Automated Clustering of Landmark Tags in Urban Images

Phil Bartie[1], William Mackaness[2], Philipp Petrenz[3], Anna Dickinson[3]

[1] School of Natural Sciences, University of Stirling, Stirling, FK9 4LA e: phil.bartie@stir.ac.uk

[2] School of GeoSciences, University of Edinburgh, Drummond St, Edinburgh, EH8 9XP

[3] Informatics Forum, University of Edinburgh, 10 Crichton Street, Edinburgh, EH8 9AB

1. Introduction

It is well known that landmarks form an important component in navigational instructions and scene descriptions (May et al., 2005; Raubal and Winter, 2002; Elias and Brenner, 2004; Duckham et al., 2010). They have properties which differ from the surrounding region and are considered to be salient according to structural, visual, and semantic components (Sorrows and Hirtle, 1999). The motivation for this work was to automate the verification of a landmark saliency model against user generated tags of landmarks in urban scenes. To do this the tags collected from an online survey need to be grouped according to the object being described, this is problematic due to the variety of supplied descriptions which varied in the level of detail (e.g. tower, clock on tower), and the vocabulary from generic types (e.g. church) to specific names (e.g. St Giles' Kirk).

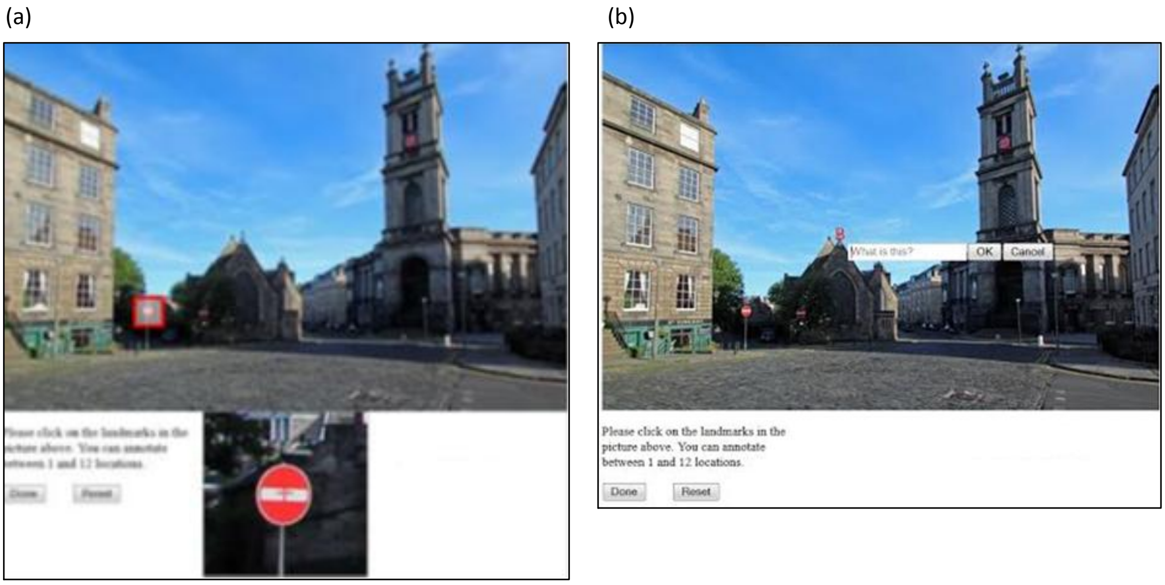
This paper focuses on the clustering technique developed to analyse the tags from 185 participants, who were asked to tag interesting objects in various urban scenes. The method developed automatically clusters the supplied tags using spatial and semantic relationships enabling the identification of the most important objects in a range of urban scenes, which at a later date will be used to verify the saliency model.

2. Web Experiment

A web based experiment was conducted to identify interesting objects in a number of urban scenes. The experiment was publicised through social media, attracting 185 participants. Users were assigned images randomly from a set of 37, and able to leave the experiment at any time but encouraged to complete as many images as possible by giving them an additional entry into a prize draw for each completed. For each task the participant would see an image of part of Edinburgh city, and be asked to identify features of interest by tagging them on the image. The user's profile and knowledge of the city was recorded as part of this process.

