Automated photogrammetry for three-dimensional models of urban spaces

Franz Leberl
Philipp Meixner
Andreas Wendel
Arnold Irschara
Graz University of Technology
Institute for Computer Graphics and Vision
Inffeldgasse 16
A-8010 Graz, Austria
E-mail: leberl@icg.tugraz.at

Abstract. The location-aware Internet is inspiring intensive work addressing the automated assembly of three-dimensional models of urban spaces with their buildings, circulation spaces, vegetation, signs, even their above-ground and underground utility lines. Two-dimensional geographic information systems (GISs) and municipal utility information exist and can serve to guide the creation of models being built with aerial, sometimes satellite imagery, streetside images, indoor imaging, and alternatively with light detection and ranging systems (LiDARs) carried on airplanes, cars, or mounted on tripods. We review the results of current research to automate the information extraction from sensor data. We show that aerial photography at ground sampling distances (GSD) of 1 to 10 cm is well suited to provide geometry data about building facades and roofs, that streetside imagery at 0.5 to 2 cm is particularly interesting when it is collected within community photo collections (CPCs) by the general public, and that the transition to digital imaging has opened the no-cost option of highly overlapping images in support of a more complete and thus more economical automation. LiDAR-systems are a widely used source of three-dimensional data, but they deliver information not really superior to digital photography.

Subject terms: photogrammetry; three-dimensional geographic information systems; virtual cities; location-aware internet; aerial cameras; streetside imaging; façade mapping; roof mapping.

1 Urban Space Models in Three Dimensions

One refers to geo-virtual environments (Geo-VE) when describing a modern geographic spatial data infrastructure, and three-dimensional city models are a central part of this environment, according to Döllner. Urban information spaces by their very nature are complex, and extending the very successful two-dimensional geographic information system (GIS) to the third dimension is a natural progression that merely depends on the appropriate innovations in information technology to become a reality. That innovation has occurred, and as a result we now experience a formidable increase in research and implementation of virtual three-dimensional cities. While initially it appeared as if the augmentation of the Internet by location was the main application, it now rather is a large assembly of applications in urban life, engineering, and planning, be it real estate and real estate value assessments, the placement of solar panels, citizen participation in planning new construction, car and pedestrian navigation, micro-climate modeling, noise propagation, and so forth.

1.1 Photogrammetry

Photogrammetric methods have long been used to map buildings and monuments in three dimensions from photographic sources. In fact, the word “photogrammetry” was coined in Berlin by Meydenbauer as early as 1867 at a time when the application to architecture was a driving force for the development of the field. However, those geometric descriptions of buildings were obtained by manual operations and modestly supported by some mechanical-optical tools. Naturally, then, the models addressed only unique buildings and were associated with significant costs for labor. The major application of photogrammetry was ultimately the mapping of large areas of the Earth’s surface. It reduced the cost of previous approaches when the only alternative was to explore the world on foot and to collect data with traditional surveying tools. Maps of buildings remained an exotic sideline for the rapidly evolving field of photogrammetry, since the topographic application did not map any building details.

1.2 Digital Cameras and the Internet

It took the advent of digital cameras to inspire the development of fully automated workflows for the creation of terrain maps. The digital camera made it possible to design end-to-end automation to achieve the transfer from a large assembly of overlapping images into a three-dimensional geometric model of an area or object of interest, with a cost advantage over previous manual methods by two orders of magnitude. It took the Internet to find an application important enough to actually implement the new technology to three-dimensional models of extended urban areas with tens of thousands of buildings, at a cost of sometimes less than $1 per building (Fig. 1). The driver is the location-aware Internet where “place” gets associated with every search (Fig. 2) and where

*This $1-estimate derives from a productivity of one city per day [or about 350 per year] with 100,000 buildings each on a budget of $35 million. This finances the purchase of the aerial source photography as well as a team of 20 people for supervising the automated work flow.
navigation is an Internet application. Initial implementations within Internet search engines have become known as Google Maps, Google Earth, and Microsoft Virtual Earth, now Bing-Maps. These capabilities have existed since 2005, as reported by Gibson.\(^5\) Car navigation as a three-dimensional application is on the horizon at companies offering car navigation systems, as reported by Strassenburg-Kleciak.\(^6\)

Search and navigation concern the users at a human-scale level of detail. This addresses objects down to a size at 10 cm and less. The source photography must be at geometric resolutions of 10 cm and even down to 3 cm. One wants to be able to read and document street signs, model very thin objects such as suspended wires, and be able to navigate merchandise and thus interact with individual shops. The vision includes a virtual reality experience with a human exploring an urban space at street level and going also inside buildings and shops. Bill Gates presented this vision in 2005 at the occasion of his 50\(^{th}\) birthday in London (for an example of a report on this event, see Ref.\(^7\)).

