Decision Tree Generation Algorithm for Image-based Video Conferencing

Yunsick Sung[1], Jeonghoon Kwak[1], Jong Hyuk Park[2]

[1] Department of Multimedia Engineering, Dongguk University-Seoul, Republic of Korea
[2] Department of Computer Science and Engineering, Seoul National University of Science and Technology, Republic of Korea
sung@dongguk.edu, jeonghoon@dongguk.edu, jhpark1@seoultech.ac.kr

Abstract

Recently, the diverse kinds of applications in multimedia computing have been developed for visual surveillance, healthcare, smart cities, and security. Video conferencing is one of core applications among multimedia applications. The Quality of Service of video conferencing is a major issue, because of limited network traffic. Video conferencing allow a large number of users to converse with each other. However, the huge amount of packets are generated in the process of transmitting and receiving the photographed images of users. Therefore, the number of packets in video conferencing needs to be reduced. Video conferencing can be conducted in virtual reality by sending only the control signals of virtual characters and showing virtual characters based on the received signals to represent the users, instead of the photographed images of the users, in real time. This paper proposes a method that determines representative photographed images by analyzing the collected photographed images of users, using K-Medoids algorithm and a decision tree, and expresses the users based on the analyzed images. The decision tree used for video conferencing are generated automatically using the proposed method. Given that the behaviors in the decision tree is added or changed considering photographed images, it is possible to reproduce the decision tree by photographing the behavior of the user in real-time. In an experiment conducted, 63 consecutively photographed images were collected and a decision tree generated by using the silhouette images of the photographed images. Indices of the silhouette images were utilized to express a subject and one index was selected using a decision tree. The proposed method reduced the number of comparisons by a factor of 3.78 compared with the traditional method that uses correlation coefficient. Further, each user's image could be outputted by using only the control image table of the image and the index.

Keywords: Communication systems, Artificial intelligence, Motion pictures

1 Introduction

Visual surveillance, healthcare, smart cities, and security in video computing are novel research domain that enrich the daily life of people. For examples, video conferencing [1-4] help people be connected with various groups of users using intelligent devices and mobile portable devices [5-6]. However, video conferencing leads to the generation of a large volume of packets in the course of transmitting and receiving videos in real time. To reduce the number of packets, approaches for capturing human behavior, subtracting the changed area in the captured photographed images, and sending the changed area have been proposed [7-14]. Representative behaviors are extracted from photographed images by Programming by Demonstration (PbD) [15-16], Support Vector Machine (SVM) [17], Recurrent Neural Network (RNN) [18], or Convolutional Neural Network (CNN) [19-21], which are utilized for recognizing human behavior [22-29]. However, the above approaches are utilized only for labeled behaviors. Therefore, automatic extraction and recognition of behaviors without any labels for video conferencing is required.

A framework for reducing the amount of data transmitted by analyzing the processes of video conferencing has also been proposed [30]. In the proposed framework, user behavior is recognized and transmitted based on the photographed images selected by a decision tree [31]. A method that automatically generates a decision tree for behavior recognition by using silhouette images has also been presented [32].

In the previous research, there is a problem that the images of the decision tree have to be manually selected. It is hard for the user to consider one image by one image in the decision tree when the user automatically generates the behaviors used for video conferencing.

To generate a decision tree necessary for behavior recognition automatically, the silhouette images and

*Corresponding Author: Jong Hyuk Park; E-mail: jhpark1@seoultech.ac.kr
DOI: 10.3966/160792642019092005019

selected representative silhouette images need to be automatically classified for the decision tree. A method of generating a decision tree considering a variety of classified images in the decision tree is required.

This paper proposes a method that automatically generates a decision tree needed for analyzing the photographed images of attendees in social computing and network environments such as video conferencing, and a control image table to express the attendees.

The contribution of this paper is as follows. First, an enhanced image comparison approach is introduced. Given that traditional approaches [30-31] analyze the collected images through correlation coefficients, only the difference between two photographed images is calculated, which is obtained by subtracting the images of different user behaviors, which are then classified in the same cluster for generating a decision tree. By dividing four sections and calculating four differences, the images of similar behaviors are classified into the same group.

Next, silhouette images for a decision tree are extracted automatically using K-Medoids algorithm, which leads to the decision tree acquiring more silhouette images automatically.

The remainder of this paper is organized as follows. Section 2 introduces related work. Section 3 describes the proposed method. Section 4 validates the proposed method. Finally, Section 5 concludes this paper.

2 Related Work

This section explains transportation optimization and human behavior analysis for video conferencing in intelligent devices. In addition, approaches that construct a decision tree using generated silhouette images are described.

2.1 Transportation Optimization in Big Data

At present, research is conducted to enable video conferencing in a variety of environments, by employing several different methods to reduce the amount of data generated during the process.

A number of researchers have designed immersive video conferencing systems to participate in video conferencing [1] that determine the video and audio quality of the participants of the video conferencing among the plurality of users from their own viewpoints. By using the video and audio of the video conferencing participants, it is possible to conduct large-scale video conferencing.

One study utilized free viewpoint in video conferencing [2], in which users participating in the video conferencing experienced a three-dimensional (3D) screen. However, in the process of composing the 3D environment, it is a necessary step to compensate for the loss caused by synthesizing images received from multiple cameras.

In a mobile environment, it is difficult to conduct video conferencing owing to limited network bandwidth [3]. This paper proposes a cloud-based transcoding framework for video conferencing in mobile environments. By optimizing utility costs in video conferencing, it can be conducted in mobile environments.

To reduce the amount of data transmitted during video conferencing, an encoding method is employed [7]. However, a more efficient method of encoding an image to reduce the amount of data in the transmitted image is being studied. Communication between users includes not only images and audio but also additional combinations, such as facial expressions and gestures [8]. For example, in a previous study, the user's behavior and a face were extracted and reconstructed as a virtual character.

