Chapter 1

Intelligent Vision Systems for Landmark-Based Vehicle Navigation

Wen Wu  
Carnegie Mellon University, USA

Jie Yang  
Carnegie Mellon University, USA

Xilin Chen  
Chinese Academy of Sciences, China

ABSTRACT

Human drivers often use landmarks for navigation. For example, we tell people to turn left after the second traffic light and to make a right at Starbucks. In our daily life, a landmark can be anything that is easily recognizable and used for giving navigation directions, such as a sign or a building. It has been proposed that current navigation systems can be made more effective and safer by incorporating landmarks as key navigation cues. Especially, landmarks support navigation in unfamiliar environments. In this chapter, we aim to describe technologies for two intelligent vision systems for landmark-based car navigation: (1) labeling street landmarks in images with minimal human effort; we have proposed a semi-supervised learning framework for the task; (2) automatically detecting text on road signs from video; the proposed framework takes advantage of spatio-temporal information in video and fuses partial information for detecting text from frame to frame.

INTRODUCTION

Navigation is the process of planning, recording, and controlling the movement of a craft or vehicle from one place to another. Navigating a vehicle in a dynamic environment is one of the most demanding activities for drivers in daily life. American people drive 12,000 miles per year in average. Studies have long identified the difficulties that drivers have in planning and following efficient routes (King, 1986).

A vehicle navigation system (also termed route guidance system) is usually a satellite navigation system designed for use in vehicles. Most systems
Intelligent Vision Systems for Landmark-Based Vehicle Navigation

typically use a combination of Global Positioning System (GPS) and digital map matching to calculate a variety of routes to a specified destination such as a shortest route. They then present map overview and turn-by-turn instructions to drivers, using a combination of auditory and visual information. A typical turn-by-turn instruction is an auditory “turn right in 0.5 mile”, accompanied by a visual right turn arrow plus a distance-to-turn countdown that reduces to zero as the turn is approached. Vehicle navigation systems generally function well although of course they are wholly dependent on the accuracy of the underlying map database and availability of GPS signal. However, from a human factors perspective, there are several potential limitations to the current design (May & Ross, 2006): mainly presenting procedural and paced navigation information to the driver, and relying on distance information to enable a driver to locate a turn.

Human drivers often use landmarks for navigation. The definition of landmark in navigation context has been studied from varying theoretical perspectives. Lynch described landmarks as external reference points that are easily observable from a distance (Lynch, 1960). Kaplan defined a landmark as "a known place for which the individual has a well formed representation”, and described two theoretical factors that lead to a object or place acquiring landmark status: the frequency of contact with the object or place, and its distinctiveness (Kaplan, 1976). Based on the human factor studies using the above attributes Burnett has further identified the top scoring landmarks in United Kingdom (UK) (Burnett, 2000) such as traffic lights, petrol station, superstore, church, street name signs, etc. It is quite evident that in the United States, we observe that common navigation-useful landmarks include (1) road signs, (2) other signs (e.g., signs of gas stations, fast food restaurants, stores, subway stations, etc) and (3) buildings (e.g., churches, stores, etc). In this chapter, we only focus on detecting text on road signs and how to semi-automatically acquire labeled image data. To learn about more relevant work on other landmarks, readers can read (Wu, 2009).

In order to learn a discriminative model of the landmark of interest for recognition, we need to first label the landmark versus its background in a given image. Manually labeling images is not only a labor intensive task, but also subject to human labeling and annotation errors. While efforts have been focused on online massive user labeling (e.g. MIT LabelMe (Russell, Torralba, Murphy, & Freeman, 2008), The ESP Game), limited attention has been paid to semi-automatically labeling objects in images or videos (Ayache & Qunot, 2007). Our proposed SmartLabel and SmartLabel-2 aim to let a user only mark a small region of interest inside the landmark (or object) on the image with simple input (e.g. dragging a rectangle), and our algorithms can then label the rest of the landmark (object) in the image (Wu & Yang, 2009). The evaluation of proposed SmartLabel-2 and comparison with other methods on a dataset of six object classes indicate that SmartLabel-2 not only works effectively with a small amount of user input (e.g., 1-5% of image size) but also achieve very promising results (macro-average F$_1$=0.84). In some cases, SmartLabel-2 even obtains nearly perfect performance.

Text on road signs carries much useful information for driving and describes the current traffic situation, defines right-of-way, provides warnings about potential risks, and permits or prohibits roadway access. Automatic detection of text on road signs can help to keep a driver aware of the traffic situation and surrounding environments by highlighting signs that are ahead and/or have been passed (Wu, Chen, & Yang 2005). We have introduced a novel framework that can incrementally detect text on road signs from video. The proposed framework takes advantage of spatio-temporal information in video and fuses partial information for detecting text from frame to frame. The