SCENE CLASSIFICATION OF URBAN AREAS EXPLOITING MULTI-VIEW HIGH RESOLUTION AERIAL IMAGES

F. Nex [a*], M. Dalla Mura [b]

[a] University of Twente, ITC Faculty, Department of Earth Observation Sciences, Enschede, The Netherlands - f.nex@utwente.nl
[b] GIPSA-lab, Grenoble Institute of Technology, Grenoble, France - mauro.dalla-mura@gipsa-lab.grenoble-inp.fr

KEY WORDS: Scene classification, multi-view aerial images, DSM, photogrammetry, urban remote sensing

ABSTRACT:
Many supervised and unsupervised algorithms for the automated and reliable classification of large regions using high resolution data have been presented in the remote sensing community in the last decades. Most of these approaches exploit a single input data: high resolution orthophotos or 3D point clouds. An increasing number of contributions has more recently exploited the combined use of orthophotos and LiDAR DSM taking advantage from the complementarity of these inputs. Nevertheless, very few applications have considered the use of overlapping multi-view images on the same area for classification. In this paper the first tests to investigate this classification architecture are presented. Different typologies of DSM and orthophoto as well as a variable number of images on the same area have been considered in the experiments. The preliminary results on the two available test areas will be shown, in order to draw the first conclusions on this approach and discuss the further developments of this research.

1. INTRODUCTION AND MOTIVATIONS

In the last decades, the improvements in the technical development of airborne image sensors and the mass production of UAVs have boosted the acquisition of extremely high resolution images for the generation of reliable and automated Digital Surface Models (DSMs) and orthoimages. On the other hand, the need to monitor, map and model urban and rural areas has led to the improvement of several algorithms for the automated and reliable classification of large regions using high resolution data (Mallet et al., 2011).
Some of the approaches presented in the literature exploit orthoimages or DSMs for performing a scene classification. More recently an increasing number of contributions combines the complementary nature of both spectral and depth information to overcome the limits of each input separately. A large variety of features from both DSMs and orthoimages have been therefore developed to make the classification process more reliable and accurate (Rottensteiner et al., 2014; Gerke and Xiao, 2014).
In all these implementations, the spectral information is only taken from the orthophoto, while the full information provided by the overlapping input images is neglected in the classification process. However, each pixel of the DSM could be easily mapped on the corresponding images thanks to the projective geometry equations, allowing the exploitation of the information from different points of view.
This paper aims at investigating the benefits for the classification process given by the use of multi-view images acquired on the same area. The conventional classification generated using (i) DSM, (ii) orthophoto and (iii) their combination has been compared to the classification provided combining together (iv) the DSM and the several images captured on the same area.
The first experiments and the first promising results will be discussed in the following sections. In the performed tests, different set of features have been adopted in each experiment and the same training sample has been used for each data configuration to compare the different results using the same input. Two different test areas have been considered in order to make the investigation independent from the input data.
The paper is organized as follow: in the following section the background on the use of photogrammetric techniques for the generation of point clouds and orthophotos will be shortly presented. Then, the adopted testing methodology will be introduced in Section 3, while the first results will be presented in Section 4. The conclusions and the future developments of our tests will be finally discussed in the last section.

2. BACKGROUND

Photogrammetry is the science of using image measurements to extract tridimensional information. Each image is ideally modelled as a central projection according to the pinhole camera model (Hartley and Zisserman, 2004): the projection center, the image point and the corresponding point in the space lie on the same line. The projective lines are then used to retrieve the position of points in the 3D space, given their 2D position on two or more images and vice-versa, to determine the position of a point on the images from its 3D coordinates in the space.
The photogrammetric process can be divided in three main steps: (i) image orientation, (ii) DSM generation and (iii) orthophoto generation.
The position (geo-referencing) and the attitude (rotation towards the coordinates system) of each acquisition is obtained by estimating the image orientation. In the dense point cloud generation, 3D point clouds are generated from a set of images, while the orthophoto are generated in the last step combining the oriented images projected on the generated point cloud, which leads to orthorectified images. The outputs from the last two steps (point clouds and orthophotos) can be directly used as input in the classification process. As a consequence, the way these products are generated and their final quality can have a direct effect on the classification of the area.
In the following sections, the dense point cloud and the orthophoto generation are reported, focusing on the possible

problems these products can have and the consequences for the final classification quality.

