International Workshop on Advanced Image Technology[IWAIT'98]

Indexing and Retrieval of Human Individuals on Video Data Using Face and Speaker Recognition

Y. Sugiyama, N. Ishikawa, M. Nishida and Y. Ariki

Department of Electronics and Informatics
Ryukoku University
sigma@arikilab.elec.ryukoku.ac.jp

Abstract: In this paper, we focus on the information retrieval of human individuals who are recorded on the video database. Our purpose is to index persons by their faces or voice and to retrieve their existing time sections on the video data. The database system can track as well as extract a face or voice of a certain person and construct a model of the individual person in self-organization mode. If he appears again at different time, the system can put the mark of the same person to the associated frames. In this way, the same person can be retrieved even if the system does not know his exact name. As the face and speaker modeling, a subspace method is employed to improve the indexing accuracy.

1 Introduction

We are getting much information from conventional media such as broadcasted TV video. They are now going to change into digital media by communication satellite and will be transmitted through an internet in near future. In this recent evolutional situation, the contents of the digital media such as TV video should be stored into the database and retrieved based on the users' interest.

In this paper, we focus on the information retrieval of human individuals who are recorded on the video database with their voice. If this retrieval is realized, following three kinds of applications will be achievable;

(1) Retrieval of the specified persons
If we specify a name of a certain person to the video database, the database system can retrieve all the frames where his face or his voice appears[1].

(2) Information retrieval of unknown persons
If we specify a face of unknown person on the video frame or specify the voice during the playback of the video data, the database system can recognize the person and retrieve the human information from the database[2].

(3) Classification of TV news article
The database system can locate the persons using their faces or voice which are already registered and can put their names to the video frames as the indices. According to these indices, the system can classify TV news articles into several topics.

The common function which lies under the above three applications is self-organization described as follows;

> *The database system can track as well as extract a face or voice of a certain person and construct a model of the individual person in self-organization mode. If he appears again at different time, the system can put the mark of the same person to the associated frames. In this way, the same person can be retrieved even if the system does not know his exact name.*

In this paper, we propose the self-organization method in face or voice extraction, tracking and modeling toward video content indexing[3]. We also describe our experiments on human individual retrieval from video database using face or voice recognition.

2 Problems to be Solved

2.1 Face indexing

In order to index the human individuals on the video data using face recognition, there are following three functions to be required;

(1) Locating human faces,

(2) Tracking human faces,

(3) Recognizing human faces.

These three functions have common difficulty that face orientations and sizes must be free.

To solve the face orientation problem, we propose a method to extract, track and recognize a facial region by projecting it onto the facial subspace. In the proposed method, two types of facial subspaces are constructed in advance. The first is constructed using facial images taken from 90 degrees right to 90 degrees left of the faces of 30 persons in order to cover horizontal orientation over 180 degrees. This facial subspace is used for extracting and tracking of facial regions regardless of the face orientation. The second type of the facial subspace is constructed for each person using the facial images with various orientations from the video data. These individual facial subspaces are used for face recognition regardless of the orientation.

2.2 Speaker indexing

In order to index the human individuals on the video data using speaker recognition, there are following two problems;

(1) How to construct the speaker models.

(2) How to deal with the overlapping of two speakers.

In the speaker indexing, we employ speaker subspaces as the speaker model constructed for each person in the same way as the face recognition. Namely the speaker subspace of the person who spoke first is constructed using his speech data. Then at every 0.5 second the input speech is judged if it belongs to the same person as speaking just before 0.5 second. If the input speech does not belong to the same person, then the system realizes that the different person appears. This method is based on the application of a speaker verification technique[4].

3 Subspace Method

Here we describe the subspace method which is commonly used in face extraction, tracking and recognition as well as speaker indexing. The advantage of the subspace method is that it can present the common features among all the persons as well as variational features so that it can recognize the object more precisely.

The first step of the subspace method is to find orthonormal bases $V_i = \{v_{i1}, \cdots, v_{ir}\}$ of training data belonging to category ω_i as shown in Fig.1 [5]. Here r is the number of dimensions of the subspace. This can be carried out by finding axes to which the total distance from the training data is minimized. It is well known that finding the orthonormal bases is equivalent with eigenvalue decomposition of the correlation matrix of the given training data.

