Collection, Processing and Augmentation of VR Cities

FRANZ LEBERL, STEFAN KLUCKNER, HORST BISCHOF, Graz

ABSTRACT

The starting point for the modeling of urban spaces was in the mid-1990s, using semi-manual aerial stereo-photogrammetric methods, also multi-view street-level image streams and at times also single images. First tangible applications of such 3D models were found in the late 1990's in the telecom-industry for the study of signal propagation of cellular and other telecommunication technologies. We now find ourselves ten years later in an evolution of Internet-search where location is both an important search criterion as well as a helpful means of visualizing search results. The Internet is becoming "location-aware". "Visualization" poses great demands on the "user-experience" to make the Internet-locations appear photorealistic and to present the human environment at great detail. Location initially was 2D, but since 2006 encompasses the 3rd dimension in various forms. While this initially is being addressed with a focus on visually pleasing photo-textures, it is rapidly evident that the applications depend on an interpretation of the urban scenes with knowledge of roads, sidewalks, trees, doors, windows, parking meters, sky lights etc. We therefore are investigating technologies and algorithms to automatically describe the urban environment in sufficient detail so that we can generate a pleasing visualization and understand the relevant objects in the scene. In this paper we show initial results using aerial photography as the main data input to find roads, green areas, trees, water bodies and buildings. Success rates are in the range of 90%, with a potential for improvements by increased image overlaps and computational methods.

1. JUSTIFICATION

Semantically interpreted 3D models of urban spaces would support interaction, search and navigation based on the elements of a city, be they windows, sidewalks, chimneys, number of floors, garages, manholes and the likes, just as we today search and navigate in alphanumeric text by means of words and their meaning. However, this needs a transition from the traditional off-line (passive) use of 3D data, as in the example of computing the cellular telephone signal propagation, towards interpreted urban models.

Urban 3D models have an approximately 15-year history, the beginnings perhaps marked by early workshops at Ascona (GRÜN, 1995; GRÜN, 1997). Providers of 2D Geographic Information Systems (GIS) of urban areas became interested in a 3D version of the GIS as the 2D data processing systems matured by the mid-1990s and new technologies began to make the use of and interaction with 3D data feasible and easy. No particular application stood out to dominate the specifications for 3D GIS. Fig. 1 is a typical 3D model used for wireless data transmission and signal propagation studies, an important application of such data since about 1998, requiring fairly generalized models of buildings or building blocks. Meanwhile, however, the developments have accelerated and 3D models of urban spaces are becoming ubiquitous, especially as an academic research field, as part of fully digital workflows in photogrammetry, and as application opportunities have evolved, for example in 3D car navigation (STRASSENBURG-KLECIAK, 2007).

1.1. Towards the Virtual City

In its simplest form, the "Virtual City" is a digital surface model (DSM) of the urban landscape with aerial photography draped over the DSM, and presented on a computer monitor for display and some limited interaction. In its most sophisticated form, each building, tree, street detail, bridge and

water body is modeled in three dimensions, details such as windows, doors, facade elements, sidewalks, manholes, parking meters, suspended wires, street signs etc. exist as separate objects. The detail is sufficiently complete with albedo, color and surface roughness for a photo-realistic visualization in a 3D immersive virtual reality environment.

When we refer to a state-of-the-art, we address the ability of creating models automatically and economically. A manual creation has been feasible since a long time and in fact was at the root of inventing early applications of photogrammetry in the late 19th century (architectural photogrammetry, GRIMM, 2007). That manual capability has evolved into a semi-manual approach to map landmarks (STRASSENBURG-KLECIAK, 2007).

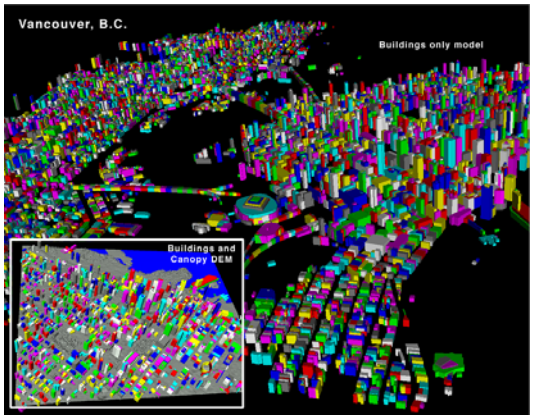

Fig. 1: 3D building models of Vancouver created at *Vexcel Corporation* (Boulder, USA) in 1998 for use in telecom signal propagation studies. Embedded is an example of the buildings placed on top of the digital surface model.

