Processing 3D Geo-Information for Augmenting Georeferenced and Oriented Photographs with Text Labels

A. De Boer, E. Dias, E. Verbree

Abstract

Online photo libraries face the problem of organizing their rapidly growing image collections. Fast and reliable image retrieval requires good qualitative captions added to a photo; however, this is considered by photographers as a time-consuming and annoying task. In order to do it in a fully automated way, the process of augmenting a photo with captions or labels starts by identifying the objects that the photo depicts. Previous attempts for a fully automatic process using computer vision technology only proved not to be optimal due to calibration issues. Existing photo annotation tools from GPS or geo-tagging services can only apply generic location information to add textual descriptions about the context and surroundings of the photo, not actually what the photo shows. To be able to exactly describe what is captured on a digital photo, the view orientation is required to exactly identify the captured scene extent and identify the features from existing spatial datasets that are within the extent. Assumption that camera devices with integrated GPS and digital compass will become available in the near future, our research introduces an approach to identify and localize captured objects on a digital photo using this full spatial metadata. It proposes the use of GIS technology and conventional spatial data sets to place a label next to a pictured object at its best possible location.

Keywords: photo annotation, object identification, label placement.

1 Introduction

The increasing availability of consumer digital cameras and integrated all-in-one devices (e.g. camera phones) enables people to capture and upload digital photos at any place and any time. The organization of these rapidly growing image collections is a major challenge for online photo libraries. Good qualitative descriptions of the content added to a photo enable easier retrieval of an image, but unfortunately, captioning photos is experienced by photographers and users as a time-consuming and annoying task (Dias et al. 2007). Those who do not caption their photos encounter problems at a later stage when users are searching for a particular photo. This issue encouraged researchers to develop tools that enable the automatic captioning of digital photos using positioning information – either automatically by a GPS device or manually by georeferencing on a map – to add descriptions about the context and surroundings (Naaman et al. 2004). However, using positioning information only, these state-of-the art photo annotation tools are limited to the adding of descriptions to a photo about its surroundings, and not about the objects that are actually pictured (Chang 2005).

Assuming that in the near future digital cameras will include a GPS chip and digital compass (to capture position and orientation), the work presented here is an approach that extends the captioning of photos and benefits from this full spatial metadata (geographic positioning, altitude, and pitch) in order to produce an abstraction of the captured scene and to identify objects on a photo.

Our process of object identification in digital photos proposes an alternative for computer vision-based image recognition and photogrammetric coordinate conversions (from pixel to terrain coordinate system). Available GIS technology and established spatial data sets are applied to identify what is visible and where it is located on a digital photo by using a perspective viewer service. This tool renders a three-dimensional model based on input view parameters (the full spatial metadata) and outputs a 2D image that is a virtual abstraction corresponding to the pictured scene. Linking the virtual scene to the three-dimensional model, attributes (i.e. street names) from the spatial data sets can be picked and associated with the objects. At this stage, the image can be augmented with captions of the objects. We go one step further: to actually label the objects in the photo with the just determined captions. To do this, constraints and rules are added to a label engine in order to place a label next to a pictured object at its best possible location. This last step of labeling photos can be especially rele-

vant in the accessibility field. It can be the basis for developing new tools used to improve accessibility for visually impaired users to "sense" digital photos using large-sized label fonts or sounds on mouse over, as objects on the photo are well-identified and well-localized.

This paper is organized as follows: Section 2 describes previous research on automatic photo annotation tools and label placement in 3D environments. Section 3 defines the collection of digital photos having full spatial metadata and the spatial data requirements for the preparation of the extrusion models. Section 4 describes our approach for object identification in digital photos. Section 5 proposes some rules that could be applied in order to find the best location to place a label on a digital photo. Finally, Section 6 provides some discussion and conclusions and recommends future research.

