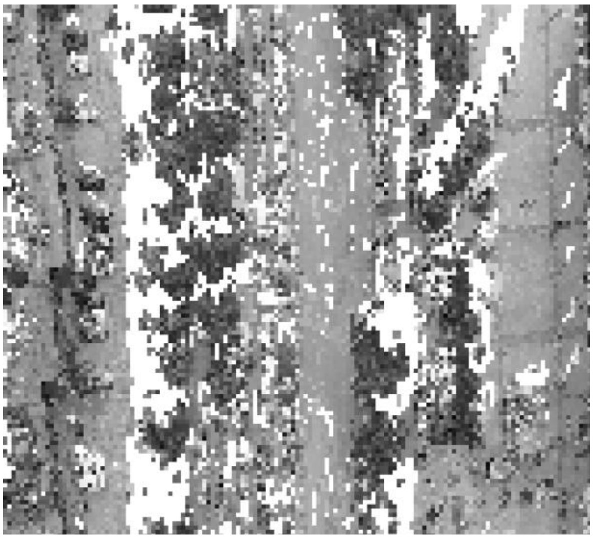

Figure 7. Detail of greyscale map focused on the indicated box.

Then, R_{class} map has been computed, as described in section 3; Figure 8 shows the portion belonging to the established boundaries, indicated in Figure 4. It is possible to notice that most of the areas are denoted with value 4, interpretable as high variability both in geometry and radiometry and both in the X and Y directions, or -2, identifying high variability in colour but null in geometry.

The starting R_{colour} and R_{geom} data are obtained using the *cv::cornerHarris* function, setting the kernel size at 7, comparable with the dimension of trees crowns and cars considering that each pixel is 0.5 m, the aperture parameter for the Sobel operator at 3, and the constant α at 0.04.

The R_{colour} and R_{geom} values considered as limit for the flat region ($|R|$ small) are 10^{-5} and 10^{-9} respectively.

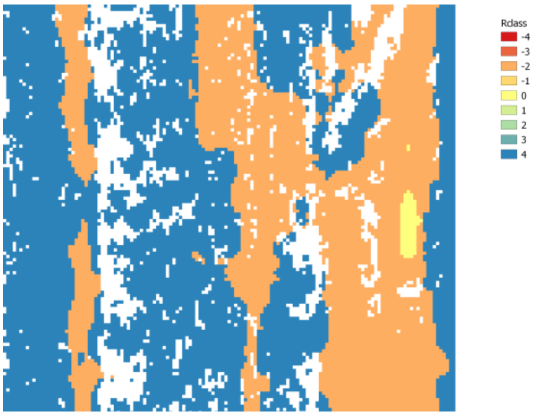

Figure 8. R_{class} map.

The preliminary segmentation, resulting from the sheer analysis of the static scenario, is depicted in Figure 9. It includes all the categories except for the moving objects (class 7). The kernel used for the static objects research has size 5×5. It is worth noting that the class of the trees is satisfying, as well as the ones of buildings, roads and low vegetation. Regarding static objects, it is recognisable the presence of the cars waiting at the crosswalk or parked on the roadside.

Moving objects are isolated subtracting the resulting projections of consecutive epochs time by time, as shown in Figure 10. The points depicted in red represent the variations between two consecutive epochs. Nevertheless, their quantity is excessive with respect to the real presence of moving objects. The development of further strategy for the outlier removal would be needed in future. In facts, some mismatches could be associated to discontinuities, which are not perfectly overlapping even if pertaining to static elements.

From the shown results, the processing outcomes are strongly affected by the quality of the input data, i.e., the 3D point cloud and the deriving geometric and radiometric DSMs. Undoubtedly, the upstream 3D reconstruction is badly influenced by the presence of many moving objects, which would require a huge time to be singularly masked in each source frame in the Agisoft Metashape© interface. Thus, it could be worthy to bring forward the moving object detection, working already on their recognition on the 2D images, so that it would be possible to automatically exclude them from the point cloud reconstruction.

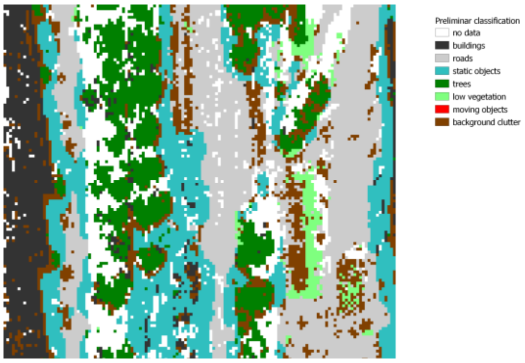

Figure 9. Preliminary segmentation of the static scenario.

