Semantic and geometric enrichment of 3D geo-spatial building models with photo captions and illustration labels

Jon Slade[*], Christopher B. Jones[†] and Paul L. Rosin[‡]

School of Computer Science & Informatics, Cardiff University, United Kingdom

March 12, 2015

Summary

Whilst the number of 3D geo-spatial digital models of buildings with cultural heritage interest is burgeoning most lack semantic annotation that could be used to inform users of mobile and desktop applications about the architectural features and origins of the buildings. As part of an ongoing project this research describes methods for enriching 3D models with generic annotation implied from images of building components and from labelled building plans and diagrams, and with the text from object-specific photo captions from social media. The work is part of a broader initiative aimed at annotating 3D models with elements of text from authoritative architectural guides.

KEYWORDS: 3D building model, semantics, enrichment, texture matching, caption extraction.

1. Introduction

3D geo-data is of significant utility in e.g. the fields of inter-visibility analysis, radio communications, visualization for urban planning, indoor navigation, augmented reality, and in cultural heritage applications relating to the built environment. The increasing prevalence of 3D city models has not however been matched by corresponding improvement in their semantic content – often the models lack any semantic content (Jones *et al.*, 2014) which limits their use for some geo-data applications. 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.

This research is developing an approach to automate the process of semantic enrichment and is currently in its early stages. The central principle of the work is that texture maps on the building models can be matched to captioned images using computer vision techniques, allowing the captions to be linked to the corresponding parts of the building model. Recent work has been focussed on image and template matching steps, described below. Future work is then outlined.

With an emphasis upon buildings with cultural heritage the research builds upon methods that used captioned photos to annotate 3D models e.g. Simon and Seitz (2008) and Russell *et al.* (2013), and complements work such as Zhang *et al.* (2013) where 3D models were employed to enhance and annotate photos. Captioned photos might be sourced from social media sites such as *Flickr*.

The lack of well-captioned photos on social media for less well-visited buildings can be addressed through the use of captioned photos in cultural heritage guides – such captions may often be superior

[*] SladeJD@cardiff.ac.uk
[†] JonesCB2@cardiff.ac.uk
[‡] RosinPL@cardiff.ac.uk

in their description of, for example, the architecture and history of the building. Such guides may also contain diagrams of the layout of buildings labelled with the names of rooms and associated spaces. It is this combination of captioned images as found in social media and authoritative guides, and annotated plans and diagrams from authoritative texts, as well as some 2D cartographic data that can be used to improve the semantic content of the building models.

The exploitation of captioned photos of buildings in order to facilitate virtual exploration of buildings was pioneered by *Photo Tourism* (Snavely *et al.*, 2008) in which the computer vision feature matching scale-invariant feature transform (SIFT) method was used to match photos of buildings, compute camera pose and generate point clouds representing the 3D geometry of buildings (referred to as structure from motion). Consequently users could view captioned photos of a building within the recreated 3D scene, in addition to which new photos could be matched and tagged with pre-existing tag content. Simon and Seitz (2008) meanwhile used a related approach to annotate 3D point clouds with *Flickr* photo captions.

Subsequently Russell *et al.* (2013) achieved finer granularity in the annotation of building components, linking text in *Wikipedia* articles about particular buildings to locations in 3D models. Again the models were created from sets of *Flickr* images tagged with the respective building name, with annotation making use of Google *Image* search based on text from the *Wikipedia* entry. As with these other methods Russell *et al*'s approach used building models comprised of point clouds rather than explicit structure. Our work differs in that we use structured 3D building models, such as those in Trimble's *3D Warehouse*[§]. A number of geo-data applications, such as navigation, can require the ability to refer to individual spatial objects at a fine level of granularity - it is argued here that only *structured* 3D building models could easily provide such capability. Moreover, the generic object annotation which acts as an aide to the latter caption-extraction step and which is described below, requires structural models.

2. Methods

2.1. Annotating 3D Building Models with Generic Objects

Annotation of rooms, doors, windows, arches, clocks etc may be assisted by the identification of their geometry within the building model. Such geometric representation in a building model may also be of use within applications which guide users around a building. Incidentally, Mayer and Reznik (2005) describe similar approaches within photogrammetry.