1.3 Range of Source Data and Platforms

A range of source data and sensors exists: conventional airplanes, driving vehicles, remotely controlled aerial microvehicles (MAVs), or indoor tripod-based scanning and manual photography. It remains unclear at this time which platforms, sensors, and type of imagery support the most efficient creation of urban two-dimensional models. For example, are point clouds from overlapping (stereo-) imagery less efficient than point clouds obtained from LiDAR measurements, be they taken from aerial platforms, driving vehicles, or stationary tripods? Particularly intriguing are the emerging CPCs on the Internet: these are images taken by us all, for example, as tourists with generic cameras and no special provisions for geometric accuracy, but at high overlaps and thus redundancy. Such CPCs offer a built-in update of aging data sets. The technology to take advantage of these sources for the creation of three-dimensional models exist and have been described on various occasions, for example, by the teams around Snavely\(^8\)–\(^10\) or Gösele.\(^11\)

1.4 Vision for Computer Models of Urban Spaces

The Internet search engines in their location-aware implementation were seen as the initial “killer” application of three-dimensional urban models. From 2006 until 2008, hundreds of cities were modeled in three dimensions and offered in support of Internet search. The user echo has been modest; monetizing of three-dimensional building models through the support of Internet search remained elusive. Therefore, the focus has shifted from the full three-dimensional models of all buildings of a city to the presentation of streetside imagery that is organized in a three-dimensional immersive visualization per street, and combined with photo-maps obtained from aerial or satellite platforms and imaging sensors. Agüera y Arcas\(^12\) has presented very persuasive evidence on behalf of Microsoft of the applicability of this work. Google’s street-view system was the first to go global with early versions of these technical approaches. While Microsoft is currently following suit in the massive Internet application, it has for a considerable time pioneered research on combining streetside images with oblique aerial imagery and the original three-dimensional models. Today it has become common practice to experience a smooth bird’s eye visualization of an urban space from the air (Fig. 3), so that the individual source images are no longer identifiable to a user.

Internet search has been a major driver of the development of urban three-dimensional mapping technology. Municipal administrations have, however, embarked on an independent evolution of the two-dimensional GIS into a three-dimensional urban GIS in the form of a data infrastructure analogous to the two-dimensional GIS. Certain municipal

![Fig. 1](https://example.com/fig1.png)

**Fig. 1** 3D model of the Vienna State Opera, produced fully automatically from overlapping aerial photography (www.bing.com/maps, October 2009).

![Fig. 2](https://example.com/fig2.png)

**Fig. 2** Searching the address “Opernring 4, Vienna” in a three-dimensional model (www.bing.com/maps).

![Fig. 3](https://example.com/fig3.png)

**Fig. 3** Microsoft’s bird’s-eye visualization of Manhattan merges multiple oblique photographs across a vast area. This is based on the use of three-dimensional models for a smooth transition from one to another bird’s eye image (www.bing.com/maps).
applications require the third dimension, such as disaster preparedness, citizen’s participation in urban planning decisions, and municipal engineering.

On the horizon is ambient intelligence as a driver toward human-scale urban three-dimensional models, both of the outdoors and in particular of the indoors. In this context “location” is an integrated component of tracking, and awareness of things and people. Once every manufactured object is encoded with a radio-enabled identifier and once these identifiers are readable everywhere, all these objects can be tracked at all times. With the object information available on the Internet and with an object’s Internet address available to “smart phones,” one has access to every such object’s location, independent of one’s own location. One can begin to think of automated interactions with every object, as proposed recently by O’Reilly and Batelle, but long ago envisioned by Mark Weiser.

1.5 Standardized Levels of Detail and Standards for Representation and Exchange

The activity with urban three-dimensional models has spread to an extent that has called for international standards in data formats and definitions of the quality and detail of such models. Some of the initiatives originate from Germany with its strong tradition in detailed urban mapping, as reflected in the biannual Photogrammetric Week. It is there where Kolbe et al. have been discussing proposals for standards in three-dimensional urban models. Five levels of detail (LOD) have been defined to extend from the 2.5-dimensional representation of a building in LOD-0 via block-type buildings with flat roofs in LOD-1 to models with the roof shape represented in LOD-2, then on to details of facades with windows in LOD-3, and finally LOD-4, which includes the interiors of buildings (Fig. 4).

