ISPRS Journal of Photogrammetry and Remote Sensing 123 (2017) 124–139

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Semantic segmentation of 3D textured meshes for urban scene analysis

Mohammad Rouhani, Florent Lafarge *, Pierre Alliez

Inria Sophia Antipolis – Méditerranée, France

ARTICLE INFO

Article history:
Received 20 July 2016
Received in revised form 29 November 2016
Accepted 1 December 2016
Available online 9 December 2016

Keywords:
Semantic segmentation
Textured meshes
Urban scenes
Markov random fields
Random forests
Joint label prediction
Contextual information

ABSTRACT

Classifying 3D measurement data has become a core problem in photogrammetry and 3D computer vision, since the rise of modern multiview geometry techniques, combined with affordable range sensors. We introduce a Markov Random Field-based approach for segmenting textured meshes generated via multi-view stereo into urban classes of interest. The input mesh is first partitioned into small clusters, referred to as superfacets, from which geometric and photometric features are computed. A random forest is then trained to predict the class of each superfacet as well as its similarity with the neighboring superfacets. Similarity is used to assign the weights of the Markov Random Field pairwise-potential and to account for contextual information between the classes. The experimental results illustrate the efficacy and accuracy of the proposed framework.

1. Introduction

The recent advances on Multi-View Stereo (MVS) imagery (Seitz et al., 2006; Vu et al., 2009) make it possible to generate in routine dense meshes from airborne images acquired on large-scale urban scenes. Several commercial solutions such as ContextCapture (Acute3D/Bentley) and Pix4Dmapper (Pix4D) generate meshes with greater geometric accuracy and completeness than the common digital surface models. Contrary to LIDAR scans, such dense meshes represent 2-manifold surfaces, and do not require the interpolation of sample points. As depicted by Fig. 1, these meshes exhibits geometric details both on roofs and facades as MVS systems can altogether deal with oblique and vertical airborne images.

Dense meshes from multiview stereo imagery are relevant 3D representations for visualization-based applications such as navigation or augmented reality. They are however too raw for applications that require additional structure and semantic information to interpret the represented scenes. There is a dire need to recover the nature of the objects composing the scenes.

We propose a dense classification algorithm that infer the class of urban objects in such dense meshes. Departing from previous work, our approach leverages both radiometric and geometric information in a supervised manner. In addition, we learn the contextual knowledge required to recover additional coherence between the urban classes of interest.

2. Related work

We review below the most recent and related works on semantic segmentation, with focus on methods dealing with meshes of urban scenes.

2.1. Classification

Many different classification approaches have been used in photogrammetry and computer vision in order to partition images or point clouds, and to identify the nature of each area. Classification approaches differ mainly in the level of supervision: Supervised algorithms require a training set to learn how to correctly classify data, whereas unsupervised approaches require tuning the model parameters. Classification approaches also differ in the use of spatial dependencies and contextual information. In Markov Random Fields (MRFs) and Conditional Random Fields (CRFs) for instance, a datum is not classified independently from the rest of the data: the classification decision relies upon non-local information that accounts for spatial consistency between neighboring areas.

2.1.1. Supervised learning

Among the many classification approaches, Texton Boost and Texton Forest can be directly applied to the input data without requiring any type of feature descriptor (Shotton et al., 2008). In

* Corresponding author.
E-mail address: florent.lafarge@inria.fr (F. Lafarge).

http://dx.doi.org/10.1016/j.isprsjprs.2016.12.001
0924-2716/© 2016 Published by Elsevier B.V. on behalf of International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS).

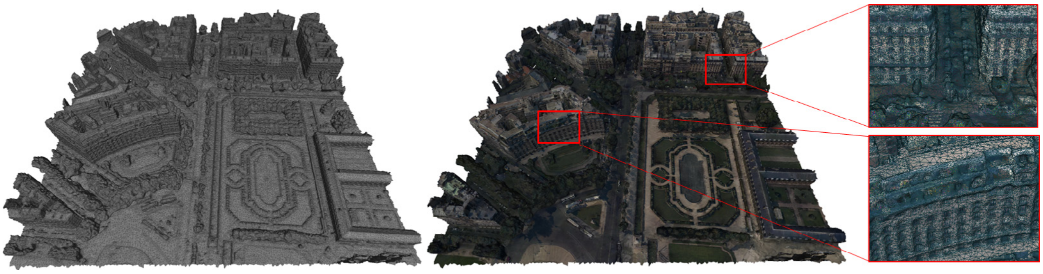

Fig. 1. Textured meshes produced from MVS imagery. Our input is a textured mesh combining geometry (left) and radiometry (right).

addition, these methods are extremely fast as they avoid computing filter-bank responses. A multi-feature variant of TextonBoost (Sengupta et al., 2013) is used for labeling each image of the stereo pair and fuse them into a scene. Both the image and geometric features have been used to train a JointBoost classifier (Valentin et al., 2013) for segmenting RGB-D images. Xiao and Quan (2009) use a series of one-vs-all AdaBoost classifiers to perform multiview semantic segmentation. In contrast, Kalogerakis et al. (2010) learn a label compatibility function between the neighboring segments of an input mesh through training a JointBoost classifier on the pairwise geometric features. Randomized decision forests (or Random Forests) are popular parametric classifiers for the segmentation and regression tasks (Zhang et al., 2010a). They are relevant for real-time applications as they generate label predictions very efficiently through performing few simple tests on the query data. Kähler and Reid (2013) employ random forests for segmenting RGB-D images; the feature vectors describe photometric and geometric information for every segment and pair of segments.