Initial work in our study has focused on a template matching approach (Kroon, 2011) whereby a template such as a tightly cropped photo of a generic window design is matched to an image of the building via normalised cross-correlation and non-maximum suppression. An alternative to template matching would be the bag of visual words (BoVW) approach (Sivic and Zisserman, 2003).

Figure 1a shows the result of template matching using a cropped window template (from another building) taken from a Google *Image* search and a texture map from the building model. Since diagrams are a suitable source of templates, we have investigated their use in template matching – Figure 1b shows the result of template matching between the same template and texture map but after both have been converted in to line drawings using the method of Kang *et al.* (2007). Kang *et al.*'s approach has been used since it is effective in extracting long, curvilinear lines while minimising the amount of clutter and noise in the output. The matches in both figures are represented by the bounding boxes in **green**. Note the false positives in Figure 1a – whilst using the line drawing approach (Figure 1b) has removed these and has detected four out of five windows, the central window on the second storey has not been detected, indicating that further refinement of the method could be conducted.

[§] https://3dwarehouse.sketchup.com

Figure 1a Template Matching with Photos
Source data: template 89x168 pixels, image 800x904 pixels. Normalised cross correlation non-maximum suppression threshold 0.73

Figure 1b Template Matching with Lines
Source data: template 95x196 pixels, image 800x904 pixels. Normalised cross correlation non-maximum suppression threshold 0.62 *

***** Blurring has been carried out in advance of matching in an attempt to consolidate scattered disjoint components of what should be solid lines, and to extend the influence of the highly localised features i.e. lines, thus increasing the tolerance of the matching to geometric differences between the template and the texture map image. The image was also padded by 100 pixels to remove erroneous detections at its edges.

2.2. Building Plan Exploitation

The text from supplementary building plans can be added to the models, following a process of spatially matching the ground or side projection in the plans with the geometry of the footprint of the model. It is envisaged that such spatial joining, via GIS conflation (Samal *et al.*, 2004), will require both geometric and semantic matching. Should there be ambiguity in the match then various machine learning techniques such as probabilistic mutual information (Walter and Fritsch, 1999) and Bayesian methods (Jones *et al.*, 1999) could be used. Incidentally, whilst most conflation works on vector data, the work of Smart *et al.* (2011) presented matching methods that entailed rasterising building geometry which lacked complete, contiguous structure before matching the result with a rasterized digital map.

2.3. Feature Matching Methods and Linking Captions to Models

Image feature matching methods such as SIFT are now widely used within computer vision (Mikolajczyk *et al.*, 2005). In this research these feature matching methods support the object detection phase, enabling the matching of captioned photos. SIFT can facilitate matching by identifying quantitative vector descriptors for distinctive local regions or key points in an image, where the similarity measures for those descriptors are then used to inform matches. Methods such as RANSAC (Fischler and Bolles, 1981) can be used to reduce outliers which may occur – this is achieved through the generation of a fundamental matrix for the perspective transformation between the two images where those matches which are inconsistent with the transform can then be removed.

