Multimed Tools Appl (2012) 59:65–88
DOI 10.1007/s11042-010-0691-z

Real-time tracking of humans and visualization of their future footsteps in public indoor environments

An intelligent interactive system for public entertainment

Ovgu Ozturk · Tomoaki Matsunami · Yasuhiro Suzuki · Toshihiko Yamasaki · Kiyoharu Aizawa

Published online: 6 January 2011

Abstract In this work, an interactive entertainment system which employs multiple-human tracking from a single camera is presented. The proposed system robustly tracks people in an indoor environment and displays their predicted future footsteps in front of them in real-time. The system is composed of a video camera, a computer and a projector. There are three main modules: tracking, analysis and visualization. The tracking module extracts people as moving blobs by using an adaptive background subtraction algorithm. Then, the location and orientation of their next footsteps are predicted. The future footsteps are visualized by a high-paced continuous display of foot images in the predicted location to simulate the natural stepping of a person. To evaluate the performance, the proposed system was exhibited during a public art exhibition in an airport. People showed surprise, excitement,

O. Ozturk (✉)
Department of Frontier Informatics, The University of Tokyo, Tokyo, Japan
e-mail: ovgu@hal.t.u-tokyo.ac.jp, ovguozturk@gmail.com

T. Matsunami · T. Yamasaki
Department of Information and Communication Engineering, The University of Tokyo, Tokyo, Japan

T. Matsunami
e-mail: matsunami@hal.t.u-tokyo.ac.jp

T. Yamasaki
e-mail: yamasaki@hal.t.u-tokyo.ac.jp

Y. Suzuki
Research Center for Advanced Science and Technology, The University of Tokyo, Tokyo, Japan
e-mail: yasusay@rcast.u-tokyo.ac.jp

K. Aizawa
Interfaculty Initiative in Information Studies, The University of Tokyo, Tokyo, Japan
e-mail: aizawa@hal.t.u-tokyo.ac.jp

curiosity. They tried to control the display of the footsteps by making various movements.

Keywords Multiple human tracking · Interactive arts · Camera and projector systems · Technology and entertainment · Real-time visualization

1 Introduction

In the last two decades the collaboration of researches in art, design and technology has increased remarkably. This provided the development of new systems such as sophisticated human-computer interaction tools, virtual reality, augmented reality, interactive entertainment and education technologies [7, 10, 13, 25]. With the advances in video, sensor systems and computer vision technologies, various intelligent systems have been developed to acquire information about objects, humans in most cases. These systems have attempted to analyze objects' motion or intended behaviour. Combined with the artistic concepts from art and design studies, sophisticated, exciting, smart, enjoyable interactive tools and environments have been created.

Most interactive systems utilize video cameras and various sensors to track humans, human body parts, and evaluate the motion to provide input or feedback to the computer or intelligent environment [3, 5–7, 13, 18, 25]. Tracking the human hands, head or the objects humans hold in their hands help to extract the gestures or locations of humans and interact easily with various multimedia devices. Examples of these applications are virtual games, interactive education systems, 3D visualization systems, etc. The other group of systems detect facial expressions or eye movements [16] to accomplish the user-system communication.

Relatively less number of researchers [5, 13, 18] build systems for simultaneous use of multiple humans. In this work, we focus on building an interactive environment that can serve many humans at the same time. Our aim is not to build a sophisticated tool, which people buy and use as a game console; or it is not to build a virtual environment, where people go and enjoy in their leisure time. The aim of this work is to build an entertainment system which naturally appears in people's daily life and becomes a part of the flow. So the technology meets people in their natural living environment, providing an interactive entertainment space. It promotes the collaboration of technology and art, introduces the technological advances to public, while letting them enjoy the results. This work introduces a low-cost real-time multiple-human tracking system by using only a single camera. It avoids the need of markers and additional sensors. To our knowledge this is the first interactive entertainment system which lets the simultaneous participation of many people (10 or more if the area permits).

The proposed system is composed of a video camera, a projector and a computer. It employs a multiple-human tracking algorithm from a single camera. People walking in an indoor environment are tracked by the camera that is located on a high place. Then, virtual footsteps of the people are created continuously and displayed in front of their feet while they keep moving. The visualized footsteps are called *future footsteps* showing their destination. It creates an effect of creating one's own future

by his/her current motions. Or, seeing one's future in front of his/her eyes might effect the present motion.

Figure 1a shows an example visualization of a future footstep. In Fig. 1b, an example input image of the scene captured by the camera is shown. The camera captures the top-view of the area. In order to track multiple humans in real-time, we use blob tracking method and associate the blobs along a sequence of frames to generate a position history data for each blob. For each blob, speed, direction of the motion are calculated and the next position is predicted. Foot shaped images are displayed in the predicted location in the direction of movement to visualize the future footsteps.

Our initial work related to the visualization of the future footsteps was presented in [12, 22]. By using the recorded data in Haneda Airport (Tokyo International Airport), which has the title of *the busiest airport in Asia*, we have done additional experiments and developed our system. In this work, we present the developed system which includes detailed analysis of the tracking results, additional experiments and applications of the system. Different than the initial work, we examined the relationship between blob area and the number of people in the blob. Multiple footsteps were displayed according to the number of people in the blob. Another analysis was held to distinguish an adult from a child when a blob contains one person and smaller size foot images were displayed for a child. This work contains one more component that accumulates the footsteps data and provides the visualization of all the footsteps once in an image. It gives us a way to describe a scene in terms of mostly followed paths.

Section 2 introduces the related work about visual tracking and interactive systems. In Section 3 the overall system is explained, and technical system details are described in Section 4. Experimental results for various situations are presented in Section 6, at the same time user studies about the system are introduced. Finally, discussions and further improvements for the system are given in Section 7, followed by the conclusions in Section 8.

(a) (b)

Fig. 1 Visualization of future footsteps in real-time

2 Related work

There are various systems utilizing tracking techniques to create interactive entertainment systems. Most systems tracks the human body parts in order to control the tools by using gestures. For example, [7] uses multiple cameras and markers, to track the hands and legs of a person to control virtual tools and games. In the work of [6], a person holds colorful objects which are detected and tracked by the system to provide the control of a game without keyboard or mouse. Welch [25] explains the history and techniques of head tracking that is used in virtual reality and augmented reality. Without going into the details of researches in 3D, which is a different area of research, a 3D motion model example for a home entertainment can be given. In research of [18], multiple cameras are used to construct a 3D motion model of a human body by tracking various body parts (head, torso, shoulders, forearms, legs), the system is proposed to be used as an automated home entertainment.

We concentrate here on the papers most relevant to our work, which tracks entire body of the human and analyzes human motions for interactive applications. There are few researches in this field. In [5] and [3], a multi-human tracking system is developed for multiple persons to interact with virtual agents in an augmented reality system. In their application, most of the time only a few humans participate and the system focuses on understanding the actions of the interacting humans and agents. Another group of researchers [13] present an entertainment system, where multiple users are involved in an interactive dance environment. They use markers and multiple cameras to capture the motion of various body parts.

Besides the scientific researches, there are many commercial applications of human motion tracking in interactive advertisements, arts or entertainment areas. Interactive floors are very popular in shopping centers or public places to attract customers. Some example products can be found in http://www.eyeclick.com/products_500.html and http://www.reactrix.com.

