

Multiple Main User Recognition

Hye-Jin S. Kim, Il-Kwon Jeong

Electronics and Telecommunications Research Institute, Korea.

marisan@etri.re.kr, jik@etri.re.kr

Abstract—We present a method for selecting main users among multiple candidates. For human-machine interaction, a machine choose a single or multiple users to interact with them. The proposed method provide an ordered user probability that indicates whether one can be a main user. We estimate various features such as distance to a user, eye detection, gaze direction, moving velocity, acceleration. Experiments show that our algorithm effectively differentiate main users from the others.

Keywords— Recognition, Multiple Candidates, Interaction

I. INTRODUCTION

In the last decade, a large range of smart HMI applications became available on the mass market [1]-[7]. This has pushed research to increase the importance of user selection, user understanding[8] and user tracking[9]-[12], especially among multiple user interaction. In case of large touchable screen, the development of display technology has lead to interact with each other on the device but the only selection strategy is the first toucher on it. Meanwhile, interaction with a robot or a remote display mostly opt for the person who has the largest face or the closest person. However, this approach is hardly able to deal with multiple people interaction. In this paper, we propose an outstanding method to select active participants in the interplay between a machine and human beings. Experimental results show that the proposed features effectively represent a user's intention.

II. MAIN USER SELECTION

Usually, human beings show their intention through action. Therefore, we assume that a main user takes a close look at the machine, stares at it and goes slower down. As a measure of human's intention, we present three features: gaze orientation, speed and distance between a person and a machine. We assume that theses features can represent a user's intention to interact with a machine and we try to find out intention's formulas showing the relations between human's actions estimated by those features and human's intention.

A. Main User Features

One's action can reveal his/her intention. We assume that important action in light of intention is human's gaze, and passing speed, and the distance to the other object. Each feature is obtained by Kinect v2 information such as user tracking, head direction, and eye detection, and depth information.

Firstly, we measure gaze direction by yaw. Among various head orientation factors such as yaw, roll, and pitch, yaw is

adopted experimentally because gazing is more relevant than roll and pitch. However, yaw itself shows weak correlation in Fig. 1. From the head orientation of Kinect v2, we obtain yaw θ_{yaw} . However, we transform it to $\hat{\theta}_{yaw}$ using Kalman filter because it is too fluctuate to be utilized. The transition matrix of Kalman Filter is (1, 1, 0, 1) and noise covariance is e^{-4} . Using eye detection, the filtered value, $\hat{\theta}_{yaw}$, becomes 0 or 1 as shown in Eq.(1). It gives us whether the person stares at the machine or not.

$$p_{gaze} = \begin{cases} 1 & \text{if } \left| \frac{\theta_{yaw}^{max}}{|\hat{\theta}_{yaw}|} \right| \leq \delta_{yaw} \text{ \& Eyes are detected.} \\ 0 & \text{,otherwise} \end{cases} \quad (1)$$

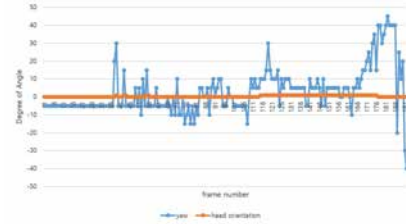


Fig. 1. Head orientation estimator: θ_{yaw} gives very fluctuate values in the blue line and its correlation of the user intention is weak. Therefore, we transform it using Kalman filter $\hat{\theta}_{yaw}$ in the orange line.

The second feature is the distance between a machine and a candidate. We measure z_t at time t based on the depth given by Kinect v2. We assume that a close person tends to have higher intention to play with than a person at along distance. Hence, the distribution of distance $p_{distance}$ is inversely proportional to the interval. In Eq.(2), δ_d is the constant of the proportion and obtained by training where ξ_z is the threshold of the close distance enough to interact each other. In particular, candidates outside the range of interest ξ_z can be excluded using $p_{distance}$.

$$p_{distance} = \begin{cases} \frac{\delta_d}{z_t} & \text{if } z_t \leq \xi_z \\ 0 & \text{if } z_t > \xi_z \end{cases} \quad (2)$$

Thirdly, we adopt a user's speed as a meaningful feature to figure out one's intention. We assume that a person who passes by the machine at a brisk pace are not willing to interact with it. When a candidate gets into close touch with the object, if his speed slows down gradually, the proposed algorithm accords high probability to p_{speed} . The speed of each person is obtained by tracking algorithms [9]-[12]. The speed can generally be calculated by $\dot{z}_t$ or $\nu = |z_t - z_{t-1}|$

where $d = \sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2}$ in a 2D image sequence. Instead, we use $z(t)$ which is a depth obtained by the range sensor. We are also tracking ν per person, and by this, we grasp progressive velocity changes $\ddot{z}_t = \nabla \nu$. Within a detectable long range ξ_{zf} , the acceleration $\ddot{z}_t$ may be negative because the user's velocity seems to decrease. We differ the coefficients δ_f and δ_c according to the far and near range threshold, ξ_{zf} and ξ_{zc} , respectively. In a very near area ξ_{zn} , we consider a person who stop at the front of the object as a main user with high probability.

