Name Identification of People in News Video by Face Matching

Ichiro IDE [*]
ide@is.nagoya-u.ac.jp, ide@nii.ac.jp

Takashi OGASAWARA [†]
toga@murase.m.is.nagoya-u.ac.jp

Graduate School of Information Science, Nagoya University; Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan

Tomokazu TAKAHASHI
ttakahashi@murase.m.is.nagoya-u.ac.jp
Japan Society for the Promotion of Science
/ Nagoya University

Hiroshi MURASE
murase@is.nagoya-u.ac.jp Graduate
School of Information Science,
Nagoya University

ABSTRACT

Recently, there is a strong demand for making use of large amounts of video data efficiently and effectively. When considering broadcast news video, people who appear in it is one of the major interests to a viewer. This is the common motivation of recent works that focus on extracting names of people that appear in news video footages. However, these works suffer a serious problem; a person is often referred to by various names depending on situations and along time. In this paper, we propose and evaluate a method that handles this problem by identifying faces together with names. Faces are extracted by face detection technology and annotated with person name candidates extracted from closed-caption text. Then, all face-name pairs are compared by face identification technology and text matching of names. As a result, different names of a same person are identified.

1. INTRODUCTION

1.1 Background

Recent advances in storage technologies have provided us with the ability to archive many hours of video streams accessible as online data. In order to make efficient and effective use of the voluminous video data, automatic analysis of contents for retrieval, browsing, and knowledge extraction is essential.

Among various types of video, we are focusing on broadcast news video, since it is a valuable record of the human society. In that sense, the main interest in news video is related to *people* who appear in them.

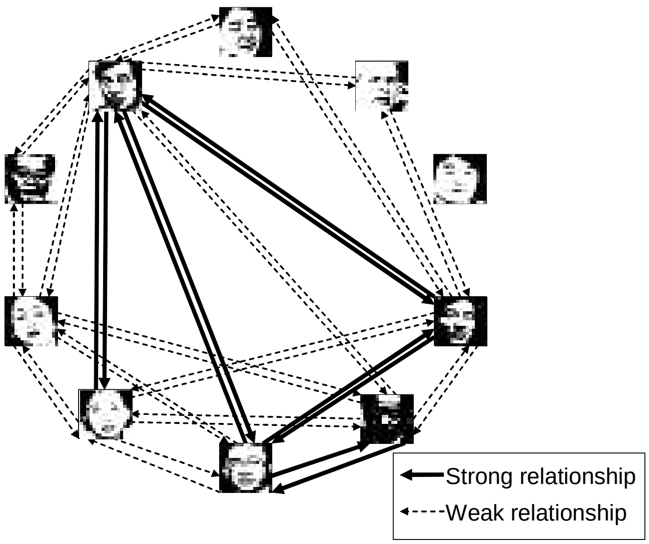

Figure 1: Example of a human relationship graph.

Previously, we have tried to extract human relationship (Fig. 1) from closed-caption texts of the news video data in an archive by counting the co-occurrences of person names in a sentence [4]. We also tried to extract the relationship from the patterns of face co-occurences in a news story [8].

As shown in Fig. 2, we implemented an interface that visually presents the obtained human relationship of a specified person (The name in the center of the circle) together with the actual news stories that the two person co-occurred in (The thumbnail icons at the bottom). The interface also lets a user track down the relationship graph structure by setting one of the person around the circle as a new person-in-focus, which is effective to understand the social network structure.

[*] Also affiliated to National Institute of Informatics.

[†] Currently at Toyota Motor Corporation.

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.

CVDB 2007 June 10, Beijing, P. R. China

Copyright 2007 ACM 978-1-59593-794-0/07/0006 ...$5.00.

