

# PERFORMANCE ANALYSIS OF VIDEO RETRIEVAL THROUGH IMAGE AND AUDIO FEATURE

<sup>1</sup> IN-KYOUNG SHIN, <sup>2</sup>HYOCHANG AHN

<sup>1</sup>Dept. of Applied Computer Engineering, Dankook University, Korea

<sup>2</sup>Dept. of Smart & PhotoVoltaic Convergence, Far East University, Korea

E-mail: <sup>1</sup>gguri81@dankook.ac.kr, <sup>2</sup>youcu92@gmail.com

## ABSTRACT

As growth of computer technology and multimedia information, not only texts but also various form of image information can be obtained and stored. Furthermore, growing hardware technology is capable to store and inquiry large size of multi data in rapid time. Traditional indexing technology was supervised by human supervisor by writing an appropriate keyword but this method is inefficient for processing large images through time factor and, indexing keyword is likely to be subjective which leads to the problem of wrong keyword. Thus the matter of context based image searching is being focused on extracting and indexing automatically from visual context of images. This paper propose a method of efficient context based image searching using a feature of face image and audio feature. Two feature vectors are extracted and weighted by similarity search method using fuzzy integration and merging techniques. Our method extracts two features, merges them, and uses similarity search as a weighting using fuzzy integral images. Objective performance accuracy and reproducibility are superior to conventional methods in experimental results with 1,000 color images.

**Keywords:** *Image Feature, Audio Feature, Video Retrieval, Image Retrieval, Multimedia Information,*

## 1. INTRODUCTION

In recent years, as the smartphone has become popular, the user of the video service have been increasing. Also users frequently can store and share video data, and often use synchronization in the cloud service environment. Besides, audio and video streaming services are becoming large capacity and high quality [1]. YouTube, the leading video streaming service, has viewed more than 4 billion videos a day and uploads 60 hours of video every minute. We need various methods to handle large capacity video data in a fast and efficient, owing to be rapidly increased the production and consumption of a large amount of video data [2]. With the development of computer technology and multimedia information, various types of image information as well as text information can be easily acquired and stored. In addition, the use of image information is increasing in various fields, but the management is becoming increasingly difficult [3].

The video information has a larger capacity than the character information when searching for specific data, and it is not easy to search quickly and efficiently since it is atypical.

From this viewpoint, a new search method for effectively managing images has become necessary.

The image information has a larger capacity than the text information when searching for specific data, and it is not easy to search quickly and efficiently since it is atypical [4,5]. Therefore, a new search method for effectively managing images is needed. Multimedia information retrieval is a new field of interest [6,7]. It is based on extraction and analysis of morphological and semantic context information of multimedia data. The field of information storage and retrieval includes a human centered user interface function through natural language and the like and the development of related technologies is being actively researched [8,9].

Multimedia information retrieval can be utilized in various fields such as medical image storage and management field, education information extraction and retrieval field, product search field of Internet cyber shopping mall, electronic commerce, and cybercrime investigation [10,11,12]. Information retrieval includes various information technology fields such as user interface technology and multimedia information retrieval technology for various user groups. Among them, image recognition, speech recognition and natural language

processing technology are techniques for effectively providing multimedia data to users [13,14]. Information building technology is used to build information infrastructure to form a wide-area distributed database and to store and manage text data and multimedia data. Also digital document technology can be flexible to use digitized documents[15].

In this paper, we proposed the efficient method of context based image retrieval to extract audio features and the face image features in the video data. We implemented a context based image retrieval system using similarity by applying weighted feature vector. Since the visual features are extracted the context based on the image and used for the retrieval, in the context based image retrieval system, video frame features corresponding the characteristics of each band in the image are extracted and merged into feature vectors for the region. In the video sequence, facial features are selected by learning the facial images through PCA (Principal Component Analysis) used in the integral image to detect facial features[16]. The voice function can collect and classify voiced unvoiced and unvoiced voices at the voice interval through frame energy and zero crossing speed. After the histogram feature vector has collected the crossover function and similarity measure, a database with a set of multiple facial images and voices can be generated by applying a fuzzy integral and assigning weights. Our method is compared with two feature vectors which are facial and audio features using weighted fuzzy integral.