Today's automated abilities are well beyond the photo-draped point clouds of urban spaces, but are yet far from this vision of a fully interpreted model. We find today an intermediate solution with the bald Earth and its vertical objects separately modeled, but with rather simple plane surfaces for building shapes, without any interpretation of the details and with total reliance for the rendering on photographic texture. Only the vegetation gets classified and is being rendered separately. An example of this state-of-the-art is the *Microsoft* system http://www.bing.com/maps (formerly *Virtual Earth*, Leberl, 2009). One can today associate alphanumeric meta-data with entire buildings and individual trees, not, however, with individual building floors, doors or windows (see Fig. 2).

Fig. 2: A state-of-the-art automated urban model with building and terrain model as separate objects (© *Microsoft*). The object shown is the Colorado Capitol in Denver, USA.

Of course, the aerial image is not the only source for such models. Satellite imagery may offer some appeal, but generally is of insufficient geometric resolution and with inappropriate overlaps to serve as a workhorse for 3D urban models. Instead, there is frequent use being made of airborne laser

scanning to directly collect the geometric surface shapes rather than to compute those shapes from overlapping imagery. And as most applications of such models require a focus on the human scale, one is investing in data acquisition from the street level, also using imagery and laser scanners carried on the roof of cars.

A german initiative since about 2002 deals with standards based on a mark-up-language *CityGML* and describes the "Virtual City" by five levels, starting from a "Level-of-Detail-0/LOD-0" of the digital terrain model (DTM) or also known as the bald Earth. LOD-1 is the simple rectangular model of buildings and city blocks without any attention to photographic realism (see Fig. 1). This gets improved by a "LOD-2" with building blocks showing generalized roof shapes. "LOD-3" is the photorealistic full model of each building (Fig. 2), LOD-4 contains sufficient detail to enter a building (http://www.citygml.org; NICKLAS, 2007; WILLKOMM, 2009).

Perhaps one should extend this classification to go beyond the buildings and focus on the many human scale details of urban spaces such as sidewalks, vegetation, manholes, street signs etc. Such details will be contributed from street-side and even indoor sensor data. And one may well expect that it will be this level of the "Virtual City" that will in the end fulfill the expectations for a wide range of applications on the Internet and elsewhere.

1.2. Two-Dimensional versus Three-Dimensional

To map-makers, the transition from two dimensions to the third is straight forward and unambiguous. The 2D GIS represents a plane into which the 3D world is projected. This may contain the third dimension as an attribute, perhaps describing the height of a building much like a color or address. Obviously then, the 3D GIS will be an XYZ-model of the environment with attributes associated with certain 3D objects. For the purpose of this discussion, we neglect the (very important) difference between fully three dimensions and the often denoted 2.5 dimensions associating with any position in the plane XY only a single Z-coordinate. To visualize an urban space in an Internet application, one needs to transfer to the user's client the XYZ-coordinates of objects and create a rendering for the user in 3D, or such renderings are being produced centrally and sent to the user.

However, we can define a hybrid that does not operate with any 3D models of urban scenes. In computer graphics, one can simulate a 3D rendering by only using 2D data. Recent examples are *Photosynth* (AGÜERA Y ARCAS, 2007) and *Photo-Tourism* (SNAVELYS, 2006). The user's position is defined in a 3D space, as is the user's viewing direction and view frustum. But the user is at all times viewing 2D images taken of the entire hemisphere (or sphere). *Google Corporation* uses this approach in its street side imagery. At each of a discrete set of street positions, a set of 2D images will be collected. Viewing is supported by changes from image to image depending on the viewing direction of the user. This may sometimes be denoted as "Bubble viewing" or "Viewing in a Box". The basic concept may have been inspired by the work done to interactively control the viewing position when viewing video streams (ZITNICK, 2004).

The advantages of the hybrid 3D viewing via 2D images are obvious: no need for the creation of 3D models, 3D data structures, transfer of 3D models in the web to a user's client, no visualization of 3D content by a projection onto a 2D viewer. The 2D images serve as "pre-computed" 3D renderings for easy presentation to the user. The client does not have to process complex 3D data on a high performance graphics card.

The disadvantages are obvious as well. Relationships between data sources cannot be made smooth, one just is viewing images, there is no interpretation of the image's contents, the illumination is frozen, and the change of a user's positions does compromise the realism of the images that were collected from another position.