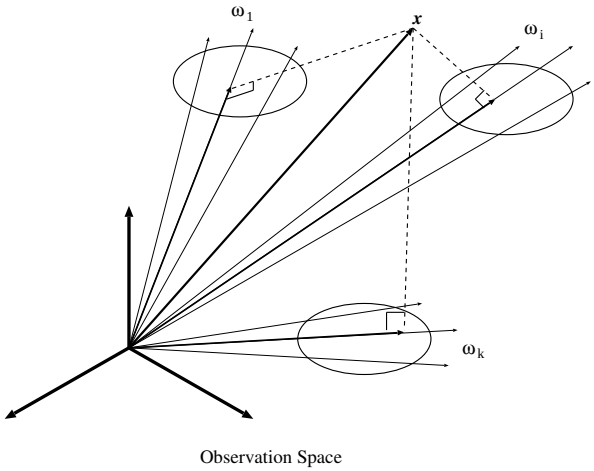

Figure 1: Subspaces in an observation space

The second step of the subspace method is to compute a distance from an input data x to the subspace V_i. The distance is presented as follows using a projection matrix P_i by which the input data x is projected to the subspace V_i:

$$Dist(V_i, x) = \| (I - P_i)x \| \tag{1}$$

The projection matrix P_i is defined as:

$$P_i = \sum_{k=1}^{r} v_{ik} v_{ik}^T = V_i V_i^T \tag{2}$$

where r is the number of the subspace dimension.

Once the projection matrices P_i are obtained for all the categories, the distance from given data x to all the subspaces can be computed. The input data is identified as belonging to category ω_i if the distance to the subspace V_i is the shortest among all the subspaces.

In practice, the squared length of the projected vector is computed as follows:

$$Project(V_i, x) = \| P_i x \|^2 = x^T P_i x \tag{3}$$

Instead of the distance, the input data x is identified as belonging to category ω_i with the maximum length of the projected vector.

The advantages of the subspace method is that it can reduce the dimension of feature space and then improve the processing time. The subspace method described above is called CLAFIC method.

4 Facial Database

We constructed a facial database to produce facial subspace as shown in Table.1, which includes facial images of 30 persons. Among them, 28 are male and 7 wear glasses. They were taken with an $8mm$ home

video camcorder, changing the face orientation by 15 degrees over 180 degrees from 90 degrees left to 90 degrees right. In total, 13 facial images were taken for each person. The person was sitting upright on a chair and the chair was rotated. All the images were taken against a homogeneous background. The size of each image is 290 × 325 pixels with 8-bit gray values.

Gray value f_{ij} of the facial images is normalized at the pixel (i, j) so as to make the mean μ of an image equal zero and the variance σ unity respectively.

Table 1: Database for facial subspace

Number of subjects	30
Male : female	28 : 2
Glasses : non-glasses	7 : 23
Maximum orientation	±90 degrees
Orientation difference	15 degrees
Gray level	256
Training images per subject	13

These facial data is used to construct one facial subspace to extract and track the facial regions as well as 30 individual facial subspaces to recognize each person.

5 Facial Region Extraction

In order to extract facial regions regardless of its orientation, one facial subspace with 180 degrees orientation range (-90 degrees to +90 degrees from the frontal position) is constructed using the facial database. The 13 facial data were collected from each person to train the facial subspace. In total, 390 facial data (the 13 images × 30 persons) were used to construct one facial subspace.

At first, facial regions are manually segmented and normalized into 30 × 30 pixel size. They are divided into 6 × 6 blocks and DCT (Discrete Cosine Transformation) is applied to each block. In each block, 4 kinds of DCT parameters are extracted as shown in Fig.2. In the figure, The DCT parameter 1 indicates the DC component. The DCT parameter 2 and 3 corresponds with horizontal and vertical AC components computed by averaging the absolute value of horizontal and vertical AC components within the numbered regions respectively. The DCT parameter 4 indicates the higher order AC component computed by averaging the absolute value of the higher order DCT components.

Therefore each facial image is presented as a 144 $(6 \times 6 \times 4)$ dimensional vector. These 144 dimensional vectors are called facial data which are free from slight changes caused by camera angle, lighting and facial expressions and are used to construct facial subspace for face extraction.