2 Related Research

Cartography is described as the graphic principles supporting the art, science, and techniques used in map making maps, which are intended to communicate a certain message to the user. The process of text insertion in maps is referred to as label placement. Label placement is one of the most difficult tasks in automated cartography (Yamamoto and Lorena, 2005). Positioning text requires that:

- overlap among texts is avoided;
- cartographic conventions and preferences is obeyed;
- unambiguous association is achieved between each text and its corresponding feature;
- a high level of harmony and quality is achieved.

Good placement of labels avoids as much as possible overlap of labels with objects and mutual labels; and is applied to provide additional information about a particular feature. Automatic label placement is therefore one of the most challenging problems in GIS (Li et al., 1998):

- optimal labeling algorithms are very computational expensive for interactive systems;
- labels compete with data objects for the same limited space.

Augmented reality (AR), considered to be part of the Multimedia Cartography research field, is an environment that includes both virtual reality

and real-world elements and forms part of the research. AR is a field of computer research which deals with the combination of real world and computer generated data. Its principle is to overlay the real world (captured by a camera device) with supplementary information (e.g. labels). It enables users to interact with their environment e.g. by hyperlinking labels inside an AR view (Cartwright et al., 2007). Interactivity is one of the key components of multimedia.

Photo labeling refers to the act of placing labels that describe the features visible on the photograph itself. Saving the labels obtained from the virtual scene to a transparent layer enables to put labels associated with an object onto an image. As such, the photo annotation issue is considered to be part of Multimedia Cartography and AR as well; visible tags in images and AR applications enable user interaction with the environment; numerous ubiquitous and/or augmented reality applications are discussed by Kolbe (2004), Toye et al. (2006) and Schmalstieg and Reimayr (2007).

As Li et al. (1998) observe, object and label placement in limited screen spaces is a challenging problem in information visualization systems. Images also have a limited screen space and therefore (particularly automatic) label placement is of concern for this research in order to avoid overlap of labels mutual and labels with objects.

Numerous researchers already examined the problem of automatic label placement in 2D maps. Recent work of Maass and Döllner (2006a), Azuma (2004) and Götzelman et al. (2006) also focused on the placement of labels in 3D landscapes and Augmented Reality views referred to as view management (Bell et al. 2001). Götzelman et al. (2006) offer complex label layouts which integrates internal and external labels of arbitrary size and shape, and real-time algorithms. Maass and Döllner (2006b) describe two point-feature dynamic annotation placement strategies for virtual landscapes including priority aspects.

Labeling is further divided into internal and external (object) annotation (Maass and Döllner 2006a). An internal annotation is drawn on the visual representation of the referenced object and partially obscures that object. An external annotation is drawn outside the visual representation of the reference object and uses a connecting element such as a line or arc to associate the annotation with the reference object.

Hagedorn et al. (2007) describe the use of a Web Perspective Viewer Service (WPVS) for the annotation of three-dimensional geographical en-

vironments (a.k.a. geo-environments). Furthermore, a three-dimensional Web View Annotation Service (3D WVAS) is proposed as an extension to a WPVS. The perspective view together with a depth image is forwarded to the 3D WVAS together with annotation definitions. This annotation technique calculates the positions of the labels, renders them into a separate image buffer, and combines the resulting image in a depth-sensitive way with the input color image (see Fig.1.).

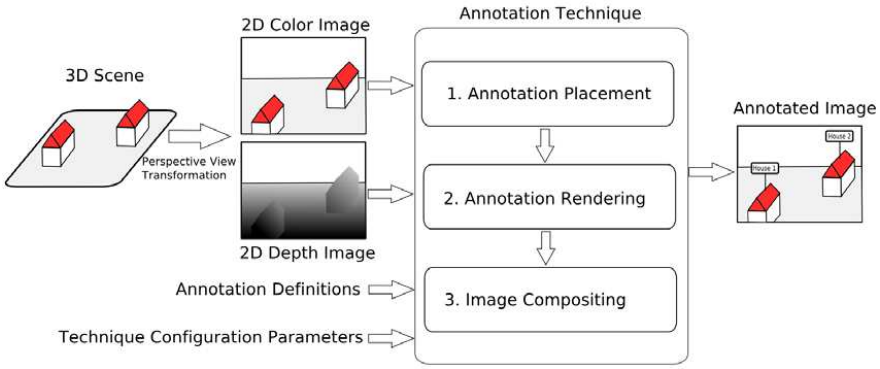

Fig. 1. Process for the annotation of 3D scenes as proposed by Hagedorn et al. (2007)