To reduce the data size, only the user domain is extracted and transmitted [9]. The 3D modeling consists of silhouette images and is used along with textures to construct 3D virtual characters for video conferencing.

The user behavior required for presentation is defined by using motion capture or by defining it in advance as a script [10]. When the operation of the virtual character is written in advance in accordance with the order of presentation, it cannot respond to the behavior of other attendees.

There is a need to define a behavior of a virtual character that can be performed by the user in advance and to generate a behavior of the virtual character by analyzing the behavior of the virtual character and the measurable data from the user. It is not possible to express the behavior of a user that is able to general behavior but is not mainly performed.

The volume of packets generated during video conferencing can be reduced. However, rather than using the actual image of the user, a virtual character is reconstructed with motion data or images measurable by the user. A method that provides real user behavior without reconstructing a virtual character based on measurable data is needed.

2.2 Human Behavior Analysis and Prediction

Several studies have extracted photographed images and behaviors for video conferencing using PbD [15]. A study on the behavior of virtual characters using PbD [16] involved the collection of the behavior of a virtual character from its predecessor and classifying the behavior by analyzing the behavior based on Bayesian probability.

A method to classify the scenes of a video involved the use of images and sounds [17]. A Support Vector Machine (SVM) classified images and sounds together. The combined use of images and sounds improved the image classification results. A study used CNN to classify the user behavior in a video [19]. The user operation was extracted using CNN.

A method for estimating the user behavior has been

proposed [22]. Other methods have used the behavior history [23], Gaussian model [24], or behavior database [26] to recognize behaviors, while the proposed method uses video and motion capture for behavior recognition [27].

For analyzing and predicting user behaviors, it is usually necessary to define user's behavior in advance. There are some disadvantages which it cannot be anticipated for undefined behaviors.

Generally, a labeled user-behavior dataset is needed to classify the users' behaviors. A method for deriving user behaviors without using labeled user-behavior dataset is required.

2.3 Decision Tree Generation Approach

A framework to control virtual characters has been proposed [30], where a decision tree is generated using the silhouette images. The first silhouette image among the silhouette images is set as the root node. Then, the next silhouette image is set as a child node. Further, the correlation coefficient between the silhouette image of the root node and the child node is calculated. The root node and child nodes are sorted in ascending order according to the calculated correlation coefficient. Thus, this method creates a decision tree by registering all of the silhouette images.

In a previous research, the decision tree was generated through the correlation coefficient [31]. The correlation was calculated using the values of the silhouette image normalized to 0 and 1. The root node determines the root node as the center value of the nodes. If the correlation coefficient of the node is large, while being similar in dimension to the right node, it is set as the left node. According to the depth of the decision tree, the longer it takes to find one of images to be displayed to the user, the longer it cannot display the user's image in real-time. The tree depth varies depending on the root node in the process of making the decision tree. There is a problem that the performance of the decision tree varies depending on how to set the root node. To solve this problem, a method to determine the root node, such that the depth of the decision tree is constant, is needed.

3 Decision Tree Generation Approach

In the proposed method, the photographed images of users captured by a camera in video conferencing are represented by indices, and the corresponding photographed images of the indices shown. By transmitting the indices only, the amount of transmission to remote devices is reduced.

3.1 Overview

To show alternate photographed images instead of showing the captured photographed images of users, the indices of the alternate photographed images of the users should be transferred to users for reducing the number of transferred packets, and then the alternate photographed images should be represented based on the transferred indices.

The processes employed to analyze and represent photographed images of the users in video conferencing are shown in Figure 1. In the Transmission Process, the Offline Phase analyzes the captured photographed images and generates a decision tree and a control image table required for analyzing the captured photographed images, and the Online Phase decides the corresponding index of each captured photographed image by applying the decision tree and the control image table. The decision tree consists of the nodes that contain the different between photographed images. The control image table consists of the pairs of an index and a photographed image. In the Reception Process, the Online Phase shows the corresponding alternate photographed images based on the received indices.

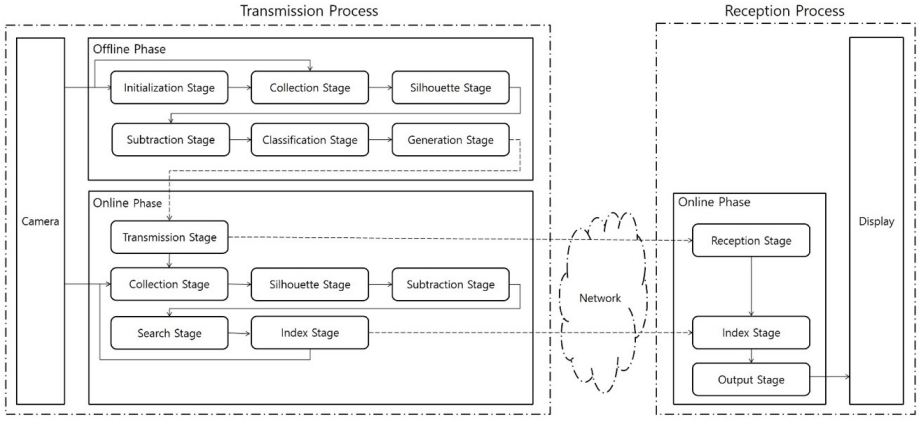


Figure 1. The Transmission Process transmits the index of the image taken by the camera of the mobile device and outputs it from the Reception Mobile Device. In the Offline Phase of the Transmission Mobile Device, the decision tree is generated by analyzing the image. In the Online Phase, the reception mobile device transmits the image of the control image table to the reception mobile device based on the decision tree. The Reception Process on the mobile devices receives the index and outputs the photographed image

3.2 Offline Phase of the Transmission Process

The Offline Phase configures a decision tree to decide and transfer the indices corresponding to the captured photographed images by indices.

