

RESEARCH

Open Access

# Development and evaluation of wireless 3D video conference system using decision tree and behavior network

Yunsick Sung<sup>1</sup> and Kyungeun Cho<sup>2\*</sup>

## Abstract

Video conferencing is a communication technology that allows multiple users to communicate with each other by both images and sound signals. As the performance of wireless network has improved, the data are transmitted in real time to mobile devices with the wireless network. However, there is the limit of the amount of the data to be transmitted. Therefore it is essential to devise a method to reduce data traffic. There are two general methods to reduce data rates: extraction of the user's image shape and the use of virtual humans in video conferencing. However, data rates in a wireless network remain high even if only the user's image shape is transferred. With the latter method, the virtual human may express a user's movement erroneously with insufficient information of body language or gestures. Hence, to conduct a video conference on a wireless network, a method to compensate for such erroneous actions is required. In this article, a virtual human-based video conference framework is proposed. To reduce data traffic, only the user's pose data are extracted from photographed images using an improved binary decision tree, after which they are transmitted to other users by using the markup language. Moreover, a virtual human executes behaviors to express a user's movement accurately by an improved behavior network according to the transmitted pose data. In an experiment, the proposed method is implemented in a mobile device. A 3-min video conference between two users was then analyzed, and the video conferencing process was described. Photographed images were converted into text-based markup language. Therefore, the transmitted amount of data could effectively be reduced. By using an improved decision tree, the user's pose can be estimated by an average of 5.1 comparisons among 63 photographed images carried out four times a second. An improved behavior network makes virtual human to execute diverse behaviors.

**Keywords:** video conferencing, chat system, virtual human, decision tree, behavior network

## 1. Introduction

Video conferencing has widely been used in public organizations and private companies. However, communication problems due to increased data traffic may occur if many users are connected simultaneously [1]. Hence, one strategy to ensure that many users are connected at the same time is to reduce the amount of data traffic.

There are at least two approaches to reduce data traffic in video conferencing. One is to extract the shape of the user when images are captured [2-4]. The shapes extracted from multiple users are then arranged in a

three-dimensional (3D) virtual environment to reconstruct a virtual conference space. A user is readily identified because actual human images are shown, as in this study [2]. However, data sent by one user are delivered to multiple other users at the same time. Hence, if more users are connected, the data traffic increases accordingly. Given that multiple images are transmitted in real time, there would be too much data to transmit on a wireless network.

Another approach to reduce the data traffic is to extract and send the physical location and features of a user and reconstruct it in the virtual environment [1,5]. This approach is advantageous in that it expresses the gestures and body language of users by using their physical location and features [1]. For example, there is a

\* Correspondence: cke@dongguk.edu

<sup>2</sup>Department of Multimedia Engineering, Dongguk University, 26, Pil-dong 3-ga, Jung-gu, Seoul 100-715, Korea

Full list of author information is available at the end of the article

method that provides distance consulting services by calculating a depth map on an image to represent a speaker in three dimensions [5]. In other studies, a speaker's body position has been used to control virtual humans by using a body-tracking system that recognizes skin color in 2D images [1]. However, in these studies, it was difficult to express the movements of a virtual character with only partial data on the speaker's body. Hence, the positions of unavailable body parts were estimated with inverse kinematics. This makes it difficult to express a speaker's motions precisely.

In studies on data reduction in video conferencing, a common problem is that of low quality of service (QoS). In particular, when multiple users are connected at the same time, the amount of data that a user can transmit simultaneously on a wireless network becomes relatively small compared to that on a physically wired network. Hence, it is necessary to improve the QoS of video conferencing.

In this article, a framework that enables video conferencing by multiple users on a wireless network is proposed. To reduce data traffic, a user's pose is first recognized through a binary decision tree and transmitted by using the markup language. Next, a method based on a behavior network is introduced to express the movements of the virtual human precisely. Subsequently, the proposed method is implemented in an experiment and verified in a mobile device. The proposed method involves multiple users communicating with each other using a mobile device. Therefore, it is applicable to various forms of communication such as chatting, gaming, and video conferencing.

The rest of the article is organized as follows. In Section 2, we introduce a method to reduce data traffic in video conferencing. In Section 3, we propose a video conferencing framework. In Section 4, we describe a series of processes to implement the proposed framework in a mobile device and control a virtual human. In Section 5, we summarize the proposed method and discuss future directions for research.

## 2. Related study

In a video conference, the amount of photographed images increases in proportion to the number of connected users. Therefore, even if the images are compressed, all images cannot be transmitted on a wireless network. To make wireless video conferencing possible, it is necessary to solve data traffic in advance. In this section, we introduce studies on video conferencing, and examine research results that could be adopted to reduce the amount of data.

The following are studies in which the aim was to extract a user's shape from an image to reduce data traffic. The Virtual Team User Environment (VIRTUE)

provides an environment in which multiple users can communicate with each other at the same time [2]. A virtual conference space was constructed and integrated with a 3D environment after receiving all users' shape images. The position of the virtual human in the virtual environment was calculated through the photographed images. In other studies, methods to improve the speed of the VIRTUE have been proposed as well [3,4]. Although not directly related to video conferencing, there have been some studies on the reconstruction of 3D shape [6,7]. In these studies, reconstruction was performed as follows. First, the background was removed from images that were photographed with multiple cameras. The objects were then extracted from the images and the 3D shape created. Lastly, the 3D shape was colored. However, to apply the photographed images and virtual environment to a wireless network at the same time, further data reduction is necessary.

Given that images increase data traffic, there have been a number of studies on the reconstruction of video conferences by extracting and transmitting only a user's features from photographed images. For example, it has been shown that medical advice can be obtained through a telemedicine system to perform an operation [5]. Here, a depth-map was extracted from the photographed images, and then, a distance user was represented. In other studies on virtual humans for video conferences, body features were extracted from photographed images [1]. After locating the body positions by identifying the hands and face, the body positions were transmitted. Then, the virtual human gestures by referencing the transmitted body positions.

However, it is difficult to express exact gestures due to insufficient data on body features. Further data are required to make the virtual human act naturally. Facial images and body features can also be extracted from photographed images. The extracted facial images are mapped onto the face of the virtual human after analyzing facial features. By extracting and transmitting only faces, data traffic is reduced. Actual human faces are used in this method. This makes it easier to distinguish real users from virtual humans.

Lastly, there have been studies on the virtual meeting room [8,9]. In these studies, the photographed images were converted into silhouettes for comparison with pre-defined models [10]. The poses were estimated, and deictic motions were expressed using the silhouette. Data traffic can be reduced because the data are in xml format. However, the problem is to define a model in person using a tool. In addition, it takes time to estimate poses if there are many models. To reduce the amount of estimation, decision tree [11] can be applied. If silhouettes which should be compared are classified, the number of comparison could be reduced.

If only data traffic is considered, studies in which the actions of a user are expressed through the virtual human by transmitting the user's features are more appropriate in a wireless network environment than the method of extracting the user's image shape. In these studies, however, the data are insufficient to describe a user's body. Therefore, it is necessary to devise a method to express the body more precisely. Moreover, to make virtual human to act naturally, behavior network [12] could be applied. Behavior network selects the behavior for virtual human considering goals and pre-executed behaviors. Therefore, virtual human can execute behaviors naturally. By defining the behaviors of the virtual human in advance and using behavior network, this article proposes a method in which body motions can be freely expressed even with little data traffic.

### 3. 3D video conference framework

To express a user's movements by a virtual human, it is necessary to devise a method to extract and transmit a user's features and then reconstruct the virtual conference by virtual humans. This section describes a method to estimate user's pose and control a virtual human with the estimated pose.

#### 3.1. Overview

The proposed virtual human-based video conference framework consists of a definition stage that predefines the data for video conferencing, a recognition stage that extracts pose data from the images, and a reconstruction stage that reconstructs the virtual conference (Figure 1). The definition stage is only performed once when a video conference is started, whereas the recognition and reconstruction stages are performed repeatedly during a video conference.

Data are determined in definition stage as follows. First, to estimate a user's pose, the necessary images must be defined. This requires the generation of images for a user's expected pose photographed by camera. The generated images are then compared with real-time photographed images of the user to estimate the user's pose. However, pose estimation will be time-consuming if the number of expected poses is excessive. Thus, the pose-estimation time can be reduced by using only a subset of expected pose images by constructing a binary decision tree (referred to as the pose decision tree)

Next, a behavior network is defined to generate the behavior that is executed by a virtual human. At first, an action is defined by a virtual human's joint angles, and a behavior is expressed by its actions. Selected behaviors are executed by virtual human. The pose images, pose decision tree, motions, actions, and consecutive action network are shown in Figure 1. These are

all defined in the definition stage. This behavior network is referred to as the consecutive action network.