In our opinion, the "simulated 3D" by series of 2D images is a transitional phenomenon that will pass as automation of 3D models improves.

1.3 From Virtual Cities towards Virtual Habitats

Microsoft called its 3D internet system *Virtual Earth* (now http://www.bing.com/maps), obviously inspired by the large Earth with its thousands of cities, each one seen as a bird would. A "Virtual City" then appears as a logical part of this "Virtual Earth".

However, one easily forgets the human scale with its experience of walking, sitting, looking, doing sports, recreating, shopping and consuming. This is broader than "the Earth" or "the city". It addresses the street side, the interiors of buildings with their rooms, perhaps objects inside rooms such as art or merchandise, and it includes recreation areas with mountains, golf courses, bike paths, children's play grounds etc. While the ambition is to be relevant across the entire Earth, the relevance is to the humans. Therefore we propose to talk about the "Virtual Habitat" as a more general concept than the "Virtual Earth" or the "Virtual City".

Associated with this specification of the human scale are the detail of the data base and the detail available from sensors. Satellite, aerial photography or LiDAR scanning, if applicable, may be the source at a level of detail in the range of 10cm to 15cm. Street-side sensors would augment that detail at a level of perhaps 2cm, in order to represent signs and text on facades and shops. As one moves indoors, the detail will further increase to a level of perhaps 0.5cm to represent various objects. An example may be museums or religious places, and the inside of shops.

2. DATA

2.1 Aerial Imagery, either Vertical or Oblique

The initial emphasis in the creation of virtual cities was put on satellite and aerial photography. Use was and is being made of this material in the form of ortho-photos at geometric resolutions that were most economical; and in the form of 3D models of buildings. The ortho-photos were thus used as a background of a digital Earth model, but for economic reasons were oftentimes not specifically created but simply re-used from existing sources. Geometric resolutions in the range of 1 m per pixel were acceptable while of course pixel sizes in the range of 2 cm would be more appealing.

Satellite imagery offers pixels of up to about 50cm; these are very useful sources for ortho-photos, but the creation of 3D models is first being obstructed by an insufficient resolution, and secondly by an insufficient redundancy for 3D reconstruction. The typical pixel size for buildings today is at 12 to 15 cm. With the inclusion of an infrared color one uses four color bands and in the process improves one's ability to deal with vegetation. One uses ten images per building for 3D reconstruction.

An interesting phenomenon was the advent of oblique aerial photography to have a quick 3D view of buildings without having to construct a model of buildings. While oblique imaging has had a

long tradition in aerial mapping to overcome the optical limitation of wide-angle views, its substitute for 3D models is a result of the application of urban mapping to the Internet. Fig. 3 illustrates two oblique views of the same area, once imaged by a tilted aerial camera, the other using the oblique view inherent in the wide viewing angle of a vertical aerial photograph.

Fig. 3: View of facades, once from a tilted aerial camera (left), once from vertical aerial imagery (*Microsoft UltraCam*), but taken from the edge of such photographs.

2.2 Aerial Laser Point Clouds

Since perhaps the mid-1990's the aerial laser scanners (LiDAR) have become a very prominent source for digital terrain models, much to the detriment of further innovation in stereo photogrammetry. The aerial laser scanners produce point clouds that then need to be analyzed and converted to a 3D model of the bald Earth and of the vertical objects on top of the bald Earth.

While aerial photography collects data in rather wide swathes, the laser scanners cover much smaller swathes and at lower air speeds. This gives the photography seemingly a large productivity advantage. However, the laser requires a limited level of post-processing whereas photography needs to be fed into an elaborate photogrammetric processing workflow. However, with such workflows being highly automated, that factor is of reduced importance (LEBERL, 2009b).

2.3 Street Side Imagery and Street Side Lasers

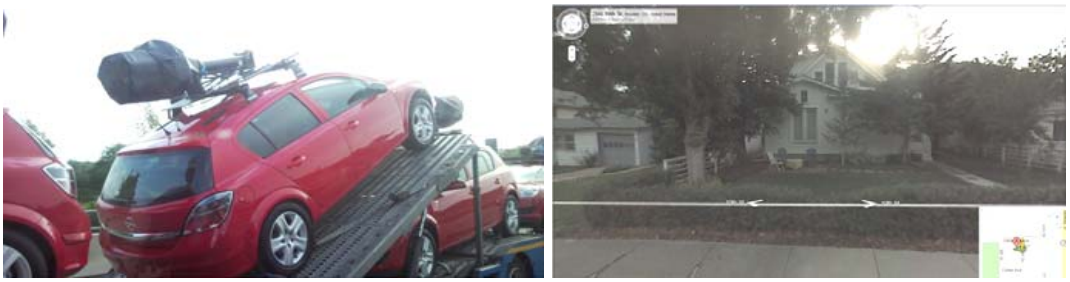

Fig. 4: Left: Snapshot of a *Google* systems for street side imaging as they are being transported on a german freeway. Right: *Google* Street-View of a house in Boulder, USA.