2.1 DSM generation and orthophoto drawbacks

Image matching represents the simultaneous establishment of correspondences between primitives (i.e. points, lines) extracted from two or more images and the estimation of the corresponding position in the 3D space using the collinearity or projective models (Remondino et al., 2014). Image matching algorithms are the bedrock for the automated generation of Digital Surface Models (DSM).
All of the proposed matching methods are based on *similarity* or *photo-consistency measures*, i.e. they compare pixel values between the images. In image space this process produce a *depth map* (that assigns relative depths to each pixel of an image) while in object space it is called *point cloud*. Anyway, the establishment of dense and accurate image correspondences is still a challenging task, and very different solutions have been conceived in the last three decades. Even if image matching algorithms have been greatly improved, they are not still able to generate error-free point clouds. Outliers are still visible in the most occluded parts of the scene. Regions with low textures or shadows can greatly increase the level of noise of the point clouds too. These problems afflict all the available solutions (Haala, 2013), producing unwanted spikes and locally wrong surface models. Of course, these problems affect in a negative way the quality of meshes and orthophotos generated from photogrammetric DSM.
Airborne (GSD <20 cm) and UAV images, with high overlaps (>70%) are more often used in the point cloud generation. The high image resolution and the high overlaps increase the quality of the achieved DSM, improving the quality in the surface reconstruction and reducing the number of outliers and mismatches in correspondence of occlusions. On the other hand, the geometric resolution (up to 40 cm) and the limited stereoscopic coverage of the satellite stereo images is nowadays still insufficient to provide DSM comparable to the aerial cases (Arefi et al., 2011).
The orthophoto is the orthogonal parallel projection of the input images where the projective distortions of the image are geometrically corrected using a 3D surface model. A Digital Terrain Model (DTM) was traditionally used for this kind of process but the extensive use of high resolution images in the last two decades has pointed out the limits of this kind of surface model on urban areas. Large artefacts or double mapping in correspondence of complex structures were often generated and the relief displacement was unacceptable for mapping purposes in correspondence of the high buildings (i.e. the position of building roofs is displaced from the correct position).
For this reason, the Digital Terrain Model is usually used in the orthogonal projection (Ahmar et al., 1998) and their output is therefore called true-orthophoto. Reliable true-orthophotos can be usually generated thanks to the use of high overlaps and high resolution images and the generally good performances of image matching algorithms.
As already mentioned, inaccurate DSMs can negatively influence the quality of the true-orthophoto, generating double mapping or wrong point projections. Then, many images with significant scene-to-scene radiometric variations can be often fused together to generate an orthophoto. This problem can be only partially mitigated using radiometric balancing and blending algorithms (such as min-cut, seamless stitching, graph-cut and watershed) to prevent/reduce patchy appearance in the final result.

3. METHODOLOGY

All the above mentioned problems usually reflect in the classification process as both DSMs and orthophotos are directly used in input. These issues become more relevant when high resolution images are used, giving misclassification in correspondence of wrong reconstructed regions of the DSM or abrupt and erroneous radiometric changes in neighbouring and homogenous regions of the orthophoto (i.e. on the same roof).
From this perspective, the use of the original images could reduce the negative impact given by these artefacts. The wrong DSMs reconstructions would be mitigated by the combined use of images from different perspectives. The different radiometric content of the images would be averaged reducing the misclassification due to inaccurate radiometric blending between images.
The developed methodology is based on the use of the original images in the attempt to make the classification more robust and less error prone in correspondence of critical areas.
For this purpose, a set of overlapping images, the generated photogrammetric DSM and the corresponding true-orthophoto are considered as input. Then, different configurations of input and different training sets are considered: the results achieved using only the DSM and only true-orthophoto as well as they combined used are initially considered. Then, the additional information provided by a variable number of images (from 1 to 4) on the same region are compared in order to assess the possible benefits given by the use of overlapping images in the classification. Each image is initially orthorectified using the available DSM: the information from different images on the same point can be therefore overlaid in the same reference system. The information provided by the images and the orthophoto is finally merged using a majority voting decision rule.
From each input data and each performed test, a minimum number of features has been considered in this stage of the research: the three bands from the orthophoto, the normalized height information and the spectral information from a variable number of images have been considered. The same training sample has been used for each data configuration to compare different results using the same input. The Random Forest algorithm has been adopted to generate the classification results. DSMs generated using different image matching algorithms has been considered in order to make our investigation software independent.
The classification process has been performed on two different areas, using different camera and different image bands in order to make our investigation independent from the considered scenario.