In the recognition stage, images are created by photographing users at certain intervals. A user's poses are estimated by comparing the photographed images with the pose decision tree. The estimated poses are then transmitted to other users through the network. In the reconstruction stage, a user's presence is expressed in a virtual human by considering the estimated pose.

#### 3.2. Framework structure

In this section, we propose a framework that expresses a user's presence through a virtual human in a video conference. The framework that handles video conferences is structured as shown in Figure 2.

The recognition stage converts the estimated pose into markup language, which is transmitted to the network as follows. The photographed images are received by the image receiver and sent to the background learner and silhouette extractor. Background learner acquires backgrounds when the user is absent and then transfers the background image to the silhouette extractor. Subsequently, the silhouette extractor extracts the shapes of users from the received images by considering the background images and transmits them to the pose estimator. The pose estimator searches the pose decision tree and estimates the poses of the received images. The estimated poses are then transmitted to the network through the message generator and message sender (the former creates messages and the latter transmits them to other users). Each message contains a user's pose and speech.

The reconstruction stage then creates the image and voice from the received messages as follows. The message receiver transmits the pose and speech to the behavior planner and speech generator, respectively. The behavior planner plans the behaviors to be executed by the virtual human. The virtual human controller then executes the planned behaviors.

##### 3.2.1. Image receiver and silhouette extractor

In the recognition stage, the images are created by photographing users at certain intervals. The image receiver receives the photographed user's images and transmits them to the silhouette extractor. The  $h$ th user-image is defined as  $i^h$ , as shown in Equation (1). The set of user-images is defined as Set  $I$ :

$$i^h \in I, I = \{i_1, i_2, \dots\} \quad (1)$$

Here, the image interval is denoted as  $\varepsilon_{\text{Interval}}$ . To estimate the poses precisely, the user-images are converted into silhouettes like silhouette extraction process [10], as shown in Figure 3.

A silhouette is an image with only a user's shape without the background. The background images are

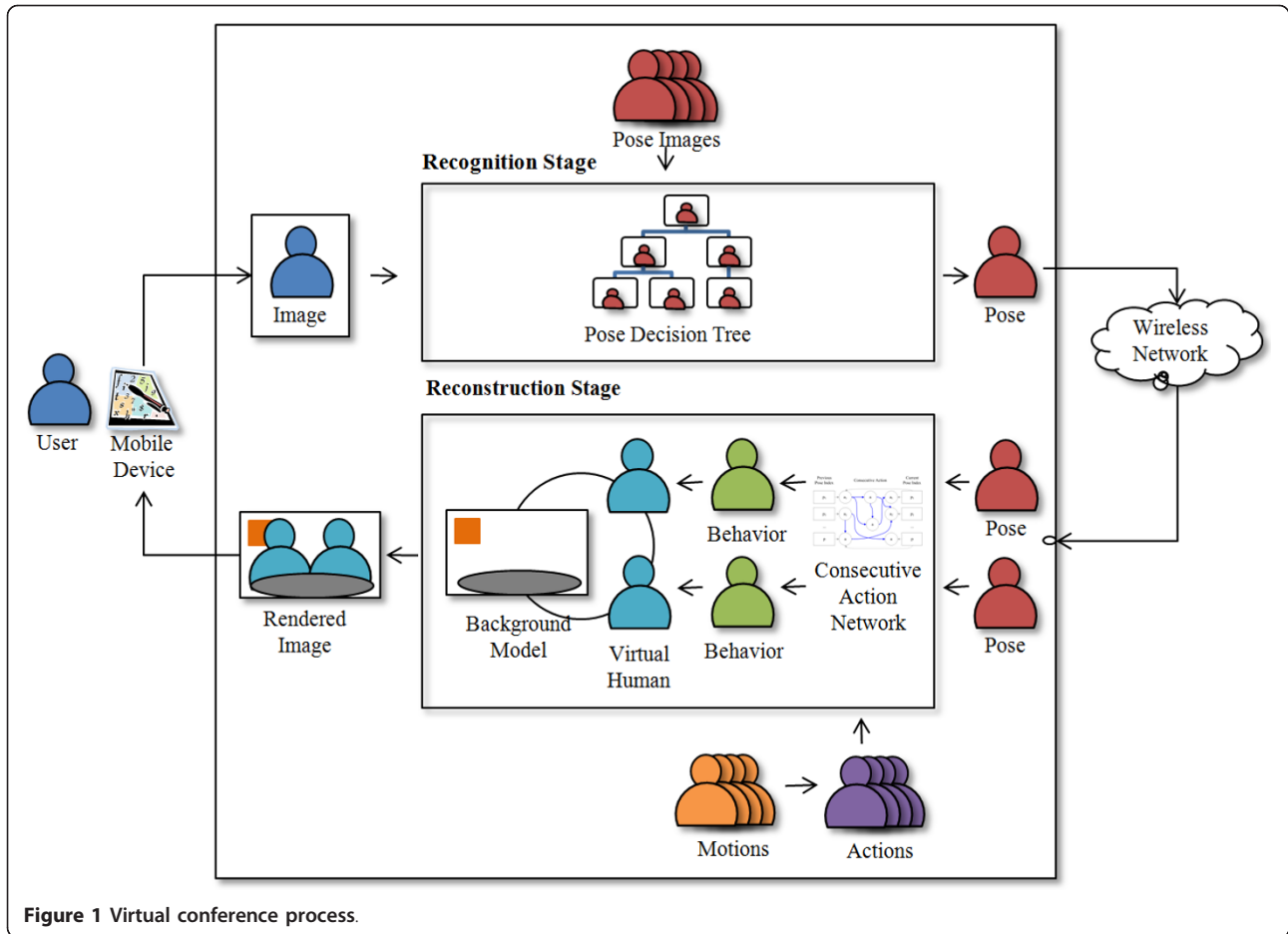


Figure 1 Virtual conference process.

recorded by the background learner in definition stage and then transferred to silhouette extractor to remove the background from the user-images. Then, the user-silhouette is extracted from the difference between the recorded background image and the user-image. The  $h$ th extracted user-silhouette from the  $h$ th user-image is defined as  $s^h$ , as shown in Equation (2). The set of user-silhouettes is defined as Set  $S$ :

$$s^h \in S, S = \{s_1, s_2, \dots\} \quad (2)$$

### 3.2.2. Pose decision tree and pose estimator

The pose estimator, which estimates poses with the extracted silhouette in the recognition stage, must recognize multiple poses in real time in a mobile environment. However, the time to estimate poses increases with the number of poses because the number of comparisons also increases. To solve this problem, we propose a pose decision tree.

In the definition stage, the expected pose images of users are predefined to construct the pose decision tree in advance. First of all, the set of all expected pose is defined as the Set  $P$ .

$$p^i \in P, P = \{p_1, p_2, \dots\} \quad (3)$$

The set of expected pose images, Set  $E$ , is defined as shown in Equation (3) to estimate the pose of the extracted silhouette.  $e^i$  is the image that is used to estimate pose  $p^i$ .

$$e^i \in E, E = \{e_1, e_2, \dots\} \quad (4)$$

The expected pose images are also converted into expected silhouettes. The set of expected silhouettes is defined as Set  $R$ , where  $r^i$  is the  $i$ th silhouette expected.

$$r^i \in R, R = \{r_1, r_2, \dots\} \quad (5)$$

The pose decision tree consists of nodes that contain each expected silhouette. The  $i$ th node  $n^i$  is defined as shown in Equation (5):

$$n^i = \langle r^i, n_{Left}^i, n_{Right}^i, m^i, v^i \rangle \quad (6)$$

where  $n_{Left}^i$  and  $n_{Right}^i$  are the left and right nodes of node  $n^i$ , respectively; and  $m^i$  and  $v^i$  are the matching value and center value of node  $n^i$  (range 0 to 1),