The remainder of this paper is organized as follows. In section 2, we describe context based image retrieval and in section 3, speech features are extracted. Implementation of image and audio functions and video search based on experimental results are shown in Section 4. Finally, Section 5 describes the conclusion and future work.

## 2. CONTEXT BASED IMAGE RETRIEVAL

The image retrieval system consists of image input process, query process and retrieval process, and original images that show the retrieval result visually to the user. The image input process extracts the features that can best represent the input image and construct the image database through the indexing process [17,18,19]. The feature of the image to be searched by the user is extracted and the same kind of feature should be extracted through the same method used in building the database for the query image. Figure 1 shows the structure of the context based image retrieval system.

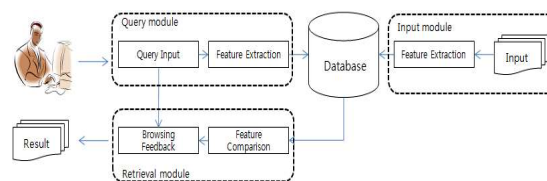


Figure 1: Diagram of the context based image retrieval system

The data processing process of the overall context based image retrieval system consists of a query interface and query processing. According to the context based image retrieval system, the method of inputting the query image is roughly divided into a query method for the example image, a query method for the sketch, and a method of inputting the feature value of the image. First, image query is one of the most commonly used methods in context - based image retrieval system. Among the many images provided by the query interface, the user selects an image most similar to the desired image and uses the selected image as a query image. In the second method, the user sketches the approximate image of the desired image by using the drawing tool provided by the query interface by the sketch, and uses the image as the query image. The method of sketching directly by the user can express various characteristics such as object shape, contour information, color, layout, and so on, so that the desired features can be freely expressed. However, the query method for the sketch has a disadvantage in that the feature expressed by the user-created image may show a lot of difference from the desired image due to the error of the user or the error in the creation process. The third method is to search quantitatively the subjective fuzziness in the process of human thinking or judgment by searching through the fuzzy integral. An ordinary set is an object which can be judged whether it belongs to the set or not. But the fuzzy set is a set of objects made up of objects whose criteria are not clearly defined as to whether they belong to it or not. A fuzzy integral is a method of choosing the best one among the worst cases

## 3. CONTEXT BASED IMAGE RETRIEVAL

In this paper, we select several facial images as one facial image by learning facial images obtained by principal component analysis from facial data through integral image to detect facial images in moving images. To detect the speech, the end point of the speech is extracted using the zero crossing rate and the frame energy, and then the noun of the

speech signal is detected. We use the database of two feature vectors to compare the similarity of facial images and speech signals using fuzzy integral weights.

### 3.1 Feature information extraction

The method of extracting the information from the image is to extract the eigencface through the eigenvectors obtained from the PCA (Principal Component Analysis) from the face data. Principal component analysis is easy to handle high dimensional data. It is a statistical technique that converts the data of a low-dimensional dimension into low-dimensional data. The facial images are linearly transformed into vectors, and then the facial image vectors are expressed by the coordinates in the orthogonal basis. The method of obtaining the eigenvectors constituting the eigenface is as follows.

The set of N face images used is defined as  $F = [F_1, F_2, F_3, \dots, F_N]$ , and the mean vector of the whole face image vector is defined as  $F = [F_1, F_2, F_3, \dots, F_N]$ . If a set of vectors obtained by subtracting the mean vector of all facial vectors from each facial image vector is  $A = [E_1, E_2, E_3, \dots, E_N]$  where is  $E_i = F_i - \Psi$ , the variance matrix of facial images is expressed by Equation (1).

$$C = \frac{1}{M} \sum_{n=1}^N E_n E_n^T = AA^T \quad (1)$$

In this case, the eigenvector matrix  $V$  of  $V$  and the diagonal matrix  $\Lambda$  whose eigenvalues are arranged in order of magnitude with respect to the diagonal elements are expressed by Equation (2).