Most of the systems introduced so far [3, 7, 13, 18], utilize multiple cameras or combine camera-sensor systems. Or they use markers, sometimes special objects for tracking. A comprehensive study of motion analysis of multiple humans with multiple cameras is given in the book [21]. Different from those, we have developed a low cost, compact system and use only a single camera to track multiple people simultaneously. Our system can be installed easily in any public place. In the literature there are various tracking systems [26], many researchers detect and track image features, such as color [15, 23], KLT [17], corners [24] or textures [4], to track humans. Sophisticated tracking algorithms have been developed to effectively deal with various situations. Among these, Kalman filter-based [11, 14] and Particle filter-based [9] algorithms or mean-shift [8] algorithms are very popular ones. Optical flow [1] is another common method which is based on calculating the motion flow of image features. The paper [19] presents a good survey of these. Basically, all these methods require two main steps, the first one is detection of image features and the second step is tracking these features in consecutive frames. Hence, they require complex calculations and they are time-consuming algorithms. The most important requirement for an interactive system is high-speed within a given range of accuracy. To be able to achieve a real-time robust tracking of multiple humans, we employ a blob tracking algorithm.

3 System overview

Figure 2a shows the placement of the system. It is composed of a video camera, a computer, a projector and a mirror. All the electronic equipments are placed inside a box and the mirror is mounted on the front shutter of the box as shown in Fig. 2b. The box is located on a high place to capture the top-view of the target area. People walking in the area are tracked by the system. By using the tracking results, locations and orientations of future footsteps are predicted. The future footsteps are projected on the floor by means of the projector and the mirror in front of the projector lens.

Tracking objects is a widely used computer vision technique, yet it depends on the object properties and it can be very challenging in different situations, such as in crowded situations. Our system is designed for tracking and visualization of the future footsteps of multiple people in real-time. It means that the tracking algorithm should be very fast and robust to pass the results, so that the footsteps can be displayed in front of a person in the right timing before he/she proceeds further. In this work, we apply blob extraction and association technique to track people. People are recognized as moving blobs in the video.

Once, people are tracked as moving blobs, the area and position history information is calculated for a short period of time. By using the history data, the orientation and position of the next footsteps are predicted, it is explained in Section 4 in detail. Foot shaped white images are displayed in the predicted positions for each existing blob in accordance with the predicted orientation by using the projector. The following sections explain the system architecture and the calibration of the camera and projector.

3.1 System architecture

In our system, a CCD camera with 640 × 480 pixels resolution is used. The camera output frame rate is selected as 6.25 fps from the camera properties, instead of default

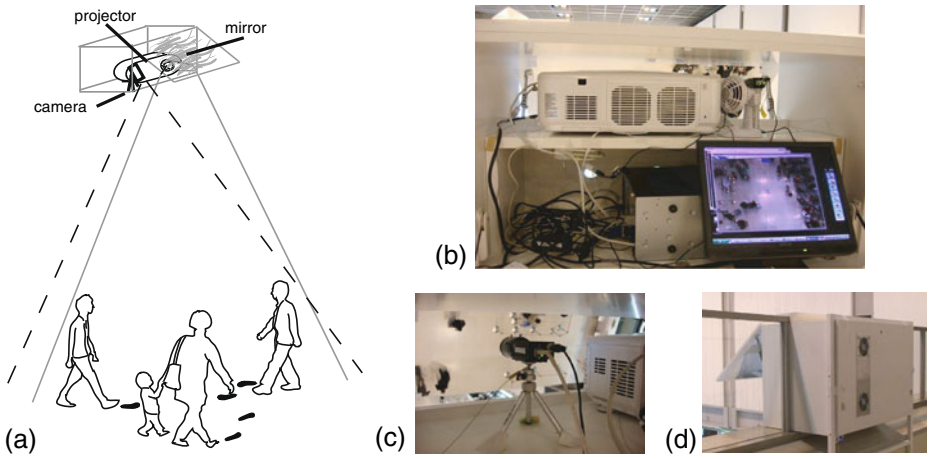


Fig. 2 System overview: **a** placement of the system **b** inside of the system box **c** video camera and mirror **d** outside of the system box

25 fps rate. This rate is enough to catch the change in motion of people and it helps to save time by reducing the number of frames to be processed. In order to enable the clear projection of the footsteps and increase the visibility on the floor, high contrast images are required. One of the newest technology projectors, DLP projector, NP 4100J, is used. To provide a simple, compact and good-looking appearance, everything is placed in a white wooden box. During long-hours operation of the system, the projector gets hot very quickly, hence, fans are mounted to ventilate inside of the box.

The proposed system is composed of three main processes as follows:

I. Real-time Tracking of Multiple Humans
II. Analysis of the Tracking Results
III. Visualization of the Future Footsteps

Figure 3 shows the system architecture. First of all, calibration is required to establish the correspondence between camera coordinates and projector coordinates. The camera and the projector view the same area from different angles. To achieve the correspondence of the coordinates between two views, calibration parameters are calculated to convert from one to another. It is explained in the following section. Once the calibration is done, processing environment is ready for the rest of the process.

The characteristic of this system is that it creates future footsteps for multiple-people in real-time. To achieve such a system, processing time is very crucial. The system should be able to track multiple people and predict their next steps. At the same time, it should be able to visualize footsteps in the predicted position before people move further and pass that position. This is done repeatedly until people leave the scene. To achieve this a simple but effective tracking algorithm has been developed. Furthermore, *tracking* and *visualization* parts are designed to work in

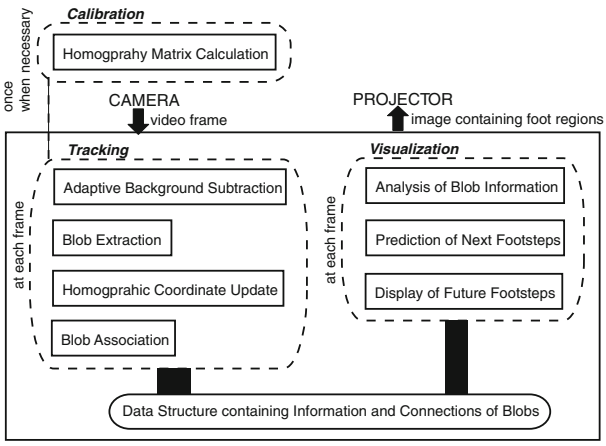


Fig. 3 Software architecture

parallel. By using the video input, humans are tracked and necessary information is stored in the designed data structure. By using data points in the data structure, *visualization* part generates future footsteps. Two parts work separately in parallel, while they access to the same data structure. This provides the necessary gain in speed while tracking a person and displaying his/her footsteps in the right timing before he/she proceeds further. However, it requires very careful synchronization to provide correct localization and timing of the footsteps. Each predicted footstep is visualized by displaying gradually disappearing foot images quickly to provide a natural appearance of a stepping foot.

3.2 Calibration

The view angle and view space of the camera and projector are different than each other. In order to provide the correspondence of coordinates between these two spaces, camera calibration is required. The camera calibration is carried out by using the calibration functions developed by the Intel's Open source computer vision libraries (OpenCV) [2]. OpenCV's *cvFindHomography()* function helps us to calculate the homography matrix between two spaces. Four reference points are chosen both in the projector screen and camera view. The four outermost corners of a chessboard image are usually chosen as reference points. The calibration step is displayed in Fig. 4. A chessboard image is projected from the projector such that it covers the entire target area. Then, from the camera view, the references points are selected by clicking on each of them. In our system, calibration is done once manually before the start-up of the system and homography matrix is stored. Homogprahy matrix is used to convert the coordinates from camera space to projector space by using *cvGEMM()* function when necessary. In the following equations, p_{src} represents a point on the floor plane in the camera space, whereas p_{dst} represents the corresponding point in the projector space and $H_{3\times3}$ represents the homography matrix. p'_{dst} is divided by the constant w to make the third element one to obtain the correct values for p_{dst} during the conversion.

