*Special issue: 3D image and video technology*

Research Paper

# Fast face clustering based on shot similarity for browsing video

Koji YAMAMOTO[1], Osamu YAMAGUCHI[2], and Hisashi AOKI[3]
[1,2,3]*Corporate Research and Development Center, Toshiba Corporation*

**ABSTRACT**

**In this paper, we propose a new approach for clustering faces of characters in a recorded television title. The clustering results are used to catalog video clips based on subjects' faces for quick scene access. The main goal is to obtain a result for cataloging in tolerable waiting time after the recording, which is less than 3 minutes per hour of video clips. Although conventional face recognition-based clustering methods can obtain good results, they require considerable processing time. To enable high-speed processing, we use similarities of shots where the characters appear to estimate corresponding faces instead of calculating distance between each facial feature. Two similar shot-based clustering (SSC) methods are proposed. The first method only uses SSC and the second method uses face thumbnail clustering (FTC) as well. The experiment shows that the average processing time per hour of video clips was 350 ms and 31 seconds for SSC and SSC+FTC, respectively, despite the decrease in the average number of different person's faces in a catalog being 6.0% and 0.9% compared to face recognition-based clustering.**

**KEYWORDS**

Video indexing, face clustering, similar shots, video clip cataloging

## 1 Introduction

Face detection enriches the user experience on entertainment PCs with television recording features. By cataloging video clips based on subjects' faces (Fig. 1), favorite scenes can be found without searching through hours of video content. In this paper, we propose a fast face clustering method to classify faces in a television title using similar shot information. Since faces are detected frame-by-frame during the recording and the same person appears in many different shots, each face data needs to be classified according to whom it belongs to. Otherwise, the same person's face will appear in the catalogue redundantly. Conventional face recognition-based clustering methods such as [1] can obtain good results for this purpose. They require, however, considerable processing time because of the large amount of calculation, leading to a long waiting time before browsing becomes available after recording. From our preliminary survey, average tolerable waiting time is 2.8 minutes per hour of recorded video clips. Moreover, about 20% of the users require it to be less than 1 minute. However, for cataloging a video clip for browsing, the accuracy of a face recognition-based method is not necessarily required. If the redundancy in the catalogue does not significantly differ, users cannot notice the difference of clustering accuracy. Therefore, processing speed is a more important issue than accuracy in our method. The main contribution of our method is that we use similarities of shots where the characters appear and the relative positions of their faces to estimate corresponding faces instead of calculating distance between each facial feature. This enables high-speed processing and, with the fastest method, a face catalog can be created as soon as the recording phase is over.

Received October 6, 2009; Revised December 14, 2009; Accepted January 5, 2010.
[1)] *koji7.yamamoto@toshiba.co.jp*, [2)] *osamu1.yamaguchi@toshiba.co.jp*, [3)] *hisashi.aoki@toshiba.co.jp*,
DOI: 10.2201/NiiPi.2010.7.7

Fig. 1 Face-based video clip cataloging application. Faces in the upper part are the subjects' faces in a video clip. Each column shows the major faces found in a short segment of the clip. The faces are aligned in time order from left to right to express the whole clip. Using these faces, it is possible to overview the whole content and pinpoint favorite scenes. Also shown are several types of information obtained with other video indexing technologies (not discussed in this paper).

## 2 Related works

Face clustering is used in video indexing, photo management, and many other fields and various applications are proposed. For example, in video indexing domain, it is used to classify the people in a news video [2], and annotate their names using closed captions [3]. It is also used to classify the characters in a drama [4] or to list up the major characters in a video [5], [6]. In photo management domain, it is used to classify and manage photos taken by a digital camera, and annotate their names [7]–[9].

Face clustering is based on face recognition or individual identification, and they have been tackled for several decades. Eigenface method uses the Karhunen-Loeve Transform (KLT) to present facial data into a low dimensional feature space for recognition [10]. Subspace method used in [2] presents facial data of individual person to different feature subspace. In [2], individual face is recognized by comparing between their feature data and the ones on a database. In [4], the face database is unnecessary because each face sequence is compared with other face sequences. Image feature based methods like eigenface tend to be sensitive against change of facial pose or expression. Therefore, like in television titles, same person's faces do not exist in narrow range in the feature space [12]. In [6], subspace is constructed not from whole faces of a person but from face sequences detected from successive frames. It clusters face sequences using a distance function that is invariant to affine transformations [5] to make it robust against transforms. In [11], [12], face sequences are divided into different facial poses before clustering. These methods are based on image features, but some methods are based on different features. In [1], facial feature points like eyes and nose are detected and normalized to make it robust against various poses and expressions. Some methods [13], [14] use SIFT [15] features which are robust against transforms. There are some other clustering or recognition method proposed using Hidden Markov Model (HMM) [16], using SVM [17] to classification between subspaces [18], and using mutual information [19].