$$CV = V\Lambda \quad (2)$$

When the size of an image is  $N \times N$  and the number of images is  $M$ ,  $C$  requires a large amount of memory and computation time, which is impossible to calculate because of the size of  $N^2 \times N^2$ , and it is difficult to learn when the number of learning images is small. However, we can obtain  $V$  and  $\Lambda$  efficiently by analyzing eigenvalues of the  $A^T A$  matrix instead of  $AA^T$  as follows. Where  $U$  and  $\Lambda$  are the eigenvectors and eigenvalues of the  $A^T A$  matrix, respectively, it can be described as the following equation (3).

$$(A^T A)U = U\Lambda \quad (3)$$

Equation (4) can be calculated by multiplying both sides of equation (3) by  $A$ .

$$A(A^T A)U = (AA^T)(AU) = (AU)\Lambda \quad (4)$$

It can be seen that eigenvectors and eigenvalues of  $C$  matrix are obtained by equation. (3) and (4). Therefore, matrix calculations with the size of  $N^2 \times N^2$  can be calculated in the space of  $M \times M$ . After the magnitudes of the eigenvalues are arranged in descending order,  $W$  is a linear transformation matrix composed of the first  $N$  eigenvectors corresponding to the magnitudes of eigenvalues in equation (5).

$$W = [Au_1, Au_2, Au_3, \dots, Au_p] \quad (5)$$

The facial image vector  $x_i$  whose size is  $N \times N$  is transformed into a vector  $y_i$  on a new space in the  $p$  dimension by the linear transformation matrix  $W$  as shown in the following equation (6).

$$y_i = W^T(x_i - \Psi) \quad (6)$$

The similarity between the value obtained by the equation (6) and the previously calculated vectors is compared.

### 3.2 Voice endpoint detection

The amplitude of the voice signal changes somewhat over time. The amplitude of the unvoiced segment is generally much smaller than the amplitude of the voiced segment. The short-term energy of the speech signal easily indicates this change in amplitude. The definition of the general short-term energy is given by the following equation (7).

$$V_n = \sum_{i=-\infty}^{\infty} [x(i)w(n-i)]^2 = \sum_{i=-\infty}^{\infty} x(i)^2 t(n-i) \quad (7)$$

where  $t(n)$  denotes  $t(n) = w^2(n)$ . In the above equation,  $x(n)$  represents a sample value and  $w(n - i)$  represents a window function. The above equation (7) has a disadvantage that it is too sensitive to a large signal level. Therefore, to solve this problem, the average magnitude function is defined as equation (8).

$$P_n = \sum_{i=-\infty}^{\infty} |x(i)| w(n - i) \quad (8)$$

Equation (8) can approximate the speech signal as a short term energy function by dividing the speech signal into frames with  $N$  samples and summing each sample in each frame. The speech signal is called  $h(n)$  and the short-term energy  $E_n$  is obtained. The calculation for this can be expressed as a short-term size  $M_n$  simply by obtaining an absolute value.

The zero crossing rate is the number of times the signal waveform intersects the zero axis within the frame interval, and occurs when the successive sample values in the discrete signal are different from each other. The zero crossing rate is the number of times the signal waveform intersects the zero axis within the analysis interval frame. It occurs when the continuous sampling values are different from each other in the discrete signal, and occurs when the continuous sampling values are different from each other in the discrete signal. In other words, the mean zero crossing rate of the sinusoidal wave having the fundamental frequency  $fa$  is  $2fa$  (*cro /rate*). The zero crossing rate is defined by equation (9), (10) and (11). The threshold value is determined as the zero difference rate when the voice is silent. If the zero crossing rate is higher than the threshold, the voice signal is unvoiced and if the zero crossing rate is low, the voice signal is voiced.

$$E_n = \sum_{i=-\infty}^{\infty} |sfn[x(i) - sfn[x(i - 1)]]| w(n - i) \quad (9)$$

$$sfn[x(n)] = \begin{cases} 1, & x(n) \geq 0 \\ -1, & x(n) < 0 \end{cases} \quad (10)$$

$$w(n) = \begin{cases} \frac{1}{2N}, & 0 \leq n \leq N - 1 \\ -1, & otherwise \end{cases} \quad (11)$$

The endpoint extraction method using zero crossing rate and frame energy is as follows. The

100ms interval of the speech interval is assumed to be silent, and the zero crossing rate and the frame energy generated in this interval are averaged to determine the threshold value for the start point and the end point. We randomly move samples by 10ms frame by frame. We can find sum energy of absolute value for each frame and compare it with the previously obtained threshold value to separate the speech interval and the silence interval.

