Joint Tracking and Transmission System for Simulating Motion of the Human Body on Android Smart Phone

Nae Joung Kwak and Teuk-Seob Song

Mokwon University, Doandong, Seogu, Dae-jeon, Korea
knj0125@hanmail.net, teuksob@mokwon.ac.kr

Abstract

This paper proposes a method of simulating the movements of the human body by automatically extracting the characteristics of an object from a camera and sending them to mobile devices. In this method, a RGB color video from a camera was converted into hue, saturation, and intensity images for extracting the silhouettes of the human body using digital subtraction of the images. Using the corner points and modeling information from the extracted silhouettes, joints are automatically detected and used as each connection points of the object. Also, block-matching algorithms were applied to extracted joints to track the characteristics, which are sent to mobile devices. On the mobile devices, body movements are simulated using the received joints. The results showed that the method automatically detects silhouettes and joints and effectively simulates the human body using the extracted joints. Also, the tracking of the joints were reflected in the mapped body, allowing for appropriate tracking the movements.

Keywords: *silhouette, object tracking, auto detection, joint, mobile device*

1. Introduction

Based on recent development of network technology and devices, surveillance systems using cameras are being introduced widely, and the technology has been studied. Existing CCTVs save and send real-time videos and use sensor devices or receive indirect information from sensors. If sensor devices are used, maintenance during monitoring is an issue. Also, because indirect information is used, it is less efficient than direct video with a supervisor and quick response in case of emergency is difficult. Surveillance systems for the elders who live alone and children need real time supervision, but the amount of the video data is enormous. Therefore, research on reducing the amount of data transmitted, as well as research on extracting and tracking objects from videos, is in progress.

Surveillance systems using videos are based on tracking and modeling of human bodies, which can be done in many ways such as using hues to divide video images and analyze contours to express and track the human body as a mass of blobs of similar colors[1-2], using template-matching on videos acquired from many cameras[3-5], and attaching sensors or markers on bodies to extract characteristics, such as the silhouettes and contours[6-8].

The characteristics that are extracted for tracking and modeling of the body include silhouette, contours, specific body parts and their connection information, and joints. Among them, joints are the connecting points of the body parts. So object tracking is able to use the information to track the movement of the extracted joints. Therefore, it is possible to save the image or avatar of the surveillance object on mobile devices and express the movements using the information on joints. Also, with the development of mobile devices, surveillance

system's monitoring can be sent to the guardian's mobile devices as well as PC. Kwak and Song proposed the method which extracts automatically joints and tracks the human body using body ratios [9-10]. And Kwak and Song proposed the method to simulate human body using the joints in PC.

In this paper, we propose the improved method that extracts the silhouette considering shadow area, transmits the joints to android smart phone, and simulates the body on the phone. The proposed method analyzes real-time video data and sends it to the guardians' or emergency centers' mobile devices, such as cell phones, PDAs, and navigation systems. The method automatically extracts joints from the videos, tracks them, sends the information to mobile devices, and simulates the human body under surveillance using the information on joints. Existing systems process the object data, compress it, and show the real video as it is. However, sending the whole video and expressing it on mobile devices can cause delays in data processing due to the massive amount of the data. The joints are text data on the location of joints of the human body. Therefore, this method can significantly decrease the amount of data.

2. Body Modeling and the Detection and Tracking of Joints

2.1. Silhouette Extraction

Extracting the silhouette is a very important pre-processing step before extracting the joints because it affects the precision of joint extraction. To extract silhouette efficiently, separating the foreground from the background is necessary and important in dividing the objects. The shadows and the changing lights are a hindrance when using subtraction method to divide the background and the foreground, and sometimes accurate detection of the object is impossible. The proposed method solves this problem by obtaining difference images of both the grayscale video and the hue video and using results from the two videos.