In this paper, we deal with television titles. Therefore, we need a clustering method robust against changes of facial pose and expression. Meanwhile, we need fast processing to avoid keeping users waiting after the recording is finished. There are few, however, methods that focus on processing speed. Especially, clustering method which is robust against changes of facial pose and expression needs a normalization phase, and this leads to long processing time. This is because most of the previous works needs high accuracy since they are used for individual identification or detailed annotation.

## 3 Face clustering using similar shots

In this paper, we propose two fast face clustering methods to catalogue a television title. They are similar shot-based clustering (SSC) methods. The first method only uses similar shots, whereas the second method uses face thumbnail clustering (FTC) as well. In the following, the two methods are called SSC and SSC+FTC.

We define the term similar shots as shots with a similar image feature. In a television title, shots with the same picture composition and camera angle appear many times and these become similar shots (Fig. 2). As shown in Fig. 3, our clustering method estimates faces are the same person's when they have similar positions and sizes in similar shots. This is because, in a television title, there is a high probability that when a composition and a camera angle are the same, the characters are also the same.

We detect similar shots by the method described in [20] as follows: 1) a feature consisting of a color histogram of the screen image and a luminance layout pattern is calculated for frames, respectively. If neighboring frames have dissimilar features, the video clip is segmented into shots by a cut point. 2) When the temporally separated shots have similar features they are considered to be similar shots. Since the similar shot detection runs during recording, its processing time is not counted as part of the face clustering.

Fig. 4 shows the regions used to extract the feature and Table 1 shows the specification. All images are stored in 16:9 aspect ratio image buffer and the features are extracted from the whole buffer except the region

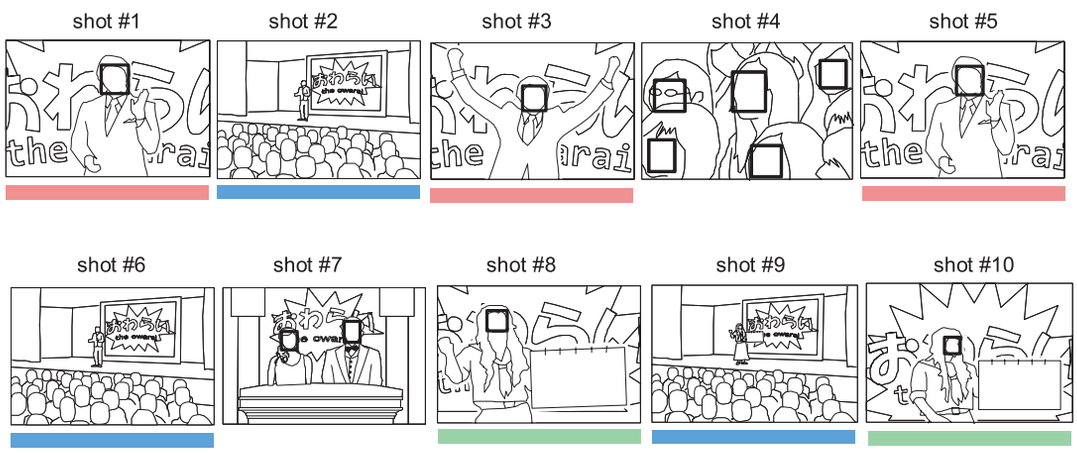



Fig. 2 Example of similar shots. Each thumbnail shows a representative thumbnail of a shot in a TV program. The thumbnails with the same color bars shows that they are similar shots.

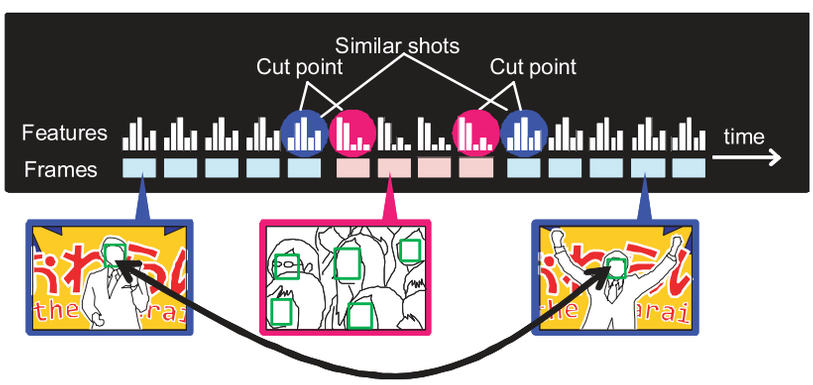



Fig. 3 Estimation of corresponding faces. Faces with similar position and size in similar shots are estimated as same person's face.

near the border. This is because today's TV programs are produced in both 4:3 and 16:9 aspect ratios. Therefore, using only the 4:3 part is more robust than using the whole image. Even if the program is aired in 16:9 aspect ratio, both sides might have irrelevant data.