The starting point  $N_1$  and the ending point  $N_2$  detected using the above frame energy. Then, the final speech interval is determined by separating voiced, unvoiced, and silent voices at the start and end points of the speech interval detected by the frame energy using the zero crossing rate. The zero crossing rate is calculated by moving 10ms ahead of the starting point extracted by the frame energy, and the zero crossing rate is calculated by moving backward in 10ms unit at the end point to determine the final voice interval.

### 3.3 Learning Method

It is very difficult to represent the whole space of non-face data with the sampling data obtained through the interface of the learned class classifier using the facial image and the non-facial image, and it is required a very large amount of data. It means that the learning time is very long in the learning of the binary classifier. In this paper, we use a single class SVM (Support Vector Machine) to study using facial image and non-facial image. Using a single class classifier that learns only with face data corresponding to the normal class, not only the learning time is much faster, but also better classification performance is obtained.

The advantages of face detection system using single class SVM are summarized as follows. First, in determining the decision boundary of the face region, the face region is expressed only by distribution of the face data without the influence of the non-face data, so that the false fraction can be greatly reduced. Second, when using a binary classifier, the amount of non-face data is tens of times larger than the amount of face data, so the learning time is very long. However, learning by face data only greatly shortens learning time. Finally, there is no additional time for collecting non-face data and no additional cost for the database to store it.

Therefore, in the face detection system, the single class SVM is used as the classifier to increase the detection rate. In addition, we use a discrete Ada-boost learning algorithm to sequentially select weak

classifiers that can complement and classify a given database, and generate strong classifiers by linear combination of them. The weak classifier selected through the iterative learning of the generated facial image responds to each input facial image. Finally, the classifier has a hierarchical structure. After selecting the same facial image and different facial images through dictionary initialization, the classifier performs learning.

### 3.4 Feature vectors merging

The face image feature vector and the voice feature vectors  $P_I$  and  $P_V$  can be calculated. The measured feature vectors may be merged using the same size region. The merging feature of  $P_I$  and  $P_V$  is called PRFV (Proceeding Region Feature Vector), and it is expressed as equation (12), where  $p_i$ ,  $h_i$ ,  $f_i$  and  $v_i$  are represented image number, value of quantization, face image value, and voice value, respectively.

$$PRFV = [p_i, h_i, f_i, v_i] \quad (12)$$

The merged feature vector is stored in the feature database through the user interface. The feature database is created using the proposed feature extraction method as color input images. The query processing extracts the feature vector of the query image from the user interface. The feature extraction method is the same as that for creating the feature database. The feature vector of the query image is compared with the feature vectors of the feature database. When the similarities are compared, the images that are most similar to the query image are displayed through the result user interface according to the priority order. The merged feature vector is extracted by extracting input images and extracting features by the proposed method. The extracted feature vectors are merged and stored in the feature database. The feature vector merging method divides the image into video and audio. Each divided region is converted into HSI color space to divide the facial image, and the image is quantized into 12 levels. At this time, the facial image component can be obtained by using the Haar-like feature mask. And it is expressed in a vector format and called PRFCV. In order to extract the speech information, the speech signal is separated by using the zero crossing rate and the frame energy, and the similarity of the vectors is discriminated using the Euclidean distance. This speech component is vector format data represented by PRVFV. The two extracted features are merged into the PRFV because the size of the region is the same, and the merged region feature vector is stored in the feature database.