All images were captured on the same day in the early morning, in an effort to minimise weather variation, and object occlusion by other city occupants. The ambition was to replicate as closely as possible the street experienced, although it is recognised from previous landscape studies (Shafer and Brush, 1977; Zube et al., 1982; Linton, 1968; Daniel and Vining, 1983), that imagery can introduce a bias in the way it is captured and displayed. In an effort to minimise these the images were captured using a wide angle lens, and due to monitors not offering the same level of visual detail as when on the street, a magnifying region was added to the web viewer, as shown in Figure 1. This allowed the participant to see a magnified portion as they moved the mouse crosshair around the main image, giving a similar level of detail to that experienced on the street, and enabling them to more easily identify and tag more distant and smaller objects.

Once the participant had clicked on the image at the location of something they considered interesting, they were presented with an input text to enter free text which described the object (Figure 1b), such as a church, pub, or no entry sign. Each participant was permitted up to 12 tags per image.



Please click on the landmarks in the picture above. You can annotate between 1 and 12 locations.

Please click on the landmarks in the picture above. You can annotate between 1 and 12 locations.

Figure 1: Web based landmark tagging experiment (a) overview and magnified region (b) adding annotation text

The number of tags per image ranged from 178 tags to 451 tags, with an average of 90 unique participants per image. The average number of tags generated per participant was 56, with a total of 10,350 tags created across the 37 images.

3. Results

An example of the output is shown in Figure 2 which displays the tagged locations and the supplied tag text.