Color histogram is a 32-bin histogram calculated from hue data. Let $h_i(k)$ be the value in the $k$th bin of the histogram; then its similarity between frame $i$ and $j$ is given as

$$Sim_{color}(i,j) = \sum_{k \in bins} \left(h_i(k) - h_j(k)\right)^2.$$

Luminance layout pattern is a small pattern in 10 x 6 pixels (*i.e.* 60 dimensions), and each pixel is an average of the luminance values of a block. Let $b_i(l)$ be the value of the $l$th block and $b_{th}$ be a threshold; then its similarity between frame $i$ and $j$ is given as

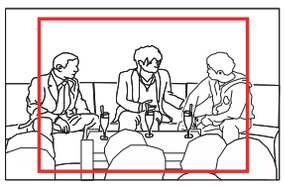
$$Sim_{luminance}(i,j) = \sum_{l \in blocks} B_{i,j}(l),$$

where

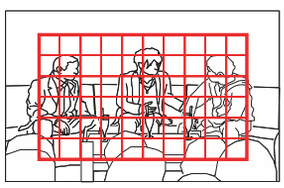
$$B_{i,j}(l) = \begin{cases} 0 & \text{if } |b_i(l) - b_j(l)| > b_{th} \\ 1 & \text{otherwise} \end{cases}.$$

We determine frame $i$ and $j$ are similar when color and luminance similarities are greater than given thresholds $C_{th}$ and $L_{th}$, respectively.

As classification only requires easy calculation of coordinates, we can process in a short time. Meanwhile, we cannot classify some faces correctly because of failures in detecting similar shots. For example, we cannot detect the shots shown in Fig. 5 as similar shots because of the difference of scale even though they have the same compositions and camera angles. In this case, there is a problem in that similar face thumbnails, with



(a) Region used to extract color histogram



(b) Region used to extract luminance layout pattern

Fig. 4 Feature extraction to find similar shots.

Table 1 Input data for similar shot detection

| | |
|---|---|
| Image size | 192×108 |
| Frame rate | 2fps |
| Region size for color histogram | 140×103 |
| Region size for luminance (Block size) | 140×84 (14×14) |

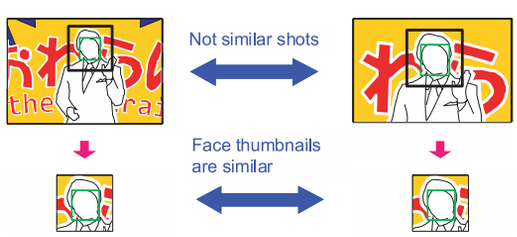



Fig. 5 Case of failure while detecting the same person's faces.

the same person with the same background image, are redundantly shown in a catalog. Redundant thumbnails of the same person are more significant if they have the same background than if they have different background as shown in Fig. 6. The second method, SSC+FTC, is to deal with this problem. It merges similar face thumbnails into one group using image features of the thumbnails.

## 4 System overview

Fig. 7 shows the overall diagram of our video indexing system. It consists of two phases: the first phase runs during the recording and the second runs after the recording. In the first phase, similar shot detection, face detection, and thumbnail extraction of detected faces are performed. The face detector we used was [21].

Fig. 6 Example of redundant face thumbnails. It is more significant if face thumbnails have similar background and picture composition such as those on the right side.

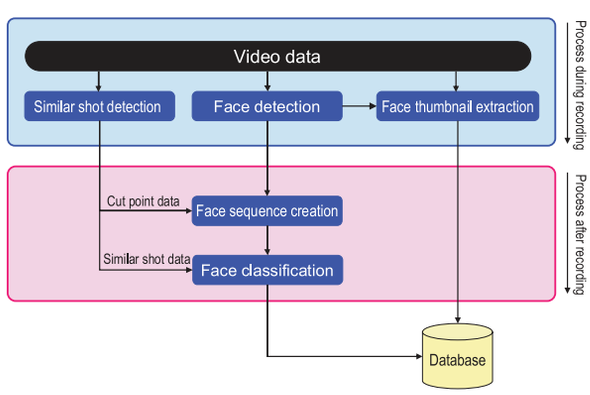



Fig. 7 System diagram.

Table 2 Input data for face detection.

| | |
|---|---|
| Image size | 768×432 |
| Frame rate | 10 fps |
| Face thumbnail size | 96×96 |

Since the first phase is a real-time phase, its processing time does not affect the waiting time and its processing time is not counted as part of the face clustering. In the second phase, which is the core part of our face clustering, creation of face sequence and classification are performed. As mentioned above, we deal with two methods for face clustering. SSC uses only similar shots information and coordinates of face regions to estimate corresponding faces, whereas SSC+FTC employs a further classification based on similarity of face thumbnails to solve the problem caused by the difference of scale. Table 2 shows the specification of the data used for face detection.

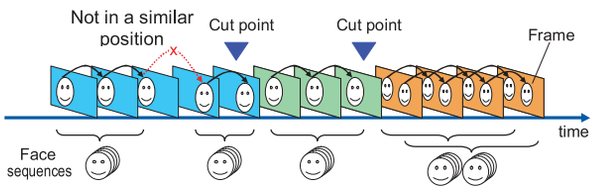



Fig. 8 Grouping faces into sequences. Consecutive frames of the same color are the same video shot.