## 4. EXPERIMENTS AND RESULTS

The process of extracting features of video and audio and retrieving is as follows. First, a face image to be used in the input image is extracted using the Haar-like feature, and the feature in the face image is extracted using the extracted face image. The Haar-like feature can detect facial images by expressing edges, lines, symmetries, and diagonal lines with features of 2-rectangles, 3-rectangles, and 4-rectangles. Then, principal components of facial data are analyzed to obtain eigenvectors, and facial images are collected through face detection using a single class SVM. In order to extract the voice information, the final voice interval is determined by dividing the voiced, unvoiced, and silent voices at the start and end points of the voice section using the frame energy and the zero crossing rate in the voice signal. The feature parameters of the speech signal are collected through LPC analysis using the separated speech information.

The face images and speech features collected by both methods are merged into one feature vector. The two feature vectors are weighted by applying a fuzzy integral to the similarity measure using the histogram intersection function. Figure 2 shows the overall system structure that weights are applied to feature information of the proposed image and voice.

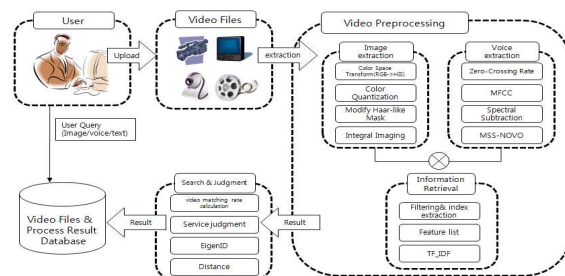


Figure 2: Diagram of the overall system structure

The video context used for the experiment were randomly collected 1,000 pieces of drama, movie trailer, and interview. The collected video context have a length of at least 25 seconds and a maximum of 3,400 seconds, and the number of people included in the experiment image is one or more than two. In addition, the face image data sets of the comparison subjects were collected by collecting 320 face images having similar appearance to face images having different angles. The speech data set was a clean acoustic model, and it was learned by 611 lecture data provided by Google Voice Search.

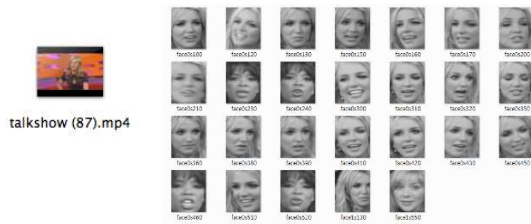


Figure 3: The result of face image detection in video

Figure 3 shows the result of the face recognition rate while changing the number of Haar-like features. At this time, the number of feature values was fixed to 50 features. As the result show, the recognition rate was constant at 98% when more than 50 features were used. However, when less than 50 features were used, the recognition rate is decreased linearly. Therefore, small number of Haar-like features cannot be used for facial recognition.

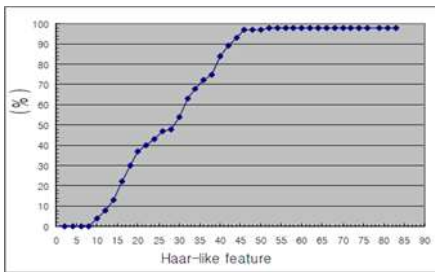
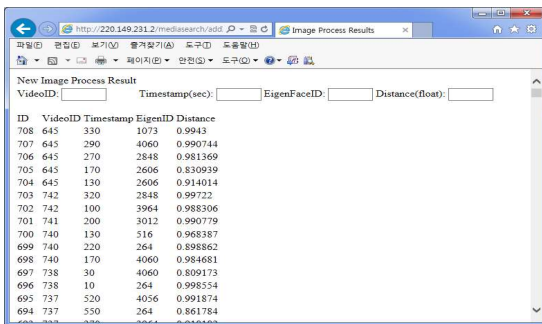