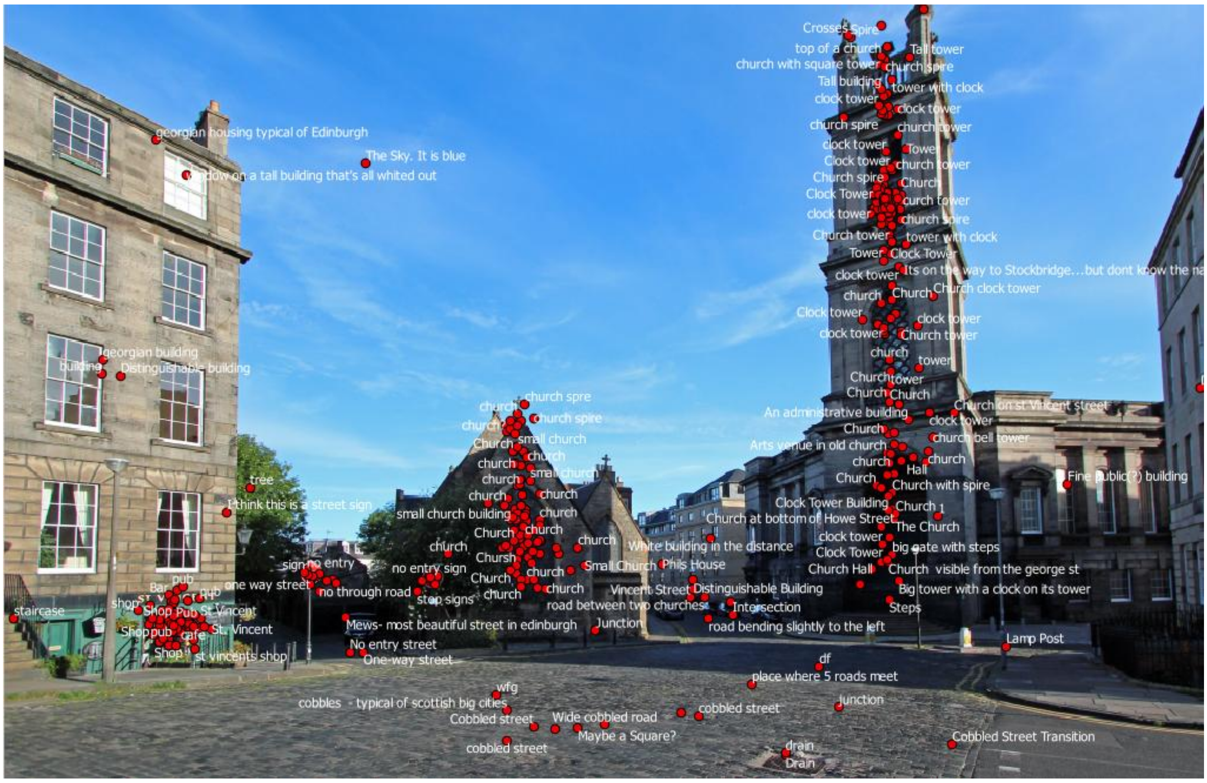


Figure 2: Example Scene with Tags from All Users Displayed

The tags for this image are shown as a word cloud in Figure 3, whereby more frequent terms are represented in a larger font. In this example the *tower* and *clock* refer to the church on the right of Figure 2, while *church* was used for both churches in the scene. Without considering the tag locations it is not possible to identify these two distinct groups, and therefore the landmarks visible.

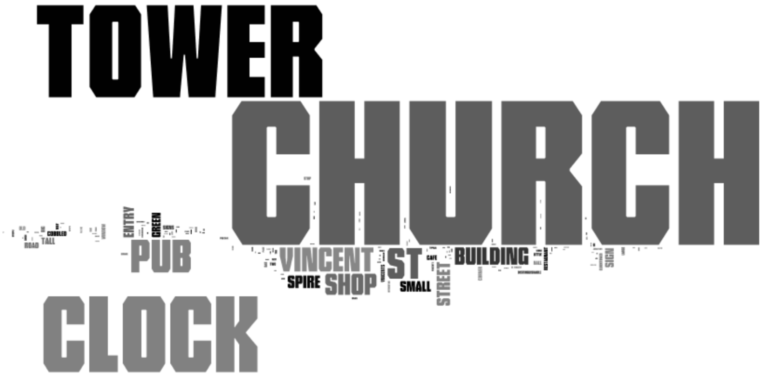


Figure 3: Word Cloud for Tags from a Single Image

The spatial pattern of the supplied tag locations may be summarised using spatial clustering, such as Kernel Density Estimation (KDE). The results for four scenes are shown in Figure 4, where red shows a dense concentration of user tags. The dense spatial concentrations are clearly noticeable for the two churches and the building on the left of the scene (a public house) in Figure 4a. In particular there is a concentration of tags around the top of the taller church tower where there is a clock face. In Figure 4b the KDE has highlighted a single cluster (group 1) where there are in fact two distinct features, which are at different viewing distances but a similar viewing angle from the observer. Figure 5 shows this in greater detail, where, due to the KDE radius, groups (i) and (ii) have been merged. This results in dense clustering not connected to a single landmark, but an artefact of two objects having a similar viewing angle. Similarly in Figure 4d, Scott Monument and Calton Hill tags have been clustered together as a single entity (group 3), and three features are clustered in group 5.

In contrast single features are presented as two distinct clusters in Figure 4c and Figure 4d (groups 2 and 4). The spatial pattern shows focal points on the façades, which although interesting fails to display the dominance of those features as single objects in the scene.