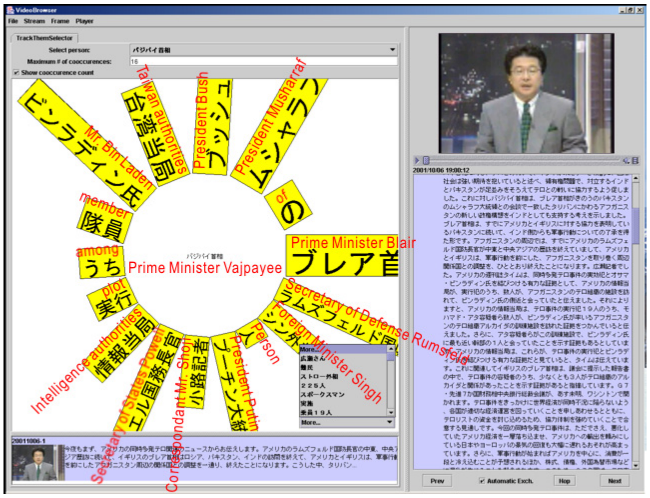

Figure 2: News video browsing based on human relationship: The *trackThem* interface. The left side is the interface, and the right side is used to play a specified video and its closed-caption text simultaneously.

These works, however, suffer a serious problem that a person is often referred to by various names depending on situations and along time. Thus, in order to improve the quality of the extracted information on human relationships, name identification is essential. In this paper, we propose a method that handles this problem by identifying faces together with names.

As a work to cluster name-face pairs, Tamara et al. proposed a method that clusters names and faces in Web news pages and their captions [1]. However, the faces and names that appear in Web news are a small portion of people who appear in the news story; they are usually people symbolic to the topic. In this paper, we aim to identify names of not only symbolic people but also people playing secondary roles who appear in broadcast videos

1.2 Variation of Person Names

Figure 3 lists how a person is referred to by different names in different situations. We classified them to the following three types:

1. **Position / Honorary titles**
   As in (b)–(g), a person is referred to by their names associated with their positions or honorary titles. In order to identify them, up-to-date knowledge on real-world affairs is needed.

2. **Synonyms**
   As in (c) and (d), there are synonyms (includes abbreviations) of the titles. In order to identify them, a thesaurus could be used.

3. **Change of states**
   As in (c), (d) and (g), a person's status may change along time. In this case, the person was first the "Minister for Health and Welfare", and later became the "Prime Minister". In order to identify them, knowledge on real-world affairs including the past is needed.

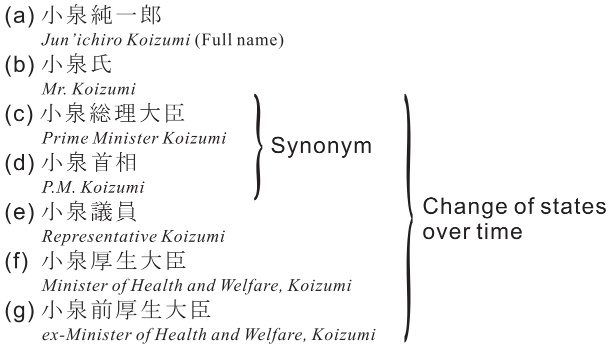
(a) 小泉純一郎
Jun'ichiro Koizumi (Full name)
(b) 小泉氏
Mr. Koizumi
(c) 小泉総理大臣
Prime Minister Koizumi
(d) 小泉首相
P.M. Koizumi
(e) 小泉議員
Representative Koizumi
(f) 小泉厚生大臣
Minister of Health and Welfare, Koizumi
(g) 小泉前厚生大臣
ex-Minister of Health and Welfare, Koizumi

Figure 3: Variation of person names.

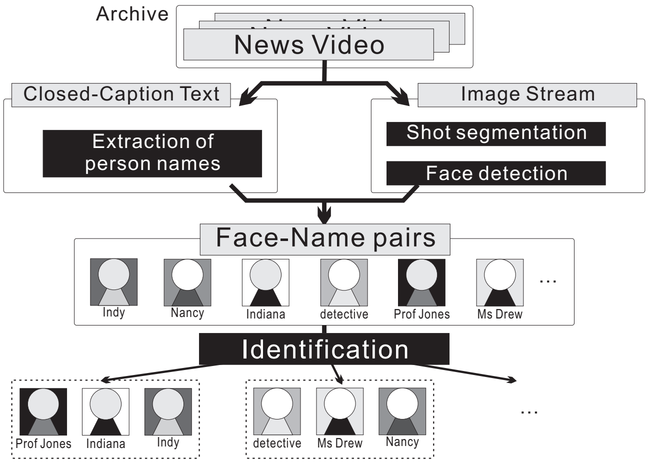

Figure 4: Process of the name identification method.

