

Accepted refereed manuscript of:

Bartie P, Mackaness W, Petrenz P & Dickinson A (2015)
Identifying related landmark tags in urban scenes using spatial
and semantic clustering, *Computers, Environment and Urban
Systems*, 52, pp. 48-57.

DOI: [10.1016/j.compenvurbsys.2015.03.003](https://doi.org/10.1016/j.compenvurbsys.2015.03.003)

© 2015, Elsevier. Licensed under the Creative Commons Attribution-
NonCommercial-NoDerivatives 4.0 International
<http://creativecommons.org/licenses/by-nc-nd/4.0/>

Identifying Related Landmark Tags in Urban Scenes using Spatial and Semantic Clustering

Abstract

There is considerable interest in developing landmark saliency models as a basis for describing urban landscapes, and in constructing wayfinding instructions, for text and spoken dialogue based systems. The challenge lies in knowing the truthfulness of such models; is what the model considers salient the same as what is perceived by the user? This paper presents a web based experiment in which users were asked to tag and label the most salient features from urban images for the purposes of navigation and exploration. In order to rank landmark popularity in each scene it was necessary to determine which tags related to the same object (e.g. tags relating to a particular café). Existing clustering techniques did not perform well for this task, and it was therefore necessary to develop a new spatial-semantic clustering method which considered the proximity of nearby tags and the similarity of their label content. The annotation similarity was initially calculated using trigrams in conjunction with a synonym list, generating a set of networks formed from the links between related tags. These networks were used to build related word lists encapsulating conceptual connections (e.g. church tower related to clock) so that during a secondary pass of the data related network segments could be merged. This approach gives interesting insight into the partonomic relationships between the constituent parts of landmarks and the range and frequency of terms used to describe them. The knowledge gained from this will be used to help calibrate a landmark saliency model, and to gain a deeper understanding of the terms typically associated with different types of landmarks.

Keywords:

Urban landmarks, scene tagging, trigram, tag clustering, mereology, graph clustering

1. Introduction

Human Computer Interaction (HCI) continues to evolve, creating more natural interfaces that increase productivity for a wider audience across a range of use environments. In particular mobile devices, used while moving, are receiving a lot of attention in the post-desktop era (Daley, 2012). As a result of this shift, and with the increase in processing power and improved statistical language models, speech recognition has grown in popularity for interacting with mobile devices. Smartphone applications such as Siri (Apple) and Cortana (Microsoft) allow the user to book diary events, look up information, or ask for directions, using only speech input.

While automatic speech recognition has improved the interaction is not entirely natural as the application is unaware of the user's surroundings and unable to refer to things as people typically do in conversation, for example to comprehend a question such as "What's that statue over there?", or to direct the user to "the café next to the bridge". To include such environmental references these devices need to model their surroundings and refer to features in common ways, so that the interface can become so natural and intuitive it is not even noticed (Weiser, Gold, & Brown, 1999).

It has been recognised for some time that further progress in mobile HCI will include expanding the machine's abilities to refer to objects in the user's surroundings, and to consider the context in which the device is being used (Bartie & Mackaness, 2006; Chen & Kotz, 2000; Long, Aust, Abowd, & Atkeson, 1996; Noh, Lee, Oh, Hwang, & Cho, 2012; Zipf, 2002). A key aspect of this link between virtual and real worlds is the use of common

50 anchor points, or landmarks, which can be recognised and referred to by both the user and the
51 machine. For example including a reference to a salient object when giving a navigation
52 instruction. There are a number of challenges in doing this, which include having access to a
53 complete dataset of objects with corresponding attribute and positional information, a method
54 to identify landmark candidates from the dataset, and the ability to select the most suitable
55 candidate for a particular task (e.g. the most suitable landmark for a turn instruction) (Kai-
56 Florian Richter & Winter, 2014).

57 As part of a wider research project looking at supplementing location based services with
58 knowledge of the user's environment, and thereby offering a greater interaction between
59 machine and place, a web based experiment was undertaken to collect data on what users
60 considered to be landmarks in urban scenes and to understand better how they describe those
61 objects. Participants were asked to identify features by viewing urban images and tagging
62 those items they considered useful in forming navigation instructions, adding a text
63 annotation to each feature that they tagged. In some cases users supplied tags for single
64 object features (e.g. a statue), while in others a label was used to represent a collection of
65 features, such as a castle with its many outbuildings and walls.

66 In order to determine the most salient objects in each scene the user generated tags first need
67 to be grouped according to the object they referred to, so that the number of unique users
68 could be calculated per landmark. The assumption was made that the more salient features
69 would be tagged by a larger number of participants who considered it a suitable landmark for
70 wayfinding. Such analysis would give a feature ranking, thus establishing the most dominant
71 landmarks in each scene, and provide a better understanding of the importance hierarchy of
72 features and sub-feature parts (e.g. the clock and the clock tower it is on). By establishing a
73 landmark ranking in each scene the various input metrics for the saliency model could be
74 adjusted so that a model's output more closely matched human landmark identification
75 choices.

76 While spatial clustering methods can be used to highlight tag concentrations across the
77 image, it did not offer adequate functionality to identify discrete objects, as tags in close
78 proximity may relate to different real world objects which appear close merely because of the
79 perspective view in the image. Therefore it was necessary to develop a clustering algorithm
80 able to group tags based on both the spatial location of the tag as well as the supplied text
81 label. The process was complicated by the range of descriptive terminology supplied in the
82 labels. For example the same landmark may be described as a *church* by one participant, and
83 as a *clock tower* by another referring to a subpart of the same structure. The algorithm
84 developed used a statistical sentence matching technique to link tags with related nearby
85 annotations, forming tag networks where nearby tags with similar content were considered to
86 have a strong relationship.

87 The paper begins by explaining the background and motivation for this research, followed by
88 a description of the web experiment conducted to collect data in Section 3, and then the
89 issues encountered with generating landmark rankings based on spatial clustering and the
90 need to develop a spatial-semantic clustering function, which is outlined in Section 5. The
91 paper concludes with suggestions for deriving other outputs from the tag data using this
92 clustering technique, and highlights some of the remaining issues which require future
93 research.

94

95

96 **2. Background and Motivation**

97 Landmarks are one aspect of the environment frequently referenced, as they assist in forming
98 mental representations of space (Hirtle & Heidorn, 1993; Tversky, 1993), and in wayfinding
99 tasks (Caduff & Timpf, 2008; Duckham, Winter, & Robinson, 2010; Lovelace, Hegarty, &
100 Montello, 1999; Werner, Krieg-Bruckner, Mallot, Schweizer, & Freksa, 1997; Winter,
101 Tomko, Elias, & Sester, 2008). Studies show that when exploring a new urban region people
102 build a mental model of the space by firstly recognising landmarks, then over time these are
103 joined together into sequences to form routes, which depending on the complexities of the
104 space may lead to a more comprehensive model of the space known as survey knowledge
105 (Hirtle & Jonides, 1985).