2 Technology Drivers

2.1 From Film to Digital Cameras

The aerial digital camera started to take over from the aerial film camera from about 2003 forward. This was late when compared to a transition from film-to-digital in consumer cameras. However, the digital cameras could not get introduced until images were available at a repeat rate and pixel-coverage similar to film cameras. This was with an 11,500 Pixels swath width and at ~100 MPixels per image. Additionally, the repeat rate of about 2 sec was needed to achieve high image overlaps along as flight line. Separately, computing and storage technologies needed to cope with the resulting demand for 3 GBits/sec data rate. Finally, digital cameras needed to be robust in a rugged aerial environment. All this delayed the advent of digital aerial cameras.

Current large-format digital cameras offer 28,000 pixels swath width, for example, in the implementation of the UltraCam-G camera system (Fig. 5 and the report by Gruber), a sensor and associated workflow focused on continental orthophoto production. For full three-dimensional mapping, the UltraCam Eagle today offers 20,000 pixels swath width, and an image size at 260 MPixels produced at a repeat rate of 1.8 sec, as reported by Wiechert et al. This increases the internal camera data rate to 3.7 GBits/sec.

Initially, the digital aerial camera was seen as just replacing the scanned digital version of a film image by a native digital image. Therefore, users were applying the same processing workflows in use with film. However, it was soon discovered that the aerial imaging paradigm was changing. First, one needed to understand that in the film era, the number of photographs for a given project was minimized due to the cost per image for film, film development, and scanning. That cost has been completely eliminated in digital sensors. The number of images is no longer relevant to a project’s cost of image collection. Second, the radiometric performance of digital images with 12 to 13 bits per color channel far surpassed what was available from film images with at times a mere 6 to 7 bits. As a result, automated matching became feasible in image areas that previously had insufficient contrast and detail. Also, weather became a somewhat less disconcerting issue when images could be taken in less than ideal light conditions. Third, film grain had disappeared, resulting in a higher stereo matching accuracy and better imagery (Fig. 6). Fourth, an infrared channel had become standard with digital cameras and four-channel imagery.

![Fig. 4](image1.png)

Fig. 4 Levels of detail of urban three-dimensional building models, as defined within CityGML (from Ref. 15).

![Fig. 5](image2.png)

Fig. 5 UltraCam-G aerial camera with swath width at 28,000 pixels. This camera is generating the source imagery for the Advanced Ortho Aerial Program AOAP to produce orthophotos with pixels at 30 cm of the entire continental United States and Western Europe within a time span of 36 months (Digital Globe and Microsoft).
was collected when previous film images only represented three color channels.

The fixed labor cost of processing an image initially invalidated the advantage of add-on images. Full automation of image triangulation and stereo matching was needed to eliminate that labor cost. Interestingly, this full automation was to be productive only when the overlaps among images were increased so that a given object was not imaged a mere two times under the “stereo” paradigm, but 10 times or more. Full automation really functions sufficiently well when manual error corrections become a minor effort and cost item. Full automation also functions well when high image overlaps are available. On the other hand, for high image overlaps to be achievable economically, full automation is needed to eliminate any variable labor cost per image. The increase in the number of images for a project is further extended by a reduction of the pixel size and improvement of the geometric resolution. Images today are taken at lower flying heights or get taken nearer to the object, so that smaller pixels are being produced than ever before.

2.2 Data Quantities Increase by Two Orders of Magnitude

Today, terabytes of imagery have become commonplace with the associated low cost of storage. When photogrammetry initially embarked on digital sensing using the first aerial large-format cameras in 2003 to 2004, the data volumes oftentimes did overwhelm the camera users. Previously, the community was used to projects with hundreds of photographs. For example, the area of the city of Graz (Austria) with about 150 km² would be imaged onto perhaps 150 aerial color photographs at a geometric resolution or GSD at 20 cm. A precision film scan with 20-μm pixels produced 132 MPixels per color photograph. With 150 photographs, the urban mapping project would use source images at a volume of 60 GB.

With digital cameras, that same surface area will be imaged with image forward overlaps increased from 60 to 80% (or even 90%), and the sideways overlap from 20 to 60% (or even 80%). In addition, one will want to get a higher geometric resolution and go from the 20-cm pixels of the film era to 8 cm in the digital era. One will now have increased the number of images to 3000 and the data set will have increased to about 2.3 TB.

In 2003, these terabytes and the associated data processing intimidated a typical aerial mapping organization. Only 8 years later, the terabyte has lost its power of intimidation. At less than €100, a terabyte has become a negligible cost item, and the size and power consumption of disk storage has become greatly more affordable. Urban mapping projects with 10,000 images have therefore become accepted.