Type 2 may be solved using a thesaurus, but it is very difficult to automatically identify Type 1 and especially Type 3, only with text information.

1.3 Overview of the Identification Method

Considering the difficulty of name identification by text, we propose a method that does not need external knowledge on real-world affairs. The method identifies a person by identifying faces obtained from the image together with names in the closed-caption text, associated with the face.

The details of the method is introduced in Sect. 2, and Sect. 3 reports the result of an experiment where the method was applied to actual news video data. Conclusions are given in Sect. 4.

2. NAME IDENTIFICATION BY FACE-NAME PAIRS

The flow of the proposed name identification process is shown in Fig. 4. In this section, we will describe each block of the process, following the definition of the terminology.

2.1 Terminology

The following are the definition of the terms that compose a broadcast video stream:

- **Frame:**
  A still image which is the minimal unit of a video stream.

- **Shot:**
  A sequence of frames that are continuous when seen as image.

- **Cut:**
  The boundary between two consecutive shots.

- **Scene:**
  A sequence of shots that are semantically continuous.

2.2 Shot Segmentation

When the contents of a shot focus on a certain person, such as in an interview or a speech at a press conference, the person usually appears largely in the center of the frame, when there are no restrictions. In addition, when the person in focus changes, the shot usually changes. Considering such characteristics related to video grammars, we defined cuts as boundaries to associate person names with a face.

Shots are segmented before all the process as follows:

- The RGB color histograms of adjoining frames are compared in order.

- When the similarity of the histograms with those of the previous frame is larger than a threshold, the gap right before the frame is detected as a cut.

The similarity $S_{\mathbf{H}_1,\mathbf{H}_2}$ between two color histograms $\mathbf{H}_1, \mathbf{H}_2$ of adjoining frames is given by calculating the histogram intersection, defined as:

$$S_{\mathbf{H}_1,\mathbf{H}_2} = \frac{\sum_{i=1}^{I} \min(H_{1,i}, H_{2,i})}{\sum_{i=1}^{I} H_{2,i}} \qquad (1)$$

$$\begin{aligned} I &: \text{Number of bins in a histogram} \\ H_{n,i} &: \text{The } i\text{–th element of } \mathbf{H}_n \end{aligned}$$

The colors in the input images are represented as a combination of 256 levels of each of the R, G, B color component.

2.3 Extraction of Names from Closed-Caption Text

Next, names are extracted from the closed-caption (CC) text corresponding to each shot. The CC text is provided from the broadcaster, and usually appears shortly behind the actual utterances of words in the audio stream. Here, we used CC texts in the archive that were already automatically synchronized to the audio stream.

Person names were extracted by applying the method proposed in [3]. The outline of the method is as follows:

Step 1. Nouns are extracted from the CC text by morphological analysis[1]

Step 2. Person names are extracted from noun compounds with specific suffices by looking up a dictionary. The dictionary contains suffices such as "Mr." "President" and "Minister" in Japanese [2].

2.4 Extraction of Faces from Image Sequences

Meanwhile, faces are extracted from the frames that compose shots. Face detection is performed by a method that uses joint Haar-like features [9, 7], which is very fast regardless of image resolution and is robust against noise and changes in illumination.

Because of the characteristics described in Sect. 2.2, at most one face should be detected from a shot. Therefore, all faces detected from a shot are considered as a sequence of the same person's face. By extracting faces as a sequence, rather than a single image, the precision of face recognition should improve. Note that even if there are several different faces in a shot, only one major one is selected by the face detection.