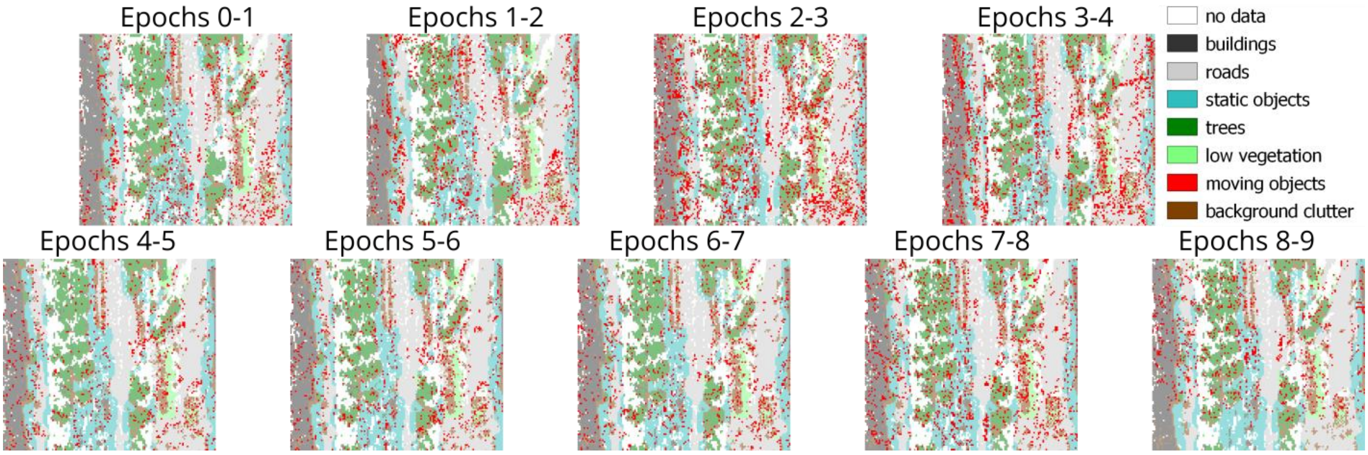

Figure 10. Final segmentation: moving objects varying in consecutive epochs are highlighted in red.

5. CONCLUSIONS AND FUTURE PERSPECTIVES

The present work shows the first experience of the authors with a segmentation strategy, based on both 2D and 3D source data and considering both the geometric and the radiometric aspects.

The produced maps point out the steps to reach the final segmentation. First, the 3D point cloud has been obtained from the provided frames of the UAVid "seq18", considered as a better choice than the extraction of input images directly from the UAV video, which highlighted a substantial distortion. The so-obtained point cloud has been used as input for the geometric and radiometric DSMs production. Then, these maps have been processed using the Harris corner detector in order to underline the image variability, according to whether the geometric or radiometric components and analysing both X and Y directions. The results of this phase are resumed in an R_{class} map, whose labels are associated to the possible combination of corner, edge and flat regions deriving from the Harris corner detector applied on the two maps R_{colour} and R_{geom}.

Starting from the DSM heights, the R_{class} values and the HSV image masking, the criteria for the static scenario classification, except for humans and cars, are applied. Further rules are established to isolate static objects that could pertain to humans or cars.

Finally, the moving objects are isolated subtracting the images referred to consecutive epochs and applying a proper threshold in order to exclude not perceivable changes. Nevertheless, a strategy to exclude outliers, i.e., points along discontinuities not perfectly overlapped, has still not been conceived. For sure a stereo camera acquisition and permanent Ground Control Points may substantially improve the obtainment of the 3D model and, as consequent result, the classification by means of the proposed strategy.

The preliminary achievements shown in the present work are useful to do a critical analysis of the proposed workflow, putting in place new prompts for the further enhancement of the procedure. In particular, this method could be improved to get more information from the primary source, represented by the acquired video and images. Moreover, the moving objects need to be treated so that they do not represent an obstacle to the processing but an opportunity to improve it.

REFERENCES

Agisoft© LLC, 2019. Agisoft Metashape Professional Software, Version 1.5.1. agisoft.com (1 March 2021).

Ajmar, A., Arco, E., Boccardo, P., Giulio Tonolo, F., Yoong, J., 2019. Updating a road network dataset exploiting the results of semantic segmentation techniques applied to street-level imagery. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2/W13, 1511–1516. doi.org/10.5194/isprs-archives-XLII-2-W13-1511-2019

Ao, W., Wang, L., Shan, J., 2019. Point cloud classification by fusing supervoxel segmentation with multi-scale features. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2/W13, 919–925. doi.org/10.5194/isprs-archives-XLII-2-W13-919-2019

Chen, X., Jiang, K., Zhu, Y., Wang, X., Yun, T., 2021. Individual Tree Crown Segmentation Directly from UAV-Borne LiDAR Data Using the PointNet of Deep Learning. *Forests*, 12(2), 131. doi.org/10.3390/f12020131

CloudCompare Development Team, 2021. CloudCompare Software, Version 2.12 alpha. cloudcompare.org (29 March 2021).

Cresson, R., 2020: Semantic segmentation of optical imagery. Deep Learning for Remote Sensing Images with Open Source Software, CRC Press

Ding, Y., Zheng, X., Xiong, H., Zhang, Y., 2019. Semantic segmentation of indoor 3D point cloud with SLENet. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2/W13, 785–791. doi.org/10.5194/isprs-archives-XLII-2-W13-785-2019

Duda, R., Hart., P., 1973. Pattern classification and scene analysis. A Wiley-Interscience publication, 271-272, https://doi.org/10.2307/2286028

Förstner, W., Wrobel, B., 2016: *Photogrammetric Computer Vision*. Springer Nature, Cham.