The captured photographed images are compared with alternate photographed images. Therefore, to calculate the difference between two types of photographed images, the proposed method compares captured photographed images with a background image and alternate photographed images with the background image, and then calculates the difference between the captured photographed images and the background image and between the alternate photographed images and the background image. However, the difference between the captured photographed images and the background image is calculated in real-time, but the difference between the alternate photographed images and the background image is calculated in advance during the Subtraction Stage and is stored in a decision tree during the Decision Tree stage. The stages of the Transmission Process are as given below.

The Initialization Stage sets the background b and the basic image c_0. The acceptable photographed image difference is defined by the image difference α. Given that the photographed image comparison results are affected by the locations in the photographed images of a user, the difference of the x-axis of the user should be considered with the x-axis difference β. Both parameters are set during this stage.

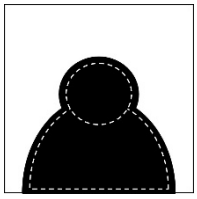
The Collection Stage collects captured photographed images, the captured image set C. The photographed image captured at tick t is the captured image c_t. The pixel $c_{t,i,j}$ is a pixel at (i, j) of the captured image c_t.

The Silhouette Stage extracts silhouettes of a user by black/white colors. The silhouette image of the captured image c_t is the silhouette s_t.

The silhouette s_t is calculated as follows. The silhouette s_t is the difference between the background b and the captured image c_t, as shown in Figure 2. Each pixel of the silhouette s_t is set to one when the difference between the corresponding pixel in the captured image c_t and the corresponding pixel in the background b is larger than the image difference α. Otherwise, each pixel is set to zero. Figure 2(a) shows a silhouette image. In the case when t is zero, the silhouette s_0 is eroded as shown in Figure 2(b) and then dilated as shown in Figure 2(c).

The Subtraction Stage creates the silhouette difference set D. The silhouette difference d_t is obtained by comparing the silhouettes s_0 and s_t, where t>0. To consider the silhouette image differences between the upper side and lower side and between the left side and right side invoked by two hands and a head, the silhouette image difference is calculated by dividing each silhouette into four sections. Thus, the silhouette difference d_t is defined by Eq. (1).

(a) A silhouette image (b) Erosion



(c) Dilation

Figure 2. Process of generation of the silhouette s_0

$$d_t = [t, d_{t,1}, d_{t,2}, d_{t,3}, d_{t,4}] \tag{1}$$

Where $d_{t,1}$ is the upper-left image difference between the silhouette s_0 and the silhouette s_t, and so on, as shown in Figure 3.

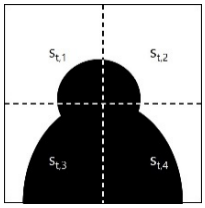


Figure 3. One silhouette image compared considering four silhouette differences with the basic image

Each silhouette image difference is calculated. The silhouette image difference d_t is processed by using $d_{t,1}$ from Eq. (2), $d_{t,2}$ from Eq. (3), $d_{t,3}$ from Eq. (4), and $d_{t,4}$ from Eq. (5).

$$d_{t,1} = \sum_{i=0}^{\frac{W}{2}} \sum_{j=0}^{\frac{H}{2}} (s_{0,i,j} - s_{t,i,j}) \tag{2}$$

$$d_{t,2} = \sum_{i=\frac{W}{2}}^{W} \sum_{j=0}^{\frac{H}{2}} (s_{0.i,j} - s_{t,i,j}) \tag{3}$$

$$d_{t,3} = \sum_{i=0}^{\frac{W}{2}} \sum_{j=\frac{H}{2}}^{H} (s_{0.i,j} - s_{t,i,j}) \tag{4}$$

$$d_{t,4} = \sum_{i=\frac{W}{2}}^{W} \sum_{j=\frac{H}{2}}^{H} (s_{0,i,j} - s_{t,i,j}) \tag{5}$$

Where $s_{t,i,j}$ is the pixel, and the width and the height of silhouette images are **W** and **H**, respectively.

After silhouette s_t is shifted to the left and the right by the x-axis difference β and is defined by silhouettes s'_t and s''_t, the silhouette image differences d'_t and d''_t of silhouettes s'_t and s''_t are calculated by comparing silhouettes s'_t and s''_t with silhouette s_0 considering each pixel in both silhouettes. The silhouette image difference d_t is set to the minimum value of the silhouette image differences, d_t, d'_t, and d''_t, as given in Eq. (6).

$$\begin{aligned} d_t = [t, & \min(d_{t,1}, d'_{t,1}, d''_{t,1}), \\ & \min(d_{t,2}, d'_{t,2}, d''_{t,2}) \\ & \min(d_{t,3}, d'_{t,3}, d''_{t,3}) \\ & \min(d_{t,4}, d'_{t,4}, d''_{t,4})] \end{aligned} \quad \textbf{(6)}$$

The Classification Stage classifies the silhouette image difference set D using the K-Medoids algorithm [32]. The K-Medoids algorithm sets the centroid by one of the elements of the silhouette image difference set D. The k-*th* centroid is μ_k. The total variance V is shown in Eq. (7).

$$V = \sum_{t=1} \sum_{k=1}^{K} \sum_{m=1}^{4} (d_{t,m} - \mu_{k,m})^2 \quad \textbf{(7)}$$

where m is the index of silhouette image differences.

The Generation Stage generates a decision tree and a control image table using the ordered centroids in order. The cluster centroids are arranged in the order of $d_k=\mu_{k,1}+\mu_{k,2}+\mu_{k,3}+\mu_{k,4}$. The p-*th* node n_p is defined by Eq. (8).

$$n_p = [\mu_p, n_p^L, n_p^R] \quad \textbf{(8)}$$

where the node n_p^L and the node n_p^R are the left node and right node of the node np, respectively. When the node np is a leaf node, the left node n_p^L and the right node n_p^R are ø.

