Semantic and geometric enrichment of 3D geo-spatial models with captioned photos and labelled illustrations

Christopher B. Jones, Paul L. Rosin and **Jonathan D. Slade**
School of Computer Science & Informatics
Cardiff University
5, The Parade, Cardiff, CF24 3AA
United Kingdom
{JonesCB2, RosinPL, SladeJD}@cardiff.ac.uk

Abstract

There are many 3D digital models of buildings with cultural heritage interest, but most of them lack semantic annotation that could be used to inform users of mobile and desktop applications about their origins and architectural features. We describe methods in an ongoing project for enriching 3D models with generic annotation, derived from examples of images of building components and from labelled plans and diagrams, and with object-specific descriptions obtained from photo captions. This is the first stage of research that aims to annotate 3D models with facts extracted from the text of authoritative architectural guides.

1 Introduction

Geographical data are used in many professional and academic applications in the environmental and social sciences and for personal applications such as navigation and local information search. For many purposes, information referenced to 2D geographical coordinates is quite adequate. There are however a number of applications for which 3D geo-data are either necessary or highly desirable. These include inter-visibility analysis, radio communications, visualization for urban planning, indoor navigation, augmented reality, and mobile and desktop applications that allow the user to be informed about cultural heritage and architectural features of detailed aspects of the built environment. While there are increasing numbers of 3D city models, and individual models of many notable buildings, for some of the 3D geo-data applications the existing models are still inadequate as they usually lack semantic annotation of any sort. There is a requirement therefore to develop effective procedures to annotate 3D building models with descriptive attributes. In a cultural heritage context these could include the materials, origins, people and events associated with them.

In this paper we describe an approach to semantic annotation of 3D building models that forms the basis of a research project that is currently in its early stages. The main premise of the project is that if captioned images and annotated plans and diagrams can be matched, using computer vision methods, to locations on 3D models of buildings, then the textual information content of the captions can be linked to the corresponding parts of the model. The project focuses upon buildings with cultural heritage and builds upon and progresses beyond previous work that has used captioned photos to annotate 3D models, e.g. Simon and Seitz (2008) and Russell et al. (2013). We aim to generate models of buildings that are semantically richer than typical existing models. The approach complements work such as Zhang et al. (2014) in which 3D models were used to enhance and annotate photos.

Social media sites such as Flickr have many freely available photographs of interesting buildings, often accompanied by captions that provide descriptions of an entire building or of parts of a building. While photos on social media are a valuable source of information for well-visited buildings, there are many buildings, particularly less visited ones, that have few or no such photographs with useful captions. There are also however many captioned photos in architectural and cultural heritage guides that describe architectural features and associated historical events with a level of detail and quality of information that is often superior to the content of social media, including that of Wikipedia. Many of these guides

contain plans and diagrams of the layout of buildings labelled with the names of rooms and associated spaces. The combination of captioned images, annotated plans and diagrams that can be found in social media and in authoritative texts provides a rich source of information that can be used to attach semantic attributes to 3D models of buildings as well as to some 2D cartographic data.

The idea of exploiting captioned photos of buildings to support virtual exploration of buildings was pioneered in the Photo Tourism application of Snavely et al (2008) in which SIFT and related methods of feature matching in computer vision were used to match photos of buildings, compute camera parameters (i.e. pose) and to generate point clouds representing the 3D geometry of buildings (referred to as structure from motion). This enabled applications in which users could view captioned photos of particular parts of a building. It also provided the possibility of matching a new photo to an existing captioned photo and hence automatically tagging the new photo with tags from the already captioned photo or with tags that users manually linked to the 3D geometry. Simon and Seitz (2008) used related methods to annotate 3D point clouds with the captions of the Flickr photos. They presented a segmentation procedure that grouped photos that contain the same 3D points (derived from the SIFT/structure from motion methods) as well as exploiting the fact that 3D points belonging to the same object could be expected to be clustered in space.

Finer granularity annotation of parts of buildings was achieved by Russell et al (2013) who linked items of text in Wikipedia articles about particular buildings to locations in 3D models. The models were created from sets of Flickr images tagged with the respective building name, but annotation of objects within the buildings, such as statues and paintings, was obtained using images found on Google Image search, with search terms obtained from extraction of prepositional and noun phrases in the Wikipedia articles. In Russell et al's application, the 3D building model is essentially a point cloud with no explicit structure, and the users' views of parts of the building consist of the Flickr photos of respective annotated objects.