2.3 All-Digital Workflow Using Novel Algorithms

Upon completion of image taking, one has to find relationships among the large number of overlapping images of a project. This is based on features found in the image overlaps at a sub-pixel accuracy and represents the computation-intensive part of any triangulation. The process consists of searches through all the images to find match points between any two or more overlaps. Once those points are found, their coordinates are submitted to a bundle adjustment. When film was used, manual labor was required at a level of about 1 h per image to get the matching accomplished. When subsequently the film was scanned and semiautomated processes could be implemented, that manual labor requirement was reduced to perhaps 10 min per image.

Today, this triangulation has been fully automated so that commercial products now exist to perform the triangulation without any manual labor. Microsoft at www.microsoft.com/ultracam may have been the first vendor of fully automated commercial triangulation software. The challenge is in the search and in succeeding to find relevant match points in slightly dissimilar images taken from slightly differing vantage points of a given object and area.

Subsequently, we need to describe the three-dimensional shape of the imaged surfaces by collecting a dense irregular cloud of three-dimensional points. This again gets accomplished by a matching process that results in homologue image points. The projection rays associated with those image points intersect in surface points. A point cloud is called dense when the spacing between neighboring surface points is near a single pixel. Previously, such points were individually found by matching two overlapping images, and spacing the matches by about 10 or 20 pixels. With high overlap images, each object point is shown on 10 or more images and we achieve an effect of “super-resolution,” supporting a match density at 1 to 2 pixels. The matches are subjected to an optimization process to directly compute a surface that satisfies smoothness constraints and employs all images of any given object. Those procedures result in surfaces without holes and spikes, and with minimum levels of noise. However, the process is computation-intensive.

This represents a transition away from the traditional two-image stereo match and toward the use of so-called global optimization approaches based on the calculus of variation, and employing as many images per object point as one has acquired. These methods have been developed in computer vision and have recently found acceptance in photogrammetry, as reflected by the contributions at the most recent Photogrammetric Week in Stuttgart in 2011, and the award presented to H. Hirschmüller for bringing the new methods into photogrammetry.
Needless to say, without the novel global optimization algorithms for multi-image surface modeling, the irregular image coverage from CPCs, also denoted as "crowd sourced modeling," would simply not be feasible. The same applies to surfaces modeled from the very large number of small images collected from unmanned MAVs.

2.4 Computing Power and the Graphical Processing Unit

The automated process to achieve both the triangulation and the dense image matches can quickly take an effort of about 1 h per photograph on current PCs. A 3000-photo project for a modestly sized city like Graz would thus be expected to run for nearly a half year on a single PC, an entirely unacceptable resource constraint.

The increase of computing power by the advent of the graphical processing unit (GPU) has been an innovation of great impact on computer vision. Throughput increases by a single GPU can reach a factor of more than 100, and multiple such GPU components can be hosted on a single PC. While this factor (~100) is possible, real throughput increases of 30 are being achieved today in aerial photogrammetry applications per GPU. It is conceivable that a 3000-photo project can now be condensed into automated processing within a single day with an appropriately designed GPU-supported computing system. Früh et al. report a GPU-supported, single-day, single-PC throughput of three million images. However, these are from CPCs and thus certainly two to three orders of magnitude smaller than large-format aerial photographs. Yet, this corresponds with the gain in throughput also achievable in urban mapping from thousands of large-format digital aerial images.

3 Sensor Data

3.1 Satellite and Aerial Photography

The traditional source of urban mapping data is the aerial photograph. Recent increases in GSD of satellite cameras are producing pixels in the 50-cm range, based on restrictions imposed by government rules. A pixel size of 30 cm may become a reality if the rules change. This resolution has given rise to applications in orthophoto-generation and has caused the occasional question whether satellite sensing will render aerial sensing obsolete.

Clearly, satellite imagery is useful in denied regions where urban imaging is politically not feasible. Outside those denied areas, sustained urban three-dimensional mapping will want, first, a geometric resolution higher than 50 cm, going toward human-scale detail in the 10-cm range and smaller. To read signs on facades, to detect suspended wires and street signs, the appetite for detail may grow to 2 to 3 cm. Second, urban mapping will also call for covering each urban object by many photographs to achieve full automation, not just one coverage for an initial visual presentation or two for a stereoscopic analysis. Finally, urban mapping will want to look at façades, thus off the vertical. These requirements do not favor the satellite-borne cameras. Figure 7 relates recent imagery at a GSD of 30 cm to another image at 8 cm, and also illustrates an image at 0.5 cm.