A decision tree is generated as below. $\mu_{\left\lfloor \frac{K}{2} \right\rfloor}$ is specified as the root node n_p. The left node is set to the left node n_p^L as an intermediate value from 1 to $\left\lfloor \frac{K}{2} \right\rfloor$. The right node n_p^R sets the value from to $\left\lfloor \frac{K}{2} \right\rfloor$ to K as the intermediate value. Nodes are added until there are no more nodes to add.

The control image table is generated with the index, and node n_p is [p, s_p] where s_p is the p-th counselor image.

3.3 Online Phase of Transmission Process

The Online Phase deducts the indices of captured photographed images by a camera in real-time to transfer the indices to the Reception Mobile Device. The Transmission Stage sends the control image table to the Reception Mobile Device, which can be the photographed image indices.

The Collection Stage receives captured photographed images, Silhouette Stage deducts silhouette images, and the Image Subtraction Stage performs comparison of the silhouette images to obtain the indices, which are transferred to the Reception Mobile Device.

The Search Stage finds the minimum silhouette image difference from the root node of the decision tree using Eq. (9), and obtains the index of the node that has the minimum silhouette image difference.

$$d_t = \sum_{m=1}^{4} d_{t,m} \quad \textbf{(9)}$$

The silhouette image difference d_t is used to compare the nodes included in the decision tree. The silhouette image difference d_p and d_t are compared, and moved to the left node n_p^L if smaller and to the right node d_p^R if larger. However, if the current node n_p of d_p is closer than the left node n_p^L of d_p^L or right node d_p^R of d_p^L, and there is no left node n_p^L or right node d_p^R, the current node n_p is returned.

The Index Stage transfers the found index. Further, it transmits the index of the in the control image table using the index of the node.

3.4 Reception Process

The Online Phase in the Reception Mobile Device receives the index from Transmission Mobile Device and represents the corresponding alternate photographed images of a user-photographed image as follows.

The Reception Stage receives the control image table. If the object to be used is the alternate photographed image, the control image table of the photographed image is received. In the case of a user-photographed image, it receives a control image table of the controls.

The Index Stage receives the index from the Transmission Mobile Device. The Output Stage prints alternate photographed images. It controls user photographed images and prints s_p.

4 Experimental Evaluation

In the experiment conducted, a decision tree and a control image table were generated, and the captured photographed images were determined using the proposed method. The proposed method and K-Medoids algorithm based on correlation coefficient were compared to generate the decision tree and control image table.

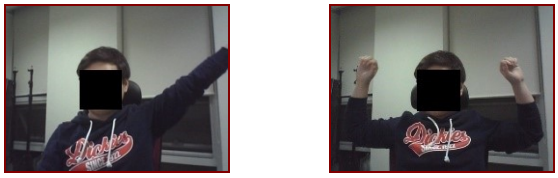
4.1 Offline Phase of the Transmission Process

During the Initialization stage, the size of the captured photographed image taken by the user using the camera was set to 160 × 120. The background and basic image in the Initialization stage are shown in Figure 4. The background b, in which the user is located, is shown in Figure 4(a), and the user's basic image is shown in Figure 4(b) for basic image c_0. The image difference α was set to 30. The x-axis difference β was set to seven. The user's behavior was defined as in [30].

(a) Background (b) Basic image c_0

Figure 4. Background b and basic Image c_0. (a) shows the environment in which the user is to be positioned from the camera. (b) shows the basic image to be taken by the user

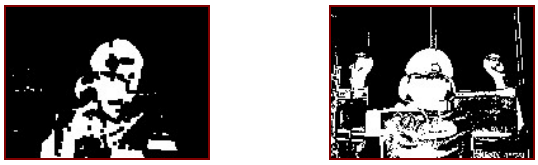
During the Collection stage, the captured photographed image of the user was collected at 5 frames per second. The user performed bowing, using right hand, using right/left hand, using two hands, using two hands to right/left, and moving their head to right/left three times. A total of 1,620 captured photographed images were collected, as shown in Figure 5. Figure 5(a) is the 772*nd* captured image, and Figure 5(b) is the 1,064*th* captured image.



(a) Captured photographed image c_{772} (b) Captured photographed image c_{1064}

Figure 5. Example of behaviors performed by user. (a) Using Left Hand, and (b) Using Two Hands

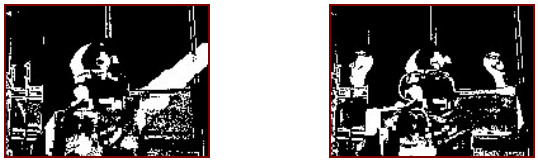
The result of the silhouette image for the Silhouette stage is shown in Figure 6. Figure 6(a) represents a change in the black and white of Figure 4(a) and Figure 4(b), followed by extraction of the user region from the silhouette. The silhouette s_t is changed to black and white.



(a) Silhouette image s_0 (b) Silhouette image s_{1064}

Figure 6. Fig. 3 and Fig. 4 show the extracted user area. (a) User region of the basic image, and (b) user area of the captured image c_{1064}

The image used to calculate the silhouette image difference during the Subtraction stage is shown in Figure 7. Subtraction was set as shown in Table 1 using the silhouette differences. The reason for dividing the image into four equal parts when calculating the subtraction is that the behavior is performed in four directions based on the x-axis when the human behavior is considered. In the case of not dividing into four equal parts, there is a problem in that the direction cannot be taken into consideration like the symmetric image. Because the user always moves to the left or the right without centering, the image in Figure 6(a) is shifted to the left or right by the x-axis difference β, so that the sum of the elements of the silhouette differences d_t, d'_t, and d''_t is minimized.



