論文の内容の要旨

論文題目：Structural Understanding of Instruction Videos by Integrating Linguistic and Visual Information
（言語情報と映像情報の統合による作業教示映像の構造的理解）

氏名：　柴田　知秀

本文

To perform real-word information processing, such as intelligent robotics, multimodal dialogue system and video processing, it is essential to integrate several media processing technique such as natural language processing, speech recognition and image analysis. From the viewpoint of natural language processing, since language in the real world is strongly depends on the scene, it is important to understand utterances in accordance with the scene.

This thesis focuses on handling video contents. Among several types of videos, in which instruction videos (how-to videos) about sports, cooking, D.I.Y., and others are the most valuable, we focus on cooking TV programs. In realizing flexible utilization/access of video contents, the crucial point is the structural understanding of their contents, which requires the interpretation of utterances based on wider contexts including the scene.

Chapter 2 describes basic linguistic analysis of cooking instruction utterances (closed caption texts). First, we perform anaphora resolution, which is inevitable to detect the discourse structure or correspond linguistic information to visual information. We build an anaphora resolution system based on the large-scale case frame. Next, we detect utterance-type of a clause of each utterance. In cooking instruction utterances, while explanations of actions are dominant, there are several types of utterances such as declaration of beginning of series of actions, tips of actions, notes, etc. We classify cooking instruction utterance and recognize utterance-type by clause-end patterns. Then, we analyze the discourse structure of instruction utterances. This analysis is performed by integrating the anaphora resolution result, utterance-type and generic discourse structure rules, which consider cue phrases and word chaining.

Chapter 3 proposes an unsupervised topic identification method integrating linguistic and visual information based on Hidden Markov Models (HMMs). Identified topics lead to video segmentation/summarization and are used for automatically acquiring the object models described in Chapter 4. We employ HMMs for topic identification, wherein a state corresponds to a topic and various features including linguistic, visual and audio information are observed. This study considers a clause as an unit of analysis and the

following eight topics as a set of states: preparation, sauteing, frying, baking, simmering, boiling, dishing up, steaming. The basic linguistic feature is a case frame, which is a generalization of utterances referring to an action, such as ``ireru(add)'' and ``kiru(cut)''. Furthermore, we incorporate domain-independent discourse features such as cue phrases, noun/verb chaining, which indicate topic change/persistence, into the case frame. We utilize visual and audio information to achieve robust topic identification. As for visual information, we can utilize background color distribution of the image. As for audio information, silence can be utilized as a clue to a topic shift.

Chapter 4 presents a method for automatically acquiring object models from large amounts of video for performing object recognition. We first collect pairs of a close-up image and a keyword. Close-up images are extracted with edge detection and, in the close-up image, region segmentation is performed and the salient region is determined considering the following points: area, center of gravity and variance of pixels in a region. A keyword is extracted from instructor's utterances when the close-up image appears. In case of cooking, objects (i.e. ingredient) change their shape/color along with the progress of cooking. Consequently, good examples for object acquisition cannot be collected from video segments whose topic is sauteing or dishing up. Therefore, a keyword is extracted only from segments whose topic, which is identified by the proposed method, is preparation. The important score of each word is calculated according to the linguistic analysis result, such as the discourse structure analysis and utterance-type detection, and the word that has the maximum score is extracted as a keyword. After collecting pairs of a close-up image and a keyword, for each keyword, its object model is acquired by summing RGB histograms in the salient region. Next, we perform object recognition based on the acquired object model and the discourse structure. We can acquire the object model of around 100 foods and its accuracy is 0.778, and the accuracy of object recognition is 0.727.

Chapter 5 describes our video retrieval system. In this system, a user can ask a query in natural language and can enjoy the search result, which is similar to the user's query. To present the accessible mean to the video, we generate a summary of the video. This analysis is based on topic segmentation, important utterances extraction, topic identification result, object recognition result.