Silhouettes are extracted as following. Background image is modeled by averaging N frames of the 24-bit color input image $I_c(x, y)$ from the camera and $B(x, y)$ and $I_c(x, y)$ converts the color images into 256 grayscale and get the difference images between both grayscale images. The obtained difference image is converted into binary image using Otsu's method and the binary image $D(x, y)$ is separated into the foreground and the background. Also, taking advantage of the fact that the area of the shadow has similar chromaticity values to the background and low intensity, hue component of the inputted video and Equation 1 were used to separate the foreground and the background. The results are combined with the object region of the binary image in order to get the final object video and the background. The results are combined with the object region of the binary image in order to get the final object video.

$$D(x,y)=\begin{cases} Background & , \left|h_c(x,y)-h_b(x,y)\right|+\left|s_c(x,y)-s_b(x,y)\right|<Th_1 \\ & \quad and\ -20<B_c(x,y)-B_b(x,y)<0 \\ Foreground & , \ otherwise \end{cases} \tag{1}$$

where h_c and h_b are the hue image of inputted image and one of background image, s_c and s_b are the saturation image of inputted image and one of background image, and B_c and B_b are the binary image of inputted image and one of background image.

The binarized image of the inputted image includes many small regions besides object regions. Hence, we need to eliminate the small areas and noise of the binarized hue image before combining the hue and grayscale binarized images to obtain the final binarized video. The contour tracking method provided by OpenCV is used on the extracted object from a video with separated foreground and background to extract the contour and the silhouette.

2.2. Body Modeling and Joints Extraction

This paper proposes a human model like in Figure 1, where fourteen joints are included based on the structure of the body to model the human body. Fourteen joints were selected by considering the points they control the body movement. The proportions of each body part based on the width and length of the face were used for extraction of each joint. The proportions were calculated by analysis of twenty randomly selected subjects. Table 1 shows the averages of the body parts according to height, and the numbers on the body parts refer to the numbering of the joints in Figure 1.

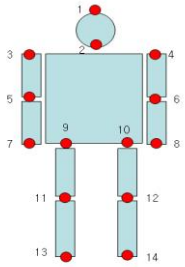


Fig. 1 Human Body Model using Joints

In Table 1, the length of the face has the similar value regardless of height. Therefore in this paper, length of the face is used to calculate the proportions for extracting the characteristics of the body, while the length of the face was calculated using the face detection technique by OpenCV. The proportions of the body parts are obtained by the following equation.

$$\beta = \frac{T_m}{f_h} \tag{2}$$

Where, T_m is the estimated value for each body part of Table 1, and f_h is the length of the face.

Also, since the proportions of the body parts are different according to height, the object's height in the inputted image, the length of the facial region, and the actual face length are used to calculate the height of the object and decide the estimated values for obtaining the location of the joints. The following is the equation that calculates the actual height of the object.

$$R_t = \frac{R_r}{f_r} \times f_h \tag{3}$$

Here, R_t is the actual height of the object, and f_r, R_r is the height of the face and the object in the extracted object in the video.

Table 1. Body Sizes According to Heights [unit:cm]

Body \ Height		140-150	150-160	160-170
Face area		15	16	16
1-2(Face length)		22	22	22
2-3(2-4)		17	19	21
3-5(4-6)		25	27	29
5-7(6-8)		23.5	23.5	23.5
29(2-10)	length	43	53	64
	area	12	14	15
9-11(9-12)		39	41	42
11-13(12-14)		33	35	38

Table 2. The Extraction Method of 1'st Joints

Joint number	x's coordinate	y's coordinate
1	upper x's coordinate of face	upper y's coordinate of face
2	lower x's coordinate of face	lower y's coordinate of face
3	$j\_2.x - f_r \times \beta$	$j\_2.y + f_r \times \beta$
4	$j\_2.x + f_r \times \beta$	$j\_2.y + f_r \times \beta$
5	$j\_3.x$	$j\_3.y + f_r \times \beta$
6	$j\_4.x$	$j\_4.y + f_r \times \beta$
7	$j\_5..x$	$j\_5.y + f_r \times \beta$
8	$j\_6..x$	$j\_6.y + f_r \times \beta$
9	$j\_2.x - face\ area \times 0.5$	$j\_2.y + f_r \times \beta$
10	$j\_2.x + face\ area \times 0.5$	$j\_2.y + f_r \times \beta$
11	$j\_9.x$	$j\_9.y + f_r \times \beta$
12	$j\_10..x$	$j\_10.y + f_r \times \beta$
13	$j\_11..x$	$j\_11.y + f_r \times \beta$
14	$j\_12..x$	$j\_12.y + f_r \times \beta$

Table 2 shows the method of joint extraction using β; face is identified in the inputted video, and the joints are extracted based on the upper and lower coordinates of the face region. j_N.x(y) is the location of x(y) for the joint number N. Because the sizes of the human body are different for each person, it is difficult to detect the exact location of the joints if it is performed using the proportions calculated from the average of the sizes. Therefore, initial joints are detected using the method in Table 2 based on the estimations in Table 1, and the locations of the joints for each person are corrected by extracting corner points of the silhouette of the extracted object and choosing the closest corner points to the initial joints in order to determine the final joints. In final joint detection, joints 1 ~ 2 are not modified, but joints 3 ~ 14 are extracted again using the corner points and the modified joints. Table 3 shows how the extraction of final points works. c.x(y) shows the location of the closest corner point x(y) to the corresponding joint.