106 Landmarks are defined as identifiable features in an environment, whose saliency may be
107 calculated by comparing scores for particular attributes (e.g. their size) and identifying those
108 which deviate from the mean (Elias, 2003; Elias & Brenner, 2004; Raubal & Winter, 2002).
109 These are the objects unlikely to be confused with others, as they appear different to their
110 surroundings (e.g. churches, statues) or are well known international brands (e.g. Starbucks,
111 McDonalds). Landmarks are particularly useful when travelling to a new destination as they
112 can be used at decision points to help orient the navigator, along routes to confirm the
113 location, and as distant landmarks (Lovelace et al., 1999). While it is common for people to
114 include landmarks in conversation, current smartphone digital assistant applications, such as
115 Apple's Siri, Microsoft's Cortana, and Samsung's S-Voice, are unaware of the user's
116 environment and therefore unable to refer to surrounding objects. As speech based interfaces
117 continue to develop it will be useful to include better context awareness which can establish
118 the user's environment and include references to visible landmarks around the user. Google's
119 Project Tango (Lee, 2014) shares a similar ambition to enrich the user experience by
120 allowing software to consider the world beyond the phone's hardware and to consider time
121 and space at a more human scale. Such an ability would allow for the generation of more
122 natural human computer interactions, helping to reduce the seam which exists between users
123 and technology (Ishii, Kobayashi, & Arita, 1994). Such situational awareness will extend the
124 range and capabilities of mobile applications, as has already been demonstrated in prototype
125 applications (Bartie & Mackaness, 2006; Mackaness et al., 2014).

126 There are two parts to the process of using landmarks in forming navigation instructions, or
127 in generating referring expressions to describe the location of city objects. These are the
128 identification of suitable candidate landmarks from all known objects, and then determining
129 the most appropriate candidate for a given task (e.g identifying the landmark which best
130 supports a turning instruction in a wayfinding task) (Kai-Florian Richter & Winter, 2014).

131 The task should determine which landmarks are selected according to the route taken rather
132 than using pre-set items from a static list of landmarks in the region (Nothegger, Winter, &
133 Raubal, 2004). Similarly when using landmarks to describe a scene or direct the user's gaze,
134 a selection process is required to determine the most suitable candidates from those in the
135 current view. The ambition is to provide no more information than is necessary, according to
136 Grice's maxim of quantity (Grice, 1975), and therefore the selection process should ensure a
137 minimal set of highly relevant landmarks are returned. This goes beyond measuring the path
138 of photons from the observer to the target feature, as it is not only a question of which are
139 physically visible, but also which are noticed by an observer at that location. For example
140 when asked to identify statues in Figure 1a people will often notice the statue of a black
141 horse in the foreground but many fail to spot the more distant statue, as highlighted in Figure

142 1b. This may be partly because it is further away, but perhaps it is also a factor of the
143 surrounding distractors in the scene of buildings and trees making it harder to separate
144 visually from background objects, or perhaps it is an artefact of the position of the statues in
145 the image frame.



146

Figure 1: Visible objects are not always noticeable, with many people failing to notice the second more distant statue in this scene

147 Existing techniques to detect landmarks from photo collections using cluster-based
148 techniques consider each image as a single object to be classified. For example Papadopoulos
149 *et al* (2010) treats each image as a node in a graph, and exploits computer vision and user
150 tag similarity metrics to find corresponding images of landmarks or events. Other approaches
151 for automated and semi-automated image clustering (Vonikakis, Jinda-Apiraksa, & Winkler,
152 2014; Wang, Ji, Tian, & Hua, 2012) or the linking to location (Ahern, Naaman, Nair, &
153 Yang, 2007) also consider the images as single objects. There are research efforts to assign
154 tags to features in images, the Tag-to-Region Assignment Problem (Liu, Hua, & Zhang,
155 2011), to correspond to the semantic region within an image but these are not yet robust. For
156 this research the images were considered as a means to portray real world features to a web
157 audience, with a goal of extracting information from the participants about the real world
158 objects portrayed in the images. Therefore the tag annotations were supplied at an object
159 feature level rather than a request to more generally describe the entire image.

160 There are arguments for imposing structured vocabularies to enable greater semantic parsing
161 of supplied annotations (Tousch, Herbin, & Audibert, 2012), however for this research
162 participants were permitted to enter any text without restrictions so that a wide range of
163 descriptors would be collected for analysis.

164

165 3. Web Experiment

166 A web based experiment was conducted in which human subjects were asked to identify
167 landmarks in a number of urban scenes. The experiment was publicised through social
168 media, attracting 185 participants. Users were assigned images randomly from a set of 37,
169 and able to leave the experiment at any time but encouraged to complete as many images as

170 possible by giving them an additional entry into a prize draw for each completed set. For
171 each task the participant saw an image of part of Edinburgh city (Scotland), and was asked to
172 identify what they considered to be landmark features by tagging them on the image. The
173 user's profile and knowledge of the city was recorded as part of this process.

174 All images were captured on the same day in the early morning over a period of ninety
175 minutes, in an effort to reduce object occlusion by other city occupants (e.g. buses,
176 pedestrians) and to minimise weather and lighting variation. The ambition was to replicate as
177 closely as possible the street experience, although it is recognised from previous landscape
178 studies that imagery can introduce a bias in the way that it is captured and displayed (Daniel
179 & Vining, 1983; Linton, 1968; Shafer & Brush, 1977; Zube, Sell, & Taylor, 1982). In an
180 effort to minimise these effects the images were captured using a wide angle lens, and as
181 computer monitors do not offer the same level of visual detail as when on the street, a
182 magnifying region was added to the web viewer, as shown in Figure 2. This allowed the
183 participant to see a magnified portion of the image as they moved the mouse crosshair around
184 the main image, giving a similar level of detail to that experienced on the street, and enabling
185 them to more easily identify and tag more distant and smaller objects.

186 Once the participant had clicked on the image at the location of something they considered
187 interesting, they were presented with an input box to enter free text which described the
188 object (Figure 2b), such as a church, pub, or no entry sign. Each participant was permitted up
189 to 12 tags per image, and asked to provide a short description for each tag. The tag limitation
190 was imposed to encourage participants to limit their tagging to the most salient objects, and
191 to then move on to the next image.

192



193
194

*Figure 2: Web based landmark tagging experiment
(a) overview and magnified region (b) adding annotation text*

195 The breakdown of contributions by age/sex is shown for user counts (Figure 3a) with a fairly
196 even balance of genders and wide range of ages. There were no statistically significant

197 difference noted between males and females in their choice of landmarks, or the number of
 198 landmarks identified per image. Figure 3b shows how well the participants thought they
 199 knew Edinburgh, with most candidates having at least some knowledge of the region. There
 200 was a trend for males with better knowledge of the city to increase annotation length
 201 marginally, and for those familiar with the city to name landmarks. The number of tags per
 202 image ranged from 178 tags to 451 tags, with an average of 90 unique participants per image.
 203 The average number of tags generated per participant was 56, with a total of 10,350 tags
 204 created across the 37 images.