Figure 2a Key point matching

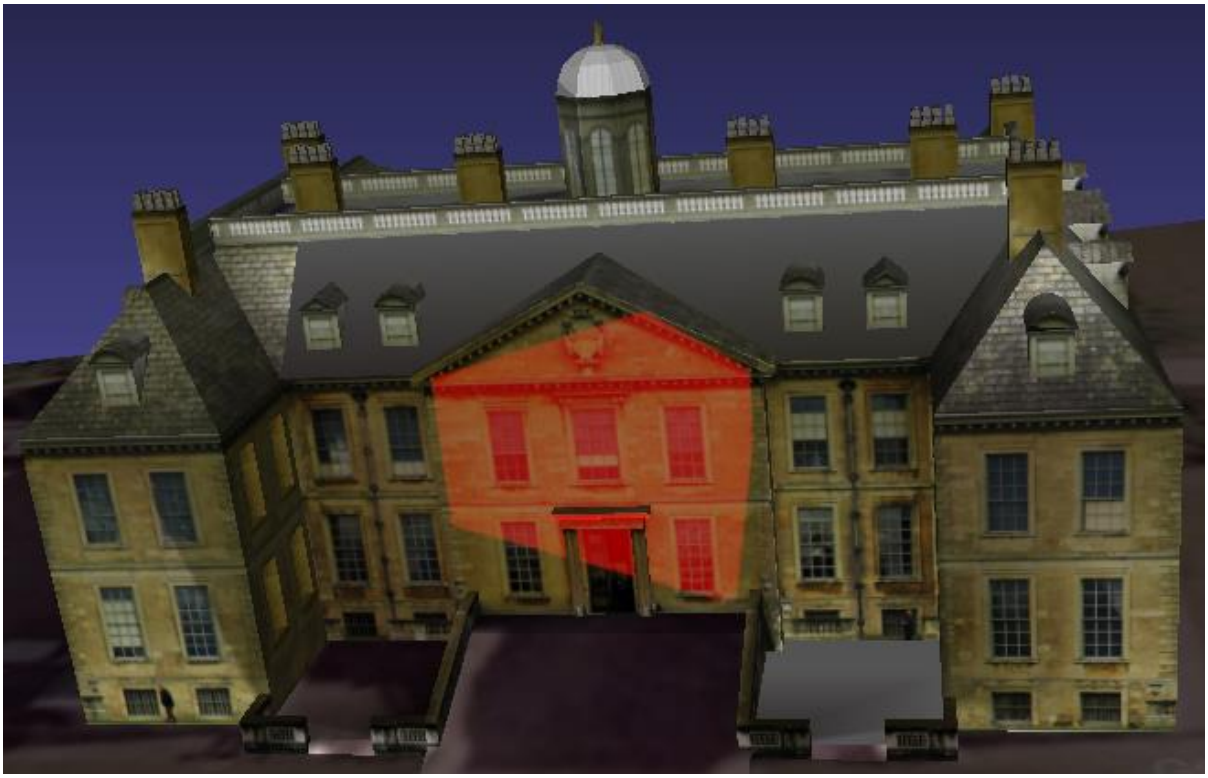

Figure 2b Convex hull on building model of matching inlier key points

(a) Key point matching between a *Flickr* image of Belton House (on the right) and the texture map imagery of a 3D model of Belton House (on the left). Inlier matches are coloured as **red** lines and outliers as dashed **cyan** lines

(b) The 3D texture-mapped model of Belton House in which the convex hull region of the matching inlier key points is highlighted in **red**.

Figures 2a and 2b shows the result of SIFT-based matching (with RANSAC outlier-removal) between a captioned photo from *Flickr* of the heritage English mansion Belton House in Lincolnshire, with a corresponding texture map from the surface of a 3D model of the same building taken from the *3D Warehouse*. Note that there is a strong cluster of matching key points on the gable front, pedimented bay above the main entrance steps (Figure 2a) i.e. the inlier matched key point pairs, represented in the figure with **red** lines. Other, false matching outlier key point 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 key points. The caption is linked, initially, to this region.

Future work may refine this matching region, perhaps using region growing from the inlier key points whereby adjacent pixels from the captioned image could be transformed to the texture map image using the estimated fundamental matrix, and retained if the (dense) SIFT descriptors of the corresponding pixels match.

2.4. Linking to Rich Text Descriptions

Russell *et al.* (2013) demonstrated the linking of descriptive texts to 3D building models, in their case from *Wikipedia*, and our work plans to build upon that by linking the models to authoritative texts. Based on their examples however, their methods are quite selective (relying on the existence of architectural components in just a *Wikipedia* entry) and can therefore fail to match interesting descriptions to respective building components. Depending as they do on the availability of well-captioned crowdsourced images on social media, their approach and the field in general would benefit from improved methods for matching texts from sources other than on social media – in turn this might pave the way for semantic enrichment of models of less popular buildings.