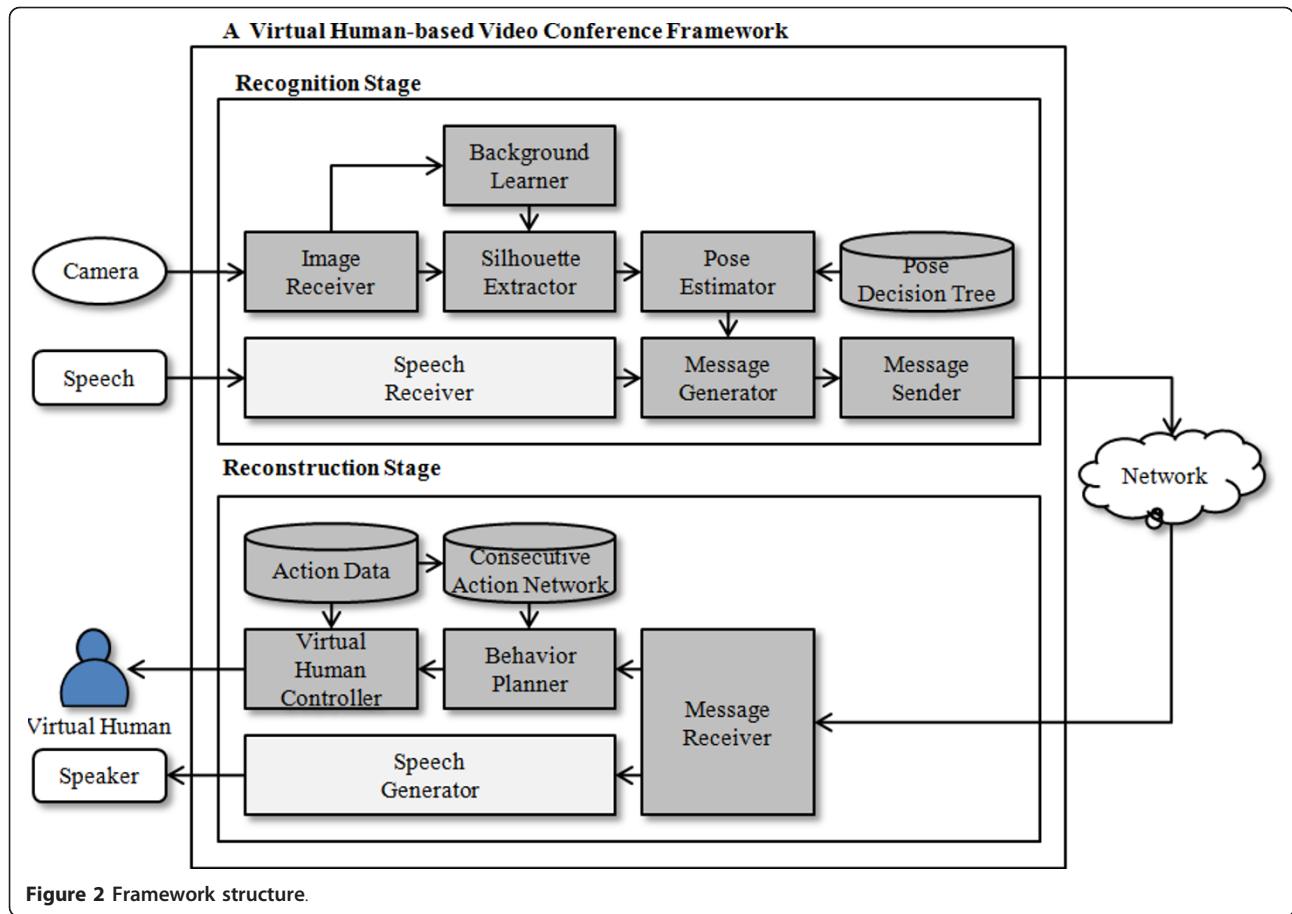


Figure 2 Framework structure.

respectively. The matching value indicates the similarity of two silhouettes. For example, if its value is 1, the silhouettes are considered identical only when they have exactly the same images. In contrast, if its value is 0, the silhouettes are considered identical regardless of their differences. The matching value is determined to estimate the pose by establishing various values. The center value, which expresses a standard based on a search of the left and right child nodes, is automatically established when the pose decision tree is constructed. As shown in Equation (6), there is a one-to-one relation between pose  $r^i$  and node  $n^i$ .

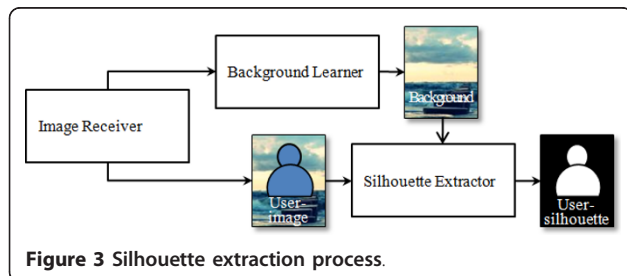


Figure 3 Silhouette extraction process.

The decision tree is constructed as follows. First, nodes are created for all expected silhouettes included in Set  $R$ . Second, node  $n^1$  is defined as the root node. The remaining nodes in Set  $R$  are then registered as the child nodes of  $n^1$  (Figure 4).

Third, the child nodes of  $n^1$  are sorted after comparing the expected silhouette  $e^1$  of the root node to that of the child node (Figure 5).

As shown in Equation (6), the comparison is expressed as a normalized value after calculating the

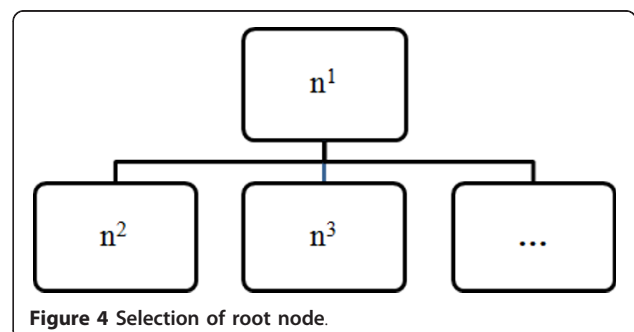


Figure 4 Selection of root node.



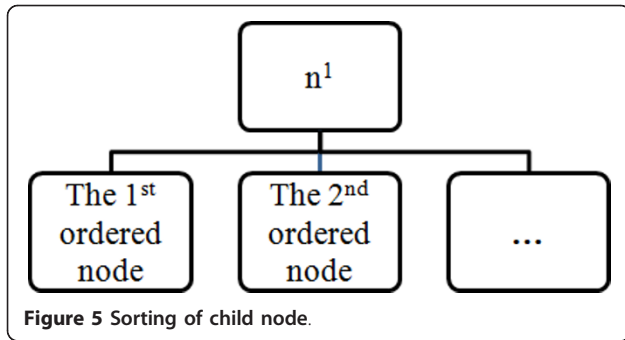


Figure 5 Sorting of child node.

correlation coefficient of the two expected silhouettes. The value ranges from 0 to 1.

$$R(r_1, r_2) = \frac{\sum_{x_1, x_2} (T(x_1, x_2) \cdot I(r_1 + x_1, r_2 + x_2))}{\sqrt{\sum_{x_1, x_2} T(x_1, y)^2 \cdot \sum_{x_1, x_2} I(r_1 + x_1, r_2 + x_2)^2}} \quad (7)$$

Fourth, when there are  $o$  children, the  $\lfloor \frac{o+1}{4} \rfloor$ th and  $\lfloor \frac{(o+1) * 3}{4} \rfloor$ th nodes are defined as the left and right nodes, respectively. The node whose index is equal to or smaller than  $\lfloor \frac{o+1}{2} \rfloor$  in terms of the sorting sequence moves to the left node, whereas the node greater than  $\lfloor \frac{o+1}{2} \rfloor$  moves to the right node (Figure 6).

Fifth, the mean of the correlation coefficients between the last node on the left and the first node on the right is set to the center value of the root node. Lastly, both left and right nodes sort the child nodes through repetitive comparisons just as in the case of the root node.

In the recognition stage, the pose decision tree is used as follows. The user-silhouette is compared to the silhouette of the root node. If the correlation coefficient of two silhouettes is equal to or greater than  $\epsilon_{PoseMatching}$ , the index of the root node is transmitted to the message generator. Otherwise, the user-silhouette is compared to

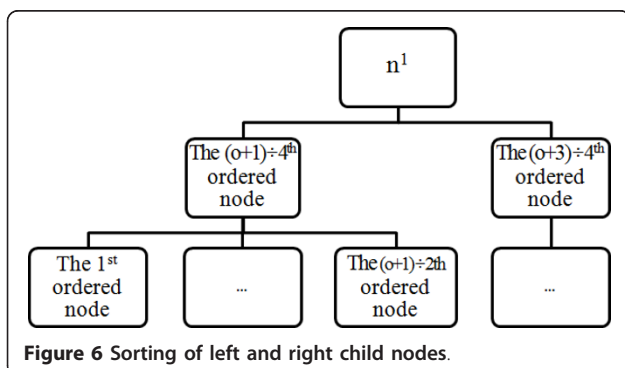


Figure 6 Sorting of left and right child nodes.

left child node of the root node. If it is greater than the center value, the user-silhouette is compared to the right child node. Therefore, the comparison of nodes continues until the correlation coefficient of the two silhouettes is over  $\epsilon_{PoseMatching}$  or the terminal node is reached. The index of the node that is ultimately reached is also transmitted to the message generator.