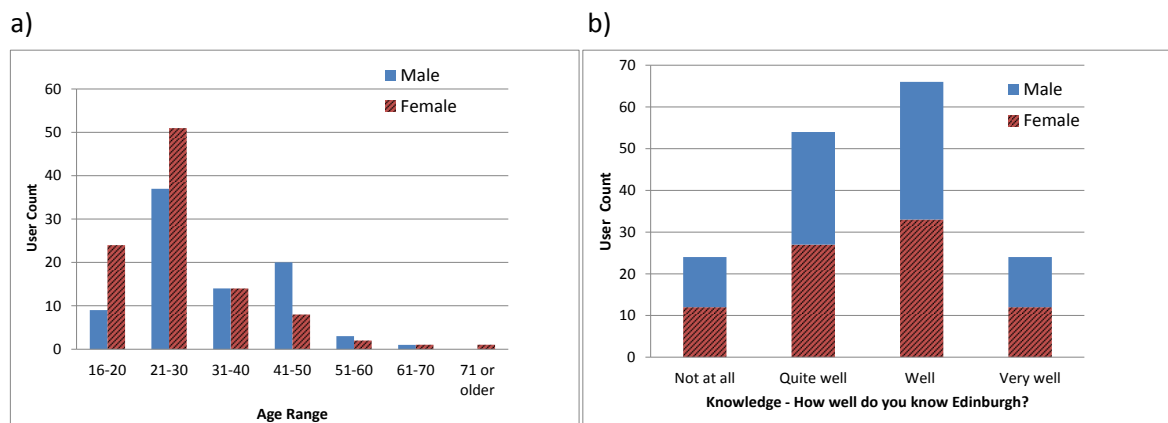


Figure 3: Breakdown of Data by Participant Group
 (a) count of users (b) knowledge of the city

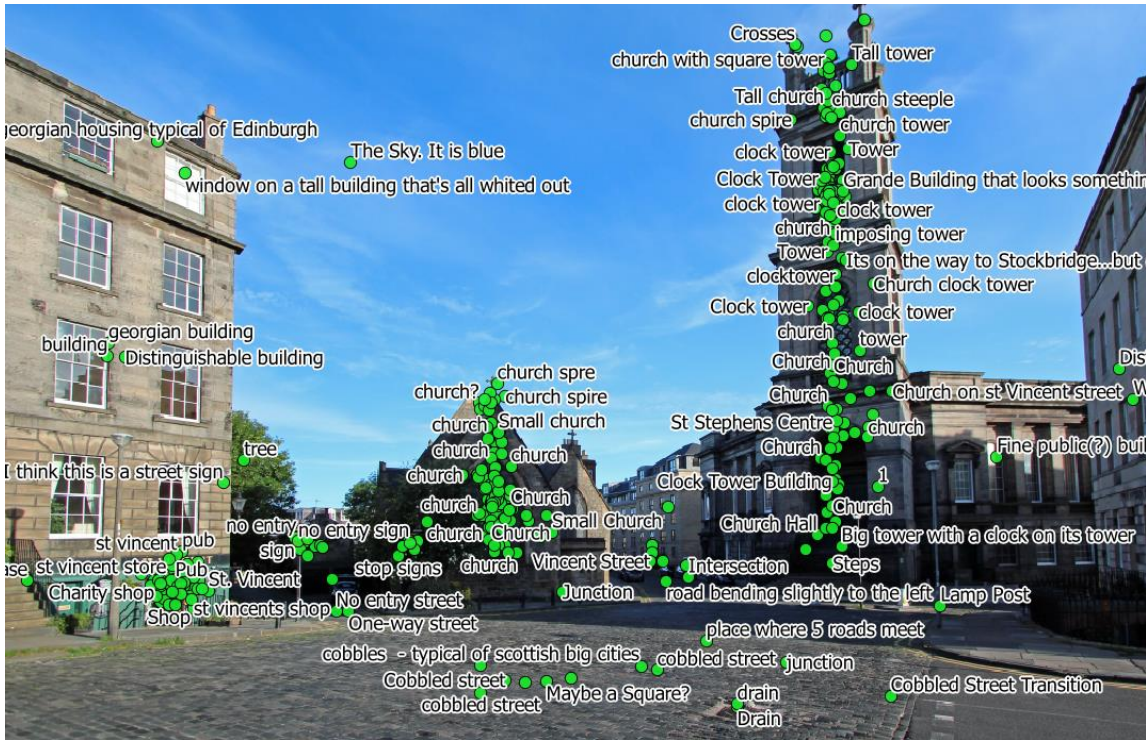
205 4. Web Data Collection Results

206 An example of the output is shown in Figure 4, which displays the tagged locations and the
 207 supplied tag text. The tags for this image are shown as a word cloud in Figure 5, whereby
 208 the more frequently occurring terms are represented in a larger font. In this example the
 209 *tower* and *clock* refer to the church on the right side of Figure 4, while *church* was used for
 210 both churches in the scene. Without considering the tag locations it is not possible to identify
 211 these two distinct groups from the term frequency (i.e. word cloud), nor therefore identify
 212 landmark dominance in the scene at a feature level.

213 Therefore analysis of the annotation text results alone was not suitable to identify landmark
 214 tag clusters because in each scene there may be multiple instances of a feature type (e.g. a
 215 small church and a large church with spire), and it would not be possible to rank individual
 216 object popularity based on term frequency.

217 The spatial pattern of the supplied tag locations may be summarised using spatial clustering,
 218 such as Kernel Density Estimation (KDE), (Silverman, 1986). The results for four example
 219 scenes are shown in Figure 6, where red shows a dense concentration of user tags. These
 220 dense spatial concentrations are clearly noticeable for the two churches and the building on
 221 the left of the scene (a public house) in Figure 6a. In particular there is a concentration of

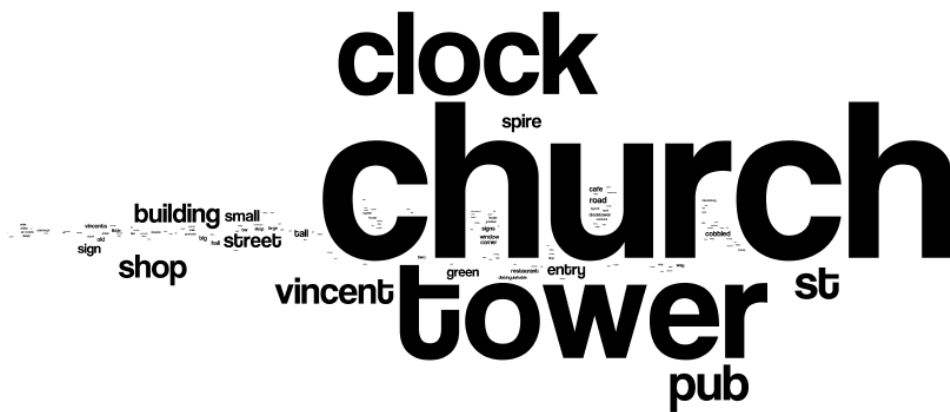
222 tags around the top of the taller church tower where there is a clock face. In Figure 6b the
 223 KDE has highlighted a single cluster (Group 1) where there are in fact two distinct features,
 224 which are at different viewing distances but a similar viewing angle from the observer.
 225 Figure 7 shows this in greater detail where, due to the KDE bandwidth setting, groups (i) and
 226 (ii) have merged resulting in a dense cluster that does not depict a single landmark but is an
 227 artefact of two objects 180 metres apart having a similar viewing angle from the observer.