(a) Silhouette image difference image d_{772} (b) Silhouette image difference image d_{1064}

Figure 7. Silhouette image difference images of silhouette s_0, silhouette s_{772}, and silhouette s_{1064}.

Table 1. Silhouette differences of the silhouettes

	T	$d_{t,1}$	$d_{t,2}$	$d_{t,3}$	$d_{t,4}$
d_1	1	742	1362	540	308
d_2	2	611	1345	580	389
d_3	3	598	1309	635	369
			…		
d_{772}	772	283	1951	233	98
d_{1064}	1,064	289	1,232	997	743
			…		
d_{1618}	1,618	442	1,059	952	577
d_{1619}	1,619	420	1,039	959	585
d_{1620}	1,620	389	1,014	950	576

During the Classification stage, in the proposed method, K-Medoids algorithm [32] is used and the silhouette difference set D is classified. The 63 possible behaviors for video conferencing in [30] while the user is sitting include 4 front-side behaviors, 26 one hand behaviors, 23 two hand behaviors, and 10 head behaviors. In order to classify the behavior into 63, K is set to 63. The cluster centroid points classified by the proposed method are shown in Table 2. The control image table of the photographed images is shown in Figure 8 with the indexes of the classified cluster centroid points. The control image table of alternate photographed images derived from the proposed method includes 15 alternate photographed images similar to the basic image, where the user's position changes slightly.

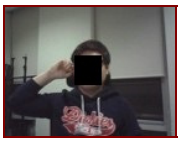
Table 2. Results of subtraction of cluster centroid

	K	t	$\mu_{k,1}$	$\mu_{k,2}$	$\mu_{k,3}$	$\mu_{k,4}$
μ_1	1	71	437	1,082	719	443
μ_2	2	37	1,532	1,540	850	478
μ_3	3	61	3,056	1,748	4,117	978
			…			
μ_{31}	31	864	462	1,173	835	513
μ_{32}	32	757	565	1,299	918	654
			…			
μ_{61}	61	1,524	510	1,247	721	471
μ_{62}	62	1,515	754	1,354	657	392
μ_{63}	63	1,502	188	922	2,231	1,284

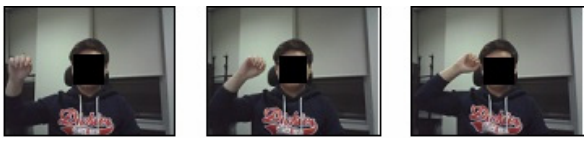
Figure 8. Control image table of alternate photographed images generated by the proposed method. Fifteen alternate photographed images similar to the basic image were created

The corresponding alternate photographed images to the cluster centroid of the 5*th* cluster and the 25*th* cluster and the indexes of the cluster members are shown in Figure 9. It contained corresponding alternate photographed images similar to the cluster centroid.

The results of classification based on the correlation coefficient are shown in Figure 10. There were 11 behaviors similar to the basic posture, and 10 images were considered to be similar, with four sets of similar images.



(a) Alternate photographed image of the 5th cluster centroid



(b) Photographed images of the 5th cluster

(c) Alternate photographed image of the 25th cluster centroid

(d) Photographed images of the 25th cluster

Figure 9. Images of the indices of the cluster subtraction classified by the proposed method. (a) Image of the index of the 5th cluster centroid point, and (b) image of the index of the members of the 5th cluster. (c) Image of the index of the 25th cluster centroid point, and (d) image of the index of the members of the 25th cluster

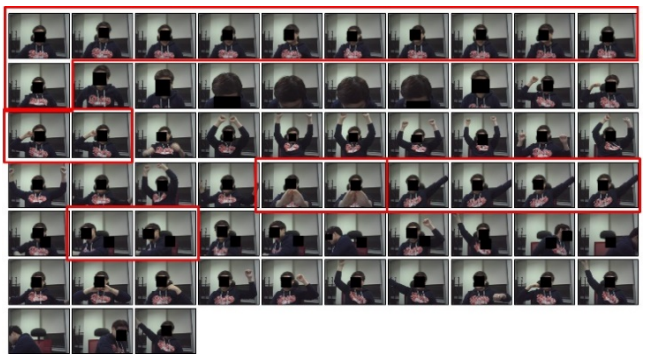


Figure 10. Correlation coefficient classification. A total of five sets of similar images were included, and five sets contained 21 images

Figure 11 shows a control image table of alternate photographed images of the result of classification into one difference value without calculating subtraction into four regions. There were 14 images similar to the basic image, three alternate photographed images, and seven alternate photographed images considered similar.

Figure 11. Result classified into silhouette image differences of the alternate photographed images. A total of five sets of similar alternate photographed images were included and four sets contained 21 alternate photographed images.

Based on the results of the proposed method, the correlation coefficient based clustering results and the image subtraction clustering results, the basic image was selected; the results are shown in Figure 12. The results of the proposed method, the correlation coefficient-based clustering results, and the image subtraction clustering results did not differ significantly, as shown in 11 to 13 and 60 to 63. However, the correlation coefficient-based clustering results, such as 19 to 20 and 29 to 31, differed significantly from the results of the proposed method and the image subtraction clustering results.

