Chapter X

Multimodal Human Localization Using Bayesian Network Sensor Fusion

David Lo, Carleton University, Canada

Abstract

In applications where the locations of human subjects are needed, for example, human-computer interface, video conferencing, and security surveillance applications, localizations are often performed using single sensing modalities. These mono localization modalities, such as beamforming microphone array and video-graphical localization techniques, are often prone to errors. In this chapter, a modular multimodal localization framework was constructed by combining multiple mono localization modalities using a Bayesian network. As a case study, a joint audio-video talker localization system for the video conferencing application was presented. Based on the results, the proposed multimodal localization method outperforms localization methods, in terms of accuracy and robustness, when compared with mono modal modalities that rely only on audio or video.
Introduction

Today’s multisensor systems are becoming more complex with an increasing number of sensors, different types of sensors and increasing complexity of the sensor. Information gathered from multiple sensors often needs to be combined to form a more complete picture of the monitored environment. The dynamics of these modern sensor systems can be very complex. Sensors can be working cooperatively, competitively, or complimentarily (Tebo, 1997). Cooperative sensors work together to collect information of the environment that neither sensor alone can provide. Competitive sensors provide similar information, hence allowing informational redundancy. Complimentary sensors do not depend on each other, but can be combined to provide a more accurate picture of the environment. The complex nature of these sensor systems makes them difficult to combine coherently. Furthermore, the large amount of raw data these sensors generate also make them very difficult to combine. In recent years, the area of data fusion has gained research interest in multisensor applications because it provides a systematic approach to combine and extract useful information from the data. This chapter starts with the high level architectural view and the basic mechanics on how a Bayesian network and its improved variant can be used to fuse data from a multimodal multisensor system. A multimodal human localization system and its implementation are given as an example in the later part of this chapter.

Multimodal Sensor Fusion

Often, a multimodal multisensor system is favored over a single sensor system. By adding more or different types of sensors, the overall system’s accuracy and robustness is improved. For example, the system’s temporal and spatial coverage can be extended by adding more sensors whereas, adding different types of sensors can improve the system’s coverage in the measurement space (Waltz & Llinas, 1990). However, in order to realize these benefits, the system has to be able to take advantage of the extra information introduced by the extra sensors. Data fusion provides a mean for doing that (Waltz & Llinas, 1990). It allows information to be systematically combined from multiple sources while refining the states the system is trying to estimate (Steinberg, Bowman, & White, 1999). Data fusion has been successfully deployed in the field of robotics (Petriu, Ionescu, Petriu, Groen, Spoelder, Yeung et al., 1996; Yeung, McMath, Petriu, Trif, & Gal, 1994) and object tracking in a variety of environments (Strobel, Spors, & Rabenstein, 2001).

Figure 1 shows the general architecture of a multimodal sensor fusion system as block diagram. The Sensor block represents any single sensor modality using either a single sensor or a cluster of similar sensors. The Data Processing block processes the raw sensor data. Often, in a multimodal sensor system, fusion happens at both the raw data level and the information level (Lo, 2004a). Therefore, the type of processing performed by the Data Processing block can range from simple data filtering at the sensor level to complex statistical analyses and features extraction at the information level. The Mapping block transforms the processed data into a common space in which all processing modules can refer to; for example, a common coordinate system or common measuring unit. The Data Fusion and
Gesture Spotting Using Fuzzy Garbage Model and User Adaptation
[www.igi-global.com/article/gesture-spotting-using-fuzzy-garbage/55996?camid=4v1a](www.igi-global.com/article/gesture-spotting-using-fuzzy-garbage/55996?camid=4v1a)