Our work links up with the previous research in chaining a Perspective Viewer Service with an Annotation Service, although the approach is more simplistic in the sense that we have chosen to use components from a commercial GIS package (i.e. ArcScene and Maplex by ESRI ArcGIS) to demonstrate the concept of internal annotation within digital photos. Our label placing strategy, concentrates in:

- linking the labels to the object they refer to;
- determining the 'free' labeling space, i.e. open sky;
- placing the labels at the best possible location.

3 Data collection and preparation

3.1 Image collection

Our concept and ideas were tested by collecting three-dimensional geo-referenced and oriented digital photo at the Market square in the historic

city centre of Delft, the Netherlands. We created two collections of test photos:

1. Low-resolution and high-spatial accuracy photos captured using a Topcon GPT-7003i © imaging total station, and
2. High-resolution and low-spatial accuracy photos were captured using a Nikon D-100© digital camera mounted with a 3-axis electromagnetic digital compass and a GPS data logger (see Fig.2.)

Fig. 2. Collage of the image collection on the Market square in Delft using the Nikon D100 camera mounted with digital compass on the hot shoe cover

The Topcon GPT-7003i distance and direction measurements are connected to a Large Scale Base Map of the Netherlands (GBKN) resulting in a position accuracy of approximately 0.5m and a directional accuracy of approximately 0.5 degrees. The camera included with the Topcon station has the disadvantage of low resolution (0.3 megapixels). On the other hand, the photos captured with the Nikon camera have high resolution (10 megapixels) but low spatial resolution. The position accuracy is around 10 meters and the compass orientation is very inaccurate due to distortion in the compass heading caused by the electromagnetic field of the camera. However, these images are particularly used to identify the misidentification of objects due to lens distortions, GPS and compass inaccuracies.

3.2 Spatial data requirements

A three-dimensional building model is required to serve as input for creating the virtual scene using a perspective viewer service (for our research we used the 3D visualization tool: ESRI ArcScene©). The three-dimensional building model was created by extruding 2D building footprints, extracted from a 1:10,000 topographic base map (the TOP10 vector dataset of the Dutch National Mapping Agency), based on the height values from a raster detailed elevation model (from an Airborne Laser Altimetry dataset, the AHN). This approach results in a gridded footprint in

which each feature is associated with an object identifier (OID) from the building footprint dataset and a height value from the elevation raster. The advantage of this approach is that height values inside the buildings are known and the building footprint geometry is preserved. After extrusion of the features and randomly coloring based on the building OID, the model of Fig.3. is obtained.