It is a reflection of the interest in the human scale that recent innovations do address the street-side, both by way of street side photography as well as by scanning lasers producing point clouds. Systems collect typically both camera images (taken in all directions to produce a hemi-sphere) and laser scans, and this gets done at intervals from a moving vehicle such that perhaps every 4m or so a new data set is acquired. This obviously will produce many images of a specific object point and helps to fill in occlusions. The laser point clouds are meant to replace the need for information

extraction from the images. Fig. 4 is a snapshot of street side imaging systems of *Google* being transported between Austria and Germany, and a view taken from such a system in Boulder, USA.

3. THREE DIMENSIONAL VISION

3.1 Triangulation

Both aerial as well as street side imagery needs to be placed into a world coordinate system. Global Positioning Systems (GPS) and Inertial Measuring Units (IMU) may not be sufficiently accurate, especially when the 3D shapes get reconstructed from overlapping images or if objects must be recognized in multiple images. In both cases one will benefit from sub-pixel matches of overlapping images. In aerial photography this will be in the range of a few centimeters, in street-side it will be at the centimeter- level.

The approach is via image triangulation using thousands of automatically collected tie points in any image overlap. This is a task of limited complexity when the image taking is in a very organized fashion, as is the case in aerial mapping. The same task becomes very difficult if one deals with amateur photography taken from any arbitrary position and into any direction, perhaps even without a well defined focal length, without any auxiliary GPS data, let alone any inertial direction information, as is typical of the *Photosynth* system (AGYERA Y ARCAS, 2007).

3.2 Dense Matching

Triangulation will typically compute the 3D coordinates of all tie points, resulting perhaps in several hundred points per photo. This does represent a “sparse point cloud” of the object. Dense matching builds on this sparse data set and densifies it to a level of a 3D point at intervals of 2 pixels or so. Per photograph one might obtain several million object points.

Dense matching employs all images taken of an object, and computes precise matches among the images so that from each image one obtains a geometric ray between image and object point. The object point gets intersected from these multiple rays. A common approach is to extract from the images a triplet and to compute a point cloud per triplet. Since many triplets can be formed, many point clouds will be computed. These then must get merged into a single seamless representation of the object by a fusion process.

3.3 DSM and DEM (Bald Earth and Vertical Objects)

The dense point cloud is an un-interpreted representation of the object surface. If this is the terrain, then we want to separate the points into those that describe the terrain surface and those that represent the buildings and vegetation. That separation can be based on image segmentation where the building roofs, the vegetation, the grass surfaces and circulation spaces get identified. Elevation above ground also is a factor entering into a process that results in a separate DTM and a DSM which comes from the points closest to the sensor, thus the aerial camera. Subtracting the DTM from the DSM will deliver the relative height information of the objects, which can be used as a discriminative feature modality for e.g. a semantic image classification task.

Fig. 5 shows a color image and relative pixel-synchronous heights, computed by using similar methods as proposed in (KLAUS, 2006) and (CHAMPION, 2006). For each pixel in the color image, a corresponding height value is extracted and directly used as additional information source.

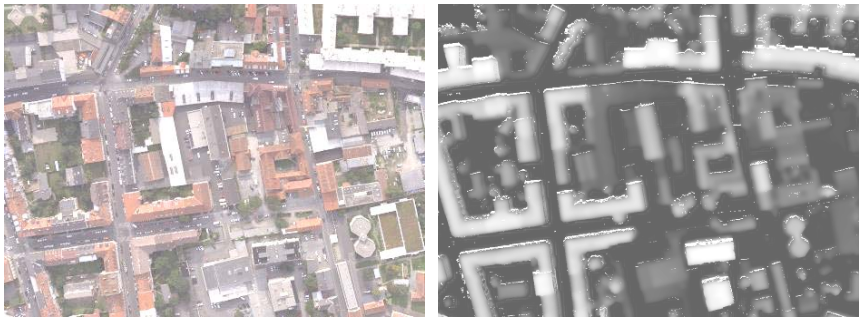

Fig. 5: A segment of an aerial color image of Graz, Austria and the corresponding relative height information, which is used as a discriminative feature channel.