2.5 Associating Names to a Face

After the processing in Sects. 2.3 and 2.4, person names that appear in a shot, if any, are associated with a face in the shot. At this point, the process does not annotate a face with a single name as in the case of a related work; the Name-It system [10]. Instead, the purpose of this process is rather to collect multiple face-name (candidate) pairs at this point, and then identify the correct name for the face later by the face-name pair-wise matching.

2.6 Name Identification

Finally, the names are identified based on the face-name pairs obtained in Sect. 2.5. All combinations of faces detected in the video archive are compared together with the associated names.

If the following two conditions are satisfied, both names are considered to represent the same person:

1. **High similarity of faces:**
   The similarity of faces is evaluated according to the method proposed in [2, 11]. An outline of the method is as follows:

   Step 1. Both eyes and the nose (strictly speaking, pupils and nostrils) are detected, and their locations are extracted as features of the face.

   Step 2. Referring to these features, the position and the size of the face are normalized, and as a result, a rectangular gray-scale image is generated.

   Step 3. The normalized faces are recognized by the constrained mutual subspace method. Note that each face is actually a sequence of faces of a same person obtained from multiple frames in a shot, which makes the method robust to changes in face direction and facial expressions. The similarity of the faces is defined as the angle between the subspaces corresponding to the two faces.

2. **Partial match of person names:**
   Since the process in Sect. 2.5 does not always associate correct names to a face, pattern matching is applied to compare the personal nouns; whether the first several characters of the names match or not [3].

[1] A Japanese morphological analysis system, JUMAN 3.61 [6] was used.

[2] In Japanese news shows, people are almost never mentioned without titles or other name-related suffices.

[3] Note that in the Japanese language, position and honorary titles are usually put at the end of the name as suffices. When applying the proposed method to other languages such as English, the pattern matching will have to be applied from the end of the name.

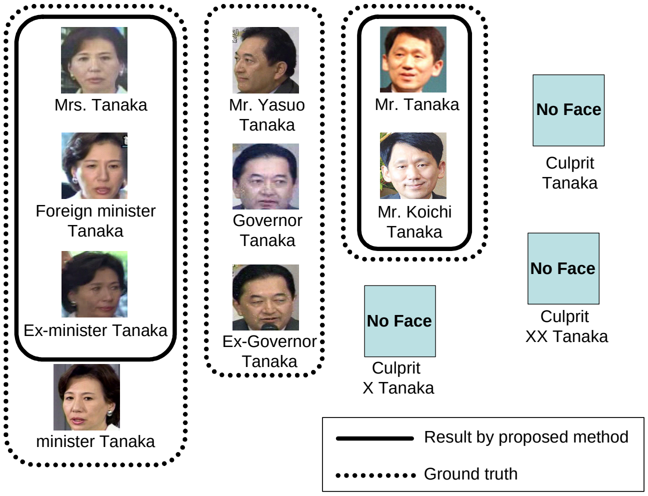

Figure 5: Sample of the identification. The actual names are in Japanese.

3. EXPERIMENT

The proposed method was applied to actual broadcast news video streams.

3.1 Conditions

The video data used in the experiment were 30 manually segmented news stories obtained from a Japanese daily news program "NHK News 7", with a total length of 120 minutes. The selected stories were mostly related to domestic politics, so that we could efficiently obtain many samples of a same person for the experiment. Anchor shots were excluded semi-automatically based on the color features of the first shot of a story, in order to avoid false association of names to the face of anchor-persons.

The parameters for face matching (threshold for the similarity in Condition 1. in Sect. 2.6) was set so that there should be no false positives; all the identified faces were correct. The ground truth was given manually.

3.2 Results

Figure 5 shows an example of the identified names. This is the result for people with 'Tanaka' as a family name [4]. Those with a label "No Face" are names who were not associated with a face. We can see that the proposed method managed to identify different names of a same person in some cases based on the face-name pairs.

The overall result is evaluated by the number of identified name groups, which resulted in 37% recall when the parameters were set so that precision should be 100%; as a result of identification, there were 27 groups, of which 10 were correct.

