News Video Classification based on Semantic Attributes of Captions

Ichiro IDE, Reiko HAMADA, Hidehiko TANAKA and Shuichi SAKAI
Graduate School of Electrical Engineering, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, JAPAN
TEL: +81-3-3812-2111 ext.7413, FAX: +81-3-5800-6922
E-mail: { ide | reiko | tanaka | sakai } @mtl.t.u-tokyo.ac.jp

Abstract

As a basis for automatic indexing to video data based on shot classification, we will present a graphical classification rule acquisition method based on semantics of accompanying natural language data *i.e.* captions. A preliminary experiment to actual television news programs showed good correspondence between graphical characteristics and semantic attributes of captions.

1 Introduction

As the amount of broadcast video data increases, it is becoming more and more important to store them in a well organized manner considering recycling and searching. Above all, television news programs are worthwhile indexing considering the importance and usefulness. Currently this process is mostly done manually, but automatic indexing is in big demand to cope with the increasing amount and to achieve sufficient precision for detailed searching.

We are trying to accomplish this task by referring to both video data and accompanying natural language data in television news programs. There are several notable attempts made to automatically index television news video from this approach such as the Informedia project's [1] News-on-Demand system [2]. Their indexing strategies are mostly based on statistics or just simple occurrences, which do not necessarily ensure the correspondence of the keyword and the image content. To avoid this problem, we have previously proposed an indexing method based on shot classification [3], which classifies shots into several graphically typical classes, and index them with keywords that match the typical image contents. Similar work is done by Nakamura *et al.* [4], but our common problem was that classification rules were given in a top-down manner, which resulted in restricting the number of classes to relatively few.

We are currently developing a method that first learns graphical classification rules from supervisory video data, and then indexes the new-coming video data with appropriate keywords for the class they belong to. The learning process employs semantic attributes of captions derived from conceptual dictionaries as classification standards. This expands the number of typical classes and even lets images without captions be vaguely indexed referring to graphical characteristics. We will introduce the basic idea of this research and present the result of a preliminary experiment of the learning process in this poster.

2 Image Classification based on Semantic Attributes of Captions

2.1 Learning Classification Rules from Supervisory Video Data

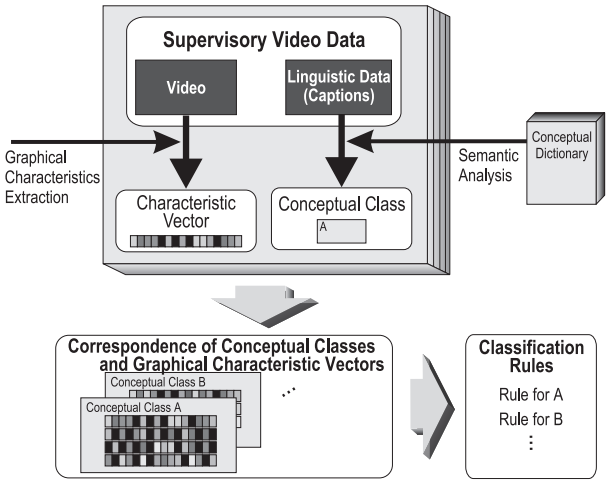

Figure 1: Learning Classification Rules from Supervisory Video Data

Figure 1 shows the outline of the learning process. As the first step, shots with captions are analyzed to learn graphical classification rules. Graphical characteristic vectors are derived from a shot, and are classified according to the semantic attributes –practically, node IDs of conceptual classes– of accompanying captions. Although the term, "rules" is used, they are actually statistic information derived from each characteristic vector in each conceptual class. After analyzing all the supervisory video data, each conceptual class should have a representative graphical characteristic vector. We are currently examining several methods such as the principal component analysis and the memory based reasoning for the acquisition of rules.

Table 1: Result of Preliminary Experiment: Correspondence between Conceptual Classes and Graphical Characteristics

Numbers of Faces	Titles of Conceptual Classes			
	Personal	Gathering	Locational	Others
None	——	'opinion, decision, investigation, acceptance'	'place name'	'counting unit'
One	'chief'; 'person's name'	——	'office, market, station'; 'place name'	——
Two	'social standing'; 'human being'	'announcement, report, rumor'	'temple, shrine, school'; 'house, inn, classroom'; 'place name'	'counting unit'; 'number'; 'principle, rule, method, custom, plan'; 'money'
Three and more	——	'speech, debate, meeting, comment, explanation'; 'promise, negotiation, approval'; 'opinion, decision, investigation, acceptance'; 'parliament'; 'gathering, presence'	'place name'	——

2.2 Preliminary Experiment

As a preliminary experiment, we have applied the learning process to 75 minutes of Japanese television news video. Only one parameter was set experimentally as an element of the graphical characteristic vector; the number of relatively large faces in the first frame of a shot. The Classified Lexical Table [5] was used to classify the captions to conceptual classes according to their semantics. Table 1 shows the correspondence of the numbers of faces and the titles of the top 30% frequent conceptual classes.

The result shows relatively good correspondence between numbers of faces and conceptual classes: (1)Conceptual classes related to human beings corresponded to one and two faces (titled 'Personal'), and (2)classes related to gatherings corresponded to two, three and more faces (titled 'Gathering'). Classes related to locations mingled in to all groups equally (titled 'Locational'), since there was no graphical characteristic parameter to classify them in this experiment. These should be classified by supplementing various characteristic parameters to the vector. The conceptual class 'opinion, decision, investigation, acceptance', related to gathering corresponded to 'no faces', since there were many 'gathering' shots taken from behind.

3 Conclusion and Future Work

We have proposed a news video classification method based on natural language information accompanying the video. The preliminary experiment showed promising results. Although, at this point, the proposed system may look somewhat similar to the Name-It system [6], further expansion of the data and graphical characteristic parameters will enhance the ability and generality.

We will also proceed with the automatic indexing phase based on the learned graphical characteristics. This should enable both advanced keyword extraction and indexing.

References

[1] "The Informedia Project"; http://www.informedia.cs.cmu.edu/.

[2] Hauptmann, A. G. and Witbrock, M. J.; "Informedia News-on-Demand: Using Speech Recognition to Create a Digital Video Library"; *Proc. AAAI'97 Spring Symposium on Intelligent Integration and Use of Text, Image, Video and Audio Corpora*, pp.120-126, May 1997.

[3] Ide, I., Yamamoto, K. and Tanaka, H.; "Automatic Indexing to Video based on Shot Classification"; *Proc. 56th National Convention IPS Japan*, Vol.2, pp.263-264, Mar 1998.

[4] Nakamura, Y. and Kanade, T.; "Semantic Analysis for Video Contents Extraction –Spotting by Association in News Video–"; *Proc. ACM Multimedia'97*, pp.393-402, Nov 1997.

[5] National Language Research Institute of Japan, The; "NLRI Natural Language Processing Data Series 5: The Classified Lexical Table (Bunrui-Goi-Hyo) [Floppy Disk Edition]"; Shuei Publishers, 1993.

[6] Satoh, S., Nakamura, Y. and Kanade, T.; "Name-It: Naming and Detecting Faces in Video by the Integration of Image and Natural Language Processing"; *Proc. IJCAI-97*, pp.1488-1493, Aug 1997.