Our work differs from these approaches in that we employ structured 3D building models, as can be found for example in 3D Warehouse[1]. We exploit the fact that many of these 3D models include texture mapped photos that have been registered to the building geometry. This enables the use of image matching methods to associate captioned photographs with specific parts of the texture mapped images, and hence via the 3D geometry to specific geometric objects on the buildings. The aim is to enrich structured 3D models both semantically and, where necessary, geometrically, using a combination of the texture-mapped imagery, captioned photos, and other annotated resources such as ground plans, that name the rooms or spaces of a building, and diagrams of the elevation (side view) of a building. In the remainder of the paper we provide a summary of the methods. We conclude with a discussion of the future challenges.

2 Methods

2.1 Annotation of 3D models with examples of generic objects

To support the objective of attaching annotation from captions, or other text, to building components such as rooms, doors, windows, arches, clocks etc, it would be useful for their presence to be recorded explicitly in the geometric model, which is not the case for many existing building models. The geometric representation of component objects would serve as a digital record of the building structure and assist applications that guide and inform users of the various aspects of a building. If these objects, such as doors and windows, are generically labelled it will also assist the process of matching caption text to the parts of the building that they describe. Thus while a caption may describe something in the image the question is exactly what part of the 3D model is being described. In some cases the caption will relate to most of the content of the photo image, if for example a particular statue or window has been photographed, but in some cases the described object may occupy just a part of the image. Simon and Seitz (2008) addressed this issue by comparing multiple (hundreds or thousands of) photos with a similar caption and identifying the region that is common to the majority of the photos, but that approach cannot be used when there are very few photos with useful captions. If the 3D model contains generic

[1] https://3dwarehouse.sketchup.com/ also known as Trimble 3D Warehouse and Google 3D Warehouse

annotation then it will be possible in some cases to infer what geometric object the caption is describing, by matching between the generic descriptor in the model and equivalent terms in the caption.

We will therefore develop automated methods to assist in enhancing the geometric detail in order to support semantic annotation. The intention is to use, in the first instance, generic object models to assist in identifying the boundary of prominent building components, such as windows, within the texture mapped imagery or within photos that have been matched to, and hence registered with, the texture maps (see Mayer and Reznik (2006; 2005) for examples of related work in photogrammetry). The vectorised geometric representation of these boundaries will then be added to the building model geometry and labelled with the object category (e.g. "door") of the corresponding object detector.

Previous work on the detection of objects in images of buildings has mostly used template-based methods that require the design of object detectors (e.g. Chen et al (2011)). These methods may not be suitable for application to texture mapped imagery due to the wide range of appearances of objects such as doors and windows as a consequence both of stylistic variation and of variation in the viewpoint, lighting, colour and size. An alternative approach is the bag of visual words BoVW (Sivic and Zisserman, 2003) which involves detecting and describing a set of keypoints (or interest points) in the image. Given a training set of images, vector quantization is applied to cluster the extracted feature vectors to form a dictionary which captures the most frequently occurring patterns (the visual words). An unseen image is represented as a vector of visual word frequencies and image classification is performed by matching the vector to prototypical vectors.

2.2 Exploitation of building plans

As indicated above, annotated plans of buildings, such as may be found in published building guides and in some web documents about specific buildings, provide a potentially useful further source of information about the nature of rooms and spaces in buildings. When 3D building models lack such annotation, as is usually the case, it will be possible to match the plans to the ground projection of the building geometry and hence label the corresponding parts of the building.

In various applications of GIS it is common practice to integrate data from multiple sources in order to generate a representation that retains the best elements of the multiple sources (see Ross et al (2009) for an example that includes 3D data) – in our case we are interested in taking labels of plans in illustrated guides and transferring the labels to the appropriate matching parts of the 3D building. The process requires geometric and semantic matching and is referred to as conflation (Samal et al., 2004), though it is most commonly applied to 2D geographical data. If there is considerable variation in the representation of the same real world objects then it can be helpful to use probabilistic methods such as mutual information (Walter and Fritsch, 1999) and Bayesian learning methods (Jones et al., 1999) that employ multiple sources of evidence for equivalence. Most conflation methods in GIS operate on vector representations. As plans in guides to buildings are typically purely image based, it may be necessary to vectorize them, for which standard GIS methods are again available, prior to application of conflation procedures. An alternative to conventional GIS methods, many of which require some interactive control, has been described in Smart et al (2011), who presented matching methods that entailed rasterizing the projection of poorly structured building geometry prior to raster based matching with a rasterized digital map.

