Motion Based Automatic Abstraction of Cooking Videos

Koichi MIURA
Reiko HAMADA
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo, 113-0033, Japan
TEL: +81-3-5841-7413
FAX: +81-3-5800-6922
miura,reiko@mtl.t.u-tokyo.ac.jp

Ichiro IDE
Nat'l Institute of Informatics
2-1-2 Hitotsubashi,
Chiyoda-ku,
Tokyo, 101-8430, Japan
ide@nii.ac.jp

Shuichi SAKAI
Hidehiko TANAKA
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo, 113-0033, Japan
sakai,tanaka@mtl.t.u-tokyo.ac.jp

ABSTRACT

In this paper, we propose a method to abstract cooking videos. We define cooking video abstraction as shrinking videos maintaining the understandability of general cooking procedures visually and intuitively. Cooking motions and appearances of foods are considered as important segments in a cooking video. A method to extract such important segments referring to the intensity of motion in an image is proposed, and effectiveness of the method is shown through evaluation experiments. We also developed a cooking video abstraction system that assembles important segments detected by the proposed method and repetitious motions that is especially important among cooking motions. The resultant abstracted videos were about 1/8 to 1/12 of the original videos in time, maintaining the understandability. The validity of the abstraction method was checked by comparing the abstracted videos with manually abstracted videos provided from a broadcasting station.

1. INTRODUCTION

Following the advance in telecommunication technology, large amount of multimedia data has become available from broadcast video. Multimedia data analysis is becoming important to store and retrieve them efficiently. However, characteristics of videos and purposes of viewers vary among different kinds of videos. Thus, it is necessory to limit the domain of the target video and refer to domain specific knowledge, for high level content analysis.

We chose cooking video as a target, and aim for building a practical system with high accuracy using domain specific knowledge[5, 8]. Cooking is familiar to daily life, and the demand for semantically structured cooking videos should grow in proportion to home (especially kitchen) automation.

In this paper, we propose a method to abstract cooking videos. A cooking program is a kind of an instruction video that people view from a practical point of view. On the other hand, cooking videos include many redundant segments such as chatting, which requires the viewers a certain amount of time to view. Therefore, to select recipes and to actually cook in daily life, a cookbook tends to be easier to browse. However, since videos contain visual information that text cannot express sufficiently, it is more effective to understand the cooking procedures.

Thus, we abstract cooking videos to easily browse through recipes. We define cooking video abstraction as shrinking videos maintaining the understandability of general cooking procedures visually and intuitively.

Various studies have been done on automatic video abstraction for news, documentary, and so on[4, 6]. These kinds of videos are relatively not redundant, and abstracted videos tend to be utilized for selecting a segment in the original videos. Works similar to ours are in sports video abstraction. A work on sports video abstraction[7] refers to an external database corresponding to video contents. This method is effective for videos that could easily refer to external data corresponding directly to their contents, which is difficult in cooking videos.

On the other hand, it is reported that abstracted videos are not so effective since split audio accompanying the abstracted images is too choppy[4]. From this point, audio naturalness is considered in abstracting videos[6, 9], and audio is usually replaced with new narrations in abstracted videos provided from a broadcasting station. However, since motion and preparation steps can be roughly understood solely from visual information in cooking videos, outline of the cooking procedure could be understood to a certain extent without audio. Thus, our abstraction does not consider audio continuity, and abstracted videos are created solely referring to image features.

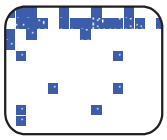
Considering the above mentioned issues, cooking video could be considered as a good target for abstraction.

2. FEATURES OF COOKING VIDEOS

2.1 Structure of cooking videos

As shown in Fig. 1, shots in cooking videos could be categorized into (a)face shot, and (b)hand shot, which appear almost alternately, as shown in Fig. 2.

In face shots, almost the whole kitchen is shown. Though a teacher or an assistant explains a cooking procedure, their motions and foods are too small to obtain visual informa-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 2002 ACM X-XXXXX-XX-X/XX/XX ...$5.00.

tion. On the other hand, hand shots are close-ups of tools and/or hands while cooking something that have rich visual information. However, each hand shot contains sub-shot segments such as cooking motion and appearance of foods that are important keys to understand the recipe, but also redundant segments between them as well.