### 3.2.3. Action, consecutive action network and behavior

In the definition stage, the actions to be executed by a virtual human are defined. Action is the movement for virtual human to express pose when pose index is received shown in Equation (8).

$$a^j = \langle p^j, d^j, c_1^j, c_2^j, \dots \rangle \quad (8)$$

where  $p^j$  is the pose that would be expressed by the  $j$ th action, and  $d^j$  is the duration of the  $j$ th action. In addition,  $c_1^j$  is the first joint angle required for the virtual human to execute the  $j$ th action  $a^j$ .

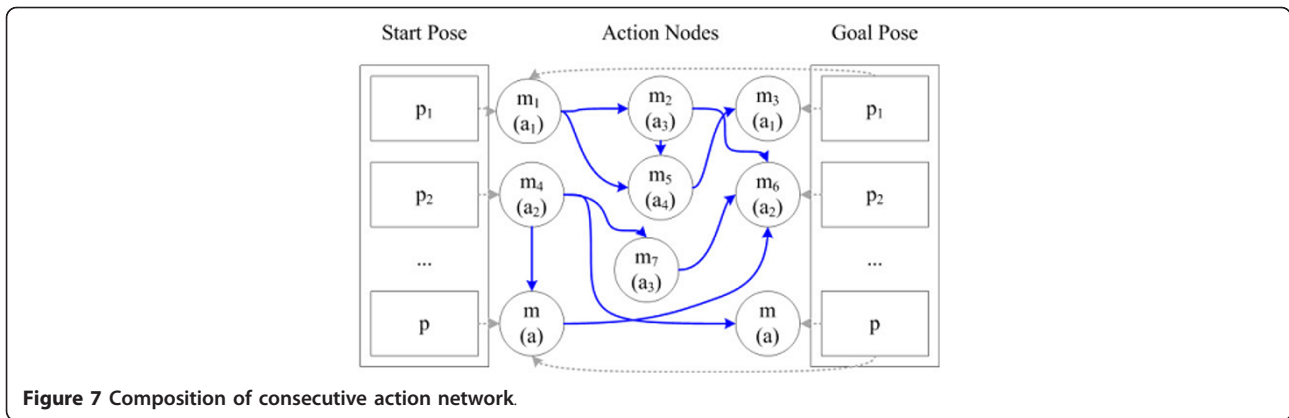
$$a^j \in A, A = \{a_1, a_2, \dots\} \quad (9)$$

If an action is defined, and based on this, then the network is defined in order to select and execute actions consecutively whenever every pose index is received. Behavior planner defines the network to execute consecutive actions in definition stage as follows. First, start-poses and goal-poses are placed. Start-poses are the starting poses to generate consecutive actions. It is the first pose for consecutive actions. Goal-poses are the targeted poses. The pose for the last action among the generated consecutive actions becomes to be the goal-pose. Hence, all the poses of Set  $P$  are placed on both sides of the network as shown in Figure 7.

Next, all the actions of Set  $A$  are placed, and the sequences of consecutive actions are expressed as a tree by using directed acyclic graph (DAG). The action nodes which contain one action of Set  $A$  are primarily defined, and it is placed between start-poses and goal-poses, as shown in Figure 7. After action node is arranged, then DAG connects each action node. To prevent tree's containing any loop, an action node which has identical action, is repeatedly defined like action  $a_1$  of Figure 7.

Next, the transition probability of all DAG is defined. The sums of probabilities transit from each action to another action are normalized to be 100.

Finally, start-pose is connected with the actions which are executable at first after receiving the pose index. Goal-pose is also connected with the actions which are executable lastly. In Figure 8, start-poses  $p_1$  and  $p_2$  are connected to each corresponding actions,  $a_1$  and  $a_2$ . Among goal-poses,  $p_1$  is connected with two action nodes,  $m_1$  and  $m_3$ , more than one. Two actions of two



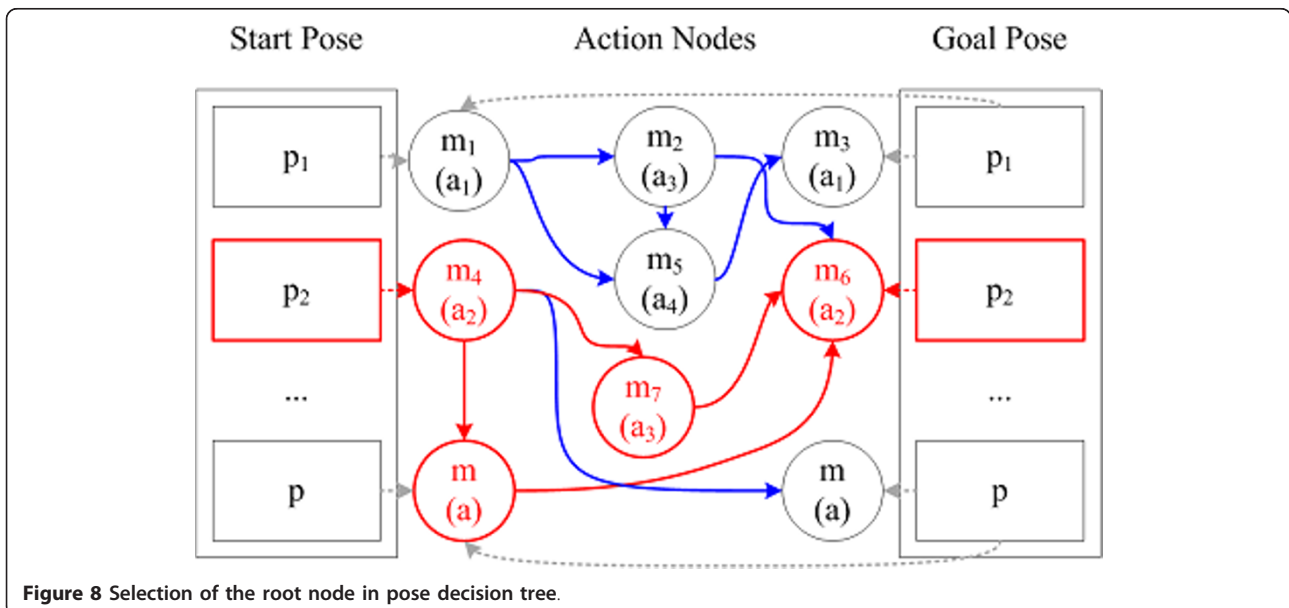
nodes can be executed lastly. Therefore, two nodes are connected to goal-pose  $p_1$ . In the case of  $p_2$ , however, one node between two action nodes containing  $a_2$  is only connected. Action node composing network is defined as show in Equation (10).

$$m^k = \langle a^j, s^k, g^k, o_1^k, o_2^k, \dots, q_1^k, q_2^k \rangle \quad (10)$$


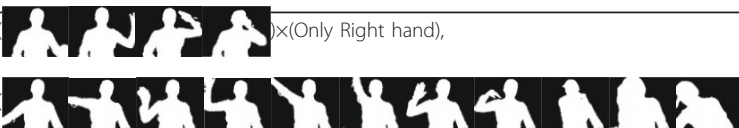


where  $m^k$  is a action node that contains  $a^j$ .  $s^k$  and  $g^k$  represent the index of start-pose and goal-pose connected with action node  $m^k$ .  $o_x^k$  means the other action nodes to which action node  $m^k$  is connected in  $x$ th.  $q_y^k$  is the probability of transition to  $o_x^k$ . The network composing of action nodes is defined as the consecutive action network.

The defined consecutive action network is used when an action is selected in the reconstruction stage. By using the pose index received in time  $t - 1$  as start-pose index and the pose index received in time  $t$  as goal-pose

index, behavior planner generates consecutive actions as follows. First, among the several numbers of start-poses, the pose in the pose index received in just time  $t - 1$  is selected. Next, among the several numbers of goal-poses, the pose in pose index received in just time  $t$  is selected. Next, all action nodes and connections which is movable from the selected start-pose to the selected goal-pose, is selected. Next, the transition probability for the selected nodes besides each action node is normalized up to be 100. Next, among the selected connections, one connection is selected through the probability. If the only one action node is connected it can be directly selected. The selection of connections is repeatedly processed until the action node which is connected with the goal-pose is reached. Finally, the consecutive actions are constructed by connecting the actions in all the visiting action nodes, and are defined as a behavior. The defined behavior is then executed.



**Table 1 Poses used in pose decision tree**

Front (4)		)(Front Direction)
One hand (26)		)(Only Right hand), )(Right +Left hand)
Two hands (23)		)(Front Direction),, )(Right +Left Direction)
Head (10)		)(Right +Left Direction)

For example, behavior is generated as shown in Figure 8 when pose index 2 is received at time  $t - 1$  and time  $t$  from Figure 7. Each  $m_4$  and  $m_2$  is activated by connecting to start-pose and to goal-pose. The other activated two action nodes,  $m_7$  and  $m$ , exist in the connection from  $m_4$  to  $m_6$ . From the consecutive action network, action nodes according to the connections from  $m_4$  are visited as the connection of  $m_4 \cdot m_7 \cdot m_6$  or  $m_4 \cdot m \cdot m_6$ . Therefore, the behaviors having the orders of  $a_2 a_3 a_2$  or  $a_2 a a_2$  are generated and executed.