2.1.2. MRFs and CRFs

MRF or CRF formulations usually rely on an energy minimization problem. The energy is commonly composed of two terms: a data term that measures the coherence of each datum with respect to a label, and a pairwise potential that favors label smoothness. A supervised classifier can be used as prediction function to model the unary data term of MRFs and CRFs models. The contextual information provides relevant clues for improving the results of semantic segmentation. Co-occurrence statistics are modeled in Ladicky et al. (2013) through a global potential function recording which pairs of classes likely to occur in the same image. Galleguillos et al. (2008) model the co-occurrence and relative locations through the following pairwise term: Four different types of relative locations are considered (*above, bellow, inside and around*) to capture the spatial context through frequency matrices that record the likelihood of two labels to appear in a relative position. Myeong et al. (2012) propose a pairwise cost function based on a similarity graph that encodes the relationship between two regions through context links. A context learning algorithm estimates the strength of a context link in the query image.

Yao et al. (2012) introduces auxiliary variables to consider different terms altogether, such as segmentation energy, object-reasoning as well as scene and class presence potentials; the relation between these potentials are formulated in a general CRF model. Global and local contexts have been modeled in Mottaghi et al. (2014) to improve both semantic segmentation and object detection. The primer considers the presence or absence of a class in the scene, while the latter refers to the classes appearing in the vicinity of the object.

Formulating the segmentation problem as MRFs makes it possible to leverage many efficient inference algorithms to find the optimal labeling. Simulated annealing (Kähler and Reid, 2013), range/swap-based approaches (Liu et al., 2015), or mean-field approximation (Krähenbühl and Koltun, 2011) are relevant inference approaches when the configuration space is large. Inferring on global co-occurrences is commonly formulated as integer programming, and solved via linear relaxation (Ladicky et al., 2013). Yao et al. (2012) relies upon a message-passing algorithm to solve a holistic CRF that includes contextual terms such as scene and class-presence potentials.

2.2. Mesh segmentation

Mesh segmentation is still a research challenge in geometry processing, robotic, and computer vision (Shamir, 2008; Chen et al., 2009; Theologou et al., 2015). In its simplest form mesh segmentation may be seen as an unsupervised clustering problem based on specific geometric criteria (Shlafman et al., 2002). Region growing (Page et al., 2003) and spectral analysis (Zhang et al., 2010b) are other instances of such deterministic approaches for mesh segmentation. The probabilistic approaches such as MRFs or CRFs provide an efficient means to enforce spatial consistency (Lafarge et al., 2010). The design of energy terms for such approaches ranges from totally unsupervised (Verdie et al., 2015) to supervised (Van Kaick et al., 2011), through semi-supervised (Lv et al., 2012).

A key aspect of the mesh segmentation approaches is the design of feature vectors encoding geometric information such as normals, curvatures or planarity. For textured meshes, additional features based on photometric information such as colors or texture histograms, provide relevant clues to design the energy terms of MRFs and CRFs. Verdie et al. (2015) design three geometric attributes to define the unary term of a MRF. In our approach we adopt a supervised approach instead to learn an effective classifier.

In MVS contexts, some methods first perform classification directly from the images before mapping it to the output 3D model (He and Upcroft, 2013). In Sengupta et al. (2013), each image is labeled by a supervised CRF before mapping and fusing the sets of labels with the mesh. Lafarge et al. (2013) propose a hybrid approach in which the output model is progressively refined while detecting regular urban entities. Kundu et al. (2014) proposes a joint CRF model defined in 3D (volumetric) space and infers altogether the semantic label and voxel occupancy. Savinov et al. (2015) and Blaha et al. (2016) define the unary cost function along rays by considering both the semantic label and the depth of the first occupied voxel along the ray. Referred to as semantic 3D reconstruction by Haene et al. (2013), such approaches are often memory intensive and require complex inference approaches. The incremental approach proposed by Vineet et al. (2015) operates in near real-time and delivers a rough reconstruction with a street-based semantics. Xiao and Quan (2009) propose a larger MRF that includes all the views, and models all connections between the associated areas. The smoothness term between two views is defined either based on the color similarity or using the number of common feature tracks between the two associated