4. SEMANTIC SCENE INTERPRETATION

We argued previously that the early work for VR Cities (which focused on photorealistic shapes of urban landscapes) is now being improved to create complete models of all the objects making up the urban landscape on a human scale. A visualization of an urban scene will then be created from the complete description of its elements. To get there, we need to detect and describe the objects based on color, texture, height etc. Here, we highlight two approaches to get towards a full semantic scene interpretation.

First, we use a car detection procedure for multiple reasons. Objects, like cars, are irrelevant for the virtual modeling of scenes since they are not a permanent feature of the terrain. Therefore, a detection mask is being used to provide information about image regions that are irrelevant for the 3D modeling. On the other hand, the car detections can be exploited as context knowledge in a later processing step e.g. to distinguish between parking lot areas and driving surface. Second, we illustrate an efficient approach for full semantic classification based on randomized forests by integrating appearance and height information.

4.1 Car Detection

The car detection problem in the aerial images is treated as a binary classification problem – with either an object being a “car” or “background”. To automate such a search one would traditionally need a large number of pre-labeled data, perhaps in the order of ten-thousand images. However since we have an on-line learning method (GRABNER, 2006), which is sufficiently fast for interactive work, we deal with the training as an interactive learning problem. The key idea is that the user has to label only those examples that are not correctly classified by the current classifier. We evaluate the current classifier on an image. The human supervisor labels informative samples, e.g. marks a wrongly labeled example which can either be a false detection or a missed car. The new updated classifier gets applied again on the same image or on a new image, and the process continues iteratively until a satisfactory detection performance is achieved. After training, the overall detection results from the exhaustive application of the trained classifier on the images.

The details of this approach are described in (NGUYEN, 2007; LEBERL, 2008). See Fig. 6 for a car detection result on a scene extracted from the dataset Graz, Austria. A quantitative evaluation on hand-labeled ground truth images gives a detection rate of approximately 85% with a false positive rate of 15%. Future work will employ more information than just a single color image.

Fig. 6: The resulting car detection is being superimposed on an aerial image of Graz, Austria.

4.2 Semantic Aerial Image Classification

A semantic classification is being achieved by means of segmentation of individual images, and then refining that segmentation by consideration of the results from overlapping images. In the following sections we highlight an efficient technique obtaining a semantic large-scale interpretation of aerial imageries, for use in virtual modeling urban environments. The classification pipeline is mainly based on previous work presented in (KLUCKNER, 2009). The work investigates an approach to compactly integrate various feature channels by using the powerful covariance descriptors (PORIKLI, 2006) within randomized forests (BREIMAN, 2001). We showed in (KLUCKNER, 2009) that this concept obtains competitive results, compared to state-of-the-art classification methods (SHOTTON, 2008; SCHROFF, 2008), on a standard evaluation dataset like the *Microsoft Research Cambridge* (MSRC) image collection.
The suggested method provides several advantages that are exploited for large-scale computations in aerial imagery: (a) Randomized forests have proven to give robust and accurate results in challenging multi-class classification tasks (SHOTTON, 2008 and SCHROFF, 2008). The forests as classifiers are very efficient at runtime since the concept is based on fast binary decisions between a small number of selected feature attributes. (b) The classifier can be trained on a large amount of data, avoids over fitting and can handle some label noise (errors in the training data), which is appropriate in case of generating large-scale training maps. (c) Since the aerial imagery consists of multiple information sources, such as color, infrared, range and panchromatic data, there is a need to reasonably combine these feature cues. We therefore use a feature representation based on covariance descriptors to compactly describe the channels including a small local neighborhood.
(d) In addition, as proposed in (PORIKLI, 2006), these descriptors can be quickly computed using integral image structures and support parallel computation techniques.

Exploiting these advantages and extending the work presented in (KLUCKNER, 2009), we apply the concept based on randomized forests with a covariance feature representation to perform a semantic classification in challenging real world aerial images. To obtain results in terms of correctly classified pixels, we manually label maps that provide the ground truth information for training and testing. Fig. 7 presents the semantic classification at the pixel level of an aerial image segment for the classes “building”, “tree”, “green area”, “road”, and “water body” using an integration of color, texture and height information.