4. TESTS

4.1 Data description

Two different datasets have been adopted in the performed experiments. The first area is on Transacqua, a small town in Italy, while the second dataset is on the city of Vaihingen, Germany. The Transacqua dataset was acquired using an amateur D3X camera installed on a helicopter. The flight was performed at 800 m height with 80% along track and 60% across track overlaps: an average Ground Sample Distance (GSD) of 9 cm was achieved on the test area. The Vaihingen data is a subset of the data used for the test of digital aerial cameras (Cramer et al., 2010) carried out by the German Association of Photogrammetry and Remote Sensing (DGPF)

and currently available thanks to the ISPRS benchmark on the 2D semantic labelling contest (http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html). The GSD is 8 cm and Red, Green and Near-Infrared bands have been stored in 11 bits radiometric resolution image. Both 8bit and 16bit images are available for participants. In the performed tests only the 8bit images have been used.

The DSM generation of the two datasets have been performed using two different algorithms. The first dataset has been processed using the open-source MicMac software (Pierrot-Deseilligny and Paparoditis, 2006), while an implementation of the semi-global algorithm (Trimble Match-T software) has been adopted in the second image block processing (Hirshmuller, 2008).

The generated point clouds have been used to ortho-rectify the images and produce a high resolution image. The first orthophoto was generated using the tool Porto of the MicMac library. The orthophoto on Vaihingen area has been processed with the commercial software Trimble INPHO OrthoVista.

A labelled ground truth has been manually generated on the first dataset. Five different classes (building, road, ground, vegetation and shadows) have been considered on this area. On the other hand, the ground truth provided by the ISPRS benchmark has been adopted on the second dataset.

4.2 First results and discussion

In the following, the first performed tests are reported. The results on the Transacqua area are reported in Table 1, while the results achieved on Vaihingen are shown in Table 2.

	5	10	50	100
ortho+ndsm	76.7 (3.3)	81.9 (3.1)	89.6 (1.6)	92.1 (0.8)
ortho	57.0 (4.5)	60.4 (5.7)	75.2 (1.9)	79.9 (1.0)
allviews+ndsm	75.3 (6.2)	81.3 (6.1)	90.1 (1.5)	92.2 (1.2)
view1+ndsm	77.6 (6.3)	83.2 (2.8)	90.2 (1.3)	92.2 (1.0)
view2+ndsm	78.5 (4.8)	83.1 (4.2)	90.6 (1.1)	92.8 (1.1)
view3+ndsm	78.8 (5.2)	83.3 (4.1)	90.4 (0.9)	92.0 (1.0)
majvot_3views	**81.8** (4.2)	**85.8** (3.2)	**91.8** (1.2)	**93.5** (0.9)

Table 1. Classification results on the Transacqua test area using different input data configurations. Best results are in bold.

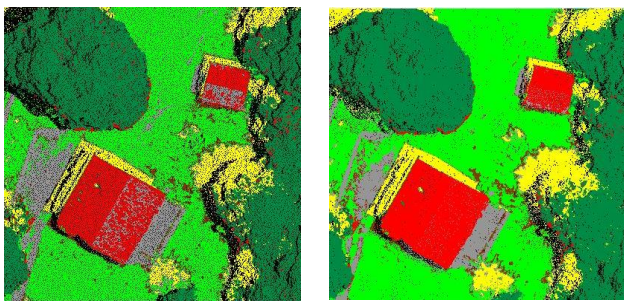
	5	10	50	100
ortho+ndsm	59.4 (6.3)	58.0 (5.7)	56.7 (1.7)	59.6 (3.1)
ortho	41.7 (12.1)	46.9 (8.3)	46.7 (0.0)	47.1 (0.9)
allviews+ndsm	55.4 (2.5)	**59.2** (3.7)	**60.0** (1.9)	**62.0** (2.4)
view1+ndsm	61.0 (6.0)	58.3 (4.4)	58.5 (2.5)	59.1 (2.2)
view2+ndsm	60.3 (9.0)	55.9 (5.5)	58.0 (3.4)	59.1 (3.8)
view3+ndsm	57.9 (5.7)	56.4 (6.6)	56.8 (5.8)	55.2 (1.8)
view4+ndsm	61.0 (7.0)	53.3 (4.7)	56.6 (1.4)	56.5 (1.0)
majvot_4views	**61.7** (5.0)	58.2 (3.2)	58.7 (1.8)	59.0 (1.2)

Table 2. Classification results on the Vaihingen test area using different input data configurations. Best results are in bold.