Table 3. The Extraction Method of Final Joints

Joint number	x's coordinate	y's coordinate
3	$c.x$	$(j\_3.y + c.y)/2$
4	$c.x$	$(j\_4.y + c.y)/2$
5	$(j\_5.x + c.x)/2$	$(j\_5.y + c.y)/2$
6	$(j\_6.x + c.x)/2$	$(j\_6.y + c.y)/2$
7	$j\_7.x - \lvert j\_7.x - c.x \rvert/2$	$(j\_7.y + c.y)/2$
8	$j\_8.x + \lvert j\_8.x - c.x \rvert/2$	$(j\_8.y + c.y)/2$
9	$(j\_9.x + c.x)/2$	$(j\_9.y + c.y)/2$
10	$(j\_10.x + c.x)/2$	$(j\_10.y + c.y)/2$
11	$(j\_9..x + j\_13.x + c.x)/3$	$(j\_11.y + c.y)/2$
12	$(j\_10..x + j\_14.x + c.x)/3$	$(j\_12.y + c.y)/2$
13	$(j\_13.x + c.x)/2$	$c.y + (j\_13.y - c.y)/2$
14	$(j\_14.x + c.x)/2$	$c.y + (j\_14.y - c.y)/2$

2.3. Joints Tracking

Existing object tracking method either uses the object's hue information to track the main object after extracting the main object or utilizes templates. In this paper, the location of the object is expressed using the location of the joints, so it is possible to track the object by tracking the movements of the region information on the joints instead of using the information on all areas of the object. We tracked the objects using the block-matching method that is used in MPEG. Also, if human objects are tracked in videos from camera, body parts do not show significant changes in movements since actions of humans don't show sudden changes except special circumstances. Thus, this paper tracked joints in the following frames based on locations of initially-extracted joints. While existing block-matching method has to track the entire area of the object, the proposed method reduces the amount of calculations by searching area around the joints. Also, considering the fact that the direction of the human body's movement is consistent, the direction of the movement was predicted by accumulating the vector values of previous movements. The predicted direction is the average of previous 3 frames and the values add to the current x-coordinate and y-coordinate. When applying block matching algorithm, the proposed method starts from the computed coordinates.

$$\begin{aligned} x &= cur_x + (m_{x1} + m_{x2} + m_{x3})/3 \\ y &= cur_y + (m_{y1} + m_{y2} + m_{y3})/3 \end{aligned} \tag{4}$$

Where cur_x and cur_y are the x-coordinate and y-coordinate of current frame of block matching algorithm and m_{xi} and m_{yi} are the moving data of the i-th previous frame.

Blocks are 20x20 for each joint, and the search window of the inputted video uses the block-matching method for horizontal/vertical directions by setting ±5 for size. The proposed method selects the point with the least error as the next location for joint by using intensity for the RGB color video. Taking this estimated value into account, the location information value used during the block-matching method was used to reduce the time required to predict movements.

3. Transmission of Joints and Simulation of Body

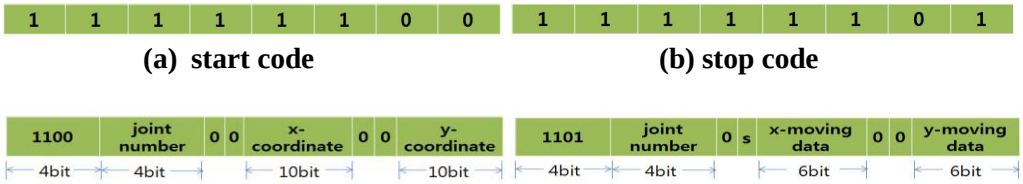
The characteristic extracted from Chapter 2 is used as location information of each connection point of the object, and the values were sent to mobile devices. For these processes, we need to set up protocol for transmission of joints. The transmitted joints are used to simulate body movement in the mobile device.