2.3 Feature matching methods and linking captions to models

In computer vision, image feature matching procedures based on SIFT (Scale Invariant Feature Transform) are now widely used (Mikolajczyk and Schmid, 2005). In our project these feature matching methods will support object detection (as explained above) and enable matching of captioned photos to the corresponding objects or regions of a building. The first step is to identify distinctive local regions, or keypoints, of an image, and compute descriptors for each keypoint. Matches between corresponding features in different images are established using measures of similarity between the keypoint descriptors. However, some of these local matches will often be inconsistent at an object (i.e. more global) level. These are eliminated by applying a RANSAC procedure (Fischler and R. Bolles, 1987) to the set of matching keypoints to robustly generate the fundamental matrix representing the perspective transformation between the two images. Those pairs of matching keypoints that are inconsistent with the

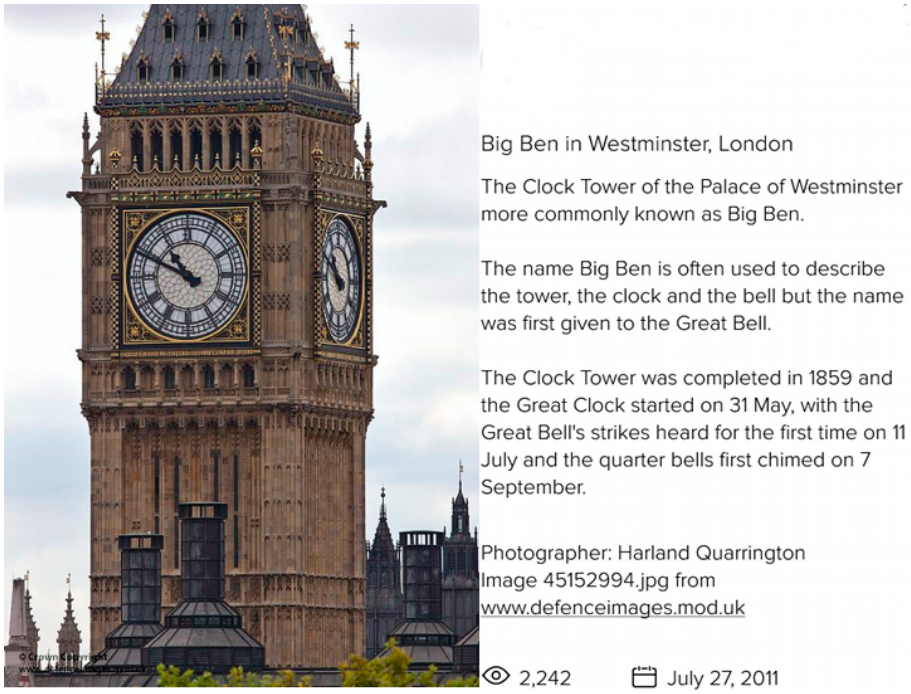


Figure 1: Flickr photo and accompanying caption of Big Ben

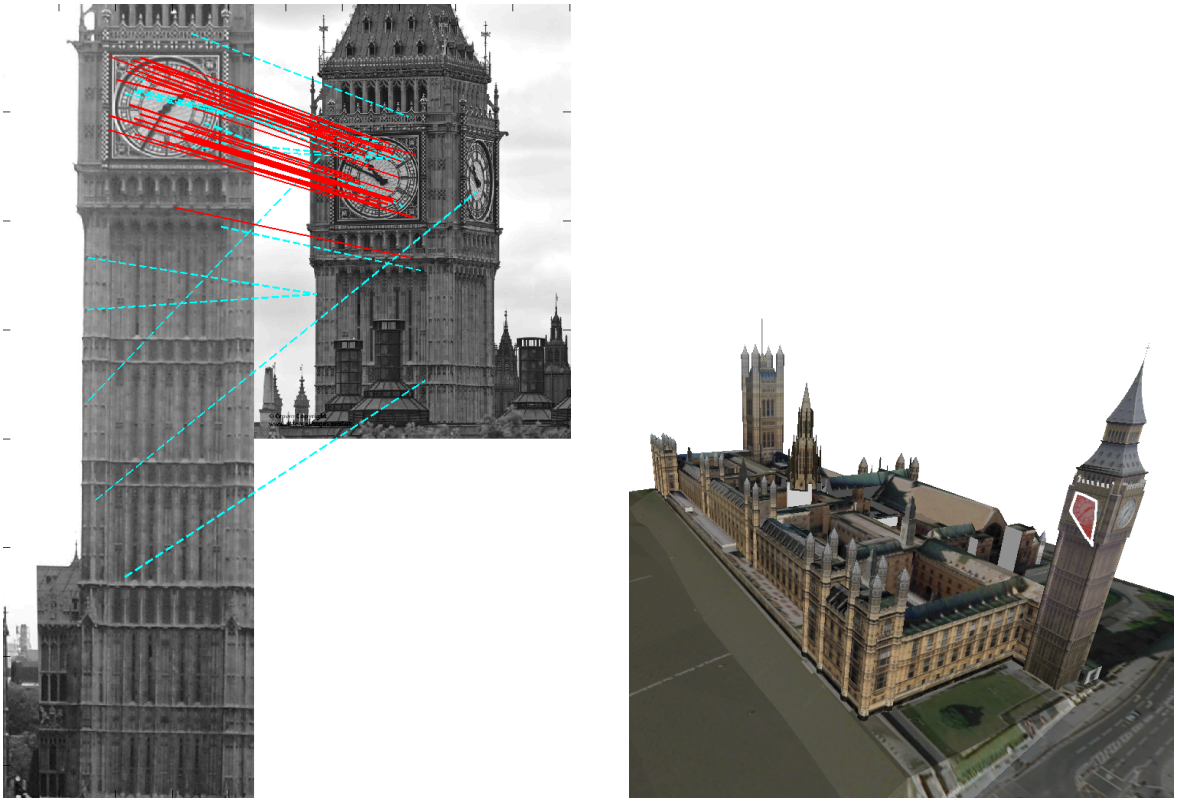
estimated transformation are identified as outliers, and removed.

Figure 1 illustrates an image and caption from Flickr of Big Ben which is part of the Houses of Parliament in London. It is of interest to match this captioned image to the corresponding part of the 3D model of the building (illustrated in Figure 2b) for purposes of annotation. Applying the SIFT-based matching procedure to compare the Flickr image with the set of texture map images for the 3D model, followed by the RANSAC procedure, results in one of the texture map images containing a strong cluster of matching keypoints for the region around the clock of Big Ben (Figure 2a), i.e. the inlier matched keypoint pairs, represented in the figure with red lines. Other, false matching outlier keypoint pairs are highlighted with dashed cyan lines in Figure 2a. The matched region on the 3D model is highlighted in red in Figure 2b and is based on the convex hull of the matching inlier keypoints. The caption becomes linked in the first instance to this region. In future work we will use other methods to refine this matching region. Initially this could use region growing from the initial set of inlier key points, whereby pixels in the captioned images adjacent to the inlier key points are transformed to the texture map image using the estimated fundamental matrix, and retained if the (dense) SIFT descriptors of the corresponding pixels match. In subsequent work the intention is to match the caption text to the corresponding generically labelled objects on the building model, which in the illustrated example would be the clock and the tower.