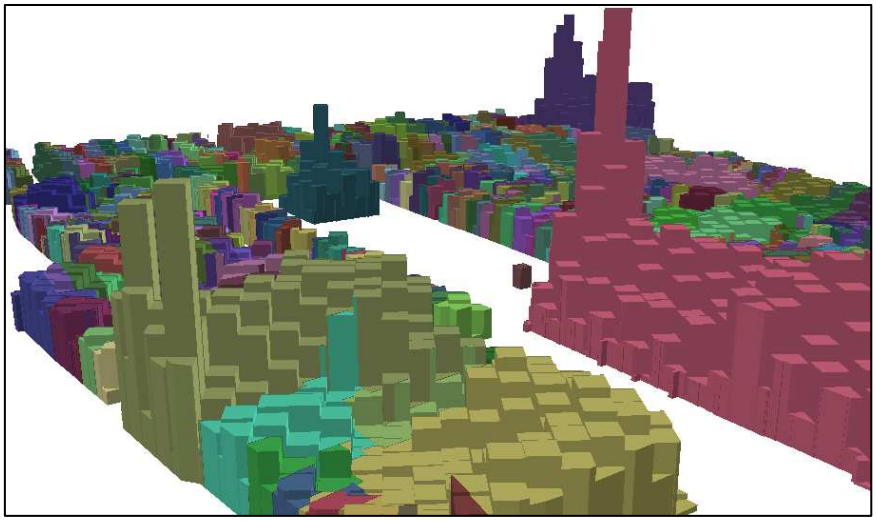

Fig. 3. 3D model created from intersecting the building footprints with the vectorized elevation raster.

Even though we believe that such a 3D model is not interesting in aesthetics terms, it was considered to be the optimal solution for our research, because it is simple to produce and reproduces the shapes of the buildings with sufficient accuracy. This was the model applied as input to create the perspective views representing the digital photo. At this step, the names of buildings and shops located around the Market square are added to the dataset based on a commercial "Points of Interest" dataset. These will be the names to pick after the objects are identified.

4 Object identification

The core of our research is dedicated to the object identification problem: "What is captured by a digital photo?", and "Where is it located in the pho-

to?" The latter question is very important in order to place a label next to an object. Nevertheless, knowing where the objects are located in a photo enables other innovative applications besides labeling, for example: hyperlinking the objects in the photos to dynamic descriptive pages or tools that help visually impaired users to understand image content by using sounds (for legally blind users) or large sized text labels (for users with low vision) on mouse-over.

As a perspective viewer service, we applied ESRI ArcScene to render our three-dimensional model based upon the view parameters (the full spatial metadata) to create a virtual abstraction that matches the pictured scene. The main issue to be solved is how to link the virtual scene to the three-dimensional model in order to pick the names. Since the virtual abstraction is returned in raster format, its coordinates are in a local pixel coordinate system so a spatial join with the three-dimensional model is not possible. The solution to this problem as proposed in our work is to color each object of the three-dimensional model based on its unique OID. Therefore, the decimal OIDs are converted to RGB color values using the relationship: OID ≡ RGBdecimal = 65536 * Red + 256 * Green + Blue

Fig. 4b shows the output of the building features from the three-dimensional model (Fig.4a) when colored with RGB color values corresponding to their OID. Subsequently, a virtual scene (Fig.4c) corresponding to the digital photo is created by using the view parameters derived from the full spatial metadata of the same digital photo. This virtual scene has to be exported to a lossless compression image format (e.g. PNG) in order to maintain a seamless color throughout the object and, in this way, avoiding additional features to intrude in the form of averaged colored pixels on the object borders.

Now we have a raster image with representation of the visible features. To further analyze this result, it was necessary to convert the virtual scene to vector features again. Therefore, first the RGB color bands of the virtual scene are summed up using the RGB-OID relationship to obtain OIDs again. Next, the vectorized virtual scene (Fig.4d) is joined with the building features of the three-dimensional model (or 2D spatial datasets) based on the OID and names are picked from these datasets (Fig.4e) to label the objects visible on the digital photo.