Fig. 4 Calibration of the camera by displaying a chessboard image from the projector

void cvFindHomography(const CvMat srcpoints, const CvMat* dstpoints, CvMat* homography);*
void cvGEMM(const CvArr src1, const CvArr* src2, double alpha, const CvArr* src3, double beta, CvArr* dst, int tABC=0);*

$$p_{src} = \begin{bmatrix} x_{src} \\ y_{src} \\ 1 \end{bmatrix}, \quad p'_{dst} = \begin{bmatrix} wx_{dst} \\ wy_{dst} \\ w \end{bmatrix} \tag{1}$$

$$p'_{dst} = H_{3\times 3} * p_{src} \tag{2}$$

$$p_{dst} = p'_{dst}/w \tag{3}$$

4 Real-time tracking of multiple humans

In our system, the camera captures the top-view of the target area and humans are tracked as moving blobs in the scene. Moving blobs are extracted by applying a background subtraction and the regions above a certain threshold are marked as foreground regions. Later, blobs in successive frames are connected with a simple method which evaluates the distance between the blob positions to decide the associations.

4.1 Background subtraction and blob extraction

Our system is designed to be used during daytime and/or night time containing various illumination changes in the environment. The change of sunlight or the lights from the surroundings can effect the scene environment. Hence, a dynamically updated background calculation algorithm is necessary to keep the best possible background scene definition. To deal with the changes in lighting conditions, parameters used in the dynamic background subtraction algorithm have been examined beforehand. Three sets of parameters are defined, which correspond to morning, noon and evening, respectively. For background subtraction, we apply an adaptive algorithm introduced in [20]. An average background is stored for the scene and it is continuously updated in time.

The intensity of a pixel in the background image is represented by I and the intensity function is modeled with the following equation:

$$I = \overline{I} + \sigma \sin(2\pi\omega t) + k\zeta \tag{4}$$

$\overline{I}$ represents the average intensity in time, while σ represents the amplitude of the intensity and ω represents the frequency. ζ symbolizes the maximum noise parameter of the camera, k is a coefficient ($-1 \leq k \leq 1$) and t is the time.

According to this model if an intensity, I, of a pixel in a given image satisfies the condition, $\overline{I} - \sigma - \zeta \leq I \leq \overline{I} + \sigma + \zeta$, then that pixel belongs to the background. Otherwise it belongs to the foreground object. $\overline{I}$ and σ in the model are updated by

using (5) and (6) for background pixels, and (7) and (8) for foreground pixels. n and m, $(m \geq n)$, are parameters representing update speeds accordingly.

$$\overline{I}' = (n-1)/n \times \overline{I} + 1/n \times I \tag{5}$$

$$\sigma' = (n-1)/n \times \sigma + 1/n \times \sqrt{2 \times (I - \overline{I})^2} \tag{6}$$

$$\overline{I} = I \tag{7}$$

$$\sigma' = (m-1)/m \times \sigma + 1/m \times \sqrt{2 \times (I - \overline{I})^2} \tag{8}$$

To extract the moving regions, the average background image is subtracted from the current video frame. Then, the regions with an area above a predefined static threshold are detected as blobs. Here, morphological opening and closing operations might be helpful to define the borderlines more clearly. However, processing time is very limited, furthermore the blobs extracted after background subtraction and thresholding is descriptive enough to define the moving regions. Each blob is defined with the following elements, (c_x, c_y): center of mass, A: area, P: preceder, $flag$: flag. So, jth blob in ith frame is defined as $B[i, j] = \{c_x, c_y, A, P, flag\}$.

Figure 5a and b show example input images from a scene in an airport. Figure 5c shows a partial region of Fig. 5b in a larger view. Figure 5d is the result after

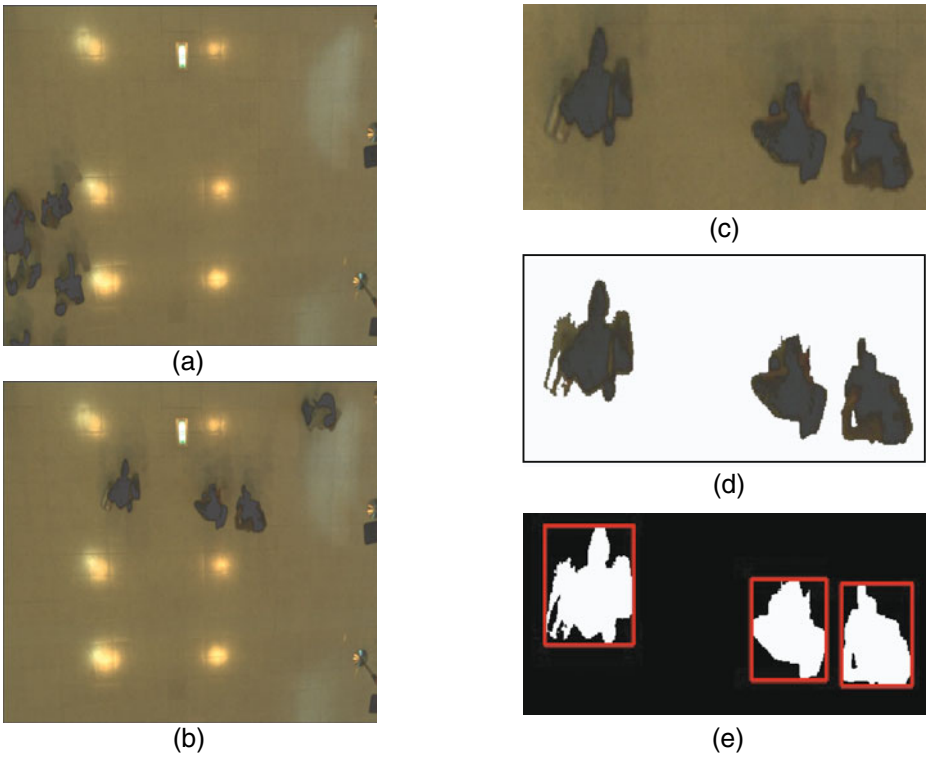


Fig. 5 Blob extraction: **a**, **b** Example input scenes **c** enlarged view of a partial area in the input scene **d** regions after background subtraction **e** detected blobs

background subtraction and Fig. 5e demonstrates the extracted blobs with bounding boxes. Moving people are very well extracted with the defined algorithm. In our method, shadow removal algorithm is not necessary, although there are shadows of objects in the scene. This is because of two reasons. The first one is, because of multiple lights in the environment, there are multiple weak shadows spread in various directions around a person. These shadows can be eliminated by the background subtraction algorithm. The second reason is that, interestingly the remaining strong shadow regions, which might be larger with the strong sunlight, can be very helpful. We are trying to predict the positions of feet of a person. If we extract the blobs without shadows, the center of mass will correspond to somewhere in the middle of a person's body. However, when we look at the extracted regions, we will see that the center of mass of each blob will slide towards the bottom part of a person's body (feet region seen from the view) with the contribution of the shadow region. Figure 6a demonstrates an example of this. The point in the square shows the center of mass excluding the shadow, the point in the circle shows the center of mass including the shadow. The point in the circle is much more closer to the feet region, supporting the aim of our system.