One such approach might involve natural language processing to detect and interpret the often vague spatial relations found in descriptions of the locations of real-world objects (Kordjamshidi *et al.*, 2011; Mani *et al.*, 2010). For example a door might be described as being in the east wall of a church – in turn identification of the respective wall of the 3D building model and hence the generically labelled geometric object within that wall, may be achieved. Moreover, terms such as right and left, and architectural descriptions such as arch or lintel may improve object location determination. Captioned photos may also be exploited to assist in the geometric annotation process.

3. Concluding Remarks

We have outlined an approach for semantic and geometric enrichment of existing 3D building models, exploiting captioned social media photos and labelled ground plans obtained from illustrated guides. Our methods differ from the current state of the art in that we employ structured 3D models, rather than 3D point clouds, and focus on enhancing geometry and generic annotation of objects within these models.

The latter, achieved through computer vision methods and GIS conflation, facilitates linking text descriptions of particular object types to corresponding components in the 3D model. This avoids the dependence of existing methods on the need for large numbers of captioned photos. Finally, we pave the way for linking rich textual descriptions in authoritative guides to corresponding geometric objects.

4. Acknowledgements

Jon Slade is a PhD candidate funded by an EPSRC Industrial CASE studentship with Ordnance Survey, Great Britain.

5. Biography

Jon Slade holds an MSc in GIScience from UCL, with a BSc in Geology from the University of Bristol. Previous roles comprise 13 years as a consultant in the commercial IT sector including work as a GIS Analyst Developer at civil engineering and mobile telecoms firms.

Chris Jones is Professor of Geographical Information Systems in Cardiff University. His research interests include multi-scale spatial database design and integration, map generalisation and geographical information retrieval with particular regard to spatio-textual indexing methods, gazetteers, vernacular place names and spatial search engine architectures.

Paul L. Rosin is Professor of Computer Vision in Cardiff University. Previous posts include Brunel University, Joint Research Centre, Italy and Curtin University of Technology, Australia. His research interests include computer vision, remote sensing, mesh processing, non-photorealistic rendering, and the analysis of shape in art and architecture.

References

FISCHLER, M. A. AND BOLLES, R. C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM* [online], 24(6): 726-740. Available from: http://dl.acm.org/ft_gateway.cfm?id=358692 [Accessed 11 June 2014].

JONES, C. B., ROSIN, P. L. AND SLADE, J., 2014. Semantic and geometric enrichment of 3D geo-spatial models with captioned photos and labelled illustrations [online]. *In:* I. Press, (Ed). *Workshop on Vision and Language (VL4)*, Dublin. Available from: http://www.aclweb.org/anthology/W/W14/W14-54.pdf#page=74 [Accessed 18 August 2014].

JONES, C. B., WARE, J. M. AND MILLER, D. R., 1999. A Probabilistic Approach to Environmental Change Detection with Area-Class Map Data. [online] *In:* P. Agouris and A. Stefanidis (Eds.) *Integrated Spatial Databases.* Springer Berlin Heidelberg. pp. 8. Available from: http://link.springer.com/content/pdf/10.1007%2F3-540-46621-5_8.pdf [Accessed 11 June 2014].

KANG, H., LEE, S. AND CHUI, C. K., 2007. Coherent line drawing [online]. *In:* M. Agrawala and O. Deussen, (Eds). *5th International Symposium on Non-photorealistic Animation and Rendering (NPAR)*, San Diego, USA, 04-05 August 2007 1274878. ACM. Available from: http://dl.acm.org/ft_gateway.cfm?id=1274878 [Accessed 05 November 2014].

KORDJAMSHIDI, P., VAN OTTERLO, M. AND MOENS, M.-F., 2011. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing* [online], 8(3): 1-36. Available from: http://dl.acm.org/ft_gateway.cfm?id=2050105 [Accessed 11 June 2014].

KROON, D.-J., 2011. *MATLAB Central - File Exchange - Fast/Robust Template Matching* [online]. MathWorks Inc. Available from: http://www.mathworks.co.uk/matlabcentral/fileexchange/24925-fast-robust-template-matching/content/template_matching.m [Accessed 01 September].