228

Figure 4: Example Scene with Some (not all) Tags from All Users Displayed

229



230

Figure 5: Word Cloud for Tags from a Single Image

231 Group 4 in Figure 6d consists of tags for ‘Scott Monument’ and ‘Calton Hill’ which have
232 been clustered together despite being more than 670 metres apart, for the same reason of the
233 viewing angle. Similarly three objects at different viewing distances are clustered in Group 6.
234 In contrast Groups 2, 3 and 5, in Figure 6b, Figure 6c and Figure 6d respectively are
235 examples of single features yet each is presented as two distinct clusters. This is interesting
236 as it shows the participants considered the object to have multiple focal points of interest,
237 nevertheless it is necessary to aggregate these tag clusters in order to determine the
238 dominance of these features as single objects in the scene.

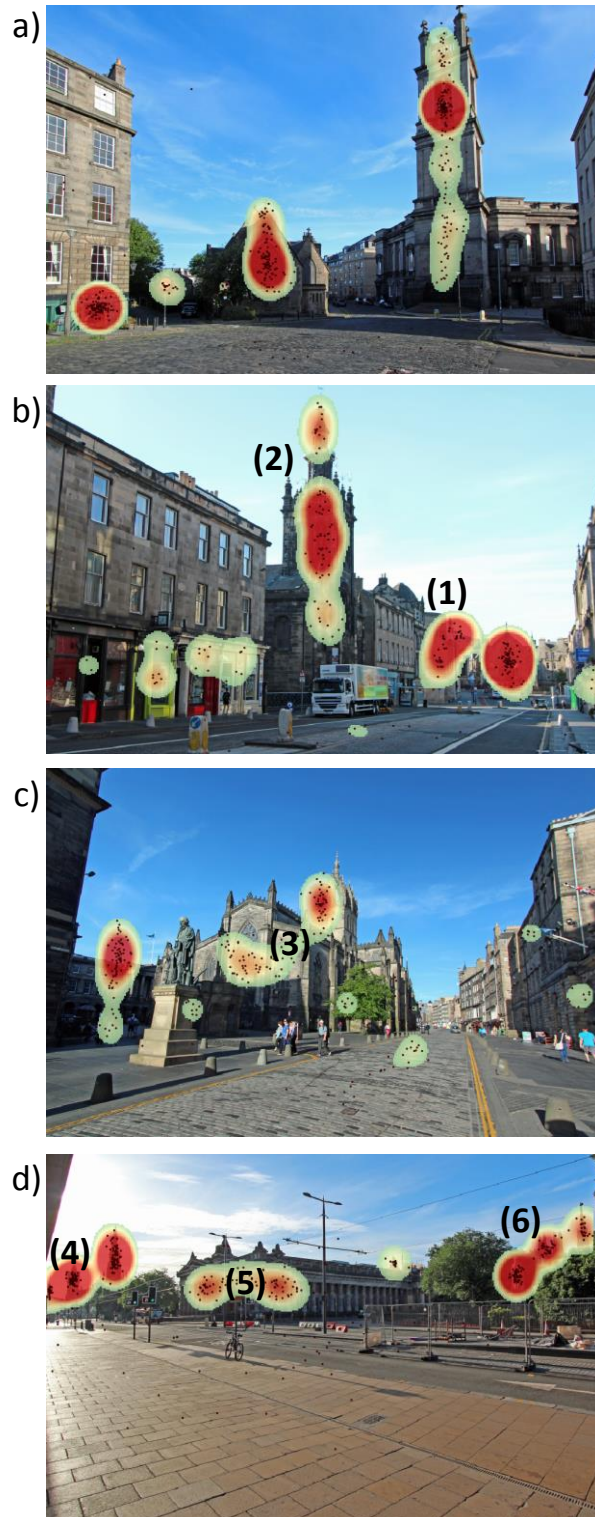


Figure 6: Kernel Density Estimation for User Tags (where red = dense clustering)

239

240

241 To improve upon this outcome a clustering technique was developed which included both
 242 spatial and semantic components, as described in Section 5. The performance of this
 243 approach is discussed in Section 6.



Figure 7: Spatial Clustering Errors due to not Modelling Distance

Group (i) and group (ii) tag locations should be considered separately as the objects are 180 metres apart

244 5. Spatial and Semantic Clustering

245 The participants supplied a text annotation for each tag, consisting of any number of words.
 246 This allowed for the creation of a more natural dataset of descriptive object terms to be
 247 collected, but added complexity in the analysis and term matching.

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

258 To improve this an enhanced matching function was developed which included access to a
 259 synonym table allowing for conceptually similar terms, such as 'street' and 'road',
 260 'cathedral' and 'church', and 'memorial' and 'statue' to be treated as identical. The results
 261 are shown in Table 1, where 'church' and 'cathedral' score an exact match of 1.0, and
 262 'church tower' and 'cathedral spire' also score an exact match. The synonym table was hand
 263 constructed by looking at the most commonly occurring words from all images. It would also
 264 be possible to populate such a table using an existing database of synonyms such as WordNet
 265 (Princeton University, 2010). However this enhanced matching function lacked the ability to
 266 model partonomic relationships (i.e. relationships between an object's parts) therefore the
 267 score for 'church' and 'clock tower' remained low (0.0556).

268 To improve on this phrase pairs were collected by processing each tag in turn, searching the
 269 corresponding image for nearby tags within a defined pixel distance, equivalent to the KDE
 270 bandwidth. The content similarity score for each tag pair was calculated and those with a
 271 score greater than 0.3, the default value for the trigram module used (Bartunov, Sigaev, &
 272 Kings-Lynne, 2014), were considered to be related. All of the tags were processed resulting
 273 in a topological network of connected content for each image.

274
 275
 276

Table 1: Word Similarity using Trigrams

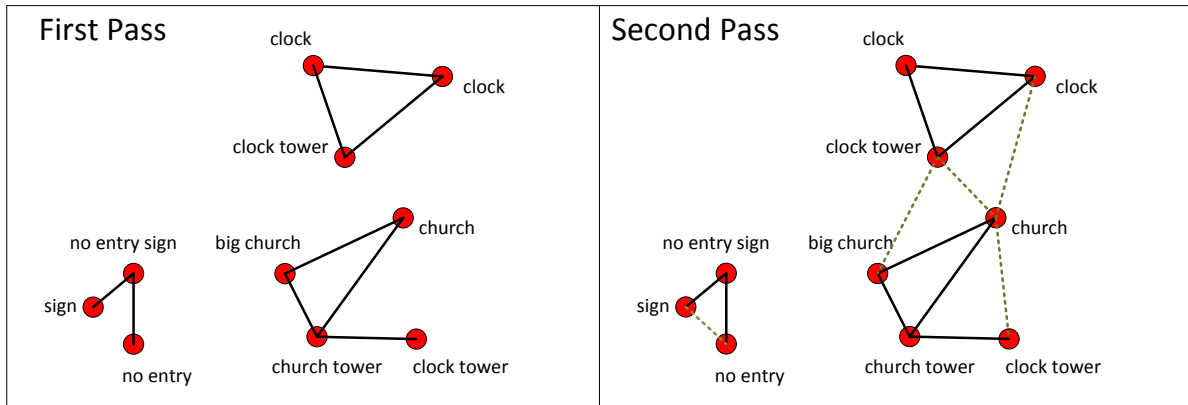
Phrase One	Phrase Two	Trigram Matching (0 to 1)	Enhanced Matching (0 to 1)
church	Churches	0.6000	0.6000
monument	momument * ¹	0.5000	0.5000
church	Cathedral	0.0625	1.0000
church tower	Clock	0.0556	0.0556
church tower	Cathedral	0.0455	0.5385
church tower	cathedral spire	0.0357	1.0000
church	kirk	0.0000	1.0000
church of St Giles	St Giles Kirk	0.3750	1.0000