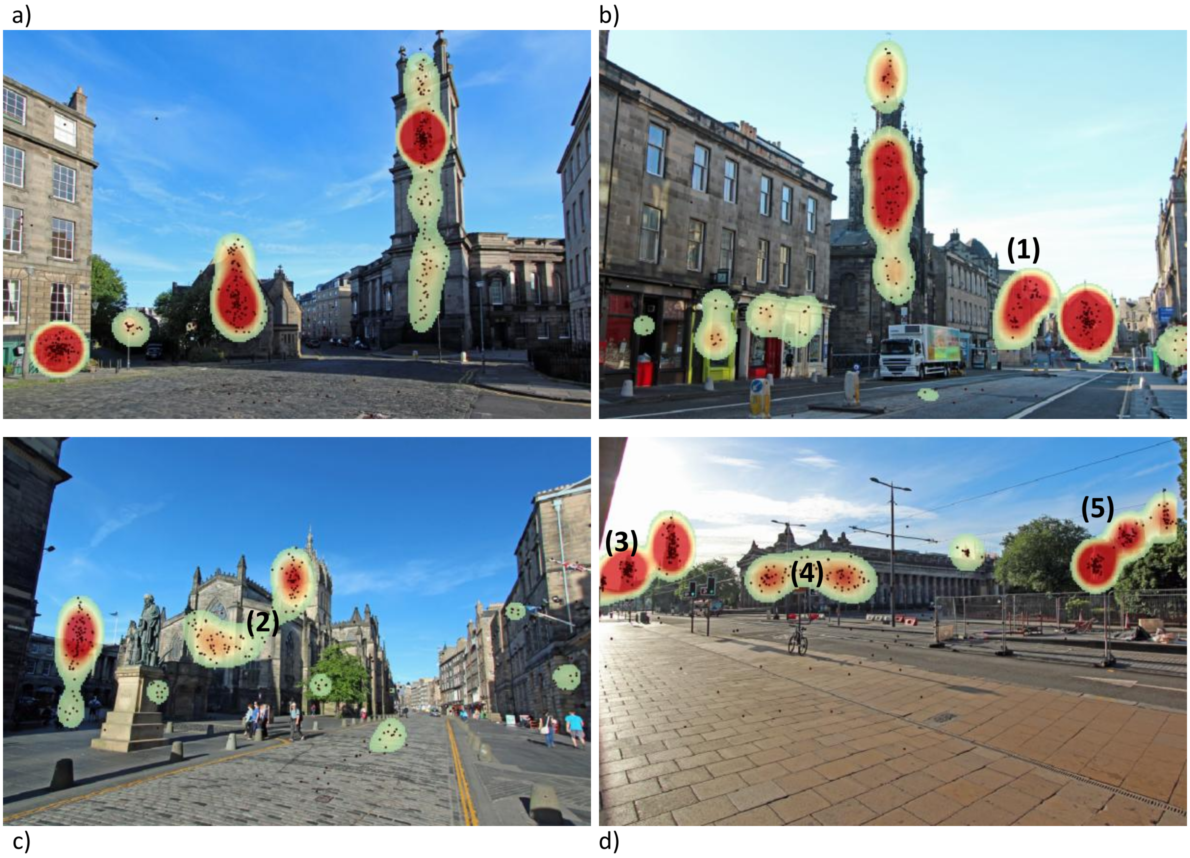


Figure 4: Kernel Density Estimation for User Tags (where red = dense clustering)

To improve upon this outcome a clustering technique was developed which included both spatial and semantic components, as described in Section 4, with its performance discussed in Section 5.

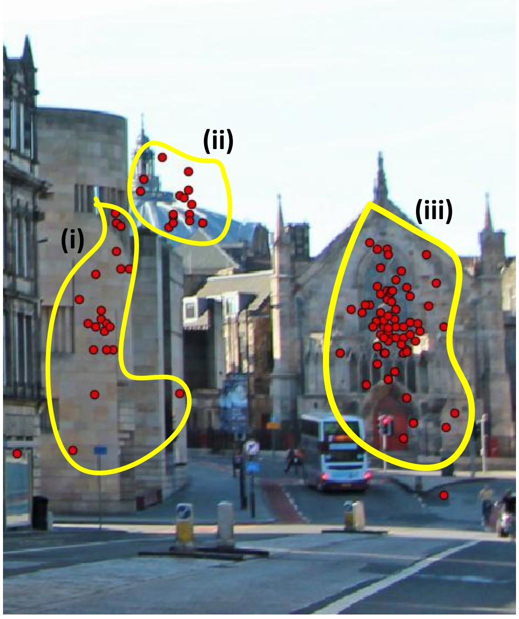


Figure 5: Spatial Clustering Errors due to Ignoring Distance

4. Spatial and Semantic Clustering

The participants supplied free text annotations for each image, consisting of any number of words. This allowed for a more natural dataset of descriptive object terms to be collected, but added complexity in analysis and term matching.