In face extraction, the size ratio of the window to search for and extract facial regions is fixed to 10:13 and the size is changed at 19 levels from 30 × 39 to 100×130. At each level and at each scanning position, the window image is divided into 6×6 blocks and 144 DCT features are extracted. The window image (144 dimensional vector) is projected to the facial subspace and projection amount is computed. The projection amount obtained from 19 size levels are summed at each position.

Fig.3 shows the bird-eye view of the projection amount summed over 19 size levels. In the figure, the bottom rectangle indicates the input image and the vertical axis to the input image indicates the projection amount. Some peaks can be seen in the figure as shown by the circles. The window images with the local peaks of the projected amount are regarded as the facial regions.

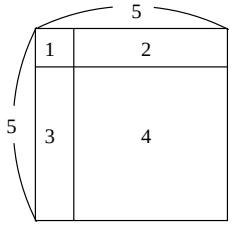

Figure 2: DCT parameters

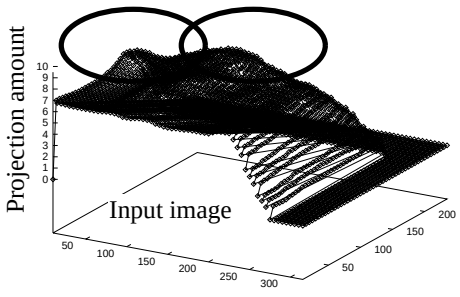

Figure 3: Three dimensional plot of the projection amount

6 Facial Region Tracking

6.1 Tracking by facial subspace

In this facial region tracking, the facial subspace is constructed by using mosaic images which are pro-

duced by dividing the facial image into 8 × 10 blocks and averaging the gray values within the block [6].

As shown in Fig.4, if there is already an extracted facial region (template) in the previous frame, the search area is set by extending the facial region on the successive frame, under the assumption that location change of facial region on input frame streams is small. The facial region extracted in the previous frame is then used as the template image to match with candidate region in the search area. As shown in Fig.4, the mosaic transformation of both the template image and the candidate region in the search area are projected onto the facial subspace.

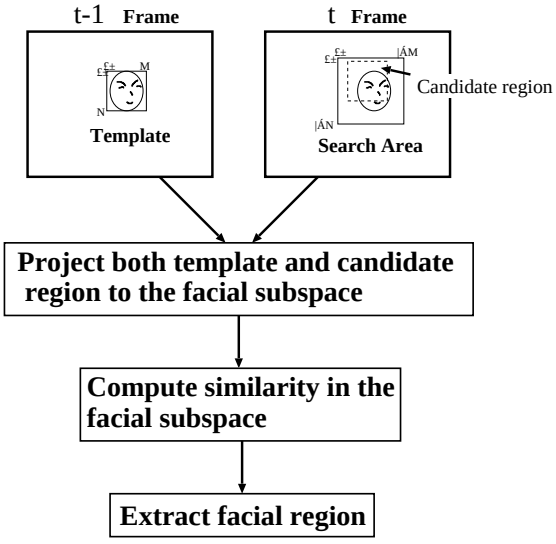

Figure 4: Facial region tracking by subspace projection

This situation is shown in Fig.5. The similarity is computed between the two vectors T_e and S_e in the facial subspace as the candidate region is shifted within the search area. The facial region is extracted as the most similar one with the projected template. If the candidate region locates on the true facial region, then the length of the projected vector is almost the same as the original one. But if not, the projected vector is changed to a small vector when projected onto the facial subspace. As a result, it is possible to compute the similarity reflecting the difference between the true facial region and the other region.

6.2 Tracking flow

Facial region tracking by subspace projection is summarized as follows;

(1) The candidate region is shifted in the search area on the frame t.

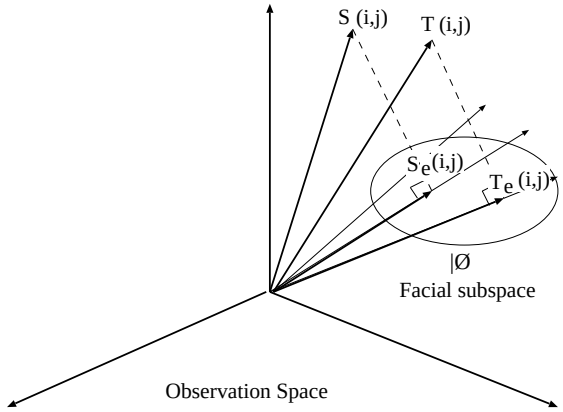

Figure 5: Subspace projection of the template and candidate regions

(2) As shown in Fig.5, the projection matrix P is multiplied by the mosaic transformed template $T(i,j)$ and mosaic transformed candidate region $S(i,j)$ respectively. The projection vector $T_e(i,j)$ and $S_e(i,j)$ are computed as shown in Eq.(4).

