eo

†

u.ac.jp

64-8601, Japan

raduate

e,

Strong relationship

Weak relationship

2. S
As i
breviation
thesaurus could b

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.

- 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 $\mathsf{H}_1, \mathsf{H}_2$ of adjoining frames is given by calculating the histogram intersection, defined as:

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

I : Number of bins in a histogram

$H_{n,i}$: The $i-$th element of H_n

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].

[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.

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].

[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.

Mrs. Tanaka

Tanaka

minister Tanaka

Mr. Yasuo Tanaka

Governor Tanaka

Ex-Governor Tanaka

Mr. Tanaka

Mr. Koichi Tanaka

No Face

Culprit X Tanaka

No Face

Culprit Tanaka

No Face

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 PaciÞc 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.