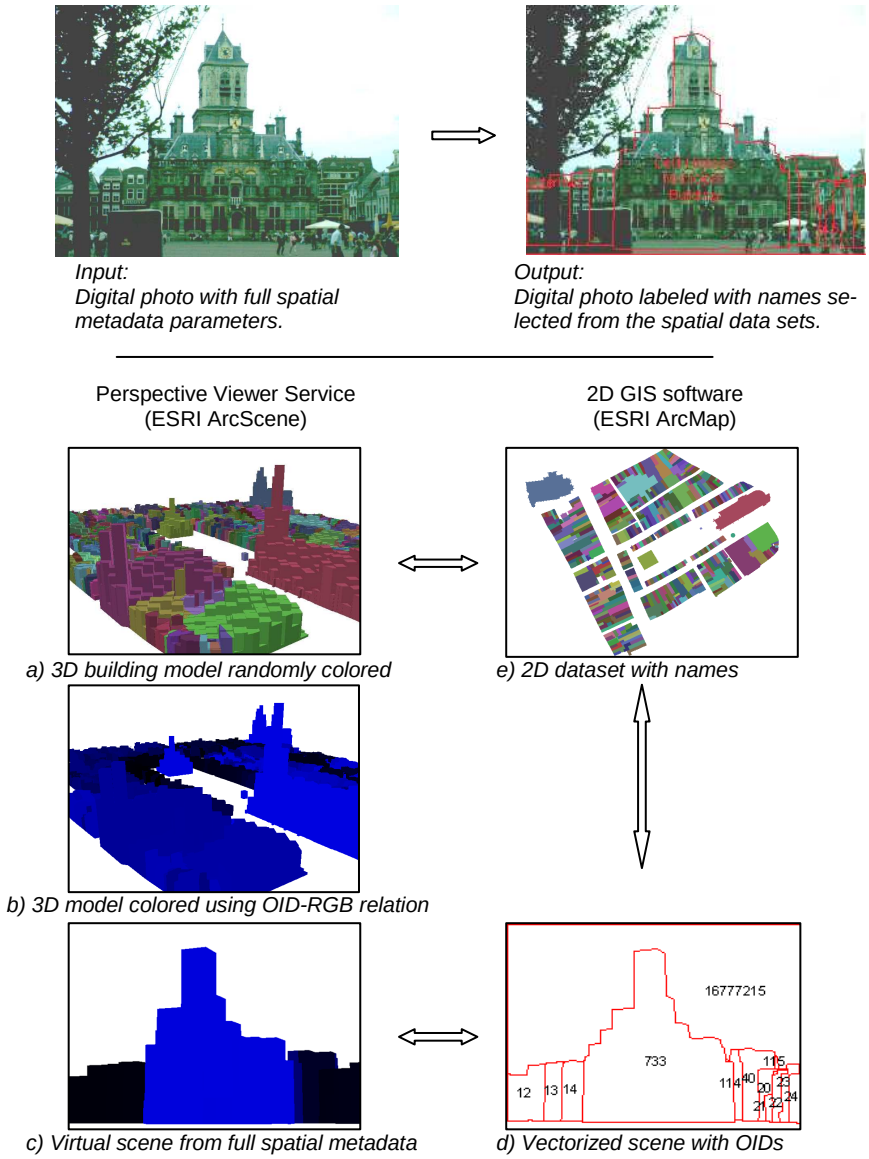

Fig. 4. Concept and process of object identification using the OID-RGB relation.

5 Label Placement

After the object identification, in which pictured objects are identified and localized on a digital photo, the next part of our research focused on label placement: proposing constraints and rules on where to place a label on a digital photo to identify a certain object. We assumed that the best location for label is at an empty area, which is defined as the area where there are no objects inside the virtual scene (or digital photo). Using ESRI ArcMap©, the virtual scene is overlaid with the digital photo and labeled. The labeling was optimized by using constraints and rules on the label engine extension Maplex©, including, among others, to avoid overlap between labels and between labels with objects, i.e. labels are placed outside visible objects using connectors (see Fig.5)