$$p_{speed} = \begin{cases} -\delta_f \ddot{z}_t & \text{if } z_t < \xi_{zf} \& \ddot{z}_t < 0 \\ -\delta_c \ddot{z}_t & \text{if } z_t \leq \xi_{zc} \& \ddot{z}_t < 0 \& \nu \neq 0 \\ \delta_c * \epsilon & \text{if } z_t \leq \xi_{zn} \& \nu = 0 \\ 0 & \text{, otherwise} \end{cases} \quad (3)$$

Fig.2 shows three features per frame: gaze orientation Fig.2 (a), speed Fig.2(c), and distance Fig.2 (b). In Fig.2 (b), the distance at 104th frame unexpectedly goes 1 because the corresponding user is placed in out of interest region. Fig.2 shows p_{gaze} , $p_{distance}$, and p_{speed} in Fig.2(a) - (c), respectively.

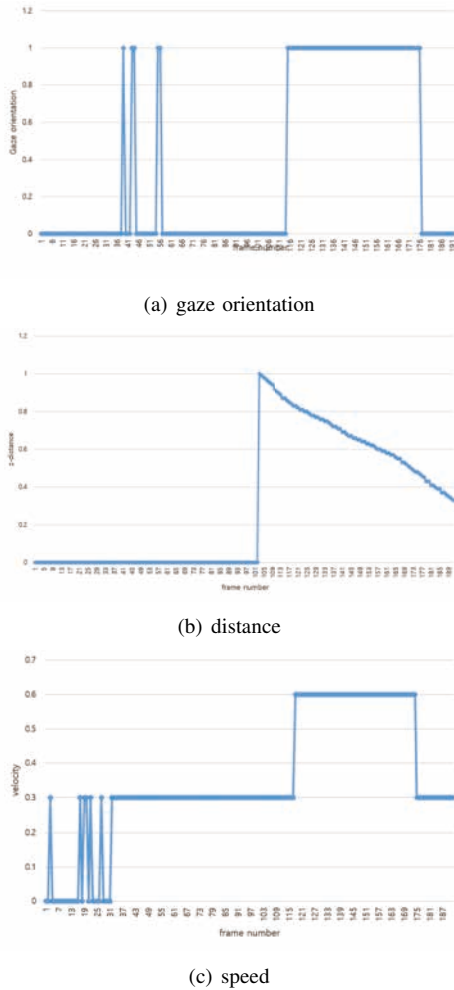


Fig. 2. Distributions of p_{gaze} , $p_{distance}$, and p_{speed} are shown. $p_{gaze} = 1$ in (a) means that a person makes eye contact with a machine. After passing by the machine, it becomes 0. shows that $p_{distance}$ increase when a user is closer to the object. We found that the highest p_{speed} in is obtained when a user is close enough and stop in front of the machine.

B. User Selection

Our strategy to select main users in order is the weighted sum of three distributions, p_{speed} , p_{gaze} and $p_{distance}$. Before applying Eq.(4), we normalize them in order to make them compatible each other. Fig.3 shows the results of a image sequence containing a single main user. At the beginning of the sequence, a candidate is located in the long distance and therefore, the confidences of gaze orientation and speed are low. The closer to the machine, gaze values grows up. The sequence scenario is like this: the main user approach to the machine, stay there for a while, and after that, she goes away. In Fig.3, the orange line refers to the ground truth of p by human annotation and the blue line is manifested in the tendency of the proposed algorithm.

$$\begin{aligned} & \text{maximize} && J(p) = [J_1 J_2 J_3]^T \\ & \text{subject to} && w_g p_{gaze} + w_d p_{distance} + w_s p_{speed} \leq 1 \\ & && J_1 = p_{gaze} \\ & && J_2 = p_{distance} \\ & && J_3 = p_{speed} \\ & && p_i \geq 0 \ (i = 1, 2, 3) \\ & && w_i \in [0, 1] \end{aligned} \quad (4)$$

In Eq.4, w_g , w_d and w_s are weighting factors for the each objective function p_i with respect to $\sum_{i=1}^m w_i = 1$ and $0 \leq w_i \leq 1$. We assumed that the weighted sum p is a convex combination of objectives due to each $0 \leq p_i \leq 1$. The Pareto front of this problem is convex and this is a general form of weighted sum optimization that helps produce well-distributed solutions[13].