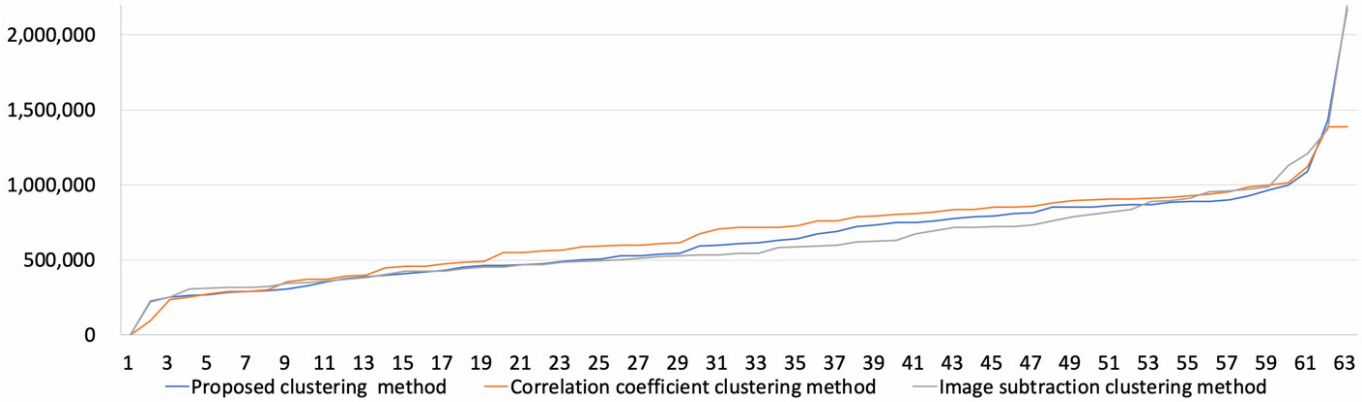


Figure 12. Result of sorting based on the silhouette image difference from the basic image selected by the user in the clustering result

The results of the Generation stage are shown in Table 3. The basic image and state are arranged in order of closeness.

Table 3. Cluster centroid aligned from the basic image

	K	t	$\mu_{k,1}$	$\mu_{k,2}$	$\mu_{k,3}$	$\mu_{k,4}$
μ_1	1	948	281	987	710	465
μ_2	2	1175	247	843	852	531
μ_3	3	109	273	998	771	501
			…			
μ_{31}	31	484	228	418	1,686	927
μ_{32}	32	450	338	1,397	996	649
			…			
μ_{61}	61	93	1,781	1,545	1,331	523
μ_{62}	62	901	2,083	1,497	1,600	671
μ_{63}	63	61	3,056	1,748	4,117	978

Figure 13 shows the result of the decision tree generated by the proposed method. The depth of the decision tree using proposed method is six and the depth of the decision tree using correlation coefficient is 43. Decision tree using correlation coefficient which is generated by using correlation coefficients between basic behavior and generated behaviors. The correlation coefficient between the basic behavior and the generated behavior was biased in one direction, and a decision tree with a depth of 43 was generated.

4.2 Online Phase of the Transmission Process

During the Collection stage, in the online phase, 815 images were taken by performing each behavior [30] once for each image taken for verification.

During the Search stage, the proposed method calculated the index to be transmitted to the Reception Mobile Device. As a result of the tree search, the proposed method was performed 4,135 times. The result of the tree search by correlation coefficient was 15,630 times, and the proposed method improved the performance by 27.68%.

Table 4 shows the results of outputting the photographed images to be displayed by the proposed method using 815 photographed images. When the arm or the body completely moved, the output was similar to the current image, but the basic index of the basic photographed image was not correctly found in the basic photographed image 1-5 times.

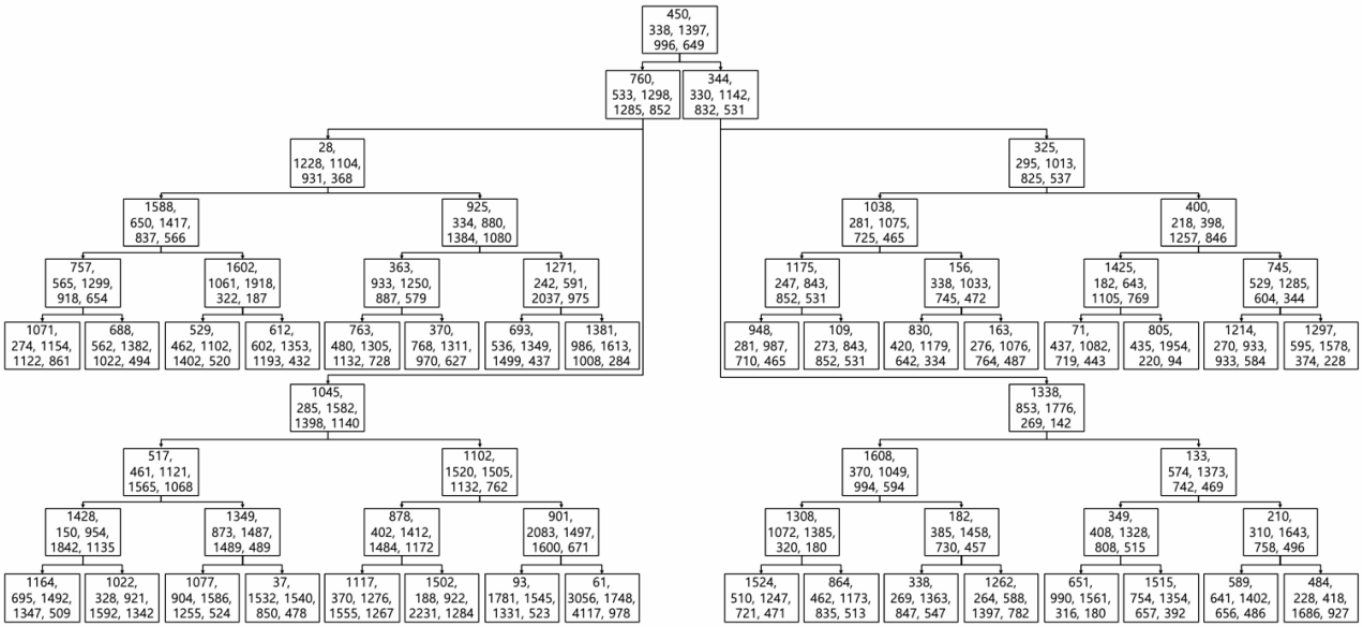
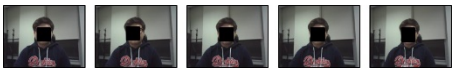
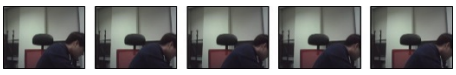
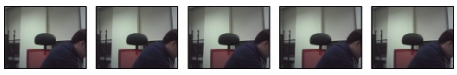
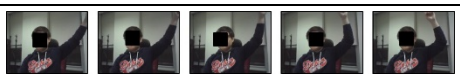
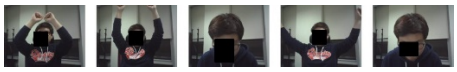
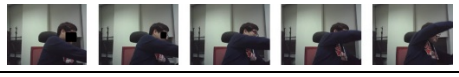
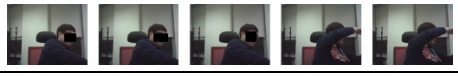
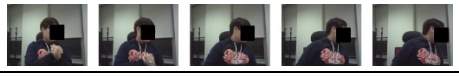