*¹ intentional spelling error based on user supplied tag

277 **5.1. Expanding the Network of Linked Tags using a Secondary Pass**

278 In some cases running the process a single time resulted in small groups of tags being left as
 279 orphan clusters. For example in Figure 8 on the ‘First Pass’ three cluster groups were formed,
 280 relating to two objects; a *no entry sign* and a *church with a clock tower*. The two groups on
 281 the right remain distinct as no synonym entry links the *church* tags with *clock* or *clock tower*,
 282 and the other *clock tower* tag was outside the search radius. This can be addressed by
 283 increasing the search distance but that could result in separate object instances being
 284 combined (e.g. two nearby churches in Figure 4 would be grouped as a single entity). Instead
 285 the data was processed a second time using the same buffer distance but the vocabulary of
 286 related terms was increased by using the word lists generated from the first pass. By doing
 287 this the conceptual links list is automatically expanded for tag groups nearby allowing for
 288 greater conceptual links, but reducing the likelihood of separate objects being merged due to
 289 the limited spatial search parameters. This is a form of query expansion (Chum, Philbin,
 290 Sivic, Isard, & Zisserman, 2007; Xu & Croft, 1996), limited by the spatial location of the
 291 supplied tags. For example a *church* node may be joined to a *clock tower* node, even though
 292 they do not share any similar terminology based on a *church tower* node elsewhere being
 293 linked to a *clock tower* through the common term *tower*. Figure 8 shows an example of this
 294 process, where initially links are made between tags forming 3 networks based on common
 295 terminology. These network phrases are used during a second pass of the data, whereby a

296 greater number of linkages may be added between groups as a result of the expanded
 297 semantic connections learned from the initial pass. The result is an expansion of the network
 298 topology through the linking of network groups, a reduction in the number of object groups,
 299 and an increase in linkages made between object parts thereby improving the partonmic
 300 modelling capability.
 301



302
303 *Figure 8: Expanding the Linked Network with Secondary Pass*

304
305

306 6. Tag Clustering Results

307 An example of the output from this process is shown in Figure 9, where colours are assigned
 308 randomly based on Cluster Group ID. There are many improvements compared to the spatial
 309 only clustering (Figure 6), as now two groups are identifiable in Figure 9b (group 1) where
 310 before there was a single cluster, and a single group identified as group 2. Figure 9d (group
 311 6) also now shows three distinct object definitions, rather than a single cluster.

312 Figure 9b (group2), Figure 9c (group 3), and Figure 9d (group 5) are now shown as single
 313 objects rather than before where the variety of focal points selected to tag the object by the
 314 participants had resulted in multiple cluster centres on these objects. The previously single
 315 group at d (group 4) is now separated into two groups, however there is also an overlap
 316 occurring (orange group connects to red group) which is due to a common use of
 317 terminology ('tower').

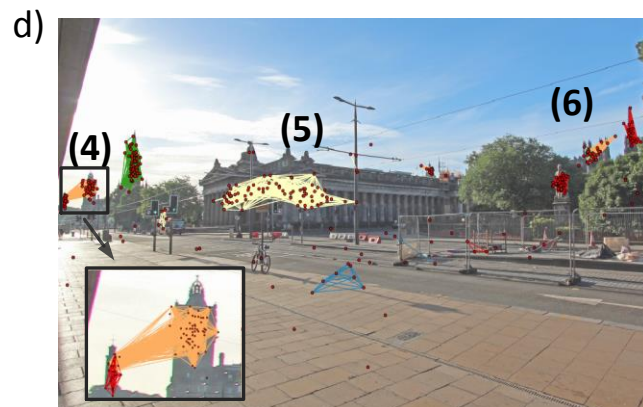
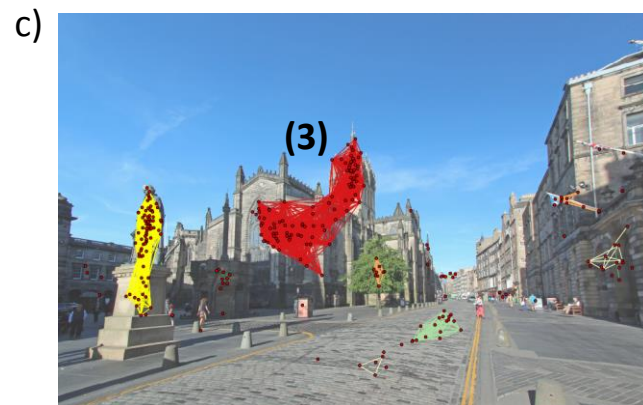
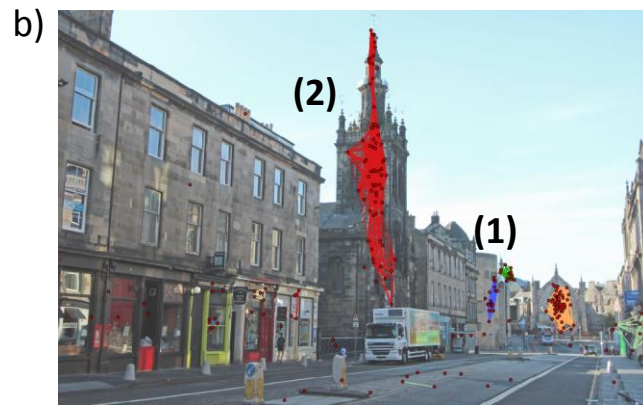
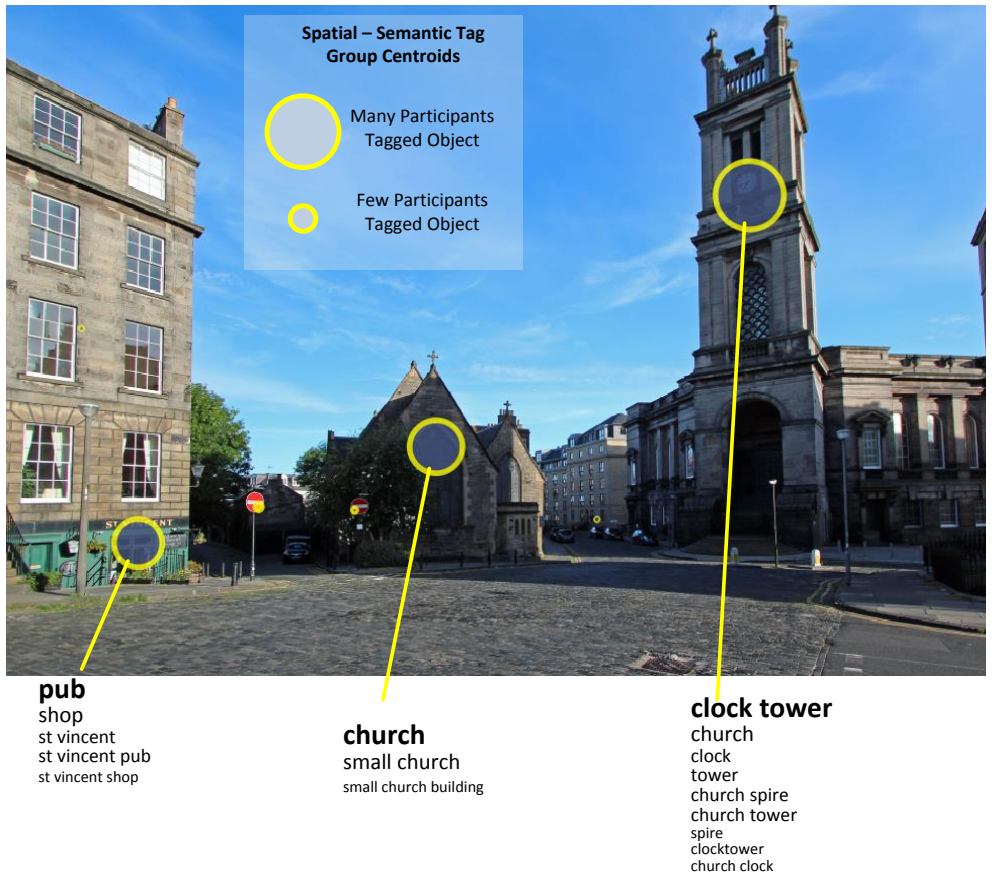