III. EXPERIMENTAL RESULTS

We evaluate our method on self-produced data set, which consists of two parts: a single main user and multiple main users. There exist human-annotated ground truth data set only for single main user data set. Hence, parameters are obtained by the single main user results and applied the parameters to multiple main user data set for the purpose of validation. The single main user data set comprises 15 video clips and among them, we use 10 clips for training and 5 for test. The multiple main user data set comprises of 10 video clips and five people move around in each clip. Some of them are just passerby and others give significant actions to show their intention for interaction. Their behaviors are varying for example, glancing, eyeing sternly, stopping, rocketing past, placing a long distance and so on.

Fig.3 presents the result of a single user. The float RGB images on each figure show key frame in that significant changes for user's intention occur. At the beginning frame, the blue solid lines swing in Fig.3(a) and Fig.3(b). Both of them are caused by sensor errors: the eye detection and object tracking was failed. Also, the blue solid lines of Fig.3(a) and Fig.3(c) tends to appear after the orange solid lines. In the case of gaze orientation, this is because we transformed yaw by applying Kalman filter and hence. Similarly, the distribution of speed is obtained by tracking and the acceleration $\ddot{z}_t$ can be calculated in a frame later.

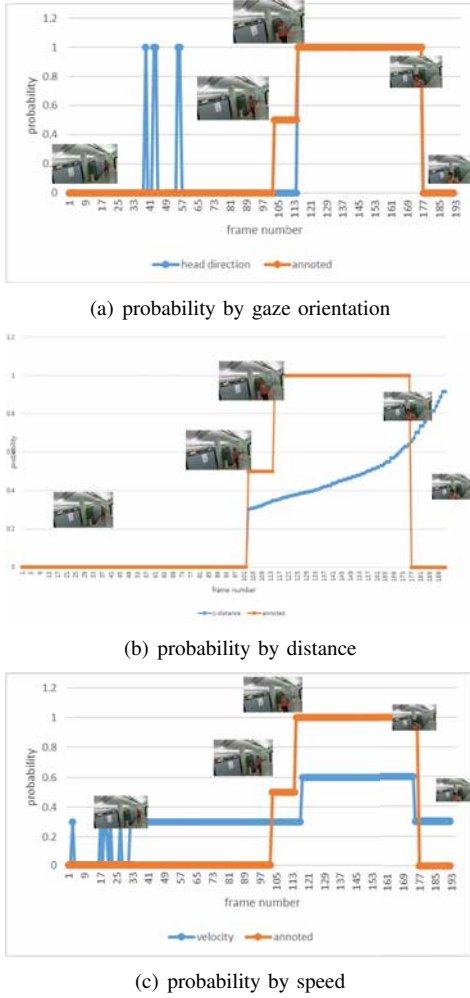


Fig. 3. Three probabilities, p_{speed} , p_{gaze} and $p_{distance}$ and the human annotation result. The floating images are placed in the corresponding frame number.

TABLE I
THE CORRELATION OF p_{speed} , p_{gaze} , $p_{distance}$ AND THE HUMAN ANNOTATION.

gaze	distance	speed	overall
0.9059	0.6089	0.8154	0.9682

$$correl(x, y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} \quad (5)$$

On the other hand, we want to investigate the proposed three distributions are meaningful. Therefore we adopt correlation formula using Eq.(5). The results are shown in Table I. Amazingly, the distribution of the highest correlation is p_{gaze} , which is very similar to human being's attitude. Moreover, we achieved the highest correlation in the final p , which proved that the proposed algorithm grasps human's intention, effectively.

The overall result, p , are presented in Fig.4. Although there were lagging information, the weighted sum p overcome it.

The parameters attained in the previous process. By applying these parameters, we attempt to catch main users among

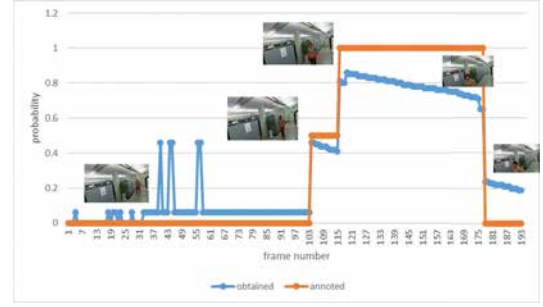


Fig. 4. Three features and their accumulation results in candidates' probabilities to be a main user.

multiple people. Our assumption is that the concept of human's intention is consistent both in the case of a single user or multiple users. The first two rows in Fig.5 shows that many candidates pass by the object naturally: sometimes they can take a look at the target or go past it beyond. The probabilities were constantly changing according to the candidates' gaze or distance, but the main user could be confirmed as a person in the yellow box. The last column of Fig.5 showed the proposed method is robust to the various conditions. The first column of the last row shows that the main user is continuously tracked even through the main user is occluded, and the middle column of the last row shows low values since there is no main user. In the last column of the last row, the person in the red box, who was the main user, changed the direction of the gaze and it caused that the main user's probability was reduced.