MANI, I., DORAN, C., HARRIS, D., HITZEMAN, J., QUIMBY, R., RICHER, J., WELLNER, B., MARDIS, S. AND CLANCY, S., 2010. SpatialML: annotation scheme, resources, and evaluation. *Language Resources and Evaluation* [online], 44(3): 263-280. Available from: http://link.springer.com/content/pdf/10.1007%2Fs10579-010-9121-0.pdf [Accessed 11 June 2014].

MAYER, H. AND REZNIK, S., 2005. Building facade interpretation from image sequences [online]. *In:* S. U, R. F and H. S, (Eds). *ISPRS Workshop on Object Extraction for 3D City Models, Road Databases, and Traffic Monitoring - Concepts, Algorithms, and Evaluation (CMRT)*, Vienna, Austria, 29-30 August 2005. Available from: http://www.isprs.org/proceedings/XXXVI/3-W24/papers/CMRT05_Mayer_Reznik.pdf

[Accessed 02 September 2014].
MIKOLAJCZYK, K., TUYTELAARS, T., SCHMID, C., ZISSERMAN, A., MATAS, J., SCHAFFALITZKY, F., KADIR, T. AND GOOL, L. V., 2005. A Comparison of Affine Region Detectors. *International Journal of Computer Vision* [online], 65(1-2): 43-72. Available from: http://link.springer.com/content/pdf/10.1007%2Fs11263-005-3848-x.pdf [Accessed 14 July 2014].
RUSSELL, B. C., MARTIN-BRUALLA, R., BUTLER, D. J., SEITZ, S. M. AND ZETTLEMOYER, L., 2013. 3D Wikipedia: using online text to automatically label and navigate reconstructed geometry. *ACM Transactions on Graphics (TOG)* [online], 32(6): 193. Available from: http://dl.acm.org/ft_gateway.cfm?id=1141964 [Accessed 20 June 2014].
SAMAL, A., SETH, S. AND CUETO, K., 2004. A feature-based approach to conflation of geospatial sources. *International Journal of Geographical Information Science* [online], 18(5): 459-489. Available from: http://www.tandfonline.com/doi/full/10.1080/13658810410001658076 [Accessed 11 June 2014].
SIMON, I. AND SEITZ, S. M., 2008. Scene Segmentation Using the Wisdom of Crowds [online]. *In:* D. Forsyth, P. Torr and A. Zisserman, (Eds). *10th European Conference on Computer Vision (ECCV)*, Marseille, France, 13-16 October 2008. Springer Berlin Heidelberg. Available from: http://grail.cs.washington.edu/pub/papers/simon08ss.pdf [Accessed 11 June 2014].
SIVIC, J. AND ZISSERMAN, A., 2003. Video Google: a text retrieval approach to object matching in videos [online]. *In:* B. Werner, (Ed). *9th IEEE International Conference on Computer Vision (ICCV)*, Nice, France, 13-16 October. IEEE Computer Society. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1238663 [Accessed 11 June 2014].
SMART, P. D., QUINN, J. A. AND JONES, C. B., 2011. City model enrichment. *ISPRS Journal of Photogrammetry and Remote Sensing* [online], 66(2): 223-234. Available from: http://www.sciencedirect.com/science/article/pii/S0924271610001231 [Accessed 11 June 2014].
SNAVELY, N., SEITZ, S. M. AND SZELISKI, R., 2008. Modeling the World from Internet Photo Collections. *International Journal of Computer Vision* [online], 80(2): 189-210. Available from: http://link.springer.com/content/pdf/10.1007%2Fs11263-007-0107-3.pdf [Accessed 11 June 2014].
WALTER, V. AND FRITSCH, D., 1999. Matching spatial data sets: a statistical approach. *International Journal of Geographical Information Science* [online], 13(5): 445-473. Available from: http://www.tandfonline.com/doi/pdf/10.1080/136588199241157 [Accessed 11 June 2014].
ZHANG, C., GAO, J., WANG, O., GEORGEL, P., YANG, R., DAVIS, J., FRAHM, J.-M. AND POLLEFEYS, M., 2013. Personal Photograph Enhancement Using Internet Photo Collections. *IEEE Transactions on Visualization and Computer Graphics* [online], 20(2): 262-275. Available from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6509872 [Accessed 11 June 2014].