Figure 13. Decision tree generated by the proposed method

Table 4. Results of index examination from the inputted filming image

t	Transmission mobile device of real time filming image	Reception mobile device of display
1-5		
101-105		
201-205		
301-305		
401-405		
501-505		
601-605		
701-705		
801-805		

5 Conclusions

In order to reduce the number of packets during video conferencing, this paper proposed a method in which each user's control image table is generated using K-Medoids algorithm and a decision tree is constructed to find a control image table of photographed images similar to the current photographed image. Further, the index of the control image table of the photographed images is searched for using the decision tree and the photographed images are outputted using the control image table of the index.

In order to verify the efficacy of the proposed method, photographed images were collected for use in video conferencing and 63 representative motions were generated using K-Medoids algorithm based on the proposed method, correlation coefficient, and difference images. The depth of the decision tree generated by correlation coefficient was 43, whereas that of the proposed method had a depth of six. The proposed method found the index of the photographed images using the generated decision tree and was faster by a factor of 3.78.

In future research, the variety of the locations of a camera and the robustness to lights should be considered more. The classification approach and clustering approach should reflect continuous video conferencing. It is required to do research to analyze and generate user's behaviors using deep learning accurately. Therefore, it is possible to provide natural images to the user by applying a method of using another character or generating an intermediate image between images without using the user's image.

Acknowledgements

"This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2018-2013-1-00684) supervised by the IITP (Institute for Information & communications Technology Promotion)".

References

[1] F. Safaei, P. Pourashraf, D. Franklin, Large-scale Immersive Video Conferencing by Altering Video Quality and Distribution Based on the Virtual Context, *IEEE Communications Magazine*, Vol. 52 No. 8 pp. 66-72, August, 2014.

[2] B. Macchiavello, C. Dorea, E. M. Hung, G. Cheung, W. Tan, Loss-resilient Coding of Texture and Depth for Free-Viewpoint Video Conferencing, *IEEE Transactions on Multimedia*, Vol. 16 No. 3, pp. 711-725, January, 2014.

[3] R. Cheng, W. Wu, Y. Lou, A Cloud-based Transcoding Framework for Real-time Mobile Video Conferencing System, *2014 2nd IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud)*, Oxford, UK, 2014, pp. 236-245.

[4] J. Chae, Y. Jin, Y. Sung, K. Cho, Genetic Algorithm-based Motion Estimation Method using Orientations and EMGs for Robot Control, *Sensors*, Vol. 18, No. 1, pp. 1-14, January, 2018.

[5] M. Jeong, S. Ahn, A Network Coding-Aware Routing Mechanism for Time-Sensitive Data Delivery in Multi-Hop Wireless, *Journal of Information Processing Systems*, Vol. 13, No. 6, pp. 1544-1553, December, 2017.

[6] B. Kim, A Distributed Coexistence Mitigation Scheme for IoT-Based Smart Medical Systems, *Journal of Information Processing Systems*, Vol. 13, No. 6, pp. 1602-1612, December, 2017.

[7] H. Fuchs, G. Bishop, K. Arthur, L. McMillan, R. Bajcsy, S. W. Lee, H. Farid, T. Kanade, Virtual Space Teleconferencing Using a Sea of Cameras, *First International Conference on Medical Robotics and Computer Assisted Surgery*, Pittsburgh, PA, 1994, pp. 161-167.

[8] A. Mortlock, P. Sheppard, D. Machin, S. McConnell, Virtual Conferencing, *IEEE Colloquium on Teleconferencing Futures*, London, UK, 1997, pp. 1-6.

[9] B. Petit, J. Lesage, C. Menier, J. Allard, J. Franco, B. Raffin, E. Boyer, F. Faure, Multicamera Real-time 3D Modeling for Telepresence and Remote Collaboration, *International Journal of Digital Multimedia Broadcasting*, Vol. 2010, pp. 1-12, August, 2009.

[10] A. Nijholt, H. v. Welbergen, J. Zwiers, Introducing an Embodied Virtual Presenter Agent in a Virtual Meeting Room, *The IASTED International Conference on Artificial Intelligence and Applications (AIA 2005)*, Innsbruck, Austria, 2005, pp. 579-584.

[11] J. Kim, D. Chung, I. Ko, A Climbing Motion Recognition Method using Anatomical Information for Screen Climbing Games, *Human-centric Computing and Information Sciences*, Vol. 7, No. 25, pp. 1-14, September, 2017.

[12] M. T. N. Truong, S. Kim, Parallel Implementation of Color-based Particle Filter for Object Tracking in Embedded Systems, *Human-centric Computing and Information Sciences*, Vol. 7, No. 2, pp. 1-13, December, 2017.