A fuzzy text matching technique based on character level trigrams was used to group similar terms (Lin, 1998; Zamora et al., 1981). This rated phrase similarity by calculating the number of shared three letter combinations found, while ignoring punctuation and letter case. To improve the matching process it was necessary to also ignore stop words such as '*of*','*the*','*a*'. The Trigram matching results are shown for a number of examples in Table 1, with values from 0 (no match) to 1 (exact match). Trigrams perform well in matching word stems ('*church*' versus '*churches*'), and misspellings ('*monument*' vs '*momument*'). However they are not able to recognise semantic similarities, for example the connection between a tag for a *church* and a *cathedral* (score of 0.0625).

An enhanced matching function was developed which included access to a synonym table allowing for conceptually similar terms, such as '*street*' and '*road*', '*cathedral*' and '*church*', and '*memorial*' and '*statue*' to be treated as identical. The results are shown in Table 1, where '*church*' and '*cathedral*' score an exact match of 1.0, and '*church tower*' and '*cathedral spire*' also score an exact match. The synonym table was constructed by looking at the most commonly occurring words from all images. This approach was straight forward to implement and quick to populate the synonym table, but lacks the ability to model partonomic relationships (i.e. relationships between an object's parts) therefore the score for '*church*' and '*clock tower*' remains low (0.0556).

Phrase pairs were collected by processing each tag in turn, searching the corresponding image for nearby tags within a defined pixel distance. The content similarity for each tag pair was calculated and those with a score greater than 0.3 were considered to be related. All of the tags were processed resulting in a topological network of connected content for each image.

Table 1: Word Similarity using Trigrams

Phrase One	Phrase Two	Trigram Matching (0 to 1)	Enhanced Matching (0 to 1)
church	Churches	0.6000	0.6000
monument	momument *[1]	0.5000	0.5000
church	Cathedral	0.0625	1.0000
church tower	Clock	0.0556	0.0556
church tower	Cathedral	0.0455	0.5385
church tower	cathedral spire	0.0357	1.0000

*[1] *intentional spelling error based on user supplied tag*

4.1 Expanding the Network of Linked Tags using a Secondary Pass

In some cases running the process a single time resulted in small tag groups being left as orphan clusters. For example in Figure 6 on the 'First Pass' three cluster groups are formed, relating to two objects of a *no entry sign* and a *church with a clock tower*. The two groups on the right remain distinct as no synonym entry links the *church* tags with *clock* or *clock tower*, and the other *clock tower* tag was outside the search radius. This can be addressed by increasing the search distance but that could result in separate object instances being combined (e.g. two nearby churches in Figure 2 are grouped being treated as a single entity). Instead the data was processed a second time using the same buffer distance but with an expanded vocabulary of related terms gained from the first pass. In doing this only nearby tag groups could be merged, reducing the likelihood of separate objects being linked, but as the related vocabulary had automatically been expanded so greater conceptual links could be made. This is a form of query expansion (Xu and Croft, 1996; Chum et al., 2007) , limited by the spatial location of the supplied tags. For example a *church* node may be joined to a *clock tower* node, even though they do not share any similar terminology based on a *church tower* node elsewhere being linked to a *clock tower* through the common term *tower*. Figure 6 shows this concept, where in the second pass a greater number of linkages have been added between groups as a result of their expanded semantic connections. The result is an expansion of the network, and reduction of groups.

