METADATA ANNOTATION THROUGH MEDIA INTEGRATION

Ichiro Ide

Graduate School of Information Science
Nagoya University
1 Furo-cho, Chikusa-ku, Nagoya
464-8601, Japan
ide@is.nagoya-u.ac.jp

Reiko Hamada

Grad. School of Info. Science & Technology
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo
113-8656, Japan
reiko@mtl.t.u-tokyo.ac.jp

1. INTRODUCTION

Metadata annotation, or indexing is the key issue in image and video retrieval. Content-based image retrieval (CBIR) has been trying to represent images by low level image features, but is suffering to bridge the so-called *semantic gap* between the graphical features and the semantics. In the real world, people called archivists are hired at libraries, archives, or even at private television broadcasting companies in order to bridge the gap manually. So far, this approach seems to be still superior to automatic indexing in quality. However, when we consider the amount of image and video data produced and stored everyday, manual indexing has a certain limit; human beings get tired, have different backgrounds, and cost much. In order to cope with the demands that exceed the limit of manual indexing, automatic indexing is an essential technology for image and video retrieval. This paper is a short summary of my works on automatic video indexing in the last decade.

2. METADATA ANNOTATION THROUGH MEDIA INTEGRATION

My interest in this field started from Rohini K. Srihari's paper titled *PICTION: A system that uses captions to label human faces in newspaper photographs* [1]. This paper introduces a system named PICTION which uses spatial and characteristic constraints in captions to identify humans in an accompanying photograph.

When I started working in this field in 1995, *Multimedia* was a big boom in the real world. The famous *Informedia* project [2] had started at Carnegie Mellon University the year before, so it was a boom in the academic world, too. Many exciting works on media integration were seen at related conferences (such as the *ACM Multimedia Conference*), but most of them seemed to just simply put the results of mono-media processing together all at the end. Since I was still a reckless master course student, I thought that these works were *pseudo*-media integration and thus I had better follow the PICTION way. So, I went out to seek for the *true* media integration; annotation of indices that reflect what is actually happening in the image or video. I named this approach recently, *WYGIWYS: What you get is what you see*.

The next Section introduces two works that annotates indices to video through media integration. The first is automatic news video indexing based on the *WYGIWYS* philosophy, and the second is cooking video indexing for cooking assistance.

3. EXAMPLES

3.1. News video indexing

News videos are records of social activities, which could be considered as an important heritage of our race. This Section introduces a work in news video indexing. It proposes an indexing method based on the *WYGIWYS* philosophy, which most other works do not necessarily consider, although it is only applied to 6 hours of video; about 1% of the data I am currently working on. Refer to [3] for details.

3.1.1. Outline

Most works on news video indexing utilizes indices simply extracted from closed-caption text (transcript of speech), regardless of the correspondence between the indices and the visual contents within the image. Considering this issue, we proposed an automatic news video indexing method that considers the correspondence between indices derived from open-caption (telop) texts and the actual visual contents.

As shown in Fig. 1, correspondence was considered separately within four attributes that represent a content; *When*, *Where*, *Who*, and *What* (*4W*). Although this may seem a rather limited correspondence, these attributes are the essential facts when understanding news, among the so-called *5W1H* (*4W* plus *Why* and *How*) attributes. Correspondence

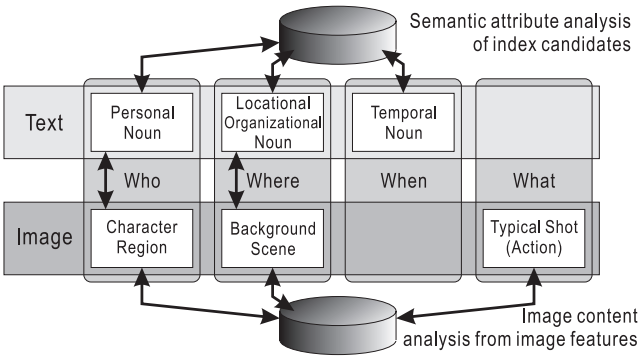

Fig. 1. Attribute-based news video indexing.

within each attribute was considered depending on the nature of the attribute. We focused especially on *Who* and *Where*; personal nouns were selected as indices for human figures, and locational/organizational nouns were checked if they match the background scene in the image. *When* is usually not visible in the image, and *What* appears seldom in the open-caption text, so we extracted them solely from text and image, correspondingly.

