Computers, Environment and Urban Systems 52 (2015) 48–57

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/compenvurbsys

Identifying related landmark tags in urban scenes using spatial and semantic clustering

Phil Bartie [a,*], William Mackaness [b], Philipp Petrenz [c], Anna Dickinson [c]

[a] School of Natural Sciences, University of Stirling, Stirling FK9 4LA, Scotland, United Kingdom
[b] School of GeoSciences, University of Edinburgh, Drummond St, Edinburgh EH8 9XP, Scotland, United Kingdom
[c] Informatics Forum, University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, Scotland, United Kingdom

ARTICLE INFO

Article history:
Received 4 August 2014
Received in revised form 13 March 2015
Accepted 14 March 2015

Keywords:
Urban landmarks
Scene tagging
Trigram
Tag clustering
Mereology
Feature graphs

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? The method developed in this research identifies related annotated tags supplied from a web based experiment in which users were asked to tag the most salient features on urban images for the purposes of navigation and exploration. The tag collections may be used to rank landmark popularity in each scene, but the challenge is in determining which tags relate 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.

1. Introduction

Human Computer Interaction (HCI) continues to evolve, creating more natural interfaces that increase productivity for a wider audience across a range of 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 as a way of 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 should 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 anchor points, or landmarks, which can be recognised and referred to by both the user and the machine. Such as including a reference to a salient object when giving a navigation instruction. There are a number of challenges in doing this, which include having access to a complete dataset of objects with corresponding attribute and positional information, a method to identify landmark candidates from the dataset, and the ability to select the most suitable

* Corresponding author.
E-mail address: phil.bartie@stir.ac.uk (P. Bartie).

http://dx.doi.org/10.1016/j.compenvurbsys.2015.03.003
0198-9715/

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

candidate for a particular task (e.g. the most suitable landmark for a turn instruction) (Richter & Winter, 2014).

The aim of the research presented here was to develop a method which could identify clusters in crowd sourced image tags, based on both the location of the tag and the text content of the tag. The paper focuses on a study of landmark saliency in urban street level scenes from a perspective viewpoint, but could also be used to determine the relatedness of crowd source attributed map points, such as identifying related Flickr tags based on their location and tag content. The identified clusters can be used to generate word lists of associated terms and to identify feature graphs. The results from the street scene study reported in this paper are being used as part of a wider research project focusing on improving HCI for location based services through increased knowledge of the user's environment.

A web based experiment was undertaken to collect data from participants who were asked to tag and annotate features in urban images that they considered to be useful in forming navigational instructions. In some cases users supplied tags for single object features (e.g. a statue), while in other cases the tag represented a collection of features, such as a castle with its many outbuildings and walls.

In order to determine the most salient objects in each scene the user generated tags needed to be grouped according to the object they referred to, so that the number of unique users could be calculated per landmark. The count of crowd-sourced tags can be used to rank feature dominance relative to the other landmarks within a scene. This provides a better understanding of the relative importance of each feature and its meronyms (sub-feature parts). For example a clock on a church tower may be tagged by many people, while the tower itself receives less attention, indicating that the clock would be considered a more dominant focal point on the church façade. The output from this study will help inform and refine the input weightings for a saliency model with the aim of more closely matching human landmark selection choices.

While spatial clustering methods can be used to highlight tag concentrations across the image, it does not offer adequate functionality to identify discrete objects, as tags in close proximity may relate to different real world objects which appear close merely because of the perspective view in the image. Therefore it was necessary to develop a clustering algorithm able to group tags based on both the spatial location of the tag as well as the annotation content. The process was complicated by the range of descriptive terminology supplied in the annotation. For example the same landmark may be described as a *church* by one participant, and as a *clock tower* by another referring to a subpart of the same structure. The algorithm developed used a statistical sentence matching technique in order to link tags with related nearby annotations, forming tag networks (a feature graph) in which nearby tags with similar content were considered to have a stronger relationship than those further away.

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

2. Background and motivation

Landmarks are one aspect of the environment frequently referenced, as they assist in forming mental representations of space (Hirtle & Heidorn, 1993; Tversky, 1993), and in wayfinding tasks (Caduff & Timpf, 2008; Duckham, Winter, & Robinson, 2010; Lovelace, Hegarty, & Montello, 1999; Werner, Krieg-Bruckner, Mallot, Schweizer, & Freksa, 1997; Winter, Tomko, Elias, & Sester, 2008). Studies show that when exploring a new urban region people build a mental model of the space by firstly recognising landmarks, then over time these are joined together into sequences to form routes, which depending on the complexities of the space may lead to a more comprehensive model of the space known as survey knowledge (Siegel & White, 1975).

Landmarks are defined as identifiable features in an environment, whose saliency may be calculated by comparing scores for particular attributes (e.g. their size) and identifying those which deviate from the mean (Elias, 2003; Elias & Brenner, 2004; Raubal & Winter, 2002). These are the objects unlikely to be confused with others, as they appear different to their surroundings (e.g. churches, statues) or are well known international brands (e.g. Starbucks, McDonalds). Landmarks are particularly useful when travelling to a new destination as they can be used at decision points to help orient the navigator, along routes to confirm the location, and as distant landmarks (Lovelace et al., 1999). While it is common for people to include landmarks in conversation, current smartphone digital assistant applications, such as Apple's Siri, Microsoft's Cortana, and Samsung's S-Voice, are unaware of the user's environment and therefore unable to refer to surrounding objects. As speech based interfaces continue to develop it will be useful to include better context awareness which can establish the user's environment and include references to visible landmarks around the user. Google's Project Tango (Lee, 2014) shares a similar ambition to enrich the user experience by allowing software to consider the world beyond the phone's hardware and to consider time and space at a more human scale. Such an ability