Aerial imaging is dependent on an aerial platform, typically a fixed-wing airplane, to fly at low heights over an urban area. Air traffic control and security concerns require the flying height to be in excess of 300 m above ground. With recent charge-coupled devices (CCD) area arrays offering a pitch of 5.2 μm, a GSD pixel size of 3 cm will be achieved from 600 m above ground, using a 100-mm focal length. It is yet to become commonplace to go after such small GSD from an aerial platform, but the sensor, storage, and computing technologies exist and legal issues do not obstruct the use of a very detailed GSD in North America and Western Europe.

A mid-sized city with its 150 km² will be covered from the air by a flight pattern consisting of flight lines totaling 250 km or so. This can be achieved in a half day, including mobilization from a home airport. The cost of an airborne project may be at about €1000 per aerial hour to include the plane, the pilot, and navigator/sensor operator, as well as the aerial sensor. The mid-sized city will get imaged for a cost of perhaps as low as €5000.

3.2 Streetside Photography

The desire for a human scale pedestrian experience of urban spaces on the Internet, in combination with an immersion in the three-dimensional urban World, have led to the advent of streetside imagery as a major source for urban mapping data. The initial impetus was for industrial systems on cars to drive through all city streets to collect images of façades. A small industry has emerged with cars and vans carrying multiple cameras on their roof. Figure 8 is an assembly of some sample industrial systems.

Imaging aims at the collection of panoramic views, and thus employs multiple centrally perspective cameras such as the LadyBug (www.ptgrey.com). Cycameleon (www.cyclemidia.com) pioneered the commercial use of 360-deg panoramic views of urban spaces and has been relying on fish-eye imaging, since this approach reduces the number of individual images being collected and therefore simplifies geometric processing. Panoramic views combining a conventional camera with a spherical mirror carry this

Fig. 7 Comparing geometric resolutions of urban aerial imagery. (a) parked plane at 30-cm pixel size at San Diego airport (Digital Globe); (b) parked plane at 8-cm pixel size at Graz airport (Microsoft-Graz); (c) skylight in Graz on UltraCam-aerial photo at 8-cm pixel size; (d) skylight on MAV photograph at 0.5-cm pixel.
simplification even further: on a single image, one can collect the entire hemisphere, a frequently proposed approach for indoor imaging employing what is called catadioptric cameras, and oftentimes applied within vision systems for robots.

The cost of imaging is a function of the kilometers driven by the vehicle. The mid-sized city with its 250,000 inhabitants may have 1000 km of streets that will nominally be driven at a velocity of 30 km/h and thus take a net 33 h. However, traffic, one-way streets, and weather constraints (no rain) will extend this to a velocity of well under 30 km/h. The total cost of coverage may exceed that for an aerial mission.

We already pointed out the somewhat sudden and therefore surprising development of CPCs and their application to urban three-dimensional modeling. This has been widely published as the “Rome-in-a-Day” paradigm by a team at the University of Washington in Seattle. The idea is to have the public collect images of street canyons and buildings, upload those with some geo-positions from GPS into a system such as FLICKR, and then process these photo collections into three-dimensional urban models. Figure 9 illustrates the approach using a result from the Wiki-Vienna project. The source data are available at no cost.

Next, the nonsystematic photo collection is subjected to a triangulation and dense matching for three-dimensional modeling of buildings. The concern over privacy and ageing of the data can be addressed very effectively if one can rely on the data users themselves and on their local geography knowledge. This is the basis for the notion of “neo-geographers,” a concept that has made it into Wikipedia and is most broadly applied in the Open Street Map initiative. The most meaningful and least intrusive assembly of imagery per urban façade and building will be obtained by such neo-geographers. While the use of CPCs is one...
approach, there also is the Photosynth paradigm, in which a set of photographs gets collected with the explicit purpose of creating a three-dimensional visualization.

Streetside imagery has become the workhorse of urban detail in Google Maps and Microsoft Bing-Maps. At this time, those images are being presented for visual orientation and pleasure. They are not being analyzed to describe the street scenes shown in the images. A combination of existing streetside Internet maps and new images from CPCs or specifically uploaded by an interested user becomes feasible. At issue is the merging or conflation of a new photograph and a pre-existing street view system. Figure 10 illustrates the result. The approach has been explained by Kröpfl et al.\(^8\)

The mid-sized city with its 250,000 inhabitants may consist of perhaps 75,000 structures, such as buildings, sheds, garages, industrial buildings, and so on. If one were to assume that each building façade gets imaged by users onto about 20 photographs, the total data set would encompass \(75,000 \times 4 \times 20 = 6\) million images. If the single photograph were to represent 10 MPixels and each pixel would consist of 3 B, the data set would approximate 180 TB.