$$T_e(i,j) = P \cdot T(i,j), \qquad S_e(i,j) = P \cdot S(i,j) \quad (4)$$

(3) As shown in Eq.(5), similarity $se(x,y)$ is computed between the projection vector $T_e(i,j)$ and $S_e(i,j)$ on the facial subspace. Here, (x,y) indicates the location of the candidate region from the top left corner of the search area.

$$se(x,y) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} T_e(i,j) \cdot S_e(i+x, j+y)}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} T_e(i,j)^2}\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} S_e(i+x,j+y)^2}} \quad (5)$$

(4) The facial region is extracted as the candidate region with the highest similarity.

7 Facial Region Recognition

The face recognition process is summarized as follows based on CLAFIC method.

[Step1] Compute the correlation matrix S_i from the mosaic images $x_{ik}(1 \le k \le N)$ of the facial images taken over 180 degrees for the person i as follows:

$$S_i = \frac{1}{N}\sum_{k=1}^{N} x_{ik} x_{ik}^T \quad (6)$$

[Step2] Decompose the correlation matrix S_i through eigenvalue decomposition as follows:

$$S_i = U_i \Sigma U_i^T \quad (7)$$

where $U_i = \{v_{i1}, \cdots, v_{in}\}$ is the orthogonal matrix and v_{ik} is an eigenvector.

[Step3] Construct the subspace $V_i = \{v_{i1}, \cdots, v_{ir}\}$ by deciding the subspace dimension r.

[Step4] Compute the projection matrix P_i from V_i as follows:

$$P_i = V_i V_i^T = \sum_{k=1}^{r} v_{ik} v_{ik}^T \quad (8)$$

[Step5] Identify the input image x as belonging to the person i by using the following rule:

For all j $(j \neq i)$, if

$$Dist(V_i, x) \leq Dist(V_j, x) \quad (9)$$

where $Dist(V_i, x) = \| (I - P_i)x \|$

8 Experimental Result

We have taken three scenes in which persons are walking in the complex background as a video data. They include variations in human face location, horizontal orientation, glasses and emotional expression. Their sequences consist of 250 frames, and they were captured at the speed of 30 frames per second. Three scenes are as follows:

[Data1] Mr.A is walking and shaking his head.

[Data2] Mr.B is walking in a room scratching his head with his hands.

[Data3] Mr.C is walking in a room changing his expression with his hands.

At first, face extraction was executed in the first frame of each video data. Next, face tracking was executed using the already extracted facial region as a template. Finally, face recognition was executed to the tracking results at every 10 frames. The dimension of the facial subspace in extraction, tracking and recognition were set to 10, 10 and 4 respective.

The results are shown in Table2. In the table, extraction rate, tacking rate and recognition rate are listed. The extraction and tracking rate of the facial region were evaluated by the overlap rate which presents how well the extracted facial region is lapped over the true facial region. For that purpose, the true regions were given manually for evaluation in advance. The recognition rate was evaluated by the ratio of the number of frames where faces were correctly recognized to the number of all frames. For that purpose, the facial region tracking and recognition were performed at every 10 frames.

Table 2: Experimental results (%)

Data	Extraction rate	Tracking rate	Recognition rate
Data1	77.5	53.1	100
Data2	87.4	87.6	100
Data3	69.0	41.5	73.0

9 Speaker Indexing

9.1 Speaker verification

Speaker verification is a technique to judge if the input speech belongs to the specified person or not[4]. Fig.6 shows the speaker verification process. When the speaker ID of speaker A and his speech are fed to the verification system, the similarity is computed between the model of the speaker A and the input. If the similarity is greater than some threshold, the speaker is accepted as the true speaker A. Otherwise the speaker is rejected. In our experiment, speaker subspace is constructed as the speaker model and the projection length of the input speech to the speaker subspace is used as the similarity.