#### 4. Experiment

To verify the proposed virtual human-based video conference framework, we carried out an experiment by establishing a conference system in iPad. We introduced a method to define the pose decision tree and behaviors for a video conference. We then verified this series of processes.

##### 4.1. Verification of definition stage

The proposed framework requires the definition of a pose decision tree and a consecutive action network. The pose decision tree requires expected poses and silhouettes for the recognition stage, and the consecutive action network requires actions and behaviors for the reconstruction stage.

##### 4.1.1. Expected poses, silhouettes, and pose decision tree

In this experiment, 63 poses were used on consideration of the comparison speed of the video conference system in iPad. Table 1 shows the expected silhouettes of the pose decision tree. Generally, a user participated in the video conference sitting down. Therefore, only the user's upper body was photographed. The silhouette poses comprising the pose decision tree included body-moving, one arm-moving, two arm-moving, and head-moving poses. Moreover, almost all poses were photographs of the front side.

The pose decision tree was constructed by using 63 expected silhouettes according to the method proposed. The node  $n^1$ , which contains the expected silhouette  $r^1$ , was defined as the root node. The rest of the nodes were added to the root node (Figure 9).

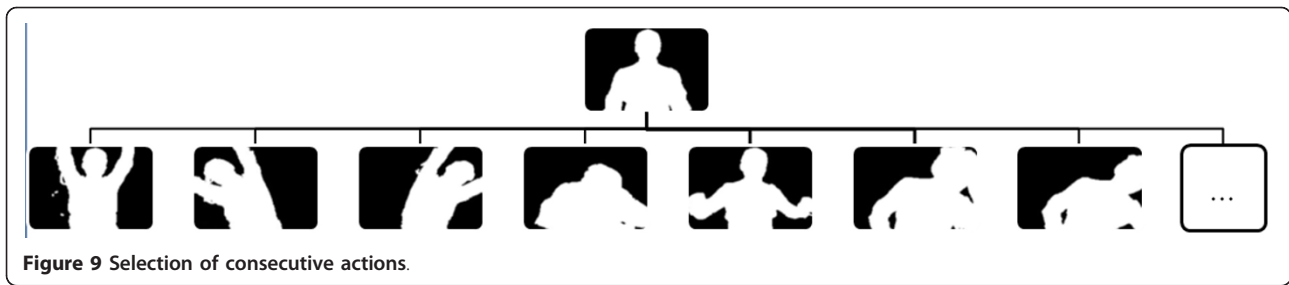
After sorting the nodes according to the correlation coefficients computed in relation to  $r^1$ , the 48th and 12th nodes were registered as the left and right child nodes of the root node, respectively. All nodes for which the correlation coefficient was equal to or smaller than that of the 31st node were then moved to the left as a child node, and the remaining nodes were registered as right child nodes. Subsequently, the center value of the root node was calculated and set to 0.640616 (Figure 10).

Each child node of the root node also constructs child nodes in the manner of the root node. The remaining nodes in each child node were also equally processed and constructed. Therefore, a pose decision tree constructed into a full binary tree is six levels high. In the pose decision tree in Figure 11, the correlation coefficient  $E_{MatchRate}$  was set to 0.93. The numerical values on each image represent the matching and center values. For example, the root node has a matching value of 0.93 and a center value of 0.640616.

After defining the pose decision tree, in recognition stage photographed images are compared with the pose of the root node. If the match rate was over 0.93, the pose of the photographed image was estimated as the pose of the first node. If the correlation coefficient was 0.93 or below and the match rate was 0.640616 or below, the pose was estimated after searching the child nodes on the left side. In contrast, the right child nodes were searched when the match rate was greater than 0.640616.

In the root node (tree level 1), the pose is estimated with one comparison. At a tree level of 2, the pose is





estimated with two comparisons, as there are two nodes. Therefore, if the pose decision tree in Figure 11 is used, 64 poses are compared about 5.1 times on average.

#### 4.1.2. Actions and consecutive actions

After defining the pose decision tree, the actions for a virtual human were described. The joint angle required for a virtual human to execute actions was defined by using motion data received with motion capture. Table 2 shows a part of the 63 actions.

In the definition stage, the behavior planner constructs consecutive action network required to generate behaviors and execute behaviors in real time during the reconstruction stage. Figure 12 shows a part of the consecutive action network applied in this experiment.

#### 4.2. Verification of recognition stage

After setting the pose decision tree, and consecutive action network, this series of processes was verified as follows. In the experiment, the image receiver photographed images ( $199 \times 144$ ) at 4 fps after setting  $\epsilon_{\text{Interval}}$  to 250 ms. A conference was then held for about 3 min using the implemented program. Table 3 shows images at a middle stage.

Using the photographed images, silhouettes were created by removing the background from the silhouette extractor. Table 4 shows the result of converting the images in Table 3 into silhouettes.

Next, the pose estimator estimated the poses of images using the constructed decision tree. The recognition results for the silhouettes in Table 4 are shown in

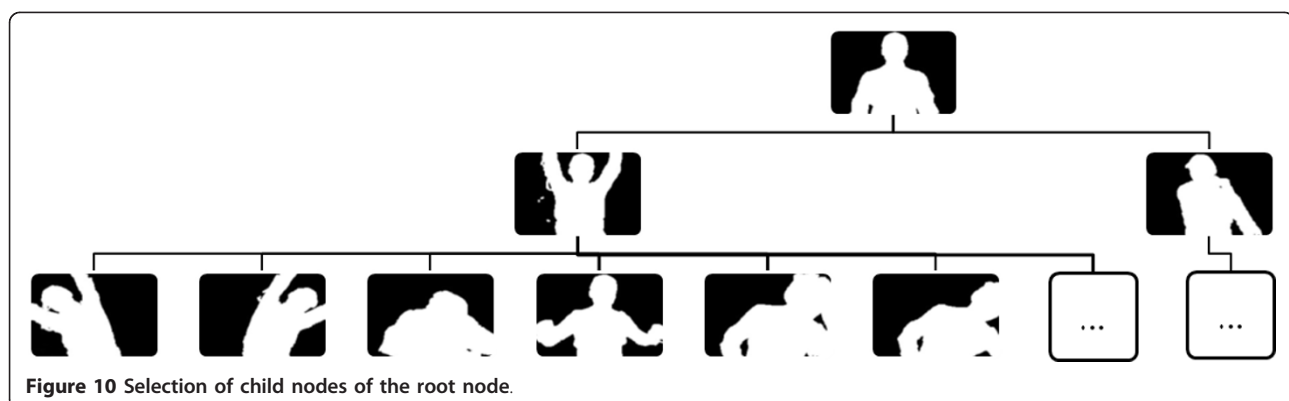
Table 5, where the number next to the image represents the pose index. In Table 5, slightly different motions were found in the eight silhouettes.

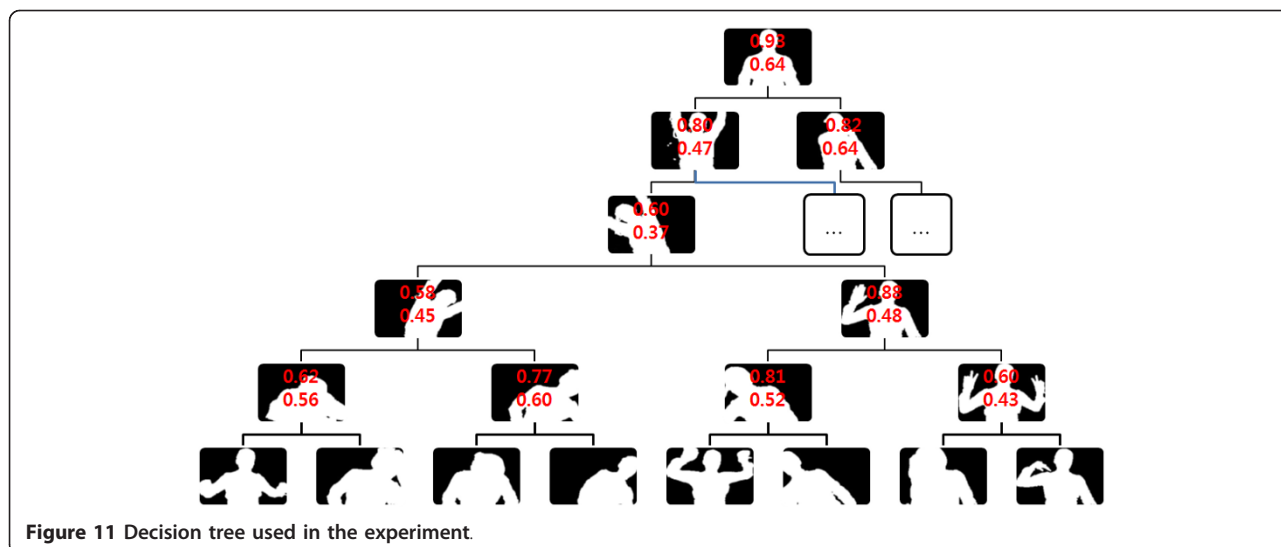
The poses perceived (4 fps) through the pose decision tree were transmitted to another iPad that was connected to the message sender. Figure 13 shows a comparison of data transmission rates when the data were sent with the proposed method and a method to extract the user's image shape [2,6,7].