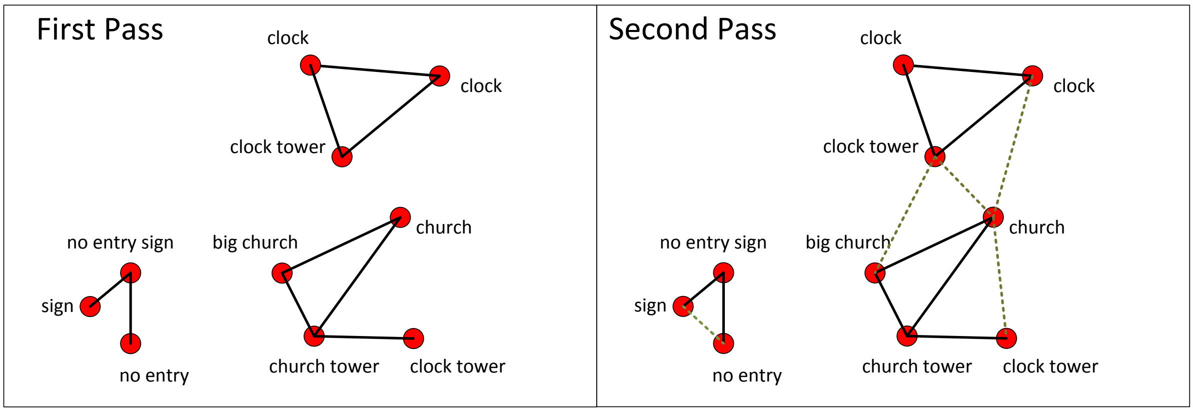


Figure 6: Expanding the Linked Network with Secondary Pass

5. Results

An example of the output from this process is shown in Figure 7, where colours are assigned randomly based on Cluster Group ID. There are many improvements compared to the spatial only clustering (Figure 4), as now two groups are identifiable in b (group 1) and a single group identified in c (group 2) and d (group 4). The previously single group at d (group 3) is now separated into two groups, however there is also an overlap occurring (orange group connects to red group).