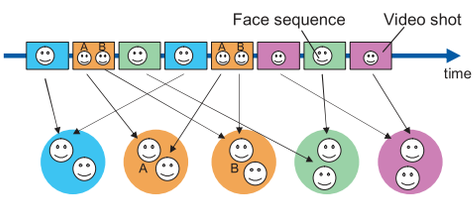
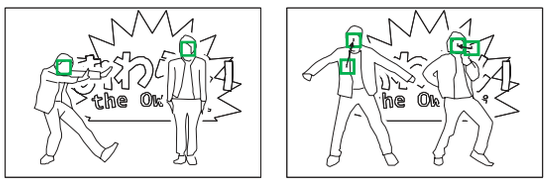
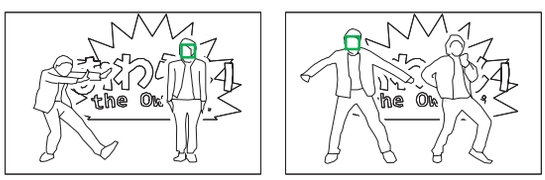



Fig. 9 Classifying face sequences with similar shots. Video shots of the same color are similar shots.



(a) Face sequences are grouped with the closest ones



(b) Even if some faces are not detected, miss of correspondence will not occur if the distance is above

Fig. 10 Correspondence of face sequences between similar shots. All of the images show one of the frame images in the shots.

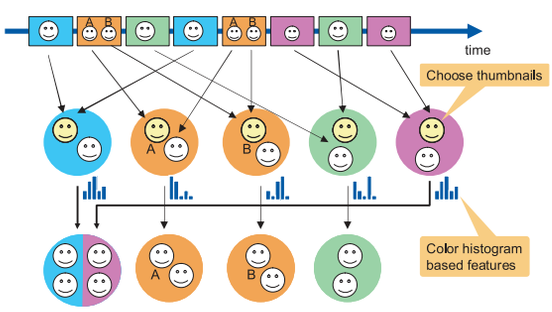



Fig. 11 Further classification with face thumbnails.

### 4.1 Similar shot-based clustering (SSC)

First, face regions that have similar positions and sizes in consecutive frames are grouped as face sequences (Fig. 8). At cut points, grouping is terminated and a new sequence is started. Likewise, when more than one person appears on the screen, they are separated as different sequences. In order to determine whether adjacent face regions have similar positions and sizes, we use area ratios between the overlapping region and respective face regions. Let $S_m^{face}$ and $S_n^{face}$ be the size of the two face regions and $S_{mn}^{overlap}$ be the size of the overlapping region of the two face regions. When the two area ratios $R_{mn}^m = S_{mn}^{overlap}/S_m^{face}$ and $R_{mn}^n = S_{mn}^{overlap}/S_n^{face}$ are above the threshold $R_{th}$ (*i.e.* $R_{mn}^m > R_{th} \wedge R_{mn}^n > R_{th}$), the face regions are judged to have similar positions and sizes.

Next, face sequences in similar shots that have similar positions are classified as the same person (Fig. 9). A distance $D_{shot}(FS_i, FS_j)$ between two face sequences $FS_i$ and $FS_j$ is given by a Euclidean distance between the centroids of one of the faces in each sequence. As shown in Fig. 10, a face sequence $FS_i$ is classified together with $FS_i^{min}$ that gives the shortest distance:

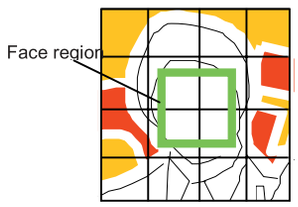
$$FS_i^{min} = \arg\min_{FS_j} D_{shot}(FS_i, FS_j)$$

unless the distance $D_{shot}(FS_i, FS_i^{min})$ is above the limit $D_{th}$. We used $D_{th} = 100$ in the experiments.

### 4.2 Classification with face thumbnails (SSC+FTC)

For each cluster obtained in section 4.1, a color histogram-based feature of a representative thumbnail is calculated, and clusters with similar features are merged (Fig. 11). The clustering algorithm used is Mean-Shift with the distance described in the following paragraph. Since face thumbnails are retrieved from the database, it is unnecessary to redecode the original video clip. Some calculation, however, is still required for feature extraction, which makes SSC+FTC slower than SSC.

A face thumbnail is extracted as a cropped image from a video frame with a face region in the center. The ratio between the face region and the thumbnail is 1/3 for both vertical and horizontal directions. Since a face region is an output of the face detector, it only covers the strict face part, *i.e.*, from the forehead to the chin, and not the whole head of a subject. A feature of a face thumbnail is extracted as a collection of color histograms. As shown in Fig. 12 (a) $96 \times 96$ pixel face thumbnail is divided into 16 blocks (4 in the row and 4 in the column), and a color histogram is calculated




(a) A face thumbnail is divided into 4 × 4 blocks

| | | | |
|---|---|---|---|
| L | L | L | L |
| L | H | H | L |
| L | H | H | L |
| H | H | H | H |

(b) Weights for the blocks (H: High, L: Low)

Fig. 12 Feature extraction of face thumbnails.

for each block in RGB color space. A distance between two thumbnails is given as a weighted sum of distances for each block. Let $FT_a$ and $FT_b$ be the thumbnails to compare, $w_k$ be the weight of the $k$th block, and $H_{a,k}(i)$ be the value in the $i$th bin of the histogram at the block; then the distance is given as

$$D_{thumbnail}(FT_a, FT_b) = \sum_{k \in blocks} w_k d_k(a, b),$$

where

$$d_k(a, b) = \sum_{i \in bins} |H_{a,k}(i) - H_{b,k}(i)|.$$

In order to make the distance less sensitive to the change of background, the weights are set high for the face region blocks and low for the background region blocks as shown in Fig. 12 (b). The values used in the following experiments are 1.5 and 0.5.