3.1.2. Text attribute analysis

Open-captions (telops) were used as the source for the indices. Only those that are noun phrases were selected for the analysis, since they briefly describe important information on the image contents.

In order to realize the method, we needed to classify the open-caption text based on their semantic attributes; (1) *personal*, (2) *locational/organizational*, (3) *temporal*, and (4) *general*. An analysis method for Japanese noun phrases developed for the purpose, which takes into account that in most cases, the suffix (last noun in a phrase) determines the semantic attribute of a noun phrase. A suffix dictionary was created for the analysis, by gathering nouns that meet certain criteria, and later expanding the vocabulary by a thesaurus.

Note that this method not only extracts named-entities with proper nouns, but also common nouns such as 'fire fighter', 'site of disaster' and so on. Refer to [4] for details.

3.1.3. Image attribute analysis

As news programs mostly focus on providing important information on human activities, there is a good chance that a considerably large human figure appears in the image. This feature leads to the importance of taking special consideration of the existence of human figures when analyzing the image contents of a news video.

In order to analyze *Who*, *Where* and *What*, we decided to segment and separately analyze the foreground figure and

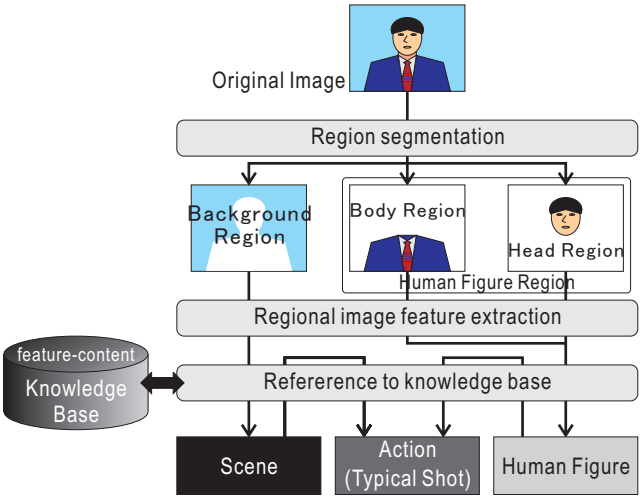

Fig. 2. Image contents analysis by character region segmentation.

the background scene, as shown in Fig. 2. First, face detection is applied to an image frame. The body and head regions are extracted according to a template, forming a figure region, and the remaining region is segmented as a background region. Next, contents are analyzed within each region by referring to a knowledgebase.

For the knowledgebase on scenes, we collected image features corresponding to typical scenes such as (1) The cabinet meeting room, (2) The parliament, (3) press conference room, (4) court, and (5) studio to estimate the location of a scene roughly. For the knowledgebase on actions, we defined typical shots such as scenery, speech, meeting, and so on based on the number of figures in an image after excluding CG and anchor shots.

Refer to [5] and [6] for details.

3.1.4. Indexing

After the text and image analyses, indexing was performed to integrate the results based on Fig. 1. The correspondence for *Who* and *Where* for each shot were realized as follows, after shot segmentation and topic boundary detection based on anchor shots:

- Noun phrase with personal attribute – Human figure existence
  First, find shots with similar image features within the current topic, and then select open-captions with a personal attribute from the similar shots as candidates for indices.