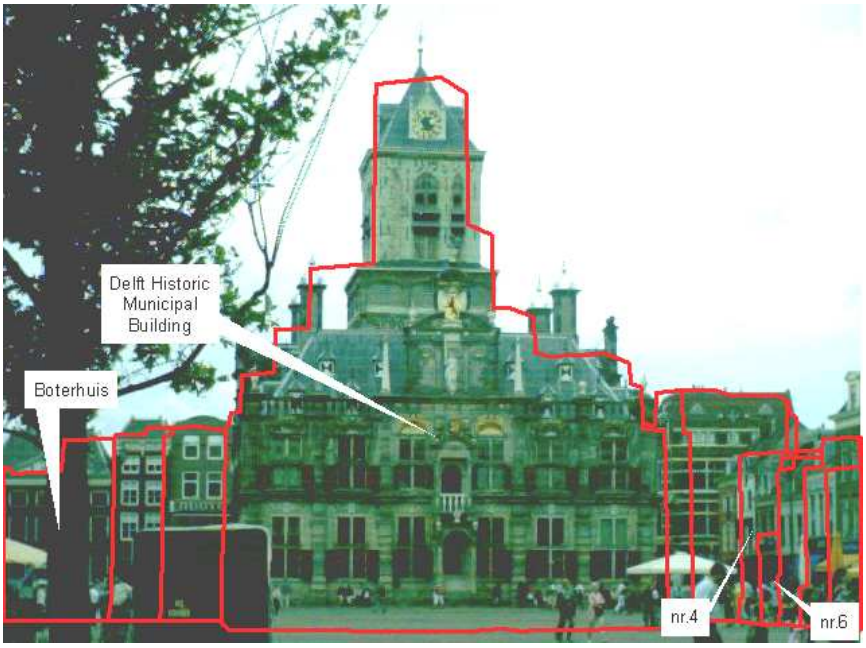

Fig. 5. Visible objects are externally annotated using connectors avoiding as much as possible overlap of labels with objects and among labels.

However, because the tree in front of our pictured scene is not included in our 3D building model, the label engine assumes that this location is a good location to place a label. But, since we do not wish to overlay objects in the picture, we needed to identify in the picture, the areas without any features. Therefore, the digital photo is reclassified to a binary image. In

our example (see Fig.6 upper left), we used the median as a cut-off value. After this, we combined the binary image derived from the virtual scene with the building model. This limits the label engine to avoid placing a label overlapping objects in the united layer. The result is shown in Fig.6.

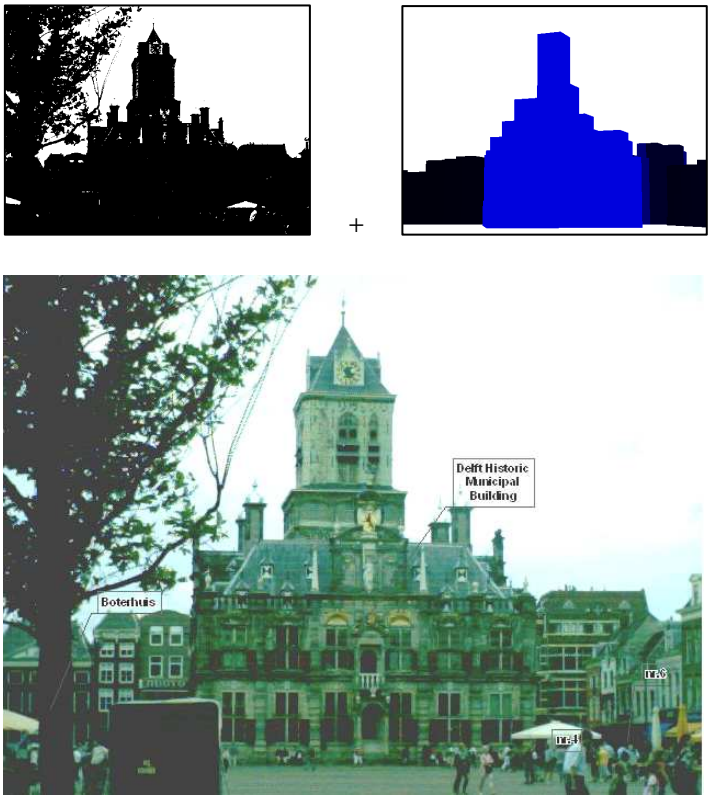

Fig. 6. A binary image and the virtual scene are used to place a label at its best possible location (assumed to be at the empty areas).