## 5 Experimental results

The first experiment was conducted to evaluate the accuracy for cataloging video clips and to compare the processing time. We used eight television titles taken from various genres. After running face clustering, clips are cataloged in the following steps: 1) each clip is segmented into groups of 5-minute clips. 2) 7 major face clusters obtained from each segment are chosen according to the number of elements. 3) The first face in each cluster is chosen as a representative thumbnail. The number of the cluster chosen in the second step is empirically determined according to the screen size and the face thumbnail size. In most cases, placing 5-10 face thumbnails in each column is suitable for a typical PC screen, and we chose 7, which is near to the average. To evaluate the accuracy, we counted the number of different person's face, same person's face with a similar background, and same person's face with a different background in the obtained catalogue. If the number of the same person's face is smaller and the number of different person's face is larger, it means there was less redundancy. Moreover, as mentioned, thumbnails of the same person with similar background are more significant errors than thumbnails with different background.

We compared SSC, SSC+FTC, and conventional face recognition-based clustering (FRC). For FRC, we used [1]. This approach extracts facial feature points first, then recognizes individuals using the mutual subspace method. Since it takes temporal sequence as an input data, it is robust against variations in facial pose and expression that are common in television titles. Its correct identification rate is 99.0% for 101 individual face data when the dimension of the subspace is 10. It is implemented using SIMD (Single Instruction Multiple Data) instruction and has adequate speed as FRC.

Fig. 13 shows the average number of faces in each segment. The blue portion shows the average number of different person's face, the red shows the same person's face with similar background, and the yellow shows the same person's face with different background. For more than half of the titles, FRC obtained the largest number of different people. The difference, however, between SSC and FRC was less than one face per segment. Moreover, the performance of SSC+FTC was close to that of FRC. The rate of decrease in overall average number of different faces among the tested 8 titles was 6.0% and 0.9% for SSC and SSC+FTC, respectively. In particular, the number of the same person with similar background was smallest with SSC+FTC in some titles. These results indicate that SSC+FTC is more robust than FRC for the titles that have drastic change in facial expressions, such as variety (stage) or drama, or titles for which extraction of facial features fails, such as swimming, because of goggles. In contrast, SSC+FTC is not robust against close-up shots with out-of-focus background taken from long distance, which are often seen in sports such as soccer. This is because the background changes drastically when the subject moves. SSC+FTC also fails in the case when a thumbnail has a complex texture in the background or the cropping area changes owing to oscillation of the face region from the face detector.

Fig. 14 shows the processing times of the three methods and Fig. 15 shows distribution of the processing time and number of different faces. As mentioned in section 4, processing times do not include the process that ran during the recording phase, such as face detection or similar shot detection. Since processing time depends on the duration of a video clip, we normalize it to processing time per hour of video clips. Note that horizontal axes are in logarithmic scale in these figures. There was no significant difference in processing times between the titles. The average times were 350ms for SSC, 31 seconds for SSC+FTC, and 10 minutes for FRC. As mentioned in section 1, the av-