A single image can be processed within two minutes at a full resolution of 11.500x7.500 pixels. Moreover, the evaluation procedure considering the labeled ground truth data reports an average rate of approx. 90% classification accuracy. A similar approach and more results of an extensive evaluation are currently under peer-review for a major computer vision conference.

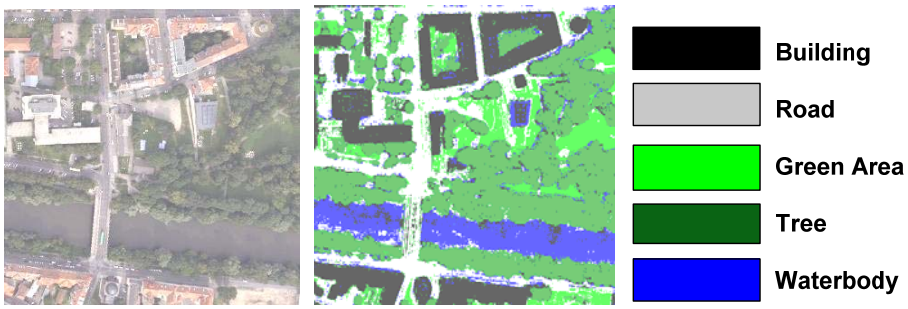

Fig. 7: Aerial color image segment and the resulting semantic classification on a pixel level by integrating color, texture and elevation. The classes are “building”, “tree”, “green area”, “road”, and “water body”.

Since our classification pipeline yields a semantic interpretation for each image in the dataset, we obtain highly redundant observations from different viewpoints for each pixel on the ground. By exploiting the available range data and camera parameters, the obtained per-image classification results are fused into an ortho-view image that includes a semantic interpretation of the full scene that is then directly used to construct the virtual 3D model. The fusion step simultaneously results in robust height information, including the DTM and the DSM data and certainly the color image.

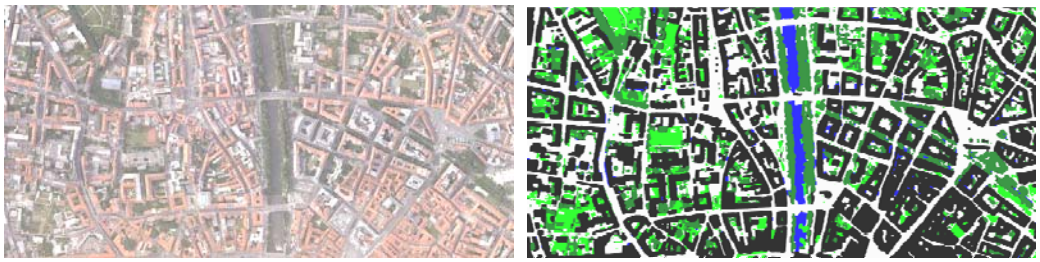

Fig. 8: The resulting large-scale classification using our fusion pipeline - the computed semantic interpretation results integrates 56 images.

Since the fused semantic classification is so far provided on pixel level, we apply an efficient technique based on conditional random fields (KOMODAKIS, 2007) to obtain a smooth final class labeling by integrating spatial neighborhood information. A quantitative investigation how a large-scale fusion of multiple per-image results influences the overall classification accuracy is part of current research; however we obtain promising visual results. Fig. 8 depicts a fused large scale classification and the corresponding color image. The result is computed within two hours on a standard PC and consists of the information of 56 single full resolution aerial images.

4.3 Towards Virtual Modeling

Taking into account the computed 2D semantic interpretation, we can construct a (yet simplistic) 3D model by additionally considering the available 3D height information. According to the obtained interpretation, specific processing steps yield a full virtual scene as shown in Fig. 9. Moreover, we use the predictions for the class “road” to automatically compute a full street-layer network, represented as an efficient graph data structure. Such graph provides important information for e.g. navigation, rendering of scenes, etc.

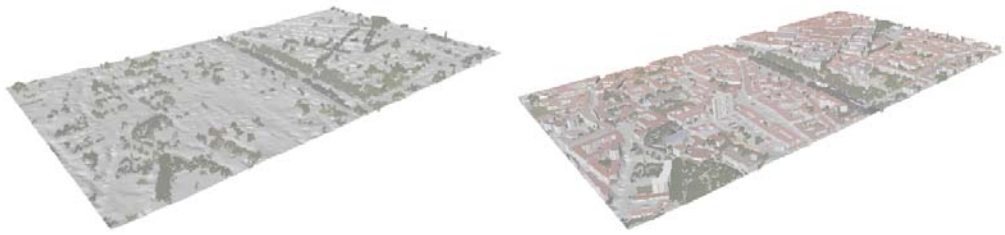

Fig. 9: A virtual 3D model of the center of Graz, Austria is generated by taking into account the semantic classification and the available 3D height information. The model on the left presents the raw DTM and the class “tree”, each tree being modeled as an individual point cloud. The right model includes the extracted building blocks, assigning to each block a mean color and height.