In Figure 13, BMP and JPG represent data rates in BMP and JPG formats, respectively, without removing the background from the images. When transmitted in BMP format, about 83K occurred four times a second (252K/s on average). If the images were compressed in JPG format, data were transmitted at 122K/s on average. To extract the user's shape, the data were transmitted in BMP or JPG format. Given that BMP remains uncompressed even when the background is eliminated, the data rates were identical even though the background was not removed. In contrast, JPB decreased by approximately 16K (57%). However, there is too much data traffic when multiple users attend the conference. In the proposed method, 111 bytes were transmitted on average at a time (i.e., 444 bytes/s). Hence, data rates were significantly less than when images were transmitted or when user's shape was extracted.

#### 4.3. Verification of reconstruction stage

In reconstruction stage, behavior is generated by receiving the pose index, and executed by virtual human.





**Figure 11** Decision tree used in the experiment

Virtual human selects the consecutive actions with consecutive action network which is defined in Figure 12. Figure 14 shows the activated action nodes and the selected connections when the pose indices of zeroth and eighth poses among start-poses and goal-poses are received. The transition is possible in various directions. Therefore, various behaviors can also be generated depending on the probability of transition.




Figure 15 shows when action nodes are selected from the action of 0th pose to the action of 43th pose. A behavior is constructed by adding three actions to the action corresponding to each pose.

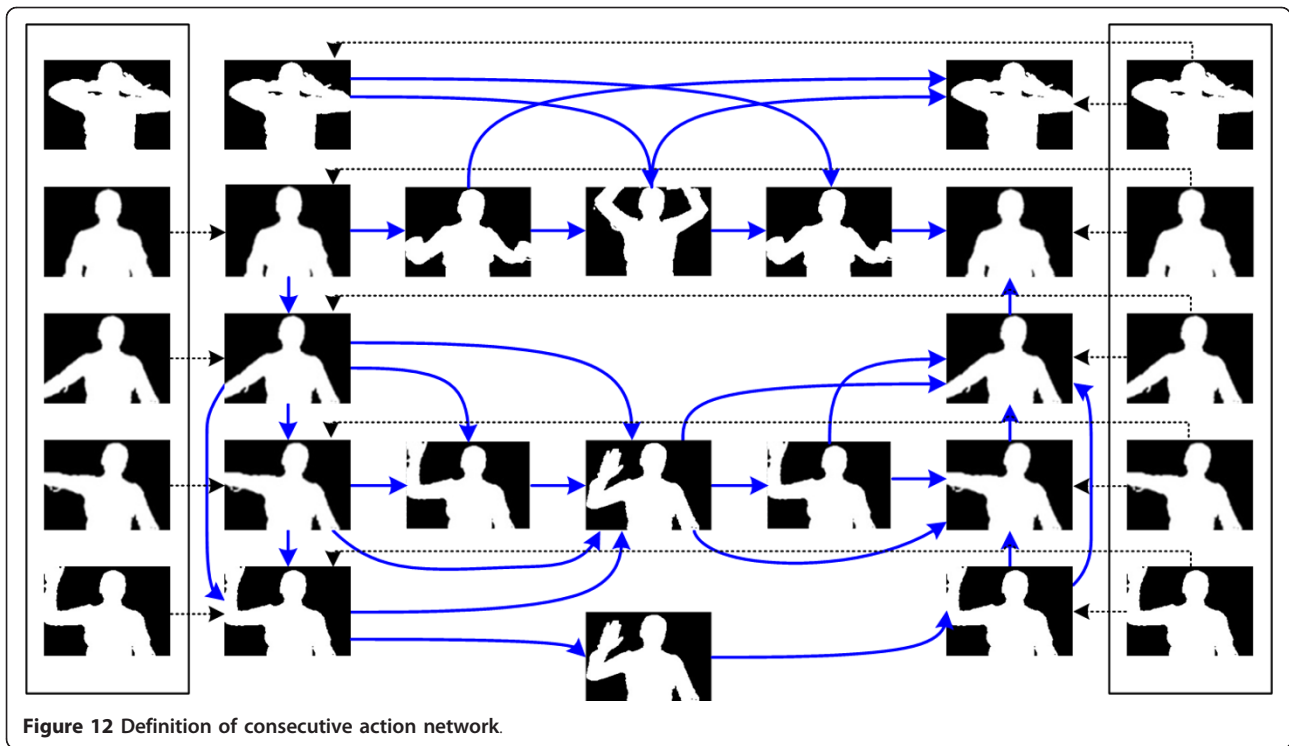
The behavior planner selected the behaviors to be executed in order after receiving the poses. Each image in Table 6 represents silhouettes of poses and actions. In

the first behavior, no previous pose was available. Therefore, no actions were constructed. Whenever receiving pose's index after the first pose index, the constructed consecutive actions are grouped as a behavior.

Once a behavior was constructed, all of its consecutive actions were transferred to the virtual human controller. The behavior was then executed. In the virtual human controller, the behavior was executed during the defined duration of each action. Therefore, the virtual human is operated after referring to actions and searching the joint angle of each action. For example, virtual human selected behaviors at 41000 and 41437.5 ms are shown in Figure 16. Then, the actions from 41000 to 41187.5 ms were executed as one behavior. From 41250 ms, the other behavior was started.

**Table 2** Action definition

Action index/ Pose index (Expected silhouette)	Virtual human's action
0/0 	-2.57,5.25,-4.12,0.55,3.55,-1.79,-0.44,-3.57,1.76,1.94,-0.71,1.56,-0.00,0.00,0.00,0.00, 0.00,0.00,3.35,-10.17,8.05,-2.85,7.04,-5.65,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00, 0.00,6.29,22.05,-1.40,-6.77,-25.64,0.51,0.00,0.00,0.00,0.00,0.00,7.25,2.37,23.87,-6.52,-5.34,-4.30,0.12,-2.41,3.07,4.67,-44.36,19.67,-58.44,4.10,30.43,13.88,6.78,-5.94, 13.11,-15.23,23.36
1/1 	11.59,94.33,-1.39,-4.09,-15.62,-1.90,3.72,15.71,0.82,-17.19,-16.47,-21.16,-0.00,-0.00,-0.00,0.00,0.00,10.91,11.62,3.12,-6.70,4.54,-3.74,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,4.23,13.51,0.51,-0.47,2.20,0.18,0.00,0.00,0.00,0.00,0.00,0.00,23.17,14. 72,12.68,-2.52,1.08,-4.05,-2.99,-15.45,-12.56,2.52,-13.81,26.56,-38.32,13.95,22.65 -13.50,-6.12,-3.80,-5.35,2.99,24.07
...(Omitted)	
62/62 	0.37,95.54,-1.21,0.25,-2.22,0.76,-0.22,2.22,-0.75,-1.28,1.63,2.92,-0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.92,-1.11,-2.19,-1.27,3.29,1.46,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,2.87,-4.64,2.62,-2.80,6.99,-3.14,0.00,0.00,0.00,0.00,0.00,0.00,-101.45,-24.16,3.75, -10.46,-5.15,-1.20,-114.09,-24.17,-47.84,-23.53,2.41,4.81,-16.77,-1.18,16.42,-2.84, 11.37,0.17,5.97,1.69,9.89



### 5. Conclusion

In this article, a framework to display a user as a virtual human in a 3D virtual conference by using pose decision tree and consecutive action network is proposed. The framework consists of a recognition stage that generates messages based on the pose decision tree and a reconstruction stage that controls a virtual human with the consecutive action network.

The recognition stage includes an image receiver, background learner, silhouette extractor, pose estimator, message generator, and message sender. The image receiver transmits the user's images to the silhouette extractor, which then extracts the user's silhouettes from user's images. The pose estimator estimates the poses of user-silhouettes based on the pose decision tree and sends the pose's index to the message generator.