- Noun phrase with locational/organizational attribute – Background scene
  First find shots with similar image features from the whole archive, and then select open-captions with a

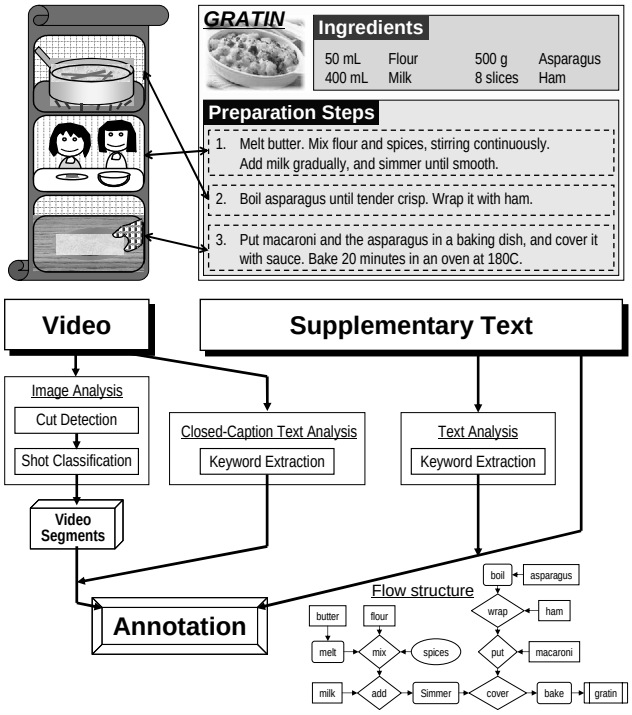

Fig. 3. Alignment of cooking steps in a cooking video and a supplementary textbook.

locational/organizational attribute from the similar shots as candidates for indices.

In both cases, the candidates are ranked according to the similarity of image features, and those with high similarity are selected as indices.

3.2. Cooking video indexing

Cooking videos are in a sense contrary to news videos since they contain contents on extremely domestic affairs. However, it concerns one of the most frequent task that we engage daily for our living, and above all, it makes our daily life more enjoyable. This Section introduces a work in cooking video indexing. It proposes an indexing method that considers ordinal constraints in recipes. Refer to [7] for details.

3.2.1. Outline

As shown in the bottom of Fig. 3, the order of preparation steps often differ between the video and the text recipe in cooking programs. This is because some steps are not interdependent, although most of them needs to be performed in correct order, where an instructor may select the order of the steps according to other conditions. In the proposed method, such dependency structures were analyzed in order to align the steps in a text recipe to corresponding video shots. The flow structure is represented as an inverted tree with nodes called *text blocks* that contain sentences with pairs of specific nouns and verbs.

On the other hand, the video stream is segmented into segments called *video scenes* based on shot and scene classifications. Clues from both image and text in the video stream are also gathered for the indexing.

At the end, *text blocks* and *video scenes* are associated under certain conditions.

(a) Board (b) Table (c) Range (1) Hand shots (2) Face Shot

Fig. 4. Shot categories in a cooking video.

3.2.2. Extraction of text blocks and their ordinal structure

In order to analyze the text recipe, we created a domain-specific dictionary which contains nouns that represent ingredients and spices, utensils and containers, and so on, and also verbs that represent certain cooking operations. The verbs are further classified into categories according to the functions that they may play in a flow structure; certain verbs represent irreversible cooking operations such as 'cut' and 'bake', while another may represent a merger or a divergence in the flow such as 'mix' or 'peel'.

The text recipe is segmented into *text blocks*, by extracting sentences that contain a series of verbs starting with a verb which is in certain relationships with certain (ingredient or container) nouns.

Next, the flow structure is analyzed based on the noun-verb modification relation and the function inherent in certain verbs, referring to the domain-specific dictionary.

Refer to [8] for details.

3.2.3. Extraction of video scenes and their classification

As shown in Fig 4, shots in a cooking video could roughly be categorized into (1) *hand shot* and (2) *face shot*, of which the former is the main interest when analyzing the video. Details on the automatic categorization could be found in [9].

Hand shots are then classified into *board*, *table*, (*range gas stove*) according to the color distribution at the bottom of the image (See Fig. 4). A *video scene* is detected by clustering adjoining *hand shots* with a similar color distribution.

3.2.4. Indexing

A *text block* in the flow structure is associated to a *video scene*. The relevance between a *text block* and a *video scene* is evaluated by the following factors:

1. Ordinal restriction
2. Matching of terms related to certain operations and background scene
3. Cooccurrences of domain specific terms

The analysis starts from the root (bottom) of the inverted tree to the leaves (top), which tries to maximize the overall relevance of the entire tree.