Gagliolo, S., 2019: High resolution orthophotos for the recognition of structural lesions with "MAGO". *Bollettino SIFET n.1 – ANNO 2019* (italian, abstract available in english)

Gagliolo, S., Passoni, D., Federici, B., Ferrando, I., Sguerso, D., 2019a: U.Ph.O and MAGO: two useful instruments in support of photogrammetric UAV survey. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XLII-2/ W13, 289-296. doi.org/10.5194/isprs-archives-XLII-2-W13-289-2019

Gagliolo, S., Federici, B., Ferrando, I., Sguerso, D., 2019b: MAGO: a new approach for orthophotos production based on adaptive mesh reconstruction, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XLII-2/W11, 533-538. doi.org/10.5194/isprsarchives-XLII-2-W11-533-2019

Harris, C., Stephens, M., 1988. A combined corner and edge detector. In Proc. of Fourth Alvey Vision Conference 1988, 147-151

Huang, S., Nex, F., Lin, Y., Yang, M. Y., 2019. Semantic segmentation of building in airborne images. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2/W13, 35–42. doi.org/10.5194/isprs-archives-XLII-2-W13-35-2019

Kempf, C., Tian, J., Kurz, F., d'Angelo, P., Reinartz, P., 2019. Local versus global variational approaches to enhance watershed transformation based individual tree crown segmentation of digital surface models from 3k optical imagery. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2/W13, 43–50. doi.org/10.5194/isprs-archives-XLII-2-W13-43-2019

Lafferty, J.D., McCallum, A., Pereira, F.C.N., 2001: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williamstown, MA, USA, 28 June–1 July 2001; pp. 282–289.

Lyu, Y., Vosselman, G., Xia, G.S., Yilmaz, A., Yang, M.Y., 2020: UAVid: A semantic segmentation dataset for UAV imagery, *ISPRS Journal of Photogrammetry and Remote Sensing*, 165, 108 – 119. doi.org/10.1016/j.isprsjprs.2020.05.009

Martinez-Soltero, G., Alanis, A.Y., Arana-Daniel, N., Lopez-Franco, C., 2020: Semantic Segmentation for Aerial Mapping. *Mathematics*, 8, 1456. doi.org/10.3390/math8091456

Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2018. Classification with an edge: improving semantic image segmentation with boundary detection. *ISPRS Journal of Photogrammetry and Remote Sensing* 135, 158-172. doi.org/10.1016/j.isprsjprs.2017.11.009

Nikhil, R.P., Sankar, K.P., 1993. A review on image segmentation techniques. *Pattern Recognition*, 26(9), 1277-1294. doi.org/10.1016/0031-3203(93)90135-J

OpenCV Development Team, 2019. OpenCV 4.1.2 Source Library. opencv.org (1 March 2021)

Ouyang, S., Li, Y., 2021: Combining Deep Semantic Segmentation Network and Graph Convolutional Neural Network for Semantic Segmentation of Remote Sensing Imagery. *Remote Sens.*, 13, 119. doi.org/10.3390/rs13010119

Pan, X., Zhao, J., Xu, J., 2020: An End-to-End and Localized Post-Processing Method for Correcting High-Resolution Remote Sensing Classification Result Images. *Remote Sens.*, 12, 852. doi.org/10.3390/rs12050852

Schmitz, M., Huang, H., Mayer, H., 2019. Comparison of training strategies for convnets on multiple similar datasets for facade segmentation. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2/W13, 111–117. doi.org/10.5194/isprs-archives-XLII-2-W13-111-2019

Sothe, C., De Almeida, C. M., Schimalski, M. B., La Rosa, L. E. C., Castro, J. D. B., Feitosa, R. Q., Dalponte, M., Lima, C. L., Liesenberg, V., Miyoshi, G. T., Tommaselli, A. M. G., 2020. Comparative performance of convolutional neural network, weighted and conventional support vector machine and random forest for classifying tree species using hyperspectral and photogrammetric data. *GIScience & Remote Sensing*, 57(3), 369-394. doi.org/10.1080/15481603.2020.1712102

Ullman, S., 1979. The Interpretation of Structure from Motion. Proceedings of the Royal Society of London Series B, 203(1153), 405-426. doi.org/10.1098/rspb.1979.0006

Zhengyu, X., Joohee, K., 2020: Mixed spatial pyramid pooling for semantic segmentation. *Applied Soft Computing*, 91, 106209. doi.org/10.1016/j.asoc.2020.106209

Zhou, K., Chen, Y., Smal, I., Lindenbergh, R., 2019. Building segmentation from airborne VHR images using mask R-CNN. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2/W13, 155–161. doi.org/10.5194/isprs-archives-XLII-2-W13-155-2019

Zoltan, K., Josiane, Z., 2012: Markov Random Fields in Image Segmentation. Collection Foundation and Trends in Signal Processing. Now Editor, World Scientific, pp.164.