5. APPLICATION

5.1 Computer-Generated Photorealism From 3d Models

Our interest in urban spaces is partly driven by the use of those spaces in Internet-applications. This recognizes the “user” of urban data as someone who is initially seeking a satisfying visual experience on the Internet. This very much drives the desire for photorealism for all 2D and 3D data used on the Internet. While this is the current state-of-the-art, we can see already that down the road the user will want to search through the urban data.

While initial photographic texture would be expected to capture shadows and sun illumination, or rain, in a specific way just for a given moment, we expect computer generated realism to render for any illumination, any sun angle, any weather or perhaps even any season.

Apart from this added flexibility, there is an advantage of smaller data quantities that need to be stored, retrieved and transmitted to a user.

Finally, creating the rendering from object models supports the idea of user-supportive generalization. Objects of importance such as schools can be visualized to a user at a level of detail and scale that differs from other objects of lesser importance to that user and the application.

5.2 Search

The look for information about any topic on the Internet has become the most dynamic development in the history of mankind. While software became a well-defined and separate industry over a time span of 30 years, starting in the mid-1970’s, the search for information on the Internet has grown into a full-blown industry in a period of perhaps only five years. This reflects the human’s hunger for information and knowledge.

Having the search associated with a location makes eminent sense. This can be because the user is looking for a location or for directions, or because there is a decision pending on which location to choose from a potentially very large choice, or because there is a location-reference to a place in another context. Is something nearest to something, or the most elegant, or the largest, or does it serve a specific purpose in lieu of a more general orientation, etc?

As search continues to grow in importance and becomes continually more sophisticated, its location data also will need to grow in flexibility and versatility, thus in the usefulness to as many different applications as possible. Why not clicking on a window and learning whatever the Internet holds about the goings on behind that window?

5.3 Navigation

Car navigation has become ubiquitous, and anyone traveling in unfamiliar areas will attest to a growing dependence on the electronic navigation aid.
Strassenburg-Kleciak (STRASSENBURG-KLECIAK, 2007) expressed that 3D car navigation could be a reality, were it not for the absence of large area 3D urban models. Personal navigation via the smart phone will follow. If there are any ambiguities in navigation, then it results from a lack of realism of the map, or from its generalization and lack of detail. Navigation will thus benefit from a more detailed, 3D model of the world and the human habitat. This justifies a transition from current symbolized street maps to realistic 3D models.

5.4 Others

Willkomm (WILLKOMM, 2009) presents an interesting overview of commercially viable applications for 3D urban data. Games, e-commerce, the Internet-of-things, ambient living all are additional applications for the “Virtual Habitat”. Real estate today can practically no longer be rented or sold without a presence in the web. This may not necessarily be using 3D; we already pointed to the simulation of 3D by series of 2D images.
Games may drift from invented, non-existing locales to actual neighborhoods modeled digitally. This represents a soft need for VR cities.

The Internet-of-Things associates with each object an Internet-visible radio-identification. This in turn needs a location to be obtained for such objects. If we describe the location of an object, we need a description of the object’s position and possibly attitude. A wallet is visible when we know that it is on top of a table in a specific room inside a known building. In analogy, ambient living is built on a model of the living space of a person (for example is a person in a bed in a room in a building).

6. OUTLOOK

At issue is the transition from the mere visual focus of shapes dressed up with photographic texture towards a model of the parts an object is composed of, with material properties and an interpretation of an object’s function.

The transition will lead to a totally different visualization of and interaction with objects and cityscapes. A “Virtual City” will get shown as a computer-generated view of its parts, built bottom-up. Searches can address buildings by properties such as height, floors and other characteristics. Searches can concern objects in a city’s streets such as parking meters, sidewalks, bicycle stands, man holes.

The work towards a fully interpreted “Virtual City” has only just begun – we are at the beginning of a long journey. That journey consists

- of technology developments to convert sensor data into urban models and into interpretations of these model’s contents,
- of the actual creation of the models of thousands of cities,
- of means to update and keep current what has previously been created,

– of ways to involve the large community of users in providing sensor data, in flagging of - errors and in correcting the errors in the "Virtual City" creation.