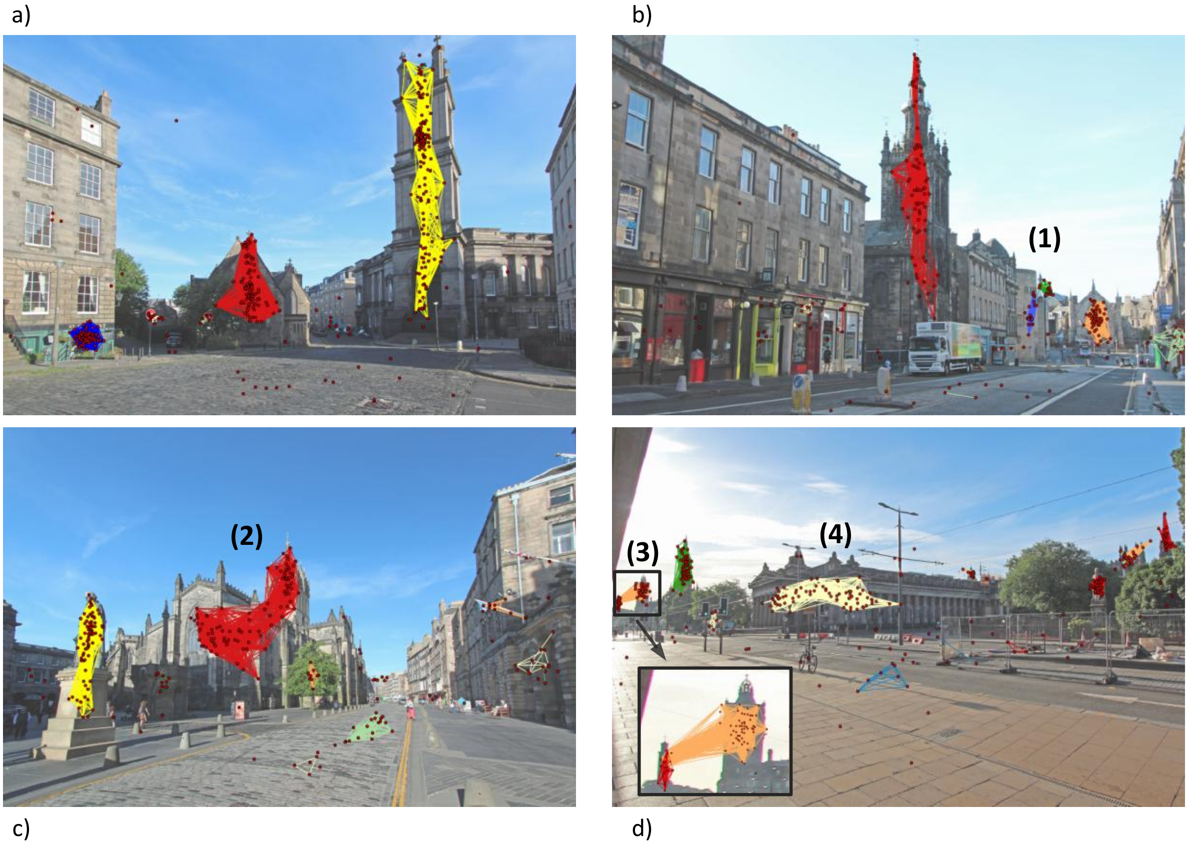


Figure 7: Spatial-Semantic Clusters

Tag groups were identified for each image making it possible to automatically generate related word lists. For example *church* is linked to *church spire*, which is linked to *church tower*, which is linked to *clock tower*, which is linked to *clock*. Tags define objects spatially and conceptually, and the frequency of each tagged phrase gives an indication of the most common term used for that object and its parts. Once tags have been linked in this way it is possible to calculate tag group centroids relating to the concept centres, for example the *clock* on the *clock tower*, which is part of the *church*. It is also possible to generate a list of the most frequent terms used per object, rather than per image, as shown in Figure 8.

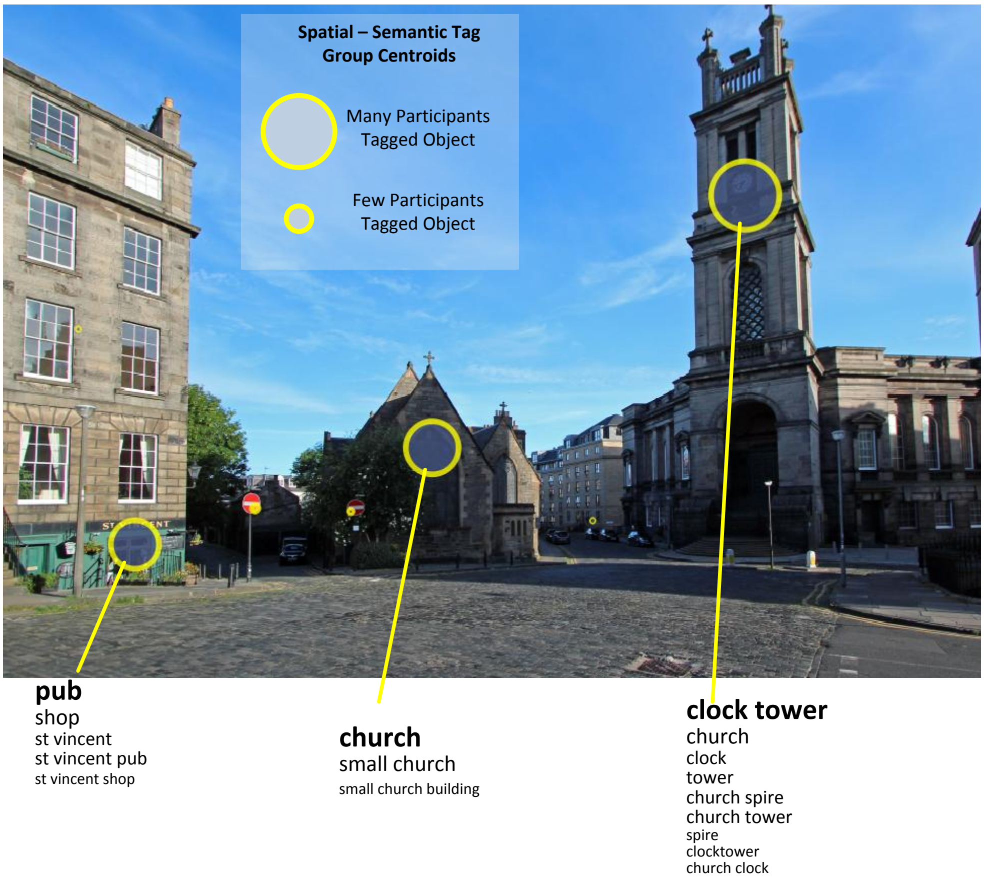


Figure 8: Phrase Ranking per Identified Object

Comparing the spatial clusters against the spatial-semantic clusters gives an insight in to objects which are interesting and easy to define versus those of interest which are hard to define.