Our second proposal is to apply a depth image (a.k.a. depth map) to vary label font sizes with the distance from the observer to the objects, maintaining the perspective view. A depth map is created (and returned) by the renderer to identify what is visible or not to build the two-dimensional abstraction of the 3D model. We created our depth (see Fig. 7 on the left). image by:

1) calculating the distance from observer to an object;
2) updating these as an attribute to the building model;
3) coloring the three-dimensional model based on the distance-attribute;
4) creating and exporting a virtual scene using the view parameters.

Finally, the digital photo is labeled using the depth image showing the varying label font sizes of Fig.7.(right hand picture).

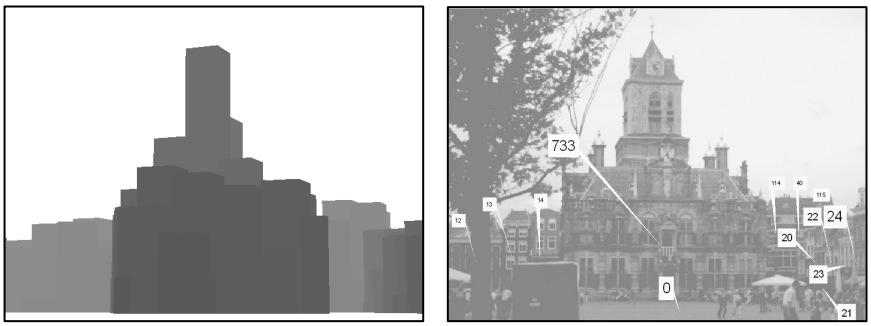

Fig. 7. Depth images as output of the perspective viewer services (left) used to vary in label font sizes depending on the distance from observer to pictured object (right).

The amount of labels that should be placed on a digital photo depends on the number of visible objects and user preferences. By adding a constraint to the label engine that only objects of a minimum specified size should be labeled (e.g. based on polygon perimeter), the amount of labels that are placed on a photo could be reduced. Fig. 8 shows an example of a digital photo labeled with names of shops located around the Market square of Delft; on the left hand side all visible objects are labeled; on the right hand side only visible object with a polygon perimeter larger than 640 m^2 are labeled.

Fig. 8. Varying the amount of labels to place: on the left hand side no constraints are added with respect to object size; on the right hand side the larger visible objects are labeled.

6. Results and conclusions

This research showed how objects on a digital photo can be identified using the photo's full spatial metadata (with position and orientation). In addition, we investigated the best location for a label to annotate an object within the photo. The results of the object identification for the photo collection with high-quality spatial metadata (acquired with the Topcon imaging total station) were very positive. There was a good match between the 3D building abstractions and the photos. The results for the photo collection with lower spatial accuracy (acquired with the Nikon camera connected to a GPS receiver and a digital compass) revealed less successful results than the Topcon, directly related with the inaccuracies of the GPS and compass devices. As expected, it was also observed that the amount of misidentification increases with increasing inaccuracy of GPS and compass and decreasing field-of-view angles.

This work proposes a methodology for object identification in digital imagery alternative from the existing methodologies: computer vision technology and photogrammetric equations. It is concluded that the use of GIS technology and spatial data to create a virtual scene as output of perspective viewer services is appropriate to apply in object identification and localization. In doing so, the problem of label placement in three-dimensional geographic environments is reduced to a two-dimensional map labeling problem. The best location of label placement was determined using constraints and rules to be applied to the virtual scene and the reclassified-to-binary image of the input photo. In addition, depth maps enable the variation of label font size depending on the object distance to the photographer.