**Table 3** Photographed images

41,000 ms	41,250	41,500 ms	41,750 ms
42,000 ms	43,250	43,500 ms	43,750 ms

**Table 4** Extracted silhouettes

41,000 ms	41,250	41,500 ms	41,750 ms
42,000 ms	43,250	43,500 ms	43,750 ms

**Table 5** Recognition results

41,000 ms	41,250	41,500 ms	41,750 ms
42,000 ms	43,250	43,500 ms	43,750 ms

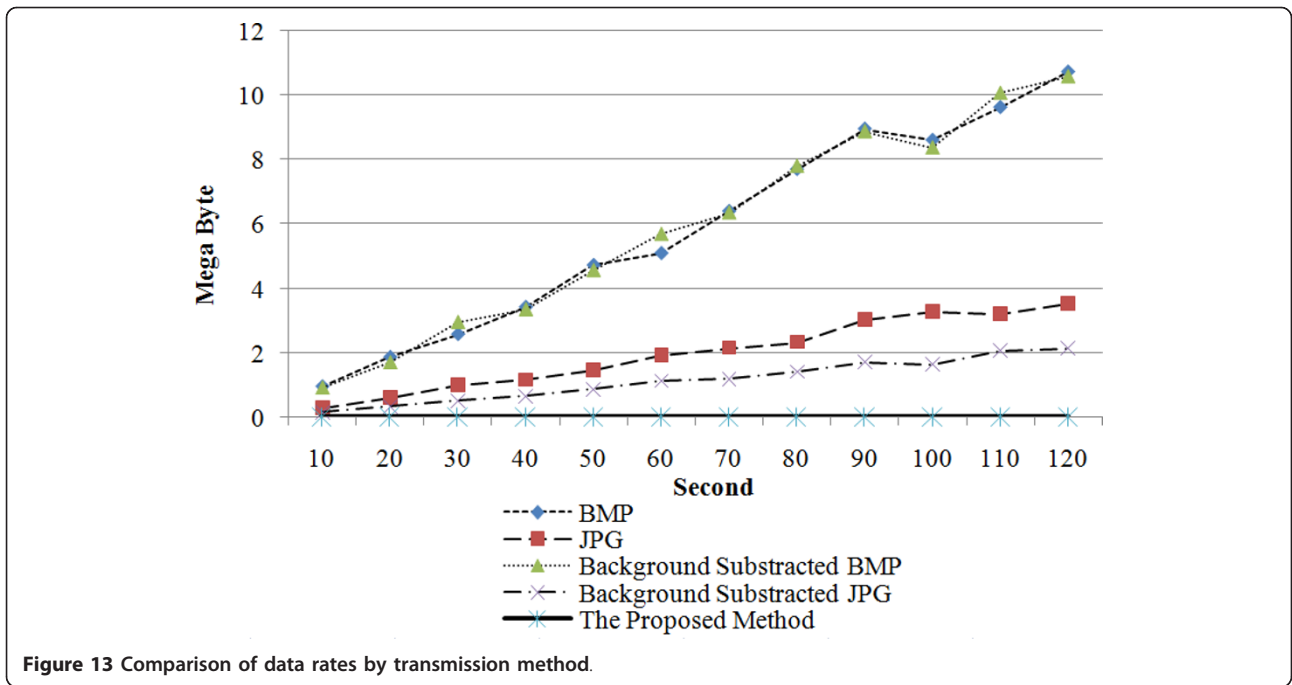


Figure 13 Comparison of data rates by transmission method.

After converting the index of the received pose into markup language, the message generator then delivers the messages to the message sender.

The reconstruction stage consists of a message receiver, behavior planner, and virtual human controller. The message receiver sends the index of the pose to the

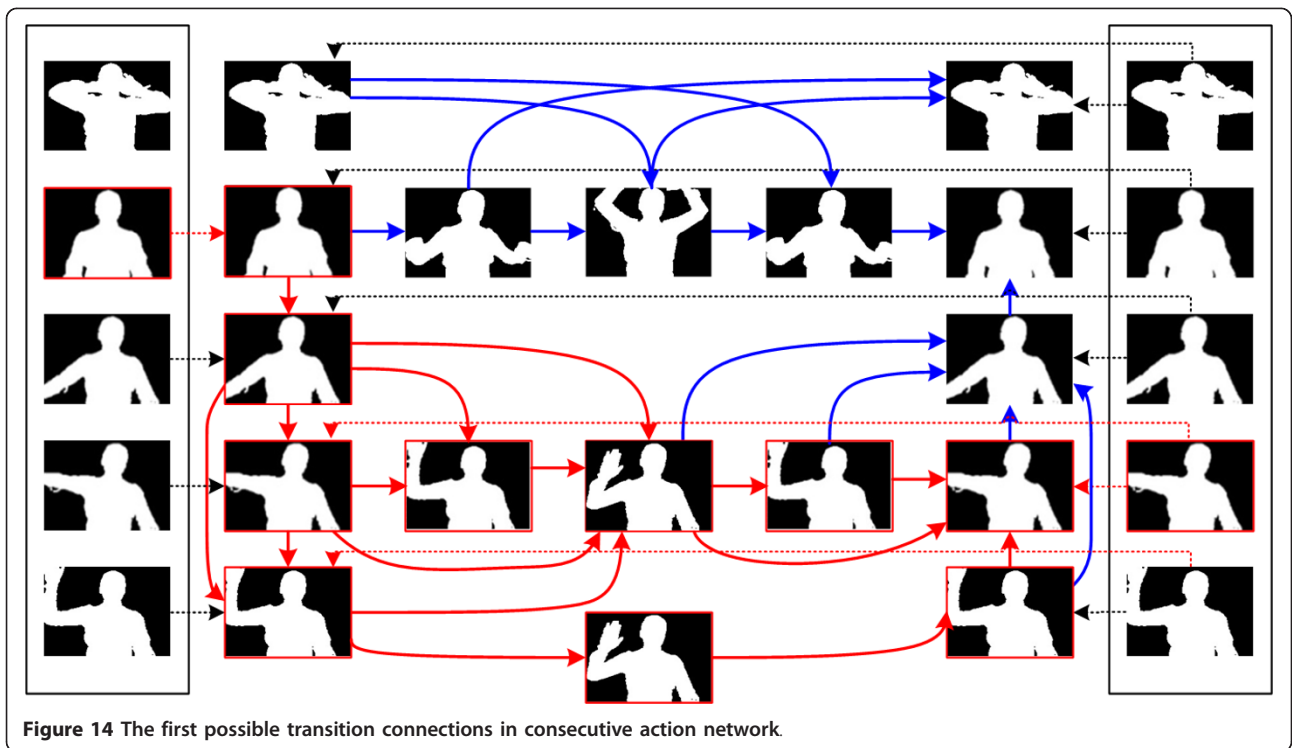
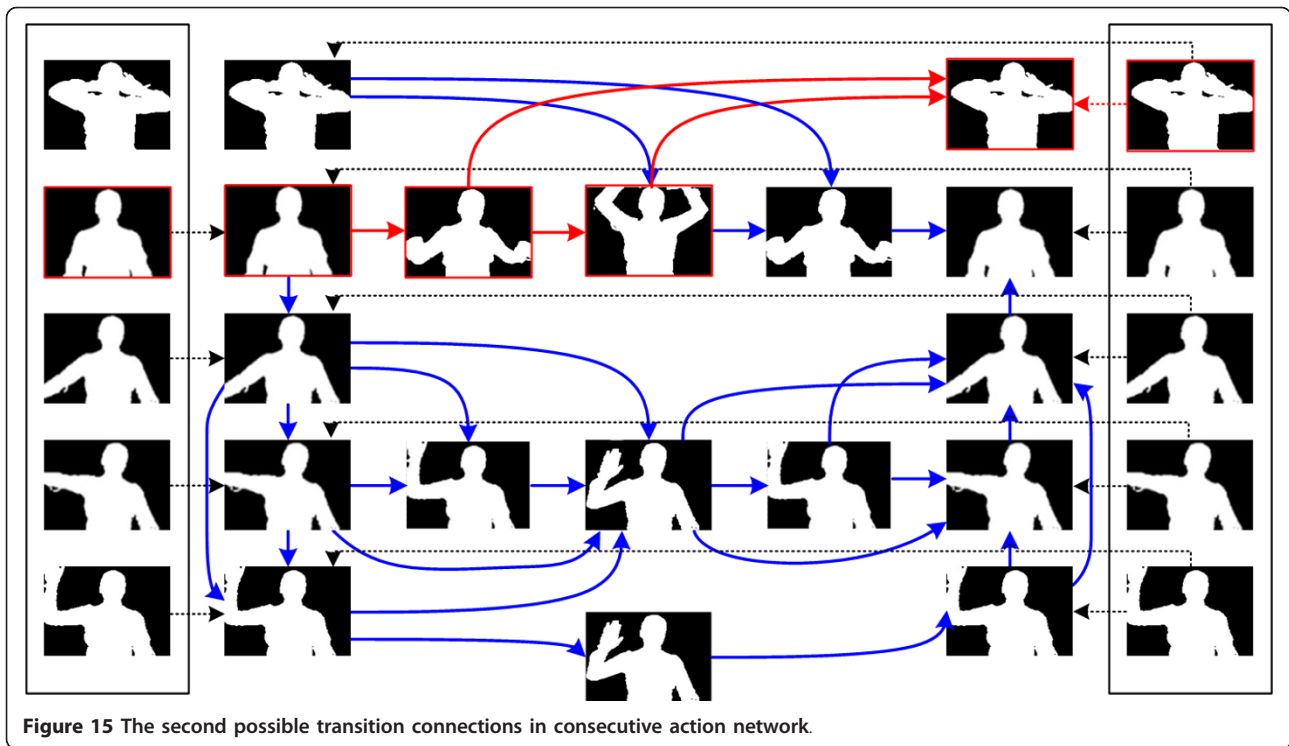


Figure 14 The first possible transition connections in consecutive action network.