3.1 Transmission Protocol

To transmit the extracted joints, transmission information is required as following.

1) Starting and stopping code of transmission
2) Number of the joints
3) x-coordinate and y-coordinates
4) The type of the transmission data : movement data/real joint coordinate

Therefore, this paper proposes the following protocol for transmission of data.



(a) start code **(b) stop code**

(c) the joint's coordinates **(d) the moving data of joint**

Fig. 2 Transmission Protocol

The control information that notifies of the start and the end of transmission is expressed in 8 bits as in (a) and (b) in Figure 2. In order to simulate the human body in mobile devices in the same position as the original, the information on the actual location of the joints is required for first frame and the information of joints of key frames. After that, frames can be expressed only with the movement changes of the already-simulated location of joints. Therefore, the information for telling whether the frame contains the real location information or the movement information, the joint number, and the x and y coordinates are used and set as shown in (c) and (d) in figure 2. (c) is the protocol to transmit the x, y coordinates of the location information of the joint and consists of 32 bits in total. (d) is the protocol to transmit the change in movements from the previous frame and consists of 24 bits in total. S is the sign bit.

3.2. Body Mapping and Simulation

The transmitted joint information and movement information are used to simulate the movements of the body on mobile devices. First, the body is expressed according to the location information of the joint. In this paper, the human body is expressed as a combination of polygons, with circle as the face, rectangle as the body and limbs.

4. Test and Result

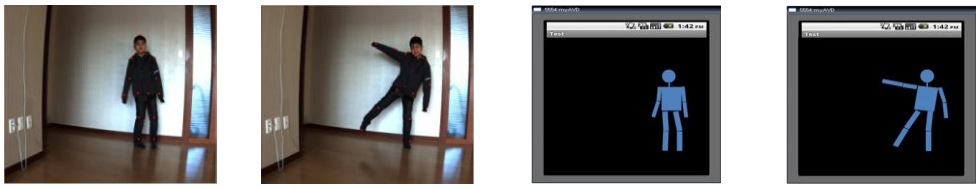
Test was done indoors to analyze the performance of the proposed method. The background video and the inputted video were analyzed real time by camera, and android emulator was used to test how the human body was restructured on mobile devices. In order to extract the body's silhouette, joints were extracted and tracked from a walking human body of a simple background. The system was implemented by using Intel cpu 2.0GHz, 1G RAM, VC++6.0 and Open CV. The resolution of the inputted video was 640X480 in 24-bit, which received ten frames per second.

Figure 3 shows the results of the proposed method. The joint image (b) shows the result extract accurately from input image. The image (c) is the tracking image from image (b) and shows the good result. The figure (d) is the moving data between (b) and (c).

The joints in the first frame were extracted, which were tracked 95 frames later in the video. Body movements that were simulated on the android emulator are shown in figure 4. (c) and (d) show how the human body was efficiently mapped by the proposed method, which extracted the joints and used them to track the object.

(a) background **(b) joint image of Input** **(c) 47th frame image** **(d) moving data between (b) and (c)**

Fig. 3 Joint Image and Tracking Image by the Proposed Method



(a) first frame **(b) 95th frame** **(c) the mapping image of (a)** **(d) the mapping image of (b)**

Fig. 4 Joint Image and Mapping Image by the Proposed Method

In this paper, the object can be expressed using the location of the joints, so it is possible to track the object by tracking the movements of the joints instead of using the information on all areas of the object. We tracked the objects to search blocks around the joints while existing block-matching method has to track the entire area of the object. If the block-matching algorithm is applied to a frame whose size is $M \times N$, block size is $l \times m$ and a search window is $p \times q$, the method needs $M \times N \times p \times q$ addition and $M \times N \times p \times q$ comparison. The proposed method can reduce $^{M \times N}/_{l \times m \times 14}$ for addition and comparison because search window and block size are small. Table 4 shows the number of operation when a frame's size is 300×240, block's size is 30×30 and a search

window is 10×10. The proposed method reduces the number of operation in comparison with the original block matching algorithm.

Table 4. The Operation Complexity of the Blocking Method and the Proposed Method

The number of operation / method	1 frame		10 frames	
	comparison	addition	comparison	addition
Block matching	$72\times10^5+3200$	72×10^5	$72\times10^6+3200$	72×10^6
Proposed method	$45\times10^4+560$	45×10^4	$45\times10^5+560$	45×10^5

Also, because the proposed method considers the direction of movement of joints and predicts the moving direction, the number of operation is less than that of table 4.