In our system, currently, a person and his luggage are considered as one blob. Since they are connected after the extraction process, they are assumed to be one

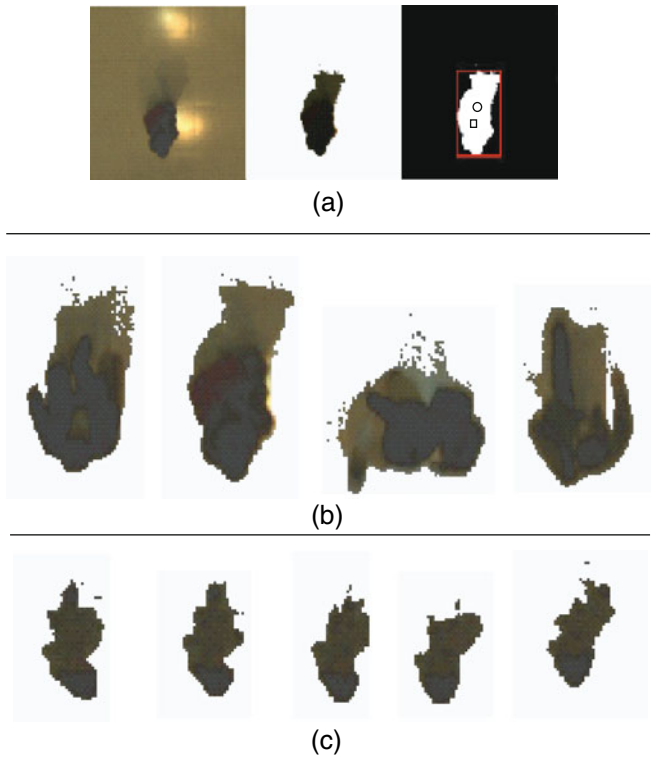


Fig. 6 Extracted regions: **a** Blob extraction for a child **b** example extracted regions of adults **c** example extracted regions of children

region and one footstep is visualized for the whole. Distinguishing and eliminating luggage in the system is left as future work. On the other hand, if the scene is very crowded and people walk very closely to each other. In that case a group of people can be extracted as connected, which means they are recognized as one blob for a long time, and one footstep is visualized for this situation. However under normal conditions, people do not walk that close during consecutive frames, and most of the time they are separately detected and tracked.

Figure 6b shows sample extracted blobs of adults and Fig. 6c shows sample blobs of a child. There is a big difference in the area. For adults, the average area size is about 3,500 pixels and for children, the average area size is around 1,200 pixels. This information is used during the visualization of footsteps and smaller footsteps are displayed for children.

4.2 Association of blobs

Considering the time requirement and the possibility of many people existing in the scene at the same time, we have developed a fast and robust blob tracking algorithm which works in real-time. Blobs are extracted at each frame with the algorithm described in the previous section. After blob extraction, the detected blobs should be assigned with their corresponding blobs in the previous frame. To achieve this, the center of mass of each blob is compared with the center of masses of the blobs in the previous frame. The one with the minimum distance and smaller than a defined connectivity threshold (C_{th}) is chosen to be the preceder to that blob. C_{th} is assigned to the radius of the current blob. A preceder, if exists, would have a distance smaller than the radius of the current blob. If there are no preceders found in the previous frame, then the current blob is defined to be the head of the tracking chain and the *flag* is assigned to -1. If the blobs are not the heads of their chain, their *flags* are assigned to 0, indicating that they have preceders. The preceder element, P, of each blob is assigned to the index of the preceder blob in the previous frame.

For each blob in the current frame:

– Find the blob in the previous frame with the minimum distance to the current blob. Find k which satisfies the following condition, where $dist(B[i, j], B[i-1, k])$ is the Euclidean distance between the center of masses.

$$Min(dist(B[i, j], B[i-1, k])) \tag{9}$$

– If $dist(B[i, j], B[i-1, k]) \leq C_{th}$, then preceder of $B[i, j]$ is $B[i-1, k]$.
$B[i, j] -> P = k$;
$B[i, j] -> flag = 0$;
– If $dist(B[i, j], B[i-1, k]) \geq C_{th}$, then there is not a preceder of $B[i, j]$.
$B[i, j] -> P = -1$;
$B[i, j] -> flag = -1$;

Blob association is calculated for 5 frames at maximum. In other words, tracking data of a person is stored only during the last five frames. It is enough to predict the speed, orientation and position of the next step. Figure 7 shows the tracked blobs for 3 consecutive frames. Blobs A, B, C, D, E exist in the first frame. A, B, D move upwards and leaves the scene after the second frame. C, E move downwards. F appears in the scene in the second frame and moves upwards. Figure 8 illustrates

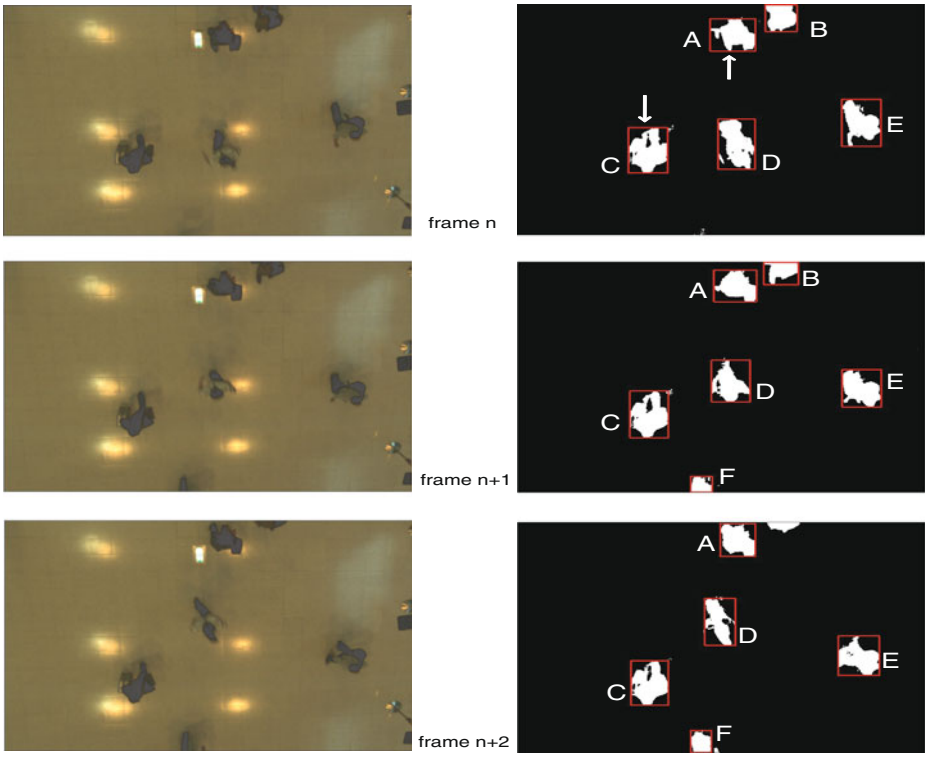


Fig. 7 Multiple human tracking by using blob tracking during three consecutive frames

the data structure for three frames in Fig. 7. Each blob is represented with a name which starts with "B" and ends with the number of the blob. The *flag* and *preceder* fields are shown below each blob. The flags indicating the existence of preceder and the number of the preceder blob in the previous frame are assigned accordingly.