### 3.3 Remotely Controlled Aerial Vehicles as Mapping Platforms

A plethora of small remotely controlled aerial vehicles exists, including airplanes as well as helicopters, quad-rotors, and octo-rotors. Unmanned small airplanes offer the advantage of fast rectilinear flight to cover a given extended area quickly and systematically. Helicopter systems would have the ability of great flexibility to explore a smaller area from multiple vantage points.

While their dimensions of up to \(1 \times 1\) m\(^2\) are reasonably small to allow simple handling, they are still able to carry payloads of 0.5 to 1 kg. Typical systems consist of a camera, some computing capability, and data storage, as well as some navigation and stabilization support in form of GPS and INS. Figure 11 illustrates a typical system, usually known as an MAV, and its operation to collect aerial photographs of an area of interest.\(^{22,23}\)

One may consider a system with remotely controlled MAVs flying along all the streets and collecting imagery from a few tens of meters above the ground. A single MAV can today collect images during a single aerial sortie for 15 min, collecting about 5 GB of data. It can proceed at a linear velocity of up to 10 m/sec, but it will have to land every 15 min to exchange batteries and memory cards. Given some good planning, and perhaps the use of multiple MAVs, the effort of image collection in an urban city environment might be very doable in terms of elapsed time and labor cost. We estimate that streetside data over a distance of 15 km could be acquired within every hour. There are, of course, legal restrictions. Having MAVs hover over urban canyons and their people will not easily get approval by authorities.

### 3.4 LiDAR

Since about 2000, a new type of sensor has become the accepted source of three-dimensional point clouds for three-dimensional mapping. LiDAR measures the distance between the sensor and the object surface directly and at accuracies in the centimeter range. The angle of the observing ray is being measured by attitude sensors. The sensor position is anchored in three-dimensional space by means of position sensors. On an aerial platform, that accuracy is defined by the satellite navigation systems, today GPS. The system accuracy is in the range of decimeters. LiDARs today are available on airborne platforms (ALS for airborne LiDAR scanning), on street vehicles, and on tripods for indoor data collection. The data product is attractive because it consists of an instant three-dimensional point cloud that can be viewed by the human user in the field (Fig. 12). The success of the technology was made possible by a
significant research effort since 2000, with the best talents focusing on LiDAR, such as Elberink and Vosselman. In addition, competing technologies based on imagery were not equally well developed or brought into a practical application, as argued by Leberl et al., largely due to the late introduction of large-format digital cameras, the lack of solid automated computer vision methods, and the demands on storage and computing for image-based solutions.

Today, it appears that no technological constraints present themselves to the fully automated analysis of images. Computing, storage, sensor resolution, and algorithmic advances have conspired to produce three-dimensional point clouds plus additional shape and semantic information about objects and scenes of interest at very competitive accuracy and cost, and with strong quality measures from high-redundancy imaging. And yet, the inertia of introduced and engrained applications keeps LiDAR-based solutions in a leading role in all three-dimensional mapping scenarios.

LiDAR systems are being augmented by imaging functions to measure reflectivity in addition to the observed distances, and the density of points is being continually increased. It recently was at 4 pts/m² from an altitude (aerial platform) of more than 1000 m. Positioning technology is advancing with the introduction of non-U.S. satellite navigation systems to improve the system accuracy of measured points. We will have to see how LiDAR will continue to compete with three-dimensional vision as sensors, systems, and computing continue to advance.

### 4 Façades

#### 4.1 From Aerial Photography

Intuitively, façade imaging is always associated with street-side photography. However, vertical aerial photographs can show useful images of façades at the edge of the fields of view, where view angles off-nadir reach 20 deg and more. Obviously, the angles under which any façades are seen will be steep and pixels on the façade will not be square. But in many cases, one will want to extract from a façade some accurate information about the number of floors, number and type of windows, or the existence of façade detail such as staircases. At issue is the ability to get this information from aerial façade images.

#### 4.1.1 Bird’s eye aerial photography

The interest in façades exists in the location-aware Internet, so that Microsoft embarked on the collection of oblique photography from the air, denoting this as bird’s eye viewing (Fig. 7). Figure 13 compares such bird’s eye aerial imagery of a façade with vertical aerial photography. The visual differences seem small. However, oblique images do suffer more from occlusions than the steeper looking vertical aerial photography.