3.3 Discussion

While the recall in the experiment is not satisfactory, the most important fact is that, although the pattern matching of person names identified too many false names, they were mostly eliminated by face matching. As a result, the overall identification ability relied mostly on the face recognition ability.

In the experiment, failures to identify names were due to the following reasons:

Reason 1. Lack of the correct name candidate
The correct name did not appear in the same shot with the face, but in the previous or the next shot.

Reason 2. Lack of face
Face for some names never appeared in the video.

Reason 3. Failure of face detection / recognition
Poor visibility of a face, pupils, or nostrils caused these problems.

Reason 1 is expected to be solved in future works by expanding the range of the face-name association. Cases like Reason 2 cannot be spared by the proposed method. Reason 3 needs to wait for improvement in face detection / recognition methods. However, the proposed method should be able to compensate for these failures by applying more hours of video data which may include better cases.

4. CONCLUSION

In this paper, we proposed a method to identify names in broadcast news video by comparing faces together with names that appear with them. Future works include filtering of name candidates and more robust face matching.

5. ACKNOWLEDGMENTS

Most of the technologies used for face detection in Sect. 2.4 and for face recognition in Sect. 2.6 were provided by Toshiba Corporate Research and Development Center through a joint research project. Video data used in the experiments were provided from the NII broadcast video archive [5] through a joint research agreement. Parts of the work were funded by the Grants-In-Aid for Scientific Research and the 21st century COE program from the Ministry of Education, Culture, Sports, Science and Technology and the Japan Society for the Promotion of Science. The method was implemented partly using the MIST library[5].

6. REFERENCES

[1] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D. A. Forsyth. Names and faces in the news. In *Proc. 2004 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pages 848–854, June–July 2004.

[2] K. Fukui and O. Yamaguchi. Face recognition using multi-viewpoint patterns for robot vision. In *Proc. 11th Intl. Symposium of Robotics Research*, pages 192–201, October 2003.

[3] 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*, pages 43–47, November 1999.

[4] I. Ide, T. Kinoshita, H. Mo, N. Katayama, and S. Satoh. trackthem: Exploring a large-scale news video archive by tracking human relations. In *Information Retrieval Technology, 2nd Asia*

[4] 'Tanaka' is one of the most popular family names in Japan.

[5] `http://mist.suenaga.m.is.nagoya-u.ac.jp/`

Information Retrieval Symposium, Procs., Lecture Notes in Computer Science, Springer-Verlag, volume 3689, pages 510–515, October 2005.

[5] N. Katayama, H. Mo, I. Ide, and S. Satoh. Mining large-scale broadcast video archives towards inter-video structuring. In *Advances in Multimedia Information Processing, PCM2004, 5th Pacific Rim Conf. on Multimedia Procs. Part II, Lecture Notes in Computer Science, Springer-Verlag*, volume 3332, pages 489–496, December 2004.

[6] Kyoto Univ. *Japanese morphological analysis system JUMAN version 3.61.*, May 1999.

[7] T. Mita, T. Kaneko, and O. Hori. Joint Haar-like features for face detection. In *Proc. 10th IEEE Intl. Conf. on Computer Vision*, volume 2, pages 1619–1626, October 2005.

[8] T. Ogasawara, T. Takahashi, I. Ide, and H. Murase. Construction of a human correlation graph from broadcasted video (in Japanese). In *Proc. JSAI 19th Annual Convention*, pages 1–4, June 2005.

[9] C. P. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Proc. 5th IEEE Intl. Conf. on Computer Vision*, pages 555–562, January 1998.

[10] S. Satoh, Y. Nakamura, and T. Kanade. Name-It: Naming and detecting faces in news videos. *IEEE MultiMedia*, 6(1):22–35, January–March 1999.

[11] O. Yamaguchi and K. Fukui. “smartface” —a robust face recognition system under varying facial pose and expression. *IEICE Trans. Information and Systems*, E86-D(1):37–44, January 2003.