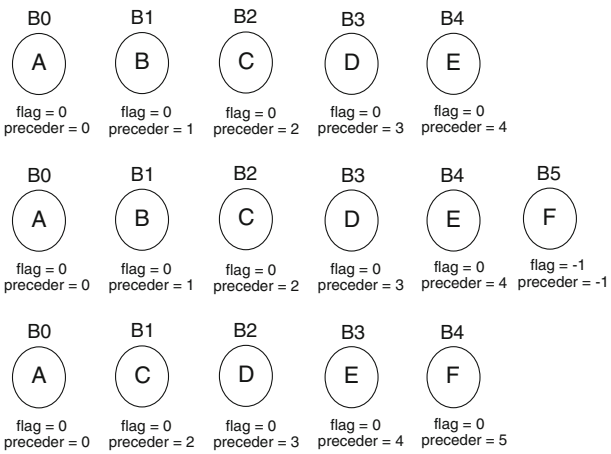


Fig. 8 Association of the blobs stored in the data structure

5 Analysis of tracking results and visualization of footsteps

In this system a person's data during the last five frames is stored, it almost corresponds to a duration of one second (6.25 fps). This data is used in three different ways. First, position history information is used to mathematically model the motion of the person and estimate his/her speed. Second, by analyzing the area information of each blob in a frame, the number of people in each blob and the total number of the people in the scene are estimated. Sometimes, blobs can contain connected group of people, in case of high density crowds. The area of the blob helps us to estimate the number of people in the blob. Third, an adult and a child can be distinguished, and footsteps can be displayed smaller for children by using this information.

By using this data, the speed and orientation of the motion of the person are calculated. By using the position data, calculated speed and orientation, position and orientation of the next step of a person is predicted by applying linear prediction.

5.1 Prediction of future footsteps

Numerically, the motion of each blob is modeled by a linear function. x and y attributes of the position of each blob are defined with two linear functions with four parameters as in (10). t represents the time. The parameters a_x, b_x, a_y, b_t are calculated by solving the equations coming from five data of the last five frames. Equation 11 shows the matrix operations. After calculating these parameters, speed of each blob is estimated by (14). Then, for each blob Δt is assigned according to the speed information. Finally, new position of the blob is estimated by using (12) and $a_x, b_x, a_y, b_t, \Delta t$. Orientation of the motion is computed by (13).

$$\begin{aligned} x(t) &= a_x t + b_x \\ y(t) &= a_y t + b_y \end{aligned} \tag{10}$$

$$\begin{bmatrix} x(t) \\ x(t-1) \\ x(t-2) \\ x(t-3) \\ x(t-4) \end{bmatrix} = [a_x b_x] \bullet \begin{bmatrix} t \\ 1 \end{bmatrix}, \quad \begin{bmatrix} y(t) \\ y(t-1) \\ y(t-2) \\ y(t-3) \\ y(t-4) \end{bmatrix} = [a_y b_y] \bullet \begin{bmatrix} t \\ 1 \end{bmatrix} \tag{11}$$

$$\begin{aligned} x_{next} &= a_x(t + \Delta t) + b_x \\ y_{next} &= a_y(t + \Delta t) + b_y \end{aligned} \tag{12}$$

$$\theta = \arctan(a_y / a_x) \tag{13}$$

$$S = \sqrt{((x(t) - x(t-4))^2 + ((y(t) - y(t-4))^2 / 4} \tag{14}$$

During prediction step, linear fitting is applied as explained. A line equation is fit to the last five measurements of the center of mass. Although the center of mass measurements are not very stable, line fitting approach helps to define an average line of motion. Figure 9 shows the tracking results for a basic Kalman filter tracking and center of mass tracking. Black circles depict the Kalman filter based tracking,

Fig. 9 Tracking results for a basic Kalman filter based tracking and center of mass tracking. *Black circles* depict the Kalman filter based tracking, whereas *red circles* depict center of mass points

whereas red circles depict center of mass points. It can be seen that, the amount of unstability in the center of mass positions is very small compared to the radius of one blob. Furthermore, the fluctuations in the center of mass are suppressed by the line fitting and do not produce a significant effect while estimating the orientation of the motion. On the other hand, this method works very well to support our aim of tracking the feet of a person, rather than tracking the person's body. Center of mass of a blob corresponds to the adjunction point of the person's body and his/her shadow, where feet of the person exist. In case of rapid changes in the blobs because of lights, variation in viewpoint, merging, splitting situations, center of mass keeps the track of feet very well. Whereas, general tracking algorithms are not sufficient to track the motion of feet in the video. These algorithms track the person's body and update itself according to the defined motion equation (linear, circular, etc...) with fewer data points(the last one or two points). It has been observed that these algorithms might be late to adapt to the change and it is difficult to adjust them to track the position of the feet.

Line fitting of center of masses approach gives good results in terms of accuracy and it has less computation time compared to other filter-based methods. To give a computational evaluation, the proposed method is compared with a Kalman filtering based approach to track humans and predict the footsteps positions. Basic parameters x,y,vx,vy, indicating position and speed are used in Kalman filter. Figure 10 shows the comparison results for various number of blobs in the scene. The experiments were run on a 32-bit Windows machine with 2GB RAM and 2 GHz processor. (The original system was running on a 64-bit machine at 2.67 GHz using 2GB RAM, and it was working with 6.25 fps). The tracking and visualization system has run for 1,000 frames and the average processing time per frame has been calculated for various number of blobs. When the number of blobs is small, there is not a big difference in processing time; but when there are many blobs in the scene, the proposed approach presents a faster solution while keeping the robustness. When the number of blobs is 8,9 or 10; around 10 msec. difference in processing time has been observed.

Estimation of number of people As stated before, the area of a blob can give information whether that person is an adult or child. Similarly, if the extracted blob region is composed of multiple persons, we can calculate the number of people in the region by looking at the area. The number of visualized footsteps for that blob

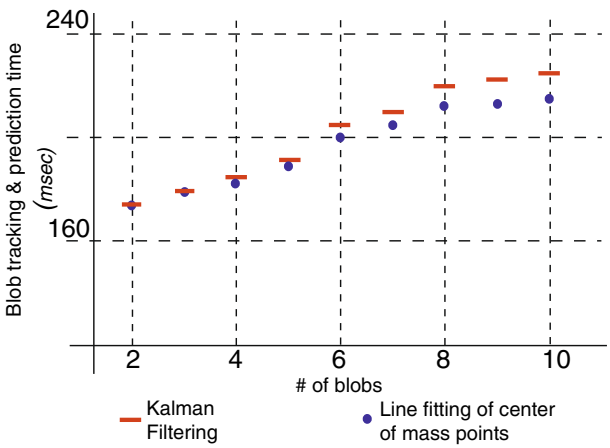


Fig. 10 Processing time for blob tracking and prediction in the system. The proposed method is compared with Kalman filtering based approach. The experiments were run on a 32-bit Windows machine with 2GB RAM and 2 GHz processor. Visualization part works in parallel and is included in the experiments

can be defined according to the area. For adults, the average area size is about 3,500 pixels. So if the area of a blob is larger than factors of this amount, then there are multiple people in the blob and the number of people is estimated by looking at the size. If the blob is far smaller than this amount, for example in the ranges of 1,200 pixels, then it is evaluated as a child. In this work, luggage can be confusing. There are many kinds of luggage, small ones, large ones. Sometimes they can be connected with the person in the blob, but most of the time they are detected as separate blobs. If the luggage is connected to the person, they are recognized as one blob and one pair of footsteps is displayed. If the luggage is detected and tracked separately, then one pair of footsteps is displayed for the luggage, too. More detailed analysis, such as distinguishing the luggage from the person when they are connected, finding the types of luggage, is left as future work.