Fig. 5. Image sequences with a large number of candidates with gaze, speed, and distance values of candidates. The above two rows show the values that are tracked in the image sequence. The first column of the last row shows various cases: the first column is when there is occlusion of the main user, the second column is when there is no candidate to become the main user, and the last column shows when the main user did not stare.

IV. CONCLUSION

In this paper, we extract various features that user candidates can represent intention to be a user and investigate the

correlation between these values and the values judged by the user. We present a method that the judgment of a main user can be represented stochastically among a large number of candidates. That is, the proposed method can effectively express intention to be a main user from the selected features (distance from the target, eye detection, gazing direction, motion speed, acceleration).

For the future work, we construct the ground truth data for multiple users and validate the results for not only qualitative analysis but also quantitative analysis.

ACKNOWLEDGMENT

This work was supported by 'The Cross-Ministry Giga KOREA Project' grant from the Ministry of Science, ICT and Future Planning, Korea.

REFERENCES

- [1] J. Lee, D. Kim, J. Kim, J.-I. Cho, and J. Sohn, "A Long-Range Touch Interface for Interaction with Smart TVs," *ETRI Journal*, Vol. 34, No. 6, pp. 932-941, 2012.
- [2] J.-H. Han, S.-H. Choi, and J.-H. Kim, "Interactive Human Intention Reading by Learning Hierarchical Behavior Knowledge Networks for Human-Robot Interaction," *ETRI Journal*, Vol. 38, No. 6, pp.1229-1239, 2016.
- [3] U. Yang, N.-G. Kim, and K.-H. Kim, "Augmented System for Immersive 3D Expansion and Interaction," *ETRI Journal*, Vol. 38, No. 1, pp. 149-158, 2016.
- [4] S. Jeong, Y.-W. Kim, and C.-H. Youn, "Personalized Healthcare System for Chronic Disease Care in Cloud Environment," *ETRI Journal* Vol. 36, No. 5, pp.730-740, 2014.
- [5] B. Jones, H. Benko, E. Ofek, and A. Wilson, "IllumiRoom: Peripheral Projected Illusions for Interactive Experiences," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 869-878, April 2013.
- [6] J. Zhou, T. Matsumaru, "Human-Machine Interaction using the Projection Screen and Light Spots from Multiple Laser Pointers," *Proceedings of the 2014 IEEE/SICE International Symposium on System Integration*, pp.16-21, Dec. 2014.
- [7] X.Cao, C. Forlines, R. Balakrishnan, "Multi-User Interaction using Handheld Projectors," *Proceedings of the 20th annual ACM symposium on User interface software and technology*, pp. 43-52, Oct. 2007
- [8] W.-H. Yun, D. Kim, C. Park, and J. Kim, "Hybrid Facial Representations for Emotion Recognition," *ETRI Journal*, Vol. 35, No. 6, pp. 1021-1028, 2013.
- [9] H.-J. Kim, D. Kim, J. Lee, I.-K. Jeong, "Uncooperative Person Recognition Based on Stochastic Information Updates and Environment Estimators," *ETRI Journal*, Vol. 37, No. 2, pp. 395-405, 2015.
- [10] J.-W. Choi, D. Moon, and J.-H. Yoo, "Robust Multi-person Tracking for Real-Time Intelligent Video Surveillance," *ETRI Journal*, Vol. 37, No. 3, pp. 551-561, 2015.
- [11] A. Andriyenko, S. Roth, and K. Schindler, "An analytical formulation of global occlusion reasoning for multi-target tracking," *ICCV Workshop*, pp. 6-13, Nov. 2011.
- [12] J. Zhang, L. L. Presti, S. Sclaroff, "Online Multi-Person Tracking by Tracker Hierarchy," *Proceeding of AVSS* pp. 379-385, Sept. 2012.
- [13] Das, I., and Dennis, J. E., "Normal-Boundary Intersection: A New Method for Generating Pareto Optimal Points in Multicriteria Optimization Problems," *SIAM Journal on Optimization*, Vol. 8, No. 3, pp. 631-657, 1998.



Hye-Jin S. Kim received her MS degree in computer science and engineering from Pohang University of Science and Technology, Pohang, Rep. of Korea, in 2003. She has been a research scientist at the Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea, since 2005. Her research interests include machine learning, humanrobot interaction, and computer vision.



Il-Kwon Jeong received his MS and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology, Daejeon, Rep. of Korea, in 1994 and 1999, respectively. Since April 1999, he has been with Electronics and Telecommunications Research Institute, Daejeon, where he was a Senior researcher, became a Principal researcher in 2010. His current research interests include computer graphics/vision, digital holography, HCI, and virtual/augmented reality.