Figure 9: Spatial-Semantic Clusters (random cluster colouring)

319 Tag groups were identified for each image making it possible to automatically generate
 320 related word lists. For example *church* is linked to *church spire*, which is linked to *church*
 321 *tower*, which is linked to *clock tower*, which is linked to *clock*. Tags define objects spatially
 322 and conceptually, and the frequency of each tagged phrase gives an indication of the most
 323 common term used for that object and its parts. Once tags have been linked in this way it is
 324 possible to calculate tag group centroids relating to the concept centres, for example the *clock*
 325 on the *clock tower*, which is part of the *church*. It is also possible to generate a list of the
 326 most frequent terms used per object, rather than per image, as shown in Figure 10.

327



328

Figure 10: Phrase Ranking per Identified Object

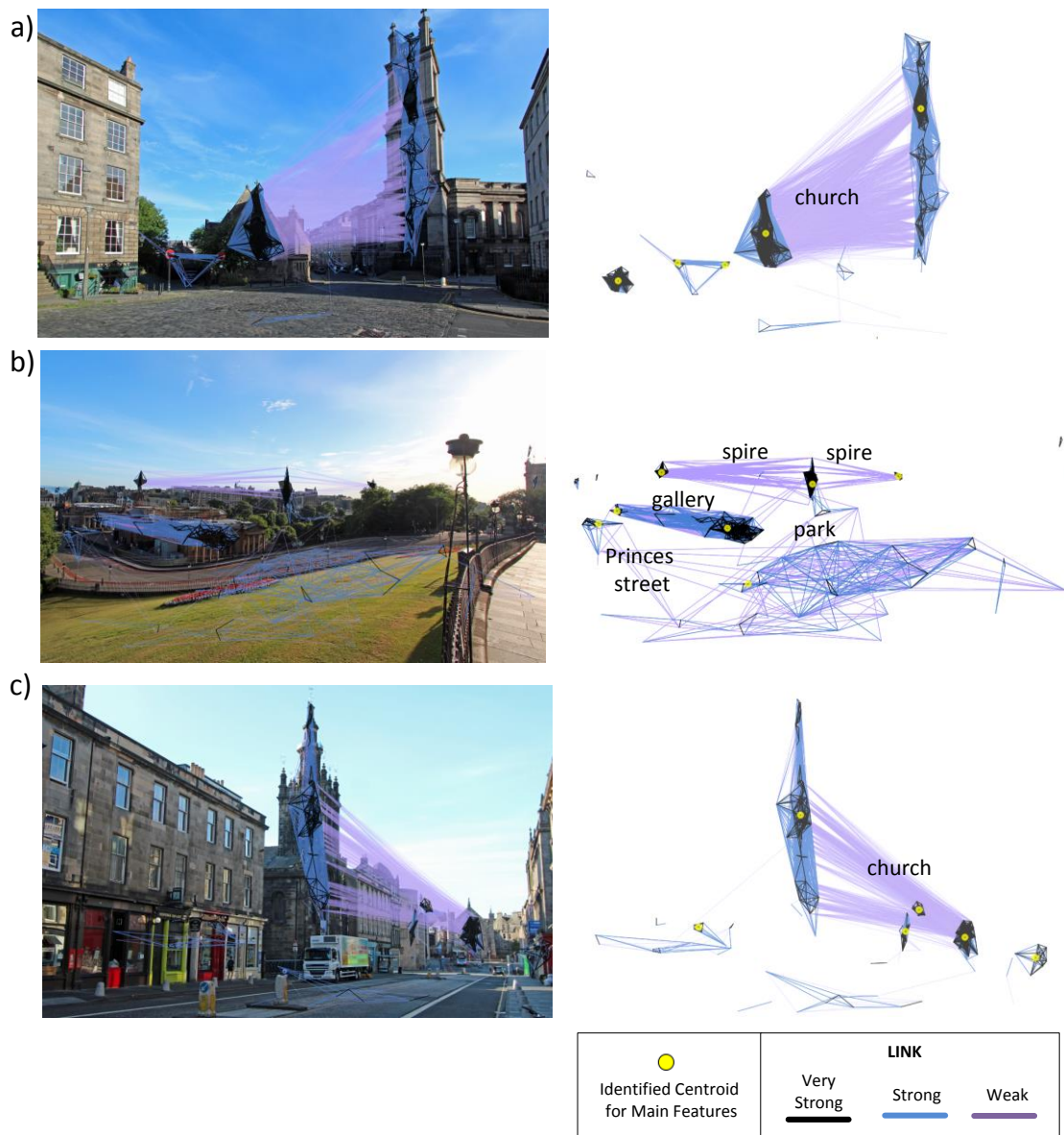
329

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

333

334 The strength of links between all tags in an image can be calculated as a function of phrase
 335 similarity and the inverse tag distance, such that similar phrases near each other receive a
 336 high value. Figure 11 visualises this for the example scenes, where very strong relationships
 337 are shown in black, strong links in blue, and weak links in purple. A low threshold was
 338 specified to remove links between very unrelated tags. This visualises the conceptual
 339 connections between features.

340 The object concept centres, as introduced in Figure 10, are displayed as yellow dots to
 341 indicate the main identified features in each example (Figure 11, right column). The weak
 342 links, shown in purple, can be considered as mapping objects which might be confused from
 343 having similar annotations. For example the link between the two churches in (a) and (c)
 344 shows their conceptual similarity and highlights the risk of a misunderstanding occurring if
 345 attempting to identify the object from the annotation alone. Also in (b) the “spire”, “gallery”,
 346 and “park” labels indicate objects which could be confused unless further details are included
 347 in a referring expression.