Figure 4: The result of face image detection in video



| ID  | VideoID | Timestamp | EigenID | Distance |
|-----|---------|-----------|---------|----------|
| 708 | 645     | 330       | 1073    | 0.9943   |
| 707 | 645     | 290       | 4060    | 0.990744 |
| 706 | 645     | 270       | 2848    | 0.981369 |
| 705 | 645     | 170       | 2606    | 0.830939 |
| 704 | 645     | 130       | 2606    | 0.914014 |
| 703 | 742     | 320       | 2848    | 0.99722  |
| 702 | 742     | 100       | 3964    | 0.988306 |
| 701 | 741     | 200       | 3012    | 0.990779 |
| 700 | 740     | 130       | 316     | 0.968387 |
| 699 | 740     | 220       | 264     | 0.898862 |
| 698 | 740     | 170       | 4060    | 0.984681 |
| 697 | 738     | 30        | 4060    | 0.809173 |
| 696 | 738     | 10        | 264     | 0.998554 |
| 695 | 737     | 520       | 4056    | 0.991874 |
| 694 | 737     | 550       | 264     | 0.861784 |

Figure 5: Similarity results of detected face images

IIM-NET was used as a basic recognizer to check the effectiveness of various noise processing techniques. The acoustic model was created through the learning process using the recorded information defined in similar phoneme units from the learning data. The training and recognition process of the

recognition unit model applied to the experiment is as follows. First, feature vectors are extracted from speech data for learning and recognition. Second, each single monophonic model is initialized by using feature vectors extracted from speech data for learning. Third, the initialized phoneme model is updated by training the speech feature vectors four times. Finally, the feature vector extracted from the recognition speech data is compared with the trained phoneme model and recognized. After obtaining the speech feature parameters from the learning video, the recognition word probabilities were calculated by using the given acoustic model and the dictionary word dictionary. The learning process and the recognition process are repeated with increasing the number of mixed phoneme models up to 20.

In order to evaluate the performance of the merging and weight search method of the proposed region feature vector, we compared the search method of Song and Saha with the method of adding voice information. Song proposed a new image representation method and similarity measurement method. First, the color image is converted into the HSV color space in order to use the features of the color image, and then the image is divided into 16 regions according to the degree of similarity. In order to extract the representative image of the divided 16 regions, the average value and the frequency of the components for each region were calculated and selected as the representative value of the similar region and stored as a feature vector. In addition, fuzzy theory was applied to calculate the distance of feature vectors in the query.

The field uses fuzzy index color for image features using coexistence matrix. After converting the image into a black and white image, the brightness is binarized. The binarized values are collected and converted into gray codes. The fuzzy index is characterized in that six important values in the HSV are used to create fuzzy membership with ideal Gaussian distribution and actual Gaussian distribution. For the same condition as the proposed method, voice information was added to each piece of information. Weights were fixed to 60% for face image and 40% for voice information. The time required for feature matching was measured by using each query method, and the time required to extract images with similar similarity to the database was measured, and the results shown in Table 1 were obtained.

Table 1: Time Required To Extract Feature Of Video From Database

|                                 | sec / 1,000 videos | sec / 1 video |
|---------------------------------|--------------------|---------------|
| Average feature extraction time | 41.923             | 0.055         |

In Table 2, 1000 videos of the whole video were used as face image and voice information. The average time obtained by searching the images of all databases was measured. Experimental results show that feature matching time required for video search is very fast.

Table 2: Feature Matching Time For Video Retrieval

| Method                | Face image           | Voice information   | Face image + Voice information |
|-----------------------|----------------------|---------------------|--------------------------------|
| Feature Matching Time | 0.0043(sec / 1 time) | 0.083(sec / 1 time) | 0.061(sec / 1 time)            |

Figure 6 shows an example of the result screen of querying "Obama". In the example shown in Fig. 1, all the images of the same category are searched from the first to the 10th rank used in the general context based search. Also, the result of searching up to 40 rankings shows that the accuracy is 97.5% and the recall rate is 39%.