2.4 Linking to rich text descriptions

The work described here is a step towards the objective of linking detailed authoritative descriptions of parts of buildings to the corresponding objects in 3D geometric models. Russell et al (2013) have demonstrated such linking between parts of Wikipedia articles to 3D models. It is clear however from their examples that their methods are quite selective and can fail to match interesting descriptions to their respective building components. Their methods depend upon crowdsourcing of suitably captioned photos that match with the authoritative text, which in turn restricts the approach to well-photographed places and objects. While that approach has benefits for popular buildings, the challenge remains to develop more effective methods for detecting references to building components in rich textual descriptions (not just in social media) and to develop methods for linking them to the geometric model that are not entirely dependent upon multiple tagged photographs.

One approach to reducing the dependancy upon captioned photos is to develop effective methods for



(a) (b)

Figure 2: (a) Key point matching between a Flickr image of Big Ben (on the right) and the texture map imagery of a 3D model of the Houses of Parliament that includes Big Ben. Inlier matches are coloured as red lines and outliers as dashed cyan lines. (b) The 3D texture-mapped model of the Houses of Parliament in which the convex hull region of the matching inlier key points is highlighted in red.

understanding localisation expressions that tell the reader where particular objects are to be found. This will involve sophisticated natural language processing to detect and interpret the often vague spatial relations that are employed in descriptions of the locations of environmental objects (Kordjamshidi et al., 2011; Mani et al., 2010). Thus a door may be described as being in the west wall of a church, in which case it may be possible to identify the respective boundary (wall) of the 3D building model and hence the generically labelled geometric object within that wall. Equally, terms such as left and right, and architectural words such as lintel, arch and column, may help to locate a described object without recourse to a captioned photo. When captioned photos are available they may however be exploited to assist in the geometric annotation process. It remains to investigate the refinement and application of these and related methods for the purpose of detailed and authoritative annotation of building models and other 3D representations of the built and the natural world.