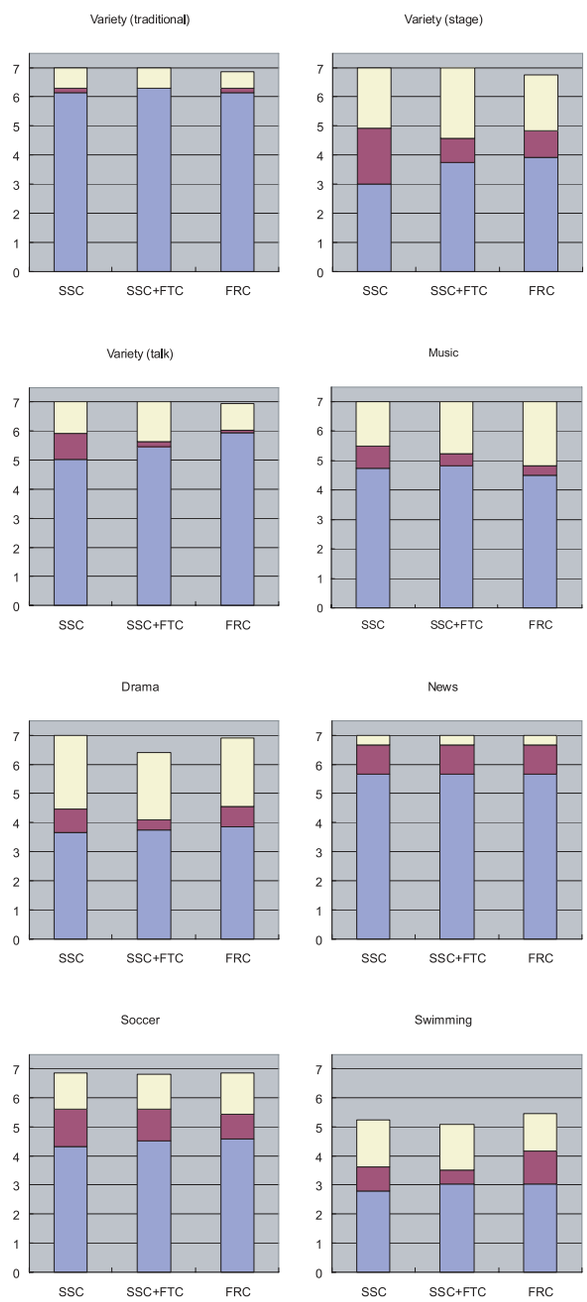



Fig. 13 Average number of people's faces in each segment (Blue: different people, Red: same people with similar background, Yellow: same people with different background).

erage tolerable waiting time is 2.8 minutes according to our survey, a condition satisfied by both SSC and SSC+FTC. Moreover, SSC satisfied the condition "less than 1 minute" for all titles for the users least inclined

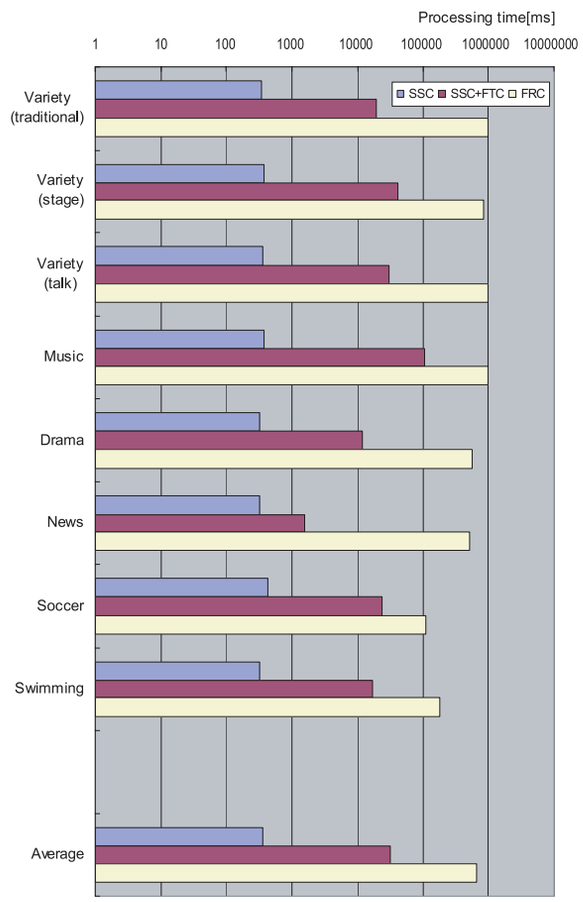



Fig. 14 Processing time per hour of video clips (Horizontal axis is in logarithmic scale).

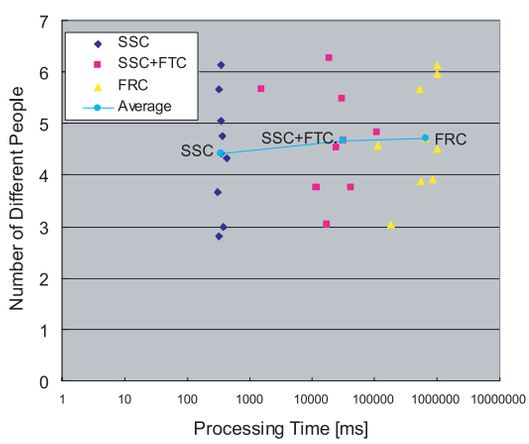



Fig. 15 Distribution of the processing time and number of different faces.

Table 3 Notation for the contingency table for comparing two partitions.

| $Class \backslash Cluster$ | $v_1$ | $v_2$ | $\cdots$ | $v_C$ | $Sums$ |
|---|---|---|---|---|---|
| $u_1$ | $n_{11}$ | $n_{12}$ | | $n_{1C}$ | $n_{1.}$ |
| $u_2$ | $n_{21}$ | $n_{22}$ | | $n_{2C}$ | $n_{2.}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | $\vdots$ |
| $u_R$ | $n_{R1}$ | $n_{R2}$ | $\cdots$ | $n_{RC}$ | $n_{R.}$ |
| $Sums$ | $n_{.1}$ | $n_{.2}$ | $\cdots$ | $n_{.C}$ | $n_{..} = n$ |

to wait. SSC+FTC exceeded this condition in the worst case, but satisfied it in most cases. FRC exceeded the average tolerable waiting time in most cases. Compared to FRC, SSC was more than 1000 times faster, and SSC+FTC was 20 times faster.