348

Figure 11: The Strength of the Linkages between Tags

349 The “Princes Street” link shown in Figure 11b has arisen from a mistake made by a number
 350 of participants who believed the foreground grassy region to be “Princes Street Gardens”,
 351 and therefore the clustering method has calculated a weak link from these tags to “Princes
 352 Street” in the distance. While it is not possible to automatically discount these incorrect tags,
 353 it is possible to validate the strength of their linkage through the annotation matching, and
 354 distance calculations. As this example shows the link strength is considered to be weak, and
 355 effectively removes these tags from the calculation of object concept centres.

356

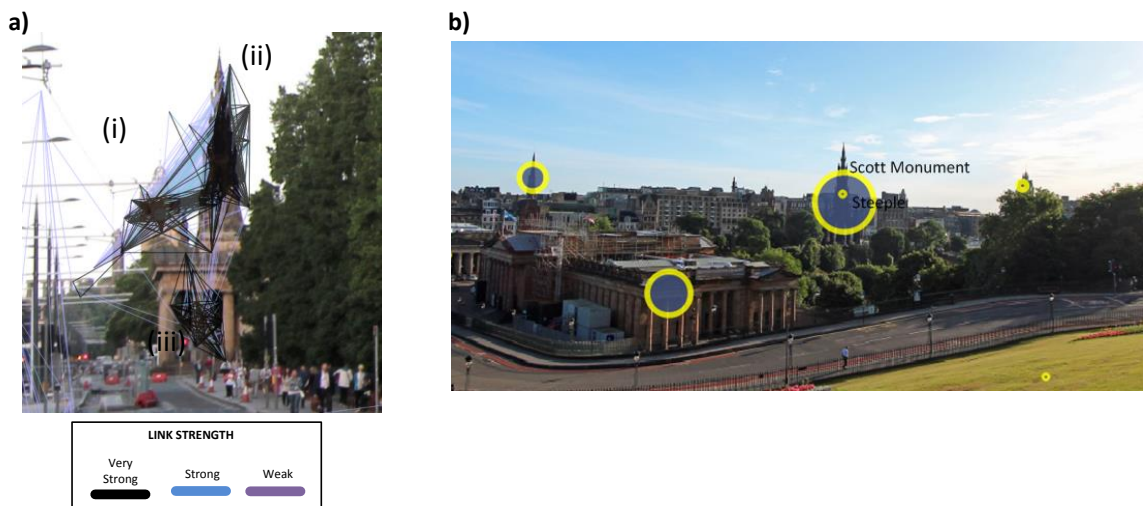
357 7. Outstanding Issues and Future Work

358 Overall the clustering technique performed well in reducing the supplied tags to related
 359 groups representing single objects in the scene, but there were cases that raised some
 360 outstanding issues. These fell into two categories: cases where two nearby objects were
 361 described in a similar way but were in fact unrelated, and where an object was described in
 362 two very different ways resulting in their being no semantic overlap.

363 Figure 12(a) shows an example of the first category, where three object clusters are
 364 identifiable but an incorrect link has been made between objects (i) and (ii). This occurred
 365 because some participants described the first object (i) as a tower, while others described
 366 object (ii) as a steeple. These words were linked together via an entry in the synonym table,
 367 and could be highlighted for checking, but to automate this disambiguation will require
 368 further work.

369 Another issue was that it was difficult to associate the proper noun with a landmark
 370 description. For example in Figure 12(b), some familiar with the city labelled “Scott
 371 Monument” while others labelled it as “Steeple”. These two concepts do not have a semantic
 372 link and therefore the output shows two overlapping groups, where there should be a single
 373 entity. These can be automatically highlighted by using spatial containment functions to
 374 produce a list of such co-occurrences to be examined in more detail and resolved by adding
 375 an entry into the synonym table to link the groups.

376



377

Figure 12: Examples of Outstanding Issues
 (a) Linkage Strengths between Tags (b) Cluster Group Centroids

378

379

380 **8. Conclusions**

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

388 The method collects nearby tags which show a correlation using trigram fuzzy matching.
389 Synonyms and stop words were used to improve the matching, and a network graph of
390 connected tags was generated for each image. This was expanded in a secondary pass by
391 using the linkages discovered on the first pass to join up orphaned tag groups. The results
392 show that it was possible to automatically identify objects of interest from the user supplied
393 tags, and that term frequencies could be discovered at an object level. The network graph
394 visualises the relationships which exist between tags, enabling the strength of the relations to
395 be inferred from the density and centrality of the graph edges.

396 This research has relevance in the context of intuitive dialogue driven systems in which rich
397 descriptions of landmarks are required to support the generation of way finding instructions
398 (Kai-Florian Richter, Tomko, & Winter, 2008) since the graph is able to provide both a
399 generic description ('the church') for the observer in the far distance, and a detailed
400 description ('the church tower with the clock') when the observer is closer. The next phase of
401 this research is to compare the results of this user experiment against a model of landmark
402 saliency, whereby the relative dominance of landmarks selected from this study will be
403 compared at an object level with the saliency model output. Where differences are noted the
404 input weightings of saliency model parameters (e.g. visible area, on the skyline, viewing
405 distance, object type) will be adjusted to more closely match these findings from this
406 experiment. Term frequencies and variations by object type and viewing distance will be
407 conducted, giving a greater understanding of how people refer to features of interest in urban
408 scenes which could then be incorporated into the natural language generation component.

409

410

411 **9. Acknowledgements**

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

414

415

416 **10. References**

417 Ahern, S., Naaman, M., Nair, R., & Yang, J. (2007). World explorer: Visualizing aggregate
418 data from unstructured text in geo-referenced collections *In JCDL* (pp. 1--10).