**Table 6** Selected consecutive actions

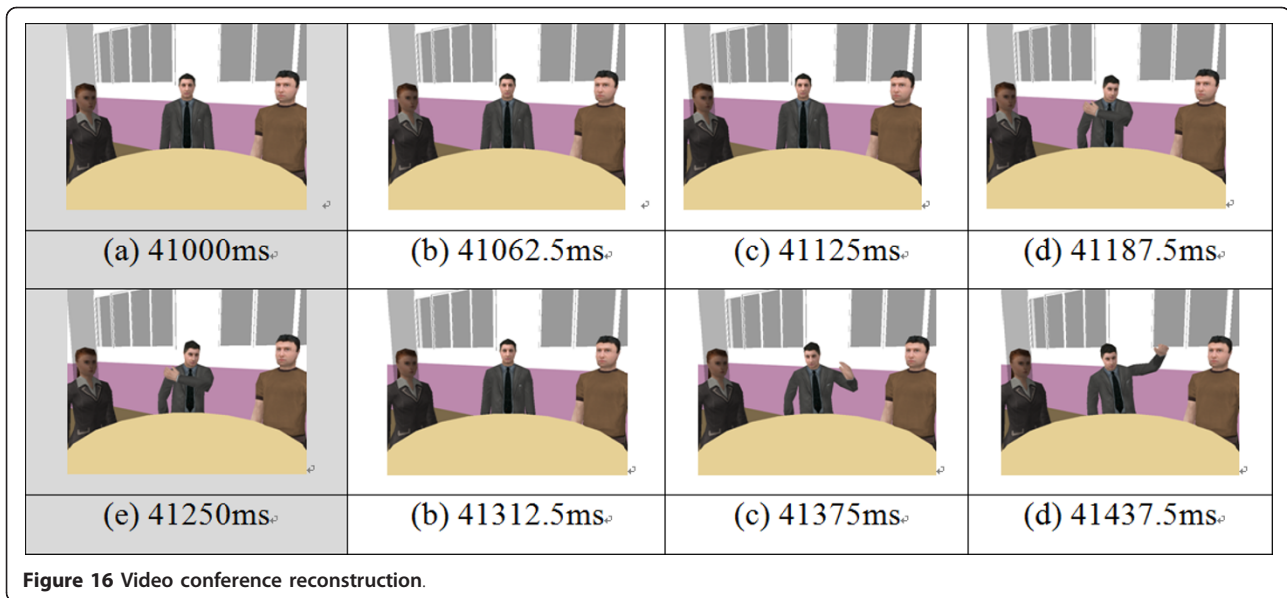
Received pose's silhouette (ms)	Generated actions
41,000	
41,250	  
41,500	 
41,750	
42,000	  
43,250	
43,500	
43,750	 

behavior planner, which in turn generates a behavior to be executed by the 3D virtual human by using a consecutive action network. The generated behavior is delivered to the virtual human controller. Finally, the virtual human controller controls the joint angle of the virtual human according to the generated behavior.

In an experiment, the proposed framework was verified for the series of processes designed to control the virtual human in a mobile device. In the definition stage, the pose decision tree was constructed by using 63 expected silhouettes. A total of 63 actions were defined. The consecutive action network was constructed on considering the 63 actions. In the recognition stage, images were photographed at 4 fps. The pose most similar to each photographed image was found by using the pose decision tree. In the recognition stage, a pose decision tree was used. As a result, 31 poses could be estimated by 5.1 expected silhouette comparisons on average. In the reconstruction stage, behaviors defined as consecutive actions are generated by consecutive action network. Therefore, even if the same pose index was received, the various behaviors could be generated and executed.

The user's images are converted to markup language by using the index of the pose. This process could allow multiple users to attend a video conference at the same time. Further studies are being planned on a method to generate the user's silhouettes for the pose decision tree





**Figure 16** Video conference reconstruction.

as well as the actions and behaviors for consecutive action network automatically.

#### Acknowledgements

This article was supported by the research program of Dongguk University.

#### Author details

<sup>1</sup>Department of Game Engineering, Graduate School, Dongguk University, 26, Pil-dong 3-ga, Jung-gu, Seoul 100-715, Korea <sup>2</sup>Department of Multimedia Engineering, Dongguk University, 26, Pil-dong 3-ga, Jung-gu, Seoul 100-715, Korea

#### Competing interests

The authors declare that they have no competing interests.

Received: 30 August 2011 Accepted: 18 February 2012

Published: 18 February 2012

#### References

1. AN Mortlock, D Machin, S McConnell, P Sheppard, Virtual conferencing. *BT Technol J.* **15**(4), 120–129 (1997). doi:10.1023/A:1018687630541
2. LQ Xu, B Lei, E Hendriks, Computer vision for a 3-D visualization and telepresence collaborative working environment. *BT Technol J.* **1**(1), 64–74 (2002)
3. BJ Lei, EA Hendriks, Real-time multi-step view reconstruction for a virtual teleconferencing system. *EURASIP J Appl Signal Process.* **2002**(1), 1067–1087 (2002). doi:10.1155/S1110865702206071
4. BJ Lei, C Chang, EA Hendriks, An efficient image-based telepresence system for videoconferencing. *IEEE Trans Circ Syst Video Technol.* **14**(3), 335–347 (2004). doi:10.1109/TCSVT.2004.823393
5. H Fuchs, G Bishop, K Arthur, L McMillan, R Bajcsy, SW Lee, H Farid, T Kanade, Virtual space teleconferencing using a sea of cameras. in *Proceeding of First International Conference on Medical Robotics and Computer Assisted Surgery*, Pittsburgh. **2**, 161–167 (22-24 September 1994)
6. LC Tai, R Jain, 3D video generation with multiple perspective camera views. *Image Process.* **1**, 9–12 (1997)
7. S Moezzi, LC Tai, P Gerard, Virtual view generation for 3D digital video. *Multimedia IEEE.* **4**(1), 18–26 (1997). doi:10.1109/93.580392
8. A Nijholt, H Welbergen, J Zwiers, Introducing an embodied virtual presenter agent in a virtual meeting room. in *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2005)*, Innsbruck 579–584 (14-16 Feb 2005)

9. H Welbergen, A Nijholt, D Reidsma, J Zwiers, Presenting in virtual worlds: towards an architecture for a 3D presenter explaining 2D-presented information, in *International Conference on Intelligent Technologies for Interactive Entertainment, (INTETAIN 2005)*, Madonna di Campiglio, pp. 203–212 (30 November-2 December 2005)
10. R Poppe, D Heylen, A Nijholt, M Poel, Toward real-time body pose estimation for presenters in meeting environments, in *13th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG 2005)*, University of West Bohemia, Plzen-Bory, pp. 41–44 (31 January-4 February 2005)
11. B Leo, JH Friedman, RA Olshen, CJ Stone, *Classification and Regression Trees*, (Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA, 1984)
12. J Yoon, S Cho, A mobile intelligent synthetic character with natural behavior generation. in *2nd International Conference on Agents and Artificial Intelligence (ICAART 2010)*, Valencia 132–133 (22-24 January 2010)

doi:10.1186/1687-1499-2012-51

**Cite this article as:** Sung and Cho: Development and evaluation of wireless 3D video conference system using decision tree and behavior network. *EURASIP Journal on Wireless Communications and Networking* 2012 **2012**:51.

**Submit your manuscript to a SpringerOpen® journal and benefit from:**

- Convenient online submission
- Rigorous peer review
- Immediate publication on acceptance
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](http://springeropen.com)