ACKNOWLEDGMENTS

This work was supported by the Austrian Science Fund Project W1209 under the doctoral program Confluence of Vision and Graphics and the APAFA Project No. 813397, financed by the Austrian Research Promotion Agency (www.ffg.at).

REFERENCES

Agüera y Arcas, Blaise (2007):
http://www.ted.com/talks/blaise_aguera_y_arcas_demos_photosynth.html

Breiman L. (2001): Random forests. Machine Learning, pp 5–32

Champion N. & D. Boldo (2006): A robust algorithm for estimating digital terrain models from digital surface models in dense urban areas. In Proc. of the ISPRS Commission 3 Symposium "Photogrammetric Computer Vision"

Grabner H. & H. Bischof (2006): On-line boosting and vision. In Proc. IEEE Conference Computer Vision and Pattern Recognition

Grimm A. (2007): The Origin of the Term Photogrammetry. In Proc. 51st Photogrammetric Week. Dieter Fritsch (ed.), Wichmann-Verlag, pp 53-60

Grün A., O. Kübler, & P. Agouris (1995): Automatic Extraction of Man-Made Objects from Aerial and Space Images. Ascona (CH), Birkhäuser Verlag Basel, Boston, Berlin.

Grün A., E.P. Baltsavias, & O. Henricsson (1997): Automatic Extraction Of Man-Made Objects From Aerial And Space Images (II). Ascona (CH), Birkhäuser Verlag Basel, Boston, Berlin.

Klaus A., M. Sormann, & K. Karner (2006): Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In Proc. of International Conference on Pattern Recognition

Kluckner, T. Mauthner, & H. Bischof (2009): Semantic image classification using consistent Regions and individual context. In Proc. British Machine Vision Conference, 2009, to appear

Komodakis N. & G. Tziritas (2007): Approximate labeling via graph cuts based on linear programming. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(8): pp 1436–1453

Leberl F., H. Bischof, H. Grabner, & S. Kluckner (2007): Recognizing Cars in Aerial Imagery to Improve Orthophotos. In Proc. 15th ACM International Symposium on Advances in Geographic Information Systems. Seattle, Washington, USA, ISBN 978-1-59593-914-2/207/11, pp. 2-10

Leberl F. & M. Gruber (2009a): 3D-Models of the Human Habitat for the Internet. In Proc. Visigrapp, Lisbon, Portugal, ISBN 978-989-8111-69-2, Volume IS, pp 7 –15

Leberl F., A. Irschara, T. Pock, P. Meixner, M. Gruber, S. Scholz, & A. Wiechert (2009b): Point Clouds from Laser Scanning and from 3D Photogrammetry. Internal Manuscript, Graz University of Technology.

Nguyen T., G. Helmut, B. Gruber, & H. Bischof (2007): On-line Boosting for Car Detection from Aerial Images. In Proc. IEEE International Conference on Research, Innovation and Vision for the Future.

Nicklas D. (2007): Nexus – A Global, Active and 3D Augmented Reality Model. In Proc. 51st Photogrammetric Week. Dieter Fritsch (ed.), Wichmann-Verlag, pp 325-334

Shotton J., M. Johnson, & R. Cipolla (2008): Semantic texton forests for image categorization and segmentation. In Proc. IEEE Conference Computer Vision and Pattern Recognition

Snavely N., S. M. Seitz, & R. Szeliski (2006): Photo tourism: Exploring photo collections in 3D. In ACM Transactions on Graphics, 25(3), pp 835-846

Strassenburg-Kleciak M. (2007): Photogrammetry and 3D Car Navigation. In Proc. 51st Photogrammetric Week. Dieter Fritsch (ed.), Wichmann-Verlag, pp. 309-314

Tuzel O., F. Porikli, & P. Meer (2006): Region covariance: A fast descriptor for detection and classification. In Proc. European Conference on Computer Vision

Willkomm P. (2009): 3D GDI - Automationsgestützte Erzeugung und Verteilung landesweiter Gebäudemodelle aus Laserdaten. In Proc. 14th Münchner Fortbildungsseminar GIS, Technische Universität München

Zitnick C.L., Sing Bing Kang, M. Uyttendaele, S. Winder, & R. Szeliski (2004): High-quality video view interpolation using a layered representation. In ACM Transactions on Graphics, 23(3), pp 600-608