6. Conclusions and Future Work

The paper outlines a method to identify clusters of tags supplied for urban scenes. A web based experiment was conducted whereby people tagged objects they considered to be interesting in the urban scene, adding free text annotations. The dataset was analysed to identify the interesting city objects in each image. Spatial clustering alone was shown to be flawed in certain cases where two objects at a similar viewing angle, but different distances away from the observer and containing different tag terms, would erroneously form a single cluster. Instead a new method was developed which combined spatial and semantic clustering techniques.

The method collects nearby tags which show a correlation using trigram fuzzy matching. Synonyms and stop words were used to improve the matching, and a network of connected tags was generated for each image. This was expanded in a secondary pass by using the linkages discovered on the first pass to join up orphaned tag groups. The results show that it was possible to automatically identified objects of interest from the user supplied tags, and that term frequencies could be discovered at an object level.

Future work will compare the results of this user experiment against a model of landmark saliency. This will compare the relative dominance ranks at an object level, and the saliency model will be enhanced from the findings. Term frequencies and variations by object type and viewing distance will be conducted as well as by user age and city familiarity, giving a greater understanding of how people refer to features of interest in urban scenes.

7. Acknowledgements

The research leading to these results has received funding from the EC's 7th Framework Programme (FP7/2011-2014) under grant agreement no. 270019 (SpaceBook project).

8. References

CHUM O, PHILBIN J, SIVIC J, ISARD M, ZISSERMAN A (2007) Total recall: Automatic query expansion with a generative feature model for object retrieval. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE

DANIEL TC, VINING J (1983) Methodological Issues in the Assessment of Landscape Quality. IN Altman, I, Wohwill, J (Eds.) In Behaviour and the Natural Environment. Plenum Press, New York

DUCKHAM M, WINTER S, ROBINSON M (2010) Including landmarks in routing instructions. *Journal of Location-Based Services* 4 28-52

ELIAS B, BRENNER C (2004) Automatic generation and application of landmarks in navigation data sets. IN Fisher, PF (Ed.) Developments in Spatial Data Handling. Springer, Berlin

LIN D (1998) An information-theoretic definition of similarity. *Proceedings of the Fifteenth International Conference on Machine Learning*. Madison, Wisconsin, USA, Morgan Kaufmann Publishers Inc.

LINTON DL (1968) The assessment of scenery as a Natural Resource. *Scottish Geographical Magazine* 84**:** 219 - 238

MAY AJ, ROSS T, BAYER SH (2005) Incorporating landmarks in driver navigation system design: An overview of results from the REGIONAL project. *Journal of Navigation* 58**:** 47-65

RAUBAL M, WINTER S (2002) Enriching wayfinding instructions with local landmarks IN Egenhofer, MJ, Mark, DM (Eds.) Second International Conference GIScience. Springer, Boulder, USA

SHAFER EL, BRUSH RO (1977) How to measure preferences for photographs of natural landscapes. *Landscape Planning* 4**:** 237-256

SORROWS M, HIRTLE S (1999) The nature of landmarks for real and electronic spaces. IN Freksa, C, Mark, D (Eds.) Spatial information theory. Springer,

XU J, CROFT WB (1996) Query expansion using local and global document analysis.

Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. ACM

ZAMORA EM, POLLOCK JJ, ZAMORA A (1981) The use of trigram analysis for spelling error detection. *Information Processing & Management* 17**:** 305-316

ZUBE EH, SELL JL, TAYLOR JG (1982) Landscape perception: research, application and theory. *Landscape planning* 9**:** 1-33

Biography

Phil Bartie is a lecturer in Geospatial Technologies at the University of Stirling.

William Mackaness is a senior lecturer in the School of GeoSciences at University of Edinburgh.

Anna Dickinson is a specialist in usability engineering and experimental methods, working at School of Informatics at the University of Edinburgh.

Philipp Petrenz is a PhD candidate also at the School of Informatics and works on automated text classification methods.

Their mutual research interests are in location based services, specifically in the context of dialogue based interaction and supporting urban spatial models.