419 Bartie, P., & Mackaness, W. A. (2006). Development of a speech-based augmented reality

- 420 system to support exploration of cityscape. *Transactions in GIS*, 10(1), 63-86. doi:
421 10.1111/j.1467-9671.2006.00244.x
- 422 Bartunov, O., Sigaev, T., & Kings-Lynne, C. (2014). Additional Supplied Modules
423 (pg_trgm). (12 May 2014). <http://www.postgresql.org/docs/9.1/static/pgtrgm.html>
- 424 Caduff, D., & Timpf, S. (2008). On the assessment of landmark salience for human
425 navigation. *Cognitive Processing*, 9(4), 249-267.
- 426 Chen, G., & Kotz, D. (2000). A survey of context-aware mobile computing research:
427 Technical Report TR2000-381, Dept. of Computer Science, Dartmouth College.
- 428 Chum, O., Philbin, J., Sivic, J., Isard, M., & Zisserman, A. (2007). *Total recall: Automatic*
429 *query expansion with a generative feature model for object retrieval*. Paper presented
430 at the Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on.
- 431 Daley, T. (2012). We're now officially in the Post-PC era. *Waste + Water Management*
432 *Australia*, Vol. 38(No 6), 27-30.
- 433 Daniel, T. C., & Vining, J. (1983). Methodological Issues in the Assessment of Landscape
434 Quality. In I. Altman & J. Wohwill (Eds.), *In Behaviour and the Natural Environment*
435 (pp. 39-83). New York: Plenum Press.
- 436 Duckham, M., Winter, S., & Robinson, M. (2010). Including landmarks in routing
437 instructions. *Journal of Location-Based Services*, 4 (1), 28-52.
- 438 Elias, B. (2003). Extracting landmarks with data mining methods. In W. Kuhn, M. F.
439 Worboys, & S. Timpf (Eds.), *Spatial information theory* (Vol. 2825, pp. 398-412).
440 Berlin: Springer.
- 441 Elias, B., & Brenner, C. (2004). Automatic generation and application of landmarks in
442 navigation data sets. In P. F. Fisher (Ed.), *Developments in Spatial Data Handling*
443 (pp. 469-480). Berlin: Springer.
- 444 Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and*
445 *Semantics 3: Speech Acts* (pp. 41-58): Academic Press.
- 446 Hirtle, S. C., & Heidorn, P. B. (1993). The structure of cognitive maps: Representations and
447 processes. *Behavior and Environment: Psychological and Geographical Approaches*,
448 170-192.
- 449 Hirtle, S. C., & Jonides, J. (1985). Evidence of hierarchies in cognitive maps. *Memory and*
450 *Cognition*, 13(3), 208-217.
- 451 Ishii, H., Kobayashi, M., & Arita, K. (1994). Iterative design of seamless collaboration
452 media. *Communications of the ACM*, 37(8), 83-97.
- 453 Lee, J. (2014). Project Tango. Retrieved 6 June 2014, from
454 <https://www.google.com/atap/projecttango>
- 455 Lin, D. (1998, July 24-27). *An information-theoretic definition of similarity*. Paper presented
456 at the Proceedings of the Fifteenth International Conference on Machine Learning,
457 Madison, Wisconsin, USA.
- 458 Linton, D. L. (1968). The assessment of scenery as a Natural Resource. *Scottish*
459 *Geographical Magazine*, 84, 219 - 238.
- 460 Liu, D., Hua, X.-S., & Zhang, H.-J. (2011). Content-based tag processing for internet social
461 images. *Multimedia Tools and Applications*, 51(2), 723-738.

- 462 Long, S., Aust, D., Abowd, G., & Atkeson, C. (1996). *Cyberguide: prototyping context-*
463 *aware mobile applications*. Paper presented at the Conference on Human Factors in
464 Computing Systems, Vancouver.
- 465 Lovelace, K. L., Hegarty, M., & Montello, D. R. (1999). Elements of good route directions in
466 familiar and unfamiliar environments. In C. Freksa & D. Mark (Eds.), *Spatial*
467 *Information Theory: Cognitive and Computational Foundations of Geographic*
468 *Information Science* (Vol. 1661, pp. 751): Springer Berlin / Heidelberg.
- 469 Mackaness, W. A., Bartie, P., Dalmas, T., Janarthanam, S., Lemon, O., Liu, X. (2014, 7-13
470 April). *Talk the Walk and Walk the talk: Design, Implementation and Evaluation of a*
471 *Spoken Dialogue System for Route Following and City Learning*. Paper presented at
472 the Annual Conference of the Association of American Geographers, Tampa Florida.
- 473 Noh, H.-Y., Lee, J.-H., Oh, S.-W., Hwang, K.-S., & Cho, S.-B. (2012). Exploiting indoor
474 location and mobile information for context-awareness service. *Information*
475 *Processing & Management*, 48(1), 1-12.
- 476 Nothegger, C., Winter, S., & Raubal, M. (2004). Computation of the Saliency of Features.
477 *Spatial Cognition and Computation*, 4(2), 113–136. doi:
478 10.1207/s15427633scc0402_1
- 479 Papadopoulos, S., Zigkolis, C., Kompatsiaris, Y., & Vakali, A. (2010). Cluster-based
480 landmark and event detection on tagged photo collections. *IEEE Multimedia*.
- 481 Princeton University. (2010). WordNet. (5 May 2014). wordnet.princeton.edu
- 482 Raubal, M., & Winter, S. (2002). Enriching wayfinding instructions with local landmarks In
483 M. J. Egenhofer & D. M. Mark (Eds.), *Second International Conference GIScience*
484 (Vol. 2478, pp. 243-259). Boulder, USA: Springer.
- 485 Richter, K.-F., Tomko, M., & Winter, S. (2008). A dialog-driven process of generating route
486 directions. *Computers, Environment and Urban Systems*, 32(3), 233-245.
- 487 Richter, K.-F., & Winter, S. (2014). *Landmarks*: Springer International Publishing.
- 488 Shafer, E. L., & Brush, R. O. (1977). How to measure preferences for photographs of natural
489 landscapes. *Landscape Planning*, 4, 237-256.
- 490 Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (Vol. 26): CRC
491 press.
- 492 Tusch, A.-M., Herbin, S., & Audibert, J.-Y. (2012). Semantic hierarchies for image
493 annotation: A survey. *Pattern Recognition*, 45(1), 333-345.
- 494 Tversky, B. (1993). *Cognitive maps, cognitive collages, and spatial mental models*. Paper
495 presented at the Spatial Information Theory: A Theoretical Basis for GIS, Italy.
- 496 Vonikakis, V., Jinda-Apiraksa, A., & Winkler, S. (2014). *PhotoCluster: A Multi-clustering*
497 *Technique For Near-duplicate Detection In Personal Photo Collections*. Paper
498 presented at the VISAPP, Lisbon, Portugal.
- 499 Wang, M., Ji, D., Tian, Q., & Hua, X.-S. (2012). Intelligent photo clustering with user
500 interaction and distance metric learning. *Pattern Recognition Letters*, 33(4), 462-470.
- 501 Weiser, M., Gold, R., & Brown, J. S. (1999). Origins of ubiquitous computing research at
502 PARC in the late 1980s. *IBM Systems Journal*, 38(4), 693-695.
- 503 Werner, S., Krieg-Bruckner, B., Mallot, H. A., Schweizer, K., & Freksa, C. (1997). Spatial

504 cognition: The role of landmark, route, and survey knowledge in human and robot
505 navigation. In M. Jarke (Ed.), *Informatik '97 GI Jahrestagung* (pp. 41–50). Berlin,
506 Heidelberg, New York. Springer.

507 Winter, S., Tomko, M., Elias, B., & Sester, M. (2008). Landmark hierarchies in context.
508 *Environment and Planning B: Planning and Design*, 35(3), 381 – 398.

509 Xu, J., & Croft, W. B. (1996). *Query expansion using local and global document analysis*.
510 Paper presented at the Proceedings of the 19th annual international ACM SIGIR
511 conference on Research and development in information retrieval.

512 Zamora, E. M., Pollock, J. J., & Zamora, A. (1981). The use of trigram analysis for spelling
513 error detection. *Information Processing & Management*, 17(6), 305-316.

514 Zipf, A. (2002). Adaptive context-aware mobility support for tourists. *Trends &*
515 *Controversies: Intelligent Systems for Tourism. IEEE Intelligent Systems*, 17(6), 57-
516 59.

517 Zube, E. H., Sell, J. L., & Taylor, J. G. (1982). Landscape perception: research, application
518 and theory. *Landscape Planning*, 9(1), 1-33.

519

520