Different combinations are reported in these tables: on the rows the input data are reported, while in the column the number of samples used to train the algorithm.

The contribution of the multiview images has been tested considering one image per time (*viewX*, with X the id of the image) or all the images together (*allviews*) and combining the results of the classifications of the single images via majority voting (*majvot* option). The results look quite similar when a single image or the true-orthophoto are considered in the classification: in these cases, no substantial improvement can be detected. On the other hand, the use of multiple images on the same area (either in the *allview* or *majvot* configurations) systematically overcomes the standard configuration (*ortho+ndsm*) in both the test areas. By using a high number of images in the classification process, the classification accuracy always improves. This positive contribution is more relevant when a reduced number of samples is used to train the data.

Analysing the classification maps, it can be noticed that the use of multiple images reduces the noise in the results, decreasing the number of ambiguous and wrong classifications in most of the analysed cases. As an example, in Figure 1 the noise on the roofs is less when the number of images increases.



(a) (b)

Figure 1. Example of classification in the Transacqua dataset using the true-orthophoto and the DSM (a) and adding to this configuration three images.

5. CONCLUSIONS AND FUTURE DEVELOPMENTS

The use of overlapping images in the classification of very high resolution data has been discussed in this paper. By means of photogrammetric techniques, the corresponding points on different images are jointly exploited to classify the scene in a more robust and efficient way.

The presented investigation is just the first step of more thorough investigations that will be completed in the near future. However, the presented results have both confirmed that the use of multi-image can improve the classification results. The classification improvement is more relevant when a reduced number of training samples is used.

New and more extensive investigations will be performed in the future with more exhaustive investigations. The use of different DSM and different set of features will be first considered. The adoption of different DSM on the same area will allow to estimate the robustness of the presented approach in relation to the quality of the 3D data, observing different behaviours in presence of noisy or wrongly reconstructed regions. Different sets of features will help to understand if the use of the overlapping images is still relevant when more complete sets of features are adopted.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the Project “Automated scene information extraction from a joint analysis of aerial

remote sensing images and their photogrammetric DSM" (in the Galileo 2013 Project framework) that supported the initial stage of this investigation.

REFERENCES

Ahmar, f., Jansa, J., Riess, C., 1998. The generation of true orthophotos using a 3D building model in conjunction with a conventional dtm, IAPRS, vol. 32, pp. 16–22.

Arefi, H., d'Angelo, P., Mayer , H., Reinartz , P., 2011. Iterative approach for efficient digital terrain model production from CARTOSAT-1 stereo images. *Journal of Applied Remote Sensing*, 2011 (5), 19 p. DOI: 10.1117/1.3595265.

Gerke, M., Xiao, J., 2014. Fusion of airborne laserscanning point clouds and images for supervised and unsupervised scene classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 87, 78-92.

Hartley, R., Zisserman, A., 2004. *Multiple view geometry in Computer Vision*. Cambridge University press ISBN 0521540518.

H. Hirschmüller, 2008. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341.

Mallet, C., Bretar, F., Roux, M., Soergel, U., Heipke, C., 2011. Relevance assessment of full-waveform lidar data for urban area classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, vol.66 (6), pp.S71-S84.

Pierrot-Deseilligny, M., Paparoditis, N., 2006. A multiresolution and optimization-based image matching approach: An application to surface reconstruction from SPOT5-HRS stereo imagery. In: *IAPRS*, vol. XXXVI, pp. 1–5.

Remondino, F., Spera, M.G., Nocerino, E., Menna, F., Nex, F., 2014. State of the art in high density image matching, *Photogramm. Rec.*, vol. 29, no. 146, pp. 144–166.