5.2 Visualization of future footsteps

Once the positions and orientations are calculated for the predicted future footsteps, an image containing foot-shaped white regions on a black background is projected on the floor with the help of a mirror as in Fig. 2. For each footstep, gradually disappearing images are displayed successively to create an effect of a foot stepping on the floor. An example group of a footstep image is shown in Fig. 11. There are three

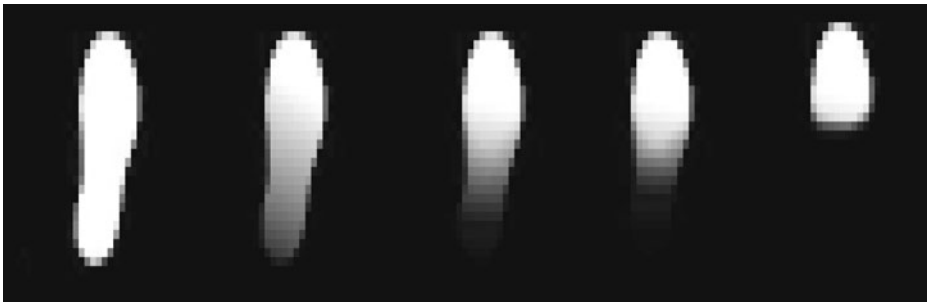


Fig. 11 Gradually disappearing images of a foot

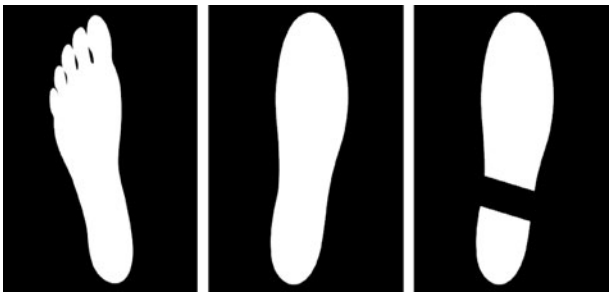


Fig. 12 Various foot shapes used in the system

kinds of foot shapes, which are shown in Fig. 12, and a shape is chosen randomly for each tracked blob. During the display of the footsteps, at each step, an image containing white foot regions on a black background is constructed and projected on the floor. White foot regions correspond to the predicted future footsteps of the existing blobs in the scene. Depending on the area, the blob is evaluated as either a child or an adult. Hence, the size of the foot image is set accordingly. Sometimes, when a person makes a turn, depending on the sharpness, some delay might occur in the direction. This system works in real-time for multiple people. To achieve fast processing, simple algorithms are developed. However, still it requires 64-bit machine working at 2.67 GHz using 2GB RAM. Illumination condition is important, less light in the environment helps to display clearer footsteps.

6 Experimental results

In this section, experimental results for the visualization of the future footsteps are presented for various situations. The reactions of people the experiencing the system is also introduced. Furthermore, a user study has been carried out to study the questions, such as what kind of people pay attention or how many people can notice the displayed footsteps in crowded situations?

6.1 Results from various situations

Figure 13 shows the visualization of the future footsteps for different people. The orientation and positions of the footsteps successfully depict the intended future step of each person. In Fig. 14, two kinds of foot shapes are illustrated. Figure 14c depicts the future footstep visualization result for a person with a luggage. In Fig. 15, the results of the future footsteps systems is presented from a top-view for four people walking in the area. The image is captured from a height of 12 m, so foot regions are a little dim, however one can notice the correct orientation and positioning of the displayed future footsteps for multiple people.

The proposed system presents an interactive public entertainment system which employs simultaneous tracking of multiple humans. Table 1 gives the comparison of existing multi-human tracking systems for interactive applications and demonstrates the advantages of the proposed system. Most of the existing systems utilize multi-sensors, they require markers on the users and they only work for a few humans (one to four). Unlike these, the proposed system offers a single camera (low-cost)

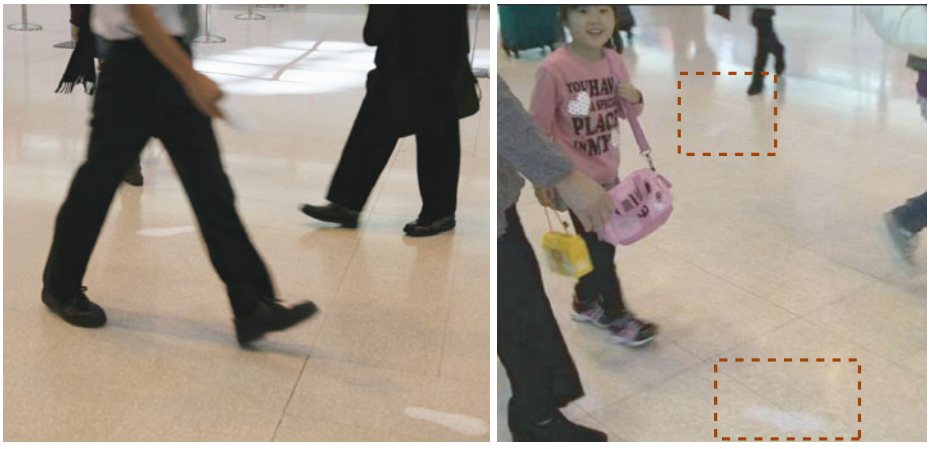


Fig. 13 Experimental results: visualization of future footsteps for various people

tracking system for many simultaneous users without having them attach markers on their bodies. From artistic point of view, "future footsteps" idea brings the concept of visualizing a person's future based on the person's movements. This arises the questions: "do we create our future? or do we follow the designated future?". One significant characteristic of this work can be described with the words "invisibility" and "natural existence". Unlike other works, such as interactive floors, visual entertainment places; the interaction area is not indicated with any lights, colors or marks. If there is no one in the region, nothing is displayed on the floor. Everything is normal as in daily life. At the beginning, there is nothing in the region, and then footsteps appear with the existence of people, move and disappear. The footsteps are born, they live and they die.

Besides the interactive entertainment application, another usage of the proposed system is the detection of the dominant motion paths in the area. For the entertainment purpose, the footsteps are displayed in real-time according to the motion of each person. When we accumulate the predicted footsteps and display all of them at the same time, we can analyze the overall motion paths. The mostly followed

Fig. 14 Experimental results: various future footsteps

Fig. 15 Experimental results: various future footsteps

paths in the scene can be recognized by looking at the resultant image. This provides an alternative visualization of the dominant paths taken by the customers in an indoor environment. It can be used by many people such as architects, social analysts, market analysts, etc. Figure 16 gives an example of a resultant image after observing 120 frames.