The second experiment was conducted to investigate the accuracy of clustering. To that end, we used the Adjusted Rand Index (ARI) [22], [23] to evaluate the similarity between a clustering result and ground truth (GT). ARI is an index that expresses a similarity between two groups of clusters in the 0 to 1 range (the larger, the better).

We briefly describe the calculation of ARI. Given a set of $n$ objects $S = \{O_1, \ldots, O_n\}$, suppose $U = \{u_1, \ldots, u_R\}$ and $V = \{v_1, \ldots, v_C\}$ represent two different partitions of the objects in $S$. Suppose that $U$ is our external criterion (GT) and $V$ is a clustering result. Let $n_{ij}$ be the number of objects in both class $u_i$ and cluster $v_j$. Let $n_{i.}$ and $n_{.j}$ be the number of objects in class $u_i$ and cluster $v_j$, respectively. The notations are shown in Table 3. Then ARI is given by the following equation:

$$ARI = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right] \Big/ \binom{n}{2}}$$

Fig. 16 shows the ARI obtained from each clustering result. Four titles are chosen from the ones used in the previous experiment according to the magnitude of motion. For all titles, FRC showed the highest accuracy.

Both variety (traditional) and variety (stage) are recorded in a studio. Characters in variety (stage) are more active and move about the stage. Positions of the characters switched in some cases when more than one person was on the stage. Variety (talk) is a complex of studio scenes and sports scenes recorded out of the studio. There are few similar shots out of the studio. In drama, there are no similar shots except in dialog scenes and the characters' facial poses and expressions change greatly. The result shows that differences between the three methods become larger when the magnitude of the activity increases and the number of similar shots decreases. The difference, however, does not greatly affect the performance of the cataloging as shown in the previous experiment.

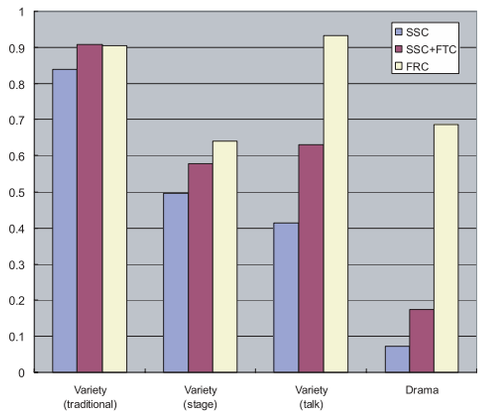



Fig. 16 Adjusted Rand Index of obtained face clusters.

## 6 Conclusions

In this paper, we proposed two face clustering methods based on similar shots that can catalogue a television title in a short time without handling facial features. The first method, SSC, uses similar shots and the second method, SSC+FTC, uses face thumbnail clustering as well. The experiment shows that the average processing time per hour of video clips was 350 ms for SSC and 31 seconds for SSC+FTC. This processing time is short enough to satisfy the average tolerable waiting time, 2.8 minutes, despite the decrease in the average number of different person's faces being 6.0% and 0.9% compared to face r ecognition-based clustering. Moreover, SSC+FTC showed better performance than face recognition-based clustering in titles with great changes of facial pose or expression or titles for which facial feature extraction was difficult. Since processing speed is the top priority of our method and accuracy remains at a high level for browsing, these results show the effectiveness of our method. Which of SSC and SSC+FTC is better depends on user preference, system configurations, or applications. If the priority is higher processing speed, SSC will be suitable, and if it is higher accuracy, SSC+FTC will be suitable. In future work, we intend to optimize FTC for speed so that it is suitable in all situations.

## References

[1] O. Yamaguchi, and K. Fukui, "“Smartface” - A robust face recognition system under varying facial pose and expression," *IEICE Trans. Inf. & Syst.*, vol.E86-D, no.1, pp.37–44, Jan. 2003.

[2] Y. Ariki, Y. Sugiyama, N. Ishikawa, "Face indexing on video data-extraction, recognition, tracking and modeling," *In Proceedings of Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp.62–69, 1998.

[3] S. Satoh, Y. Nakamura, and T. Kanade, "Name-It: Naming and detecting faces in news videos," *IEEE MultiMedia*, vol.6, no.1, pp.22–35, 1999.

[4] S. Satoh, "Comparative evaluation of face sequence matching for content-based video access," *In Proceedings of Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pp.163–168, 2000.

[5] A. W. Fitzgibbon and A. Zisserman. "On affine invariant clustering and automatic cast listing in movies". *European Conference on Computer Vision* (ECCV), vol.3, pp.304–320. Springer-Verlag, 2002.

[6] A. W. Fitzgibbon and A. Zisserman. "Joint Manifold Distance: a new approach to appearance based clustering," *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR '03*), vol.1, pp.26–36, 2003.

[7] J. Cui, F. Wen, R. Xaio, Y. Tian, and X. Tang, "Easyalbum: An interactive photo annotation system based on face clustering and re-ranking," *Proceedings of the SIGCHI conference on Human factors in computing systems* (*CHI '07*), pp.367–376, 2007.