3 Concluding Remarks

In this paper we have summarised an approach to semantic and geometric enrichment of 3D building models that exploits a mix of captioned photos from social media with captioned photos and labelled ground plans obtained from illustrated guides. The methods differ from the current state of the art for annotating 3D models in that we employ structured 3D models, rather than 3D point clouds, and we focus on enhancing the geometric representation and generic annotation of objects within these models. Generic annotation, achieved through a combination of computer vision methods and GIS conflation procedures, facilitates the linking of text descriptions of particular types of object to the corresponding components of the 3D model. This will avoid the sometimes prohibitive dependence of existing methods on the presence of multiple captioned photos of all objects of interest. It also paves the way for linking

rich textual descriptions in authoritative guides to the corresponding geometric objects.

Acknowledgements

Jonathan Slade is funded by an EPSRC Industrial CASE studentship with Ordnance Survey, GB.

References

Z. Chen, Y. Li, and S. T. Birchfield. 2011. Visual detection of lintel-occluded doors by integrating multiple cues using data-driven MCMC. *Robotics and Autonomous Systems*.

M. Fischler and R. R. Bolles. 1987. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. In *Readings in computer vision: issues, problems, principles, and paradigms,*.

C.B. Jones, J.M. Ware, and D.R. Miller. 1999. A probabilistic approach to environmental change detection with area-class map data. In *ISD '99 International Workshop on Integrated Spatial Databases, Digital Images and GIS*, volume 1737 of *Lecture Notes in Computer Science*, pages 122–136, Berlin. Springer.

P. Kordjamshidi, M. van Otterlo, and M.F. Moens. 2011. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing*, 8(3).

I. Mani, C. Doran, D. Harris, et al. 2010. SpatialML: annotation scheme, resources, and evaluation. *Language Resources and Evaluation*, 44:263–280.

H. Mayer and S. Reznik. 2005. Building façade interpretation from image sequences. In U Stilla, F Rottensteiner, and S Hinz, editors, *International archives of photogrammetry, remote sensing and spatial information sciences*, volume XXXVI. Joint workshop of ISPRS and DAGM.

H. Mayer and S. Reznik. 2006. MCMC Linked With Implicit Shape Models and Plane Sweeping for 3D Building Facade Interpretation in Image Sequences. In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information*, volume XXXVI. ISPRS Commission III.

K. Mikolajczyk and C. Schmid. 2005. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10):1615–1630.

L. Ross, J. Bolling, J. Döllner, and B. Kleinschmit. 2009. Enhancing 3d city models with heterogeneous spatial information: Towards 3d land information systems. In *Lecture Notes in Geoinformation and Cartography*, pages 113–133.

B.C. Russell, R. Martin-Brualla, D.J. Butler, S. M. Seitz, and L. Zettlemoyer. 2013. 3d wikipedia: Using online text to automatically label and navigate reconstructed geometry. *ACM Transactions on Graphics (SIGGRAPH Asia 2013)*, 32(6).

A. Samal, S. Sheth, and K. Cueto. 2004. A feature-based approach to conflation of geospatial sources. *International Journal of Geographical Information Science*, 18(5):459–489.

I. Simon and S.M. Seitz. 2008. Scene segmentation using the wisdom of crowds. In *European Conference on Computer Vision (ECCV)*, pages 541–553.

J. Sivic and A. Zisserman. 2003. Video google: A text retrieval approach to object matching in videos. In *Int. Conf. Computer Vision*, pages 1470–1477.

P.D. Smart, J.A.Quinn, and C.B. Jones. 2011. City model enrichment. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(2):223–234.

N. Snavely, S.M. Seitz, and R. Szeliski. 2008. Modeling the world from internet photo collections. *Int. J. of Computer Vision*, 80(2):189–210.

V. Walter and D. Fritsch. 1999. Matching spatial data sets: A statistical approach. *International Journal of Geographical Information Science*, 13(5):445–473.

C. Zhang, Gao J., Wang O., Georgel P., Yang R., Davis J., Frahm J.M., and Pollefeys M. 2014. Personal photograph enhancement using internet photo collections. *IEEE Transactions on Vision and Computer Graphics*, 20(2):262–75.