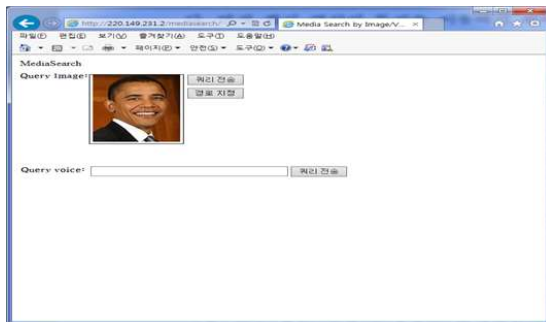


Figure 6: Example Of A Search Screen Shot Of Querying "Obama"

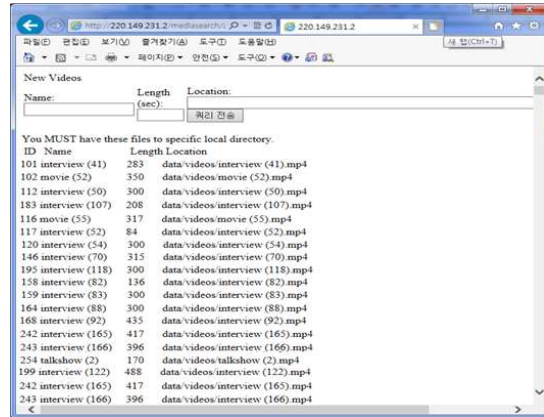


Figure 7: Example of a search result of querying "Obama"

Table 3: Average Search Time Of Feature Vector

| Method                               | Feature extraction time (sec) | Search time (sec) |
|--------------------------------------|-------------------------------|-------------------|
| Face Image                           | 0.382                         | 0.147             |
| Voice                                | 0.438                         | 0.219             |
| Proposed Method (Face Image + Voice) | 0.497                         | 0.097             |

As shown in Table 3, the effectiveness of the retrieval system was confirmed by comparing the feature retrieval time and the average retrieval time of the feature vector for the proposed method and the conventional retrieval method. The feature extraction time using the proposed method is 0.115 seconds slower than when only facial images were detected, and 0.059 seconds slower than when only voice information was detected. However, this problem is caused by detecting face image and voice information at once. It is more efficient than the time for successively detecting the features of the face and the voice. Since the search time has the feature vector as index information in the database, it can be seen that it is processed faster than the conventional method.

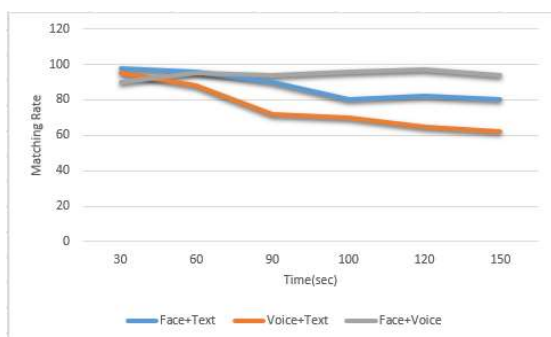


Figure 8: Various Search According To Length Of Image

Also, the image length is 30 seconds, 60 seconds, 90 seconds, 100 seconds, 120 seconds, and 150 seconds. As the length of the image increases, the degree of match of the image shows that it is a little more efficient than the existing methods of searching, face, text, voice and text. The longer the length of the image, the higher the degree of agreement than the search method of voice and text.

## 5. CONCLUSIONS

The proposed algorithm is designed as a Java applet so that it can be processed through a web browser which can be easily accessed by users. In order to detect the facial image in the moving image, several facial images are selected as one facial image by learning the facial image obtained through principal component analysis from the facial data through the integral image. To detect the speech, the end point of the speech is extracted using the zero crossing rate and the frame energy, and then the noun of the speech signal is detected. We use the database of two feature vectors to compare the similarity of facial images and speech signals using fuzzy integral weights.

By extracting two feature vectors and experimenting with 1,000 color images using the weighted similarity search method using the merging method and the fuzzy integral, the objective performance, accuracy and recall rate are superior to those of the conventional methods. In addition, subjective visual comparison shows better results than the conventional methods.

Future research projects should improve the parts of the face image that have a high search weight and use the voice information features more accurately. This will make it easier to apply weights using the fuzzy integral. In addition, the fuzzy measure of the fuzzy integral is passively set by experiment according to the image, and there is a need to study

the part that automatically sets this part, and it is necessary to study not only face image but also other image information.