6.2 User study

When people see their footsteps, they express great excitements. Some people try to make interesting movements to see what will happen to the footsteps. Some people speed up to catch and step on the footstep but they can never do it. Some others

Table 1 Comparison of real-time multiple human or body parts tracking systems for interactive applications

Method	Markers on the users	# of cameras	# of max participants	Tracking part	Prediction
Cho et al. [5]	No	Single	Two	Whole body	No
Chae et al. [3]	No	Camera, laser scanner	One	Whole body	No
POSTRACK [7]	5 markers	4 cameras (IR)	One	Hands, legs, torso	No
Chung et al. [6]	Color objects	Single	One	Hands	No
James et al. [13]	Markers	Multiple	Four	Whole body	No
Michoud et al. [18]	No	Four	One	Head, torso arms, legs	No
Our system 2010	No	Single	Around 10 (or more)	Whole body	Yes

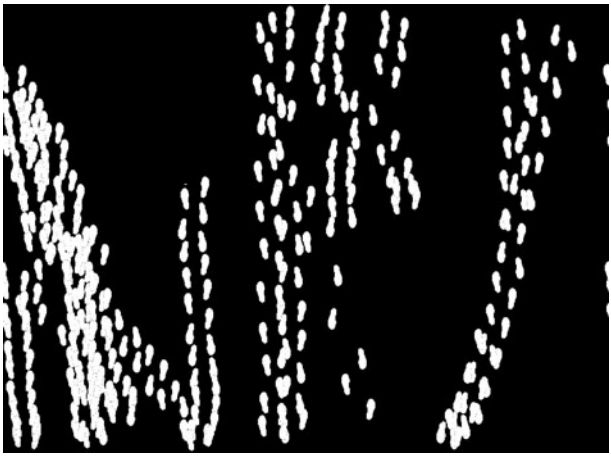


Fig. 16 Visualization of mostly followed paths from top-view

try to find where the footsteps are coming from by searching around. The woman in Fig. 17 jumps from right to left, left to right and tries to control the future footsteps. In Fig. 18, a little girl plays with the footsteps, tries to catch them exploring the area. As a result, we can say that this system serves our purposes: to awaken an interest in technology, to entertain, to make people think and to show the recent progress in technology.

During one hour of a study, approximately 900 people passed through the target area in the airport. Around 50 people recognized the existence of visualized footsteps. They came usually in groups, such as couples, university students or tour groups. When someone in a group noticed the footsteps, he/she showed it to the friends, and more people noticed it. Some groups came with flyers in their hands, they tried to find the footsteps visualization area by using the map on the flyer. The total number of the people who came alone and noticed the footsteps while looking around randomly is five.

The display frequency and distance of the predicted footsteps from the person's current position were highly dependent on the person's speed. When a person moves slowly proceeding short distances, then the footsteps are displayed close to the person's current position. When a person moves very fast taking long distances, then footsteps are displayed farther away from the current location. Some users noticed this and played with the footsteps accordingly. They were observing the difference between walking slowly and walking fast. Sometimes, they were asking questions such as: why are the shapes of the footsteps different?, what is the reason that

Fig. 17 User reaction: a woman is jumping right and left to play with the displayed footsteps

Fig. 18 User reaction: a little girl is exploring and trying to step on the footsteps

sometimes right foot and sometimes left foot are displayed? When the frequency and order of displayed footsteps did not match with their own stepping foot (left, right), people became curious and wanted to learn the details of the system. If a person walks smoothly, (without interrupting, moving back or taking sharp turns), then the system predicts the footsteps in front of them correctly in real-time. However, when a user takes a sharp turn or suddenly stops and starts to walk back, the prediction comes with delay. Another comments came from the users which mentioned about the delay, for example they made zigzag movements or they walked backwards while facing forward.

7 Future work

We have developed a system which employs tracking of multiple people for an interactive entertainment application. The most challenging task was to achieve the real-time tracking process and to synchronize it with the visualization part. Hence, we chose blob tracking which is fast and robust for multiple people case. However, our aim is to build a system for indoor environments, an airport building in the current case. It is very likely that people will carry luggage. It is important to distinguish people from luggage. Some line detection algorithm can be used to detect the objects with straight lines(luggage has straight lines) and eliminate them.

In our system, most of the time linear prediction works very well. As long as people make soft direction changes in their movements, the predicted footsteps are still displayed in the correct place with a correct direction of motion estimation. However, when people take sharp turns, the footsteps are displayed with delays. To improve this, another prediction method is planned to be developed considering the faster movements and variety in human motion.

Displaying all predicted footsteps during a given period of a video at once is good to describe the movements in the scene. Social analysts or public area designers can benefit from this kind of visualization to store statics. Further analysis of the overall motion can be added to the system, such as finding the dominant motion flows in a scene, etc.

8 Conclusions

This paper presents an interactive entertainment system for simultaneous use of multiple humans. The system tracks people walking freely in an indoor environment. It continuously visualizes their predicted future footsteps in front of them while they keep moving. A real-time multiple human tracking algorithm has been developed and combined with a visualization process. A video camera and a projector are located high above the target area. Humans walking through the area are captured and tracked by the camera-computer system. Then, by using the tracking results, their next locations are predicted by analyzing direction and speed of their motion. Foot-shaped images are displayed in the predicted location in front of them by using a projector. This provides people to see their destination created by themselves. It gives the feeling that they control their future.

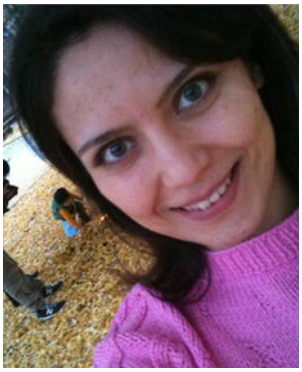
The system can be installed in any indoor place easily. It does not affect the natural flow of life in the sense that it does not affect the movements of people until they notice the displayed interactive foot shapes. When they notice, people show surprise, excitement, astonishment. They try to discover where and why the foot shapes are coming from. They play with the system by making various movements. Sometimes they try to step on the visualized images, sometimes they observe the foot images very carefully. As a result, this interactive entertainment system becomes a part of the daily life, brings technology into people's life by presenting technology with artistic concepts.

Acknowledgements A part of this work was exhibited as a part of Digital Public Art project in Haneda Airport. The authors are deeply thankful to Aizawa Yamasaki Laboratory members for their continuous support during the exhibition. Special thanks go to Chaminda de Silva for his comments and support throughout the project.