Two main limitations were identified in this approach. The first is that using current consumer devices (GPS receiver and digital compass) to acquire the geographical spatial metadata resulted in increased misidentification of features owing to inaccuracies of the sensors, when compared to the high-accuracy professional device. The other limitation found relates to the performance when handling a large amount of features in the extrusion model (from the vectorized elevation model). The tools we used for this study create a single 3D model that owing to its large size limits the spatial extent of the area to analyze. To solve this issue, it is proposed to implement this process in the form of a webservice that uses the data stored on dedicated spatial databases. In this way the service can create on-the-fly the 3D model based on only the relevant area for the photo, making the

area extent not a limitation since we eliminate the need for a unique 3D model. Only the availability of the data for any region is the limitation.

Further research is recommended to evaluate with real users the constraints and rules for the label algorithm, since the strategies to place the labels should be user driven or based on user preferences.

Acknowledgements

The authors gratefully acknowledge all who contributed to this research, in particular Peter van Oosterom. Parts of this work were supported by the IST FP6 project TRIPOD (045335), see: www.projecttripod.org.

References

Azuma R (2004) Overview of augmented reality. International Conference on Computer Graphics and Interactive Techniques ACM SIGGRAPH 2004 Course Notes, Los Angelos.

Bell B, Feiner S, Höllerer T (2001) View Management for Virtual and Augmented Reality. Proceedings of the 14th annual ACM symposium on User interface software and technology 2001 pp 101-110.

Cartwright W, Gartner G, Peterson MP (2007) Multimedia Cartography Second Edition. Springerlink, Berlin Heidelberg New York.

Chang EY (2005) EXTENT: Fusing Context, Content, and Semantic Ontology for Photo Annotation. ACM SIGMOD CVDB Workshop, Baltimore.

Dias E, de Boer A, Fruijtier S, Oddoye JP, Harding J, Matyas C, Minelli S (2007) Requirements and business case study. Project deliverable D1.2. TRIPOD: TRI-Partite multimedia Object Description. EC-IST Project 045335 (www.projecttripod.org).

Li J, Plaisant C, Schneiderman B (1998) Data object and label placement for information abundant visualizations. Proceedings of the 1998 workshop on New paradigms in information visualization and manipulation, Washington D.C., pp 41-48

Götzelman T, Hartman K, Strothotte T (2006) Agent-based annotation of Interactive 3D Visualizations. 6th International Symposium on Smart Graphics, Vancouver, pp 24-35.

Hagedorn B, Maass S, Döllner J (2007) Chaining Geoinformation Services for the Visualisation and Annotation of 3D Geovirtual Environments. 4th International Symposium on LBS and Telecartography, Hong Kong.

Kolbe TH (2007) Augmented Videos and Panoramas for Pedestrian Navigation. 2th International Symposium on LBS and Telecartography, Vienna.

Naaman M, Harada S, Wang QY, Garcia-Molina H, Paepcke A (2004) Automatic Organization for Digital Photographs with Geographic Coordinates. Proceedings of the Fourth ACM/IEEE-CS Joint Conference on Digital Libraries, pp 53-62.

Maass S, Döllner J (2006a) Dynamic Annotation of Interactive Environments using Object-Integrated Billboards. Proceedings 14-th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, WSCG'2006, Plzen, pp 327-334.

Maass S, Döllner J (2006b) Efficient View Management for Dynamic Annotation Placement in Virtual Landscapes. 6th Int. Symposium on Smart Graphics, Vancouver, pp 1-12.

Schmalstieg D, Reitmayr G (2007) The World as a User Interface: Augmented Reality for Ubiquitous Computing. Location Based Service and TeleCartography, Springer, pp 369-391.

Toye E, Sharp R, Madhavapeddy A, Scott D, Upton E, Blackwell A (2006) Interacting with mobile service: an evaluation of camera-phones and visual tags. Personal and Ubiquitous Computing, pp 1 – 10.

Yamamoto M, Lorena LAN (2005) A Constructive genetic approach to point-feature cartographic label placement. Metaheuristics: Progress as Real Problem Solvers, Springerlink, New York.