#### 4.1.2 Façades seen as a single plane

We have developed a workflow to segment the image areas representing façades in vertical aerial photography. This is based on a dense point cloud of each building, and a classification of the building’s roofs based on color and texture. Auxiliary data may exist from cadastral property records and help to separate buildings from one another, should they be connected. A façade is defined in two independent steps. First is as a vertical plane from the roof’s edge to the Earth. Second is a computation of vertical surfaces from the point clouds. The vertical façade surface patches get projected in to the aerial photographs available for a building, and the largest area for a given façade is chosen to analyze its details. Multiple images may show one façade multiple times, and any results obtained from one image can be compared to results from overlapping images.

When the assumption of a plane façade surface is valid, we are dealing with a two-dimensional image analysis problem. Windows can be detected using an approach by Lee and Nevatia. The success of counting floors and the number of windows is summarized in Table 1 for a test area in Graz (Austria). Errors in window detection may be caused first by a façade that consists of more than one plane, or by occlusions and vegetation, or by shadows.

#### 4.1.3 Complex façades

The side of a building may be structured (for example, by extruding staircases). The façade then will not be a two-dimensional object, as illustrated in Fig. 14. Instead, one will have to model it in three dimensions. A point cloud from aerial photography may be sufficient to model the three-dimensional structure of such a complex façade. The forward overlap of aerial photographs will show a

![Figure 13](image)

**Fig. 13** Sample images from Graz (Austria) featuring segments from an oblique aerial imagery (a) for comparison with the edge of a vertical aerial image (b). The specially acquired oblique image does not seem to offer an advantage over the standard vertical aerial mapping photography.

---

**Table 1** Floor and window count on 104 buildings in Graz (Austria) based on vertical aerial photography at a pixel size of 10 cm (from Ref. 31).

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Detected</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buildings</td>
<td>104</td>
<td>104</td>
<td>100</td>
</tr>
<tr>
<td>Floors</td>
<td>397</td>
<td>369</td>
<td>93</td>
</tr>
<tr>
<td>Windows</td>
<td>2646</td>
<td>2276</td>
<td>86</td>
</tr>
</tbody>
</table>
given façade on multiple images, and a three-dimensional structure can be found from those points. One approach is to perform a plane sweep resulting in multiple plane surfaces fit to the sparse cloud of points (Fig. 14 explains).

4.2 From Streetside Photography

The majority of literature about façades addresses streetside imagery as the data source. An issue is the segmentation of the images into building sides versus sky, streets, and vegetation or other occlusions. Once building sides are segmented, one will want to identify separate and individual façades. Figure 15 illustrates a semantic segmentation of raw streetside images, and the subsequent separation of individual façades, using repeated patterns on a building’s sides to separate one façade from another.34,35

4.3 Streetside LiDAR

Industrial data collection vehicles often include one or several LiDAR ranging systems to collect façade points (Fig. 16). Segmentation of those point clouds into façade surfaces versus windows or doors has been studied.36 The density of LiDAR point clouds is at 5000 points per square meter, versus a photography-based point cloud at 50,000 pixels per square meter.30

4.4 Façade Visualization

One may separate the rendering of a façade from the source of a façade’s geometry. While the number of floors and location and size of windows may be satisfactorily extracted from vertical aerial photography, photo texture is not visually pleasing due to the large pixels and their nonsquare shape. However, texturing the façade with streetside imagery may suffer from occlusions, especially by vegetation (Fig. 17).

In this context, an interesting alternative is procedural modeling of façades being developed at ETH-Zürich by P. Müller et al.37 Given separated façade images, the structure of a façade is analyzed. Once the hierarchical subdivisions of the façade are known, it can be described by a shape...
grammar\textsuperscript{38} and it is thus possible to replace architectural elements by instances of a predefined set of parameterizable models (including, for example, different window styles and synthetic textures). Such a representation improves not only the visual quality of the rendering, it also provides the ability to change the point of view and the scale while still looking at physically correct renderings. Additionally, semantic meanings are assigned to individual parts of the façade, which is important if one wants to locate, for instance, the entrance of a building. Finally, the data necessary for visualization is significantly reduced by transmitting only a predictable number of parameters instead of textures. Figure 18 illustrates the results of this approach.

5 Roofs and Roof Detail

5.1 Roofs from Aerial Photography

Most recent literature on roof mapping is concerned with the analysis of point clouds from airborne LiDAR systems, as typified by Jochem et al.\textsuperscript{39} However, a natural data source about building roofs is, of course, aerial photography. At the core of an analysis workflow is the point cloud of a segmented roof. The density of the point cloud from aerial photography depends on the pixel size and image overlaps. Typically, one will obtain a point every 1 to 2 pixels provided that 10 or so images overlap and that therefore a super resolution is achieved. With 10-cm pixels of the source images, this results in point clouds at a density of 25 to 100 points per square meter. If one were to go to 3-cm pixels, that density would increase to 225 to 1000 points. A typical ALS in operation today will deliver 4 native points per square meter from an aerial platform.