[8] L. Zhang, Y. Hu, M. Li, W. Ma, and H. Zhang, "Efficient propagation for face annotation in family albums," *In Proceedings of ACM Multimedia*, pp.716–723, 2004.

[9] E. Ardizzone, M. La Cascia, F. Vella, "Mean shift clustering for personal photo album organization," *In Proceedings of 15th IEEE International Conference on Image Processing* (*ICIP 2008*), pp.85–88, 2008.

[10] M. Turk and A. Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, vol.3, no.1, pp.71–86, 1991.

[11] P. Huang, Y. Wang, and M. Shao, "A New Method for Multi-view Face Clustering in Video Sequence," *Proceedings of the 2008 IEEE International Conference on Data Mining Workshops* (*ICDMW*), pp.869–873, 2008.

[12] J. Tao and Y. P. Tan, "Face Clustering in Videos Using Constraint Propagation," *IEEE International Symposium on Circuits and Systems* (*ISCAS*), Seattle, WA, pp.3246–3249, 2008.

[13] A. Asthana, R. Goecke, N. Quadrianto, and T. Gedeon, "Learning Based AutomaticFace Annotation for Arbitrary Poses and Expressions from Frontal Images Only," *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (CVPR 2009), pp.1635–1642, June 2009.

[14] P. Antonopoulos, N. Nikolaidis, and I. Pitas, "Hierarchical Face Clustering using SIFT Image Features," *In Proceedings of IEEE Symposium on Computational Intelligence in Image and Signal Processing* (*CIISP 2007*), pp.325–329, 2007.

[15] D. G. Lowe, "Object recognition from local scaleinvariant features," *In Proceedings of International Conference on Computer Vision* (*ICCV*), pp.1150–1157, 1999.

[16] S. Eickeler, F. Wallhoff, U. Iurgel, and G. Rigoll, "Content based Indexing of Images and Videos using Face Detection and Recognition Methods", *IEEE Int. Conference on Acoustics, Speech, and Signal Processing* (*ICASSP*), Salt Lake City, UT, 2001.

[17] V. N. Vapnik, "The Nature of Statistical Learning Theory," Springer Verlag, 1995.

[18] Z. Li and X. Tang, "Bayesian Face Recognition Using Support Vector Machine and Face Clustering," *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR 2004*), pp.374–380, 2004.

[19] N. Vretos, V. Solachildis, I. Pitas, "A Mutual Information based Face Clustering Algorithm for Movies," *In Proceedings of IEEE International Conference on Multimedia and Expo*, pp.1013–1016, 2006.

[20] H. Aoki, S. Shimotsuji, and O. Hori, "A shot classification method of selecting effective key-frames for video browsing," *In Proceedings of ACM Multimedia '96*, Boston, MA, pp.1–10, 1996.

[21] T. Mita, T. Kaneko, and O. Hori, "Joint Haar-like Features for Face Detection," *In Proceedings of 10th IEEE International Conference on Computer Vision* (*ICCV*), vol.2, pp.1619–1626, 2005.

[22] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol.2, pp.193–218, 1985.

[23] K. Yeung and W. Ruzzo, "Details of the adjusted rand index and clustering algorithms. supplement to the paper "an experimental study on principal component analysis for clustering gene expression data" ", *Bioinformatics*, vol.17, no.9, pp.763–774, 2001.

**Koji YAMAMOTO**

Koji YAMAMOTO received his B.E. degree in information and communication engineering and M.E. degree in electrical engineering from the University of Tokyo, Japan, in 1996 and 1998, respectively. He joined Toshiba Corporation in 1998. He is currently a Research Scientist at Multimedia Laboratory, Corporate Research and Development Center. His research interests include multimedia content analysis and retrieval.

**Osamu YAMAGUCHI**

Osamu YAMAGUCHI received his B. E. and M. E. degrees from Okayama University, in 1992 and 1994, respectively. In 1994, he joined Toshiba Corporation. He is currently a senior research scientist at Multimedia Laboratory, Toshiba Corporate Research and Development Center. He is a member of IPSJ, IEICE and IEEE.

**Dr. Hisashi AOKI**

Hisashi AOKI joined Toshiba Corporation in 1993 and is currently a Senior Research Scientist at Multimedia Laboratory, Corporate R&D Center. He is engaged in research on multimedia content understanding. He has been a visiting researcher at MIT Media Laboratory (1998-1999), a part-time lecturer at the University of Tokyo (2005-2006) and at Chuo University (2007-2010). He has been a secretariat of SIGs Human Interface (2005-2007) and Human-Computer Interaction (2007-2009) of Information Processing Society of Japan. Dr. Aoki organized the program committee of Interaction 2009 Symposium as a program co-chair, and is an editor-in-chief for special issue of "Technology, Design and Application of Interaction" of IPSJ Journal (published in 2010). He received the IPSJ Best Paper Award in 2001 and the IPSJ Nagao Makoto Special Researcher Award in 2006.