## REFERENCES:

- [1] Bagri, Neelima, and Punit Kumar Johari. "A Comparative Study on Feature Extraction using Texture and Shape for Content Based Image Retrieval." *International Journal of Advanced Science and Technology* 80 41-52, 2015.
- [2] Miyazawa, Yuta, Yukiko Yamamoto, and Takashi Kawabe. "Context-Aware Recommendation System Using Content Based Image Retrieval with Dynamic Context Considered." *Signal-Image Technology & Internet-Based Systems (SITIS), 2013 International Conference on. IEEE, 2013.*
- [3] K. Hirata and T. Kato, "Query by visual example," in *Proceedings EDBT conference*, pp. 56-71, Mar. 1992.
- [4] V. Castelli and L. D. Bergman, *Image databases*, WILEY press, 2002.
- [5] Y. Rui, T. S. Hang, and S. Fu Chang, "Image retrieval : Current technique, Promising directions, and open issues," *Journal of Visual Communication and Image Representation*, vol. 10, no. 4, pp.39-62, April 1999.
- [6] K. Wong, K. Cheung and L. Po, "MIRROR : An Interactive Content Based Image Retrieval System," *IEEE International Symposium on Circuits and Systems*, vol. 2, pp. 1541-1544, May 2005.
- [7] R. Schettini, G. Ciocca and S. Zuffi, *A Survey of Methods for Colour Image Indexing and Retrieval in Image Databases*, L. W. MacDonald and M. R. Luo, Editors, *Color Imaging Science : Exploiting Digital Media*, Wiley, J & Sons Ltd, 2001.
- [8] S. Chang, A. Eleftheriadis, and R. McChlintock, "Next-generation content representation, creation and searching for new-media applications in education," in *Proceedings of the IEEE*, vol. 86, no. 5, pp. 884-904, May 1998.
- [9] Y. Rui and T. Huang, "Image Retrieval : Current Techniques, Promising Directions, and Open Issues," *Journal of Communication and Image Representation*, vol. 10, pp. 36-62, 1999.
- [10] C. H. Kuo, T. C. Chou, N. L. Tsao, and Y. H. Lan, "Can Find-asemantic image indexing and retrieval system," *Proceedings of ISCAS 2003*, vol. 2, pp. 25-28, 2003.



- [11] J. R. Smith and S. F. Chang, "Visually searching the web for content," *IEEE Multimedia Magazine*, vol. 4, no. 3, pp. 12-20, July 1997.
- [12] J. K. Wu, "Content-based indexing of multimedia databases," *IEEE Trans. on Knowledge and Data Engineering*, vol. 9, no. 6, pp.978-989, 1997.
- [13] G. D. Guo, A. K. Jain, W. Y. Ma, and H. J. Zhang, "learning similarity measure for natural image retrieval with relevance feedback," *IEEE Reans. Neural Networks*, vol. 13, pp. 811-820, 2002.
- [14] D. Zhang, "Review of shape representation and description techniques," *Pattern Recognition*, vol. 37, no. 1, pp. 1-19, Jaun 2004.
- [15] H. Noda and M. Niimi, "Colorization in YCbCr Color Space and Its Application to JPEG Images", *IEEE International Conference on Image Processing*, vol.4, no.12, pp.3714-3720, Sep. 2007.
- [16] D. W. Kim, D. J. Kwon, N. J. Kwak, and J. H. Ahn, "A content-based image retrieval using region based color histogram," *Proceedings of ICIC2005*, 2005.
- [17] V. Perlibakas, "Distance measures for PCA-based face recognition," *Pattern recognition Letters*, vol. 25, pp. 771-724, 2004.
- [18] Liu, Guang-Hai, and Jing-Yu Yang. "Content-based image retrieval using color difference histogram." *Pattern Recognition* 46.1, 188-198, 2013.
- [19] Ho, Jan-Ming, et al. "A novel content based image retrieval system using K-means with feature extraction." *Systems and Informatics (ICSAI), 2012 International Conference on. IEEE*, 2012.