We are interested in defining the main roof shape, as represented by planes that ignore the details extruding from, or intruding into, those roof planes. In Meixner et al.,\textsuperscript{40} smoothing of the roof’s point cloud is applied to suppress local noise. Then planes get matched to the points so that many plane segments are being obtained. These get merged into the major and secondary planes defining the main shape of the roof and the smaller dormers and other roof detail. Experiments with aerial photography at a GSD at 10 cm in an area of 400 × 400 m\textsuperscript{2} in Graz (Austria) addressed a total of 186 buildings and 446 major roof planes. Of those, 92% were correctly mapped. Given those roof planes, it now becomes possible to classify each roof by its architectural style, such as hip or gabled. This has been achieved at a success rate of 78%.

5.2 Roof Detail

We define “roof detail” as the small objects that extrude from a main roof plane or intrude into it, such as chimneys, elevator shafts, air conditioning units, antennas, terraces, and dormers.\textsuperscript{41} Of interest are also skylights as separate objects, even though they may not geometrically violate the roof’s planes. The roof planes and plane segments that are too small to be a roof are input to an analysis of roof details. The characterization of the small roof elements and color enter into a feature vector for a classification of the superstructures into dormers, chimneys, and skylights.

Fig. 18 Procedural modeling of façades from Ref. 38. Separated façades (a) are analyzed and described by shape grammars (b). Replacement of individual parts by parameterizable models leads to a high quality reconstruction (c).
with the experimental 186-roof data set in Graz shows a total of 1312 dormer windows, chimneys, and skylights. Of those, 1024 were correctly mapped, representing a success rate of 78%.

Figure 19 illustrates the results of the automated separation into a roof and its roof details achieved from aerial photography and associated point clouds in the Graz test area.

5.3 LiDAR and Photography

The LiDAR-literature has not addressed the models of roof details, only general roof shapes. Figure 7 illustrates the quality of photography of roof details with GSD values at 8 and 0.5 cm. Mapping of roof details requires a density of at least 25 (preferably 100) pts/m². Specifications of currently flown airborne LiDARs do not yet meet that requirement. Therefore, the roof detail is often not mapped satisfactorily from ALS. Aerial photography shows promise to solve this issue due to its superior point density and availability of edge and line information, as well its ability to support texture- and color-based segmentation.

6 Outlook

Geo-virtual environments present urban spatial information in an immersive, accurate, and complete form. They are a new phenomenon. Denoting these also as three-dimensional virtual cities, they inspire a location-aware Internet²⁴ and many aspects of urban life, planning, and engineering. Such spatial data get extracted from satellite and aerial photography, where high overlaps have been found to improve the degree of automation and the level of detail along the concept of super-resolution. Airborne LiDARS are also being operated to develop three-dimensional point clouds, which get converted into geometric building models.

We are experiencing a rapid evolution of the relevant technologies for urban three-dimensional mapping. Sensors become more efficient and diversify, source data become less costly, and the Internet offers opportunities for source data becoming entirely cost-free. Image storage improves in efficiency and is quickly becoming a non-cost issue.

Computing is improving by means of the GPU to a formidable throughput so that one at times can hear the comment that “Computer vision now no longer has any significant hardware restrictions.” Finally, the development of novel algorithms for multi-image matching, segmentation, and classification are evolving rapidly as well.

Façades have been a topic of research and assume that the sources are streetside photographs, sometimes also streetside LiDAR. However, it turns out that vertical aerial photography does present sufficient detail and geometric rigor to provide useful models of façades with less effect of occlusions than what one needs to accept in streetside imagery. Façades are shown to not only be plane objects subject to two-dimensional analysis, but also potentially three-dimensional assemblies of multiple planes, and must therefore be modeled in three dimensions. A very promising data source will be the combination of aerial imagery and street-level images from CPCs.

Roofs and roof details are of interest in describing a building and understanding its use. While aerial photography is a most natural source of roof information, airborne LiDARS are prominently being studied as an alternative. We have shown that roof planes and roof detail can be extracted from aerial photographs and three-dimensional point clouds at success rates of 78% to 92%, depending on the specific task at hand, and relying on aerial images with 25 to 100 pixels/m².

References

27. J. Bennett, Open Street Map—Be your Own Cartographer, p. 252, Packt Publishing (2010).