5. Conclusion

This paper proposes a system that automatically extracts silhouettes and joints of the human body in real time video from a camera and transmits extracted joints to mobile devices in order to simulate the human body and express the movements of the body.

The method uses the proportions of the parts of the body to extract the joints to model the human body. The modeling information of body is used to automatically extract joints from the video. Also, the silhouettes and the joints were extracted by using information from a single camera. The extracted joints from the first frame were used to track the body utilizes and the block-matching algorithm on joints was applied in order to track the movement of the joints and to predict the direction of the movements. This reduced the amount of calculation and increased efficiency, as it does not use the whole video but instead uses the local information and the accumulated movement vectors for predicting the direction. Extracted joints are sent to mobile devices. To do this, agreement on transmission of information of control and joint was set to enable simulation of the human body on mobile devices. The joint information of the proposed method is in text form, so it can greatly reduce the amount of video data that is transmitted in existing surveillance systems. Applying the new method, it was possible to simulate the human body that is inputted to the camera on android emulators, and the body movements were adequately expressed according to the transmission of movement information.

References

[1] G. Mori and J. Malik, "Estimation Human Body configurations using Shape Context Matching", in Processings of ECCV, **(2002)**, pp. 666-680.

[2] C.Wren, A. Azarbayejani, T. Darrell and A. Pentland, "Pfinder: Real-time tracking of the human body", IEEE trans. on PAMI, vol. 19, no. 7, **(1997)**, pp. 780-785.

[3] S. Iwasawa, J, Ohya, K. Takahashi, T. Sakaguchi, S. Kawato, K. Ebihara and S. Morishima, "Real-time 3D extimation of human body postures from triocular images", in Processings of Workshop on modeling people, **(1999)**, pp.3-10.

[4] T. E. de Campos, D. W. Murray, "Regression- based Hand Pose Estimation from Multiple Cameras", CVPR, vol. 1, **(2006)**, pp. 782-789.

[5] Q. Delamarre and O. Faugeras, "3D articulated models and multi-view tracking with sillouettes", Proc. ICCV, **(1999)**, pp. 716-72.

[6] E. Murphy-Chutorian and M. Trivedi, "Head Pose Estimation in Computer Vision: A survey", IEEE Trans. on Pattern Anaysis and Machine Intelligence, vol. 31, no. 4, **(2009)**, pp.607-626.

[7] T. Horptasert, I. Haritaoglu, C. Wren, D. Harwood, L. Davis and A. Pentland, "Real time 3D motion capture", in Processings of Workshop on perceptual user interface , **(1998)**.

[8] Pengfei Zhu and Paul M.Chrlian, "On Critical Point Detection of Digital Shapes", IEEE Transaction on Pattern Analysis and Machine Intelligence, vol.17, no.8, **(1995)**.

[9] N. J. Kwak and T. S. Song, "Automatic Detecting of Joint of Human Body and Mapping of Human Body using Humanoid Modeling", KIMICS, vol. 15, no. 4, **(2011)**, pp. 851-859.

[10] N. J. Kwak and T. S. Song, "Automatic Detecting and Tracking Algorithm of Joint of Human Body using Human Ratio", The Korea Contents Association, vol. 11, no. 4, **(2011)**, pp. 215-224.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0005121).

Authors

Nae-Joung Kwak She received the B.S. in February 1993 and M.S. in February 1995, Ph.D in February 2005 from the Department of Computer and Communication Engineering, Chungbuk National University. She currently teaches at Mokwon University, Hanbat University, and Chungnam National University in Korea. Her research interests include multimedia communication, multimedia signal processing, video surveillance system, and MPEG. She is a member of the Korea Information Science Society, The Korea Contents Association, and the Institute of Electronic Engineers of Korea

Teuk-**Seob Song** He received his Ph.D Degree in Computer Science in 2006 and Ph.D Degree in Mathematics from Yonsei University in 2001, respectively. He is currently an Assistant Professor in the Department of Computer Engineering at Mokwon University in Korea. His research interests include 3D Virtual Environment, Web3D, Annotation Technology and Structured Document Transcoding. He is a member of the Korea Information Science Society, the Korea Information Processing Society, and the Korea Multimedia Society