References

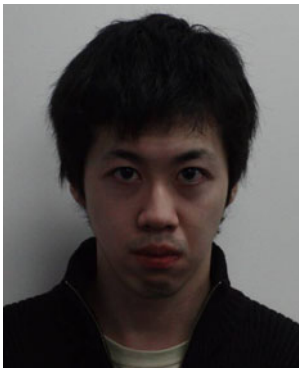
1. Barron JL, Beauchemin SS, Fleet DJ (1992) Performance of optical flow techniques. In: IEEE conf. on computer vision and pattern recognition, pp 236–242
2. Bradski G, Kaehler A (2008) Learning OpenCV. O'Reilly, pp 370–403
3. Chae YN, Kim Y, Choi J, Cho K, Yang HS (2009) An adaptive sensor fusion based objects tracking and human action recognition for interactive virtual environments. In: Proc. of int. conf. on virtual reality continuum and its applications in industry, pp 357–362
4. Chan AB, Liang ZSJ, Vasconcelos N (2008) Privacy preserving crowd monitoring: counting people without people models and tracking. In: IEEE conf. on computer vision and pattern recognition, pp 1–7
5. Choi J, Cho Y, Cho K, Bae S, Yang HS (2008) A view-based multiple objects tracking and human action recognition for interactive virtual environments. Int J Virtual Real 7(3):71–76
6. Chung J, Shim K (2006) Color object tracking system for interactive entertainment applications. IEEE Int Conf Acoust Speech Signal Process 12:5355–5358
7. Chung J, Kim N, Kim GJ, Park C (2001) POSTRACK: a low cost real-time motion tracking system for VR application. In: Proc. of the int. conf. on virtual systems and multimedia, pp 383–392
8. Comaniciu D, Ramesh V, Meer P (2000) Real-time tracking of non-rigid objects using mean shift. IEEE Conf Comput Vis Pattern Recognit (2):142–149

9. Doucet A, De Freitas JFG, Gordon NJ (2001) Sequential Monte Carlo methods in practice. Springer, New York
10. Fox J, Arena D, Bailenson JN (2009) Virtual reality: a survival guide for the social scientist. J Media Psychol: Theor Methods Appl 21(3):95–113
11. Han Z, Ye Q, Jiao J (2008) Online feature evaluation for object tracking using Kalman filter. In: Int. conf. on pattern recognition, pp 1–4
12. Hirose M et al (2010) Digital public art in Haneda airport. Bijutsu Shuppan Ltd., Tokyo, pp 1–160
13. James J, Ingalls T, Qian G et al (2006) Movement-based interactive dance performance. In: Proc. of ACM int. conf. on multimedia, pp 470–480
14. Kalman RM (1960) A new approach to linear filtering and prediction problems. Trans ASME–J Basic Eng 82(D):35–45
15. Kong S, Sanderson C, Lovell BC (2007) Classifying and tracking multiple persons for proactive surveillance of mass transport systems. In: IEEE conf. on advanced video and signal based surveillance, pp 159–163
16. Lee S, Kim GJ, Choi S (2007) Real-time tracking of visually attended objects in interactive virtual environments. In: Proc. of ACM symposium on virtual reality software and technology, pp 29–38
17. Lucas BD, Kanade T (1981) An iterative image registration technique with an application to stereo vision. In: Proc. of int. conf. on artificial intelligence, pp 674–679
18. Michoud B, Guillou E, Bouakaz S (2007) Real-time and markerless 3D human motion capture using multiple views. LNCS. Springer, Heidelberg, pp 88–103
19. Moeslund TB, Hilton A, Kruger V (2006) A survey of advances in vision-based human motion capture and analysis. Comput Vis Image Underst 104(2):90–126
20. Morita S, Yamazawa K, Terazawa M, Yokoya N (2005) Networked remote surveillance system using omnidirectional image sensors. Trans Inst Electron Inf Commun Eng (5):864–875
21. Ohya J, Utsumi A, Yamato J (2002) Analyzing video sequences of multiple humans: tracking, posture estimation and behavior recognition. In: Int. series in video computing. Springer, pp 1–160
22. Ozturk O, Matsunami T, Suzuki Y, Yamasaki T, Aizawa K (2010) Can you see your "future footsteps"? In: Proceedings of int. conf. on virtual reality
23. Snidaro L, Micheloni C, Chiavedale C (2005) Video security for ambient intelligence. IEEE Trans Syst Man Cybern Part A Syst Humans 35(1):133–144
24. Shi J, Tomasi C (1994) Good features to track. In: Int. conf. computer vision and pattern recognition, pp 593–600
25. Welch GF (2009) HISTORY: the use of the Kalman filter for human motion tracking in virtual reality. PRESENCE: Teleoperators and Virtual Environments 18(1):72–91
26. Yilmaz A, Javed O, Shah M (2006) Object tracking: a survey. ACM J Comput Surveys 38(4):1–45



Ovgu Ozturk received her B.S. degree in Electrical and Electronics Engineering Department from Middle East Technical University in Ankara, Turkey, in 2003. She came to the University of Tokyo to proceed her career with the Japanese Government Scholarship, and received her M.Eng. degree from

the same university in Tokyo, Japan, in 2006. She has experience in both image processing hardware and software algorithms/system design, development. Her research interests are in the area of object recognition, human tracking, pose estimation and machine learning for computer vision. Currently, she is pursuing her PhD studies in the Department of Frontier Informatics, the University of Tokyo.



Tomoaki Matsunami received his B.S. degree from the Department of Information and Communication Engineering, The University of Tokyo in Japan, in 2009. He is interested in media forensics such as image detection and visual surveillance systems. Currently, he is pursuing his M.Eng. degree in the same department in the University of Tokyo.

Yasuhiro Suzuki was born 1979 in the Shizuoka Prefecture, Japan, Graduated from the Department of Design of Tokyo Zokei University and in 2001. That year he won the NHK Digital Stadium Grand Prix Award for his Perspective of Globe-Jungle, an installation utilizing a revolving, spherical jungle gym for children, Globe-Jungle, as a screen. After this success, he was invited to a number of exhibitions and arts festivals. Many of his works apply an afterimage phenomenon. In Blinking Leaves (2003), he drew open and closed eyes on each side of a great number of paper leaves and scattered them. Property of Water (2005) expressed water by swaying rays of light on a drop-shaped revolving screen in darkness. He is currently involved in the research of basic technology for creating digital public arts at The University of Tokyo.

Toshihiko Yamasaki received the B.S. degree in Electronic Engineering, the M.S. degree in Information and Communication Engineering, and the Ph.D. degree from The University of Tokyo in 1999, 2001, and 2004, respectively. He is currently an Associate Professor at Department of Information and Communication Engineering, Graduate School of Information Science and Technology, The University of Tokyo. His current research interests include 3D video processing, 3D computer graphics, analog VLSI design, and so on. Dr. Yamasaki is a member of IEICE, IEEE, ACM, ITE, and so on.

Kiyoharu Aizawa received the B.E., the M.E., and the Dr.Eng. degrees in Electrical Engineering all from the University of Tokyo, in 1983, 1985, 1988, respectively. He is currently a Professor at the Department of Information and Communication Engineering and Interfaculty Initiative of Information Studies of the University of Tokyo. He was a Visiting Assistant Professor at University of Illinois from 1990 to 1992. His research interests are in image processing and multimedia, and he is currently engaged in multimedia lifelog and three dimensional video. He received the 1987 Young Engineer Award and the 1990, 1998 Best Paper Awards, the 1991 Achievement Award, 1999 Electronics Society Award from IEICE Japan, and the 1998 Fujio Frontier Award, the 2002 and 2009 Best Paper Award from ITE Japan. He received the IBM Japan Science Prize in 2002. He is currently the Editor in Chief of Journal of ITE Japan, and an Associate Editor of IEEE Trans. Image Processing and is on Editorial Board of ACM TOMCCAP and Journal of Visual Communications and Image Processing. He served as an Associate Editor of IEEE Trans. CSVT and IEEE Trans. Multimedia, too. He has also served a number of international and national conferences; he was the General co-Chair of MMM2008 and SPIE VCIP99. Program Co-Chair of ACM CIVR2008 and Short Paper Track of ACM 2005 etc. He is a Member of IEEE, ACM, IEICE, ITE.