[13] M. Naslcheraghi, S. A. Ghorashi, M. Shikh-Bahaei, FD Device-to-Device Communication for Wireless Video Distribution, *IET Communications*, Vol. 11, No. 7, pp. 1074-1081, May, 2017.

[14] Y. Sani, M. Isah, C. Edwards, A. Mauthe, Experimental Evaluation of the Impact of Mobility Management Protocols on HTTP Adaptive Streaming, *IET Networks*, Vol. 6, No. 6, pp. 186-192, November, 2017.

[15] A. Fod, M. J. Matarić, O. C. Jenkins, Automated Derivation of Primitives for Movement Classification, *Autonomous Robots*, Vol. 12, No. 1, pp. 39-54, January, 2002.

[16] C. Thurau, T. Paczian, C. Bauckhage, Is Bayesian Imitation Learning the Route to Believable Gamebots, *GAME-ON North America*, Montreal, Canada, 2005, pp. 3-9.

[17] Y. Zhu, Z. Ming, SVM-Based Video Scene Classification and Segmentation, *International Conference on Multimedia and Ubiquitous Engineering*, Busan, South Korea, 2008, pp. 407-412.

[18] Y. Song, I. Kim, Deep Act: A Deep Neural Network Model for Activity Detection in Untrimmed Videos, *Journal of Information Processing Systems*, Vol. 14, No. 1, pp. 150-161, February, 2018.

[19] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale Video Classification with Convolutional Neural Networks, *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, USA, 2014, pp. 1725-1732.

[20] J. Li, J. Li, Prompt Image Search with Deep Convolutional Neural Network via Efficient Hashing Code and Addictive Latent Semantic Layer, *Journal of Internet Technology*, Vol. 19, No. 3, pp. 949-957, May, 2018.

[21] Y. Sung, Y. Jin, J. Kwak, S. Lee, K. Cho, Advanced Camera Image Cropping Approach for CNN-Based End-to-End Controls on Sustainable Computing, *Sustainability*, Vol. 10, No. 3, pp. 1-13, March, 2018.

[22] M. B. Holte, C. Tran, M. M. Trivedi, T. B. Moeslund, Human Pose Estimation and Activity Recognition from Multi-view Videos: Comparative Explorations of Recent Developments,

IEEE Journal of Selected Topics in Signal Processing, Vol. 6, No. 5, pp. 538-552, May, 2012.

[23] D. Weinland, R. Ronfard, E. Boyer, Free Viewpoint Action Recognition Using Motion History Volumes, *Computer Vision and Image Understanding*, Vol. 104, No. 2, pp. 249-257, November–December, 2006.

[24] S. Y. Cheng, M. M. Trivedi, Articulated Human Body Pose Inference from Voxel Data Using a Kinematically Constrained Gaussian Mixture Model, *Computer Vision and Pattern Recognition (CVPR) 2nd Workshop Evaluation of Articulated Human Motion Pose Estimation*, Los Alamitos, USA, 2007, pp. 1-11.

[25] Y. Sung, R. Choi, Y. Jeong, Arm Orientation Estimation Method with Multiple Devices for NUI/NUX, *Journal of Information Processing Systems*, Vol. 14, No. 4, pp. 980-988, August, 2018.

[26] B. Hwang, S. Kim, S. Lee, A Full-body Gesture Database for Automatic Gesture Recognition, *7th International Conference on Automatic Face and Gesture Recognition, Southampton*, UK, 2006, pp. 1-6.

[27] L. Sigal, A. O. Balan, M. J. Black, HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion, *International Journal of Computer Vision*, Vol. 87, No. 1-2, pp. 4-27, March, 2010.

[28] C. Li, J. Zhang, Y. Luo, Cloud-based Mobile Service Provisioning for System Performance Optimisation, *International Journal of Ad Hoc and Ubiquitous Computing*, Vol. 29, No. 3, pp. 193-207, October, 2018.

[29] M. Saad, An Improved Hybrid Genetic Algorithm for Multi-user Scheduling in 5G Wireless Networks, *International Journal of Internet Protocol Technology*, Vol. 11, No. 2, pp. 63-70, June, 2018.

[30] Y. Sung, K. Cho, Development and Evaluation of Wireless 3D Video Conference System using Decision Tree and Behavior Network, *EURASIP Journal on Wireless Communications and Networking*, Vol. 5, No. 1, pp. 1-14, December, 2012.

[31] Y. Sung, K. Cho, Data Generation and Representation Method for 3D Video Conferencing using Programming by Demonstration, *Multimedia Tools and Applications*, Vol. 67, No. 1, pp. 71-95, November, 2013.

[32] H. Park, C. Jun, A Simple and Fast Algorithm for K-medoids Clustering, *Expert Systems with Applications*, Vol. 36, No. 2, pp. 3336-3341, March, 2009.

Biographies

Yunsick Sung is an Assistant Professor of Department of Multimedia Engineering, Dongguk University-Seoul, Republic of Korea. He is a managing editor in Journal of Information Processing Systems from 2017 and is a managing editor in Human-centric Computing and Information Sciences from 2015.

Jeonghoon Kwak is a doctoral student in Dept. of Multimedia Engineering, Dongguk University-Seoul, Republic of Korea. He received the BS degree in Game Mobile Contents from Keimyung University, Republic of Korea, in 2015 and the MS degree in Computer Engineering from Keimyung University, Republic of Korea, in 2017.

Jong Hyuk (James J.) Park is currently a Professor with the Department of Computer Science and Engineering and the Department of Interdisciplinary Bio IT Materials, Seoul National University of Science and Technology, Republic of Korea. He is a member of the IEEE Computer Society, KIPS, and KMMS.