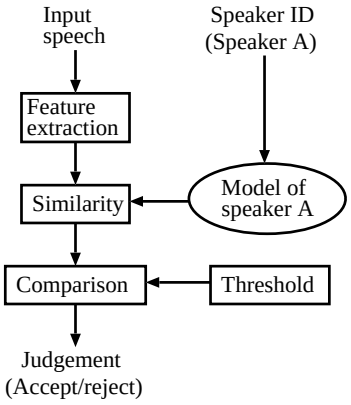

Figure 6: Speaker verification

9.2 Speaker indexing flow

The process flow of the speaker indexing is as follows;

(1) Averaged power is computed at every 0.5 second on the input speech. If it is lower than some threshold it is regarded as silence. The speech section is extracted between two silences.

(2) On the extracted speech section, a speaker subspace is constructed using 5 seconds speech. This speaker subspace corresponds to the model of the speaker A shown in Fig.6.

(3) On the successive speech, at every 0.5 second, the similarity is computed between the input speech and the model. If the similarity is lower than some threshold (namely speaker is rejected) more than three times (1.5 seconds), it is judged that the speaker A finishes his speech and new speaker, say, B is speaking. Otherwise it is judged that speaker A is still speaking. The three times judgement plays a role of reducing the mistakes.

(4) If the speaker is judged as the speaker A, the subspace of the speaker A is retrained including the newly obtained speech. Otherwise the speaker subspace is newly constructed for the speaker B.

The threshold θ to accept or reject the speaker is set as follows in advance using μ (mean) and σ (standard deviation);

$$\theta = \mu + \frac{\sigma}{2} \tag{10}$$

10 Experimental Result

10.1 Experimental condition

TV video data in which five persons are talking for 7 minutes was used for speaker indexing. The dimension of speaker subspace was set to 5 after preliminary experiment. The evaluation of the speaker indexing was carried out by verification rate, recall rate and precision rate which are defined as follows;

$$\text{Verification rate} = \frac{\text{Number of the correct verification}}{\text{Number of verification at 0.5 second}} \tag{11}$$

$$\text{recall rate} = \frac{\text{Number of correctly verified boundaries}}{\text{Number of speaker boundaried}} \tag{12}$$

$$\text{Precision rate} = \frac{\text{Number of correctly verified boundaries}}{\text{Number of extracted speaker boundaries}} \tag{13}$$

10.2 Experimental result

The speaker indexing result is shown in Table3. The verification rate is 95.5% high. However, the recall rate and precision rate is 75.0% and 60.0% respectively. The reason is that 4.5% error of the verification rate causes the wrong extraction of the speaker boundaries so that the recall and precision rate are not so high at present.

Table 3: Experimental result

	Number of verification	%
Verification rate	640 / 670	95.5
	Number of boundaries	%
Recall rate	6 / 8	75.0
Precision rate	6 / 10	60.0

11 Conclusion

In this paper, we proposed face indexing and speaker indexing for video data using face recognition and speaker verification techniques. In the face indexing, facial region extraction, tracking and recognition by the subspace method were proposed. In the speaker indexing, the speaker verification technique is extended to automatic model construction. Further work is planned in applying this face and spaker indexing to the TV news articles for the classification and human information retrieval.

References

[1] Y.Ariki, N.Ishikawa:"Integration of Face and Speaker Recognition by Subspace Method", ICPR'96, 1996.

[2] Y.Ariki, N.Ishikawa and Y.Sugiyama, "Human Information Retrieval by Face Extraction and Recognition on TV News Images by Subspace Method", ACCV97, to appear, 1998.

[3] Y.Sugiyama and Y.Ariki: "Facial region tracking and Recognition by Subspace Projection", VSMM96, pp.225-230, 1996.

[4] T.Matsui and S.Furui: "Comparison of text independent speaker recognition methods using VQ distortion and discrete/continuous HMMs", Proc.ICASSP, Vol.II, pp157-160, 1992.

[5] E.Oja:"Subspace Methods of Pattern Recognition", Research Studies Press, England, 1983.

[6] Makoto Kosugi:"Human-Face Search and Location in a Scene by Multi-Pyramid Architecture for Personal Identification", Trans. of IEICE, D-II, Vol.J77-D-II, No.4, 1994.