4. CONCLUSION

The two works introduced in this paper has already reached the first stage, and are now proceeding to the next stage.

For news video indexing, I am working on analysis and knowledge extraction from inter-topic relations in a large-scale archive with more than 800 hours of daily video [10, 11], rather than indexing individual shots / topics precisely. This project aims to utilize implicit information inherent in the relation of the contents, rather than extracting explicit information present in the contents.

Cooking video indexing is now in the stage of using the indices for a real-world application. We are working on an interface that assists the cooking activity in kitchen [12]. This project aims to utilize the indexed contents as a multimedia knowledge base, in order to assist daily cooking tasks.

In the future, we will continue pursuing a method to utilize every available data source for better understanding of multimedia contents.

5. ACKNOWLEDGMENTS

Most of the presented works are collaborations with the current and past members of Professors Hidehiko Tanaka and Shuichi Sakai's laboratory at the University of Tokyo. Especially credits to the cooking video indexing go primarily to Dr. Reiko Hamada and Mr. Koichi Miura.

Most part of the work in Section 3.2 was funded by a Grant-in-Aid for Scientific Researches from JSPS (#14380173), and collaborated under a Joint Research Program at NII.

6. REFERENCES

[1] R. K. Srihari, "PICTION: A system that uses captions to label human faces in newspaper photographs," in *Proc. 9th National Conf. on Artificial Intelligence (AAAI-91)*, July 1991, pp. 80–85.

[2] School of Computer Science Carnegie Mellon University, *Informedia Digital Video Library*, http://www.informedia.cs.cmu.edu/.

[3] I. Ide, R. Hamada, S. Sakai, and H. Tanaka, "An attribute based news video indexing," in *Proc. ACM Multimedia 2001 Workshops –Multimedia Information Retrieval–*, Oct. 2001, pp. 70–73.

[4] I. Ide, R. Hamada, S. Sakai, and H. Tanaka, "Semantic analysis of television news captions referring to suffixes," in *Proc. 4th Intl. Workshop on Information Retrieval with Asian Languages*, Nov. 1999, pp. 37–42.

[5] I. Ide, R. Hamada, S. Sakai, and H. Tanaka, "Scene analysis in news video by character region segmentation," in *Proc. ACM Multimedia 2000 Workshops*, Nov. 2000, pp. 195–200.

[6] I. Ide, K. Yamamoto, and H. Tanaka, *Automatic video indexing based on shot classification*, vol. 1554 of *Lecture Note in Computer Science*, pp. 87–102, Springer-Verlag, Jan. 1999.

[7] R. Hamada, K. Miura, I. Ide, S. Satoh, S. Sakai, and H. Tanaka, *Multimedia integration for cooking video indexing*, vol. 3332 of *Lecture Note in Computer Science*, pp. 657–664, Springer-Verlag, Dec. 2004.

[8] R. Hamada, I. Ide, S. Sakai, and H. Tanaka, "Structural analysis of preparation steps on supplementary documents of cultural tv programs," in *Proc. 4th Intl. Workshop on Information Retrieval with Asian Languages*, Nov. 1999, pp. 43–47.

[9] K. Miura, R. Hamada, I. Ide, S. Sakai, and H. Tanaka, "Motion based automatic abstraction of cooking videos," in *Proc. ACM Multimedia 2002 Workshop on Multimedia Information Retrieval*, Dec. 2002.

[10] I. Ide, H. Mo, N. Katayama, and S. Satoh, *Topic threading for structuring a large-scale news video corpus*, vol. 3115 of *Lecture Note in Computer Science*, pp. 123–131, Springer-Verlag, July 2004.

[11] I. Ide, T. Kinoshita, H. Mo, N. Katayama, and S. Satoh, *trackThem: Exploring a large-scale news video archive by tracking human relations*, vol. 3689 of *Lecture Note in Computer Science*, pp. 510–515, Springer-Verlag, Oct. 2005.

[12] R. Hamada, J. Okabe, I. Ide, S. Satoh, S. Sakai, and H. Tanaka, "Cooking Navi: Assistant for daily cooking in kitchen," in *13th ACM Intl. Conf. on Multimedia*, to appear in Nov. 2005.