Long-Range Hand Gesture Interaction Based on Spatio-temporal Encoding

Jaewon Kim, Gyuchull Han, Ig-Jae Kim, Hyounggon Kim, and Sang Chul Ahn

Imaging Media Research Center
Korea Institute of Science and Technology (KIST)

Abstract. We present a novel hand gesture interaction method which has a long-range working space (1m∼5m) overcoming conventional approaches' limitations in cost-performance dependency. Our camera-free interaction system is composed of a pair of lighting device and an instrumented glove with photosensor markers. The lighting devices spatiotemporally encode user's interaction space via binary infrared light signals and markers' 3D position at fingertips is tracked at high speed (250 Hz) and fair accuracy (5mm at 3m working distance). Each marker consisting of a photosensor array allows a wide sensing range and minimizes fingers' self-occlusion. Experiment results demonstrate various applications where hand gestures are recognized as input commands to interact with digital information mimicking natural human hand gestures toward real objects. Our system has strengths in accuracy, speed, low price, and robustness comparing with conventional long-range interaction techniques. Ambiguity-free nature in marker recognition and little cost-performance dependency are additional advantages of our method.

1 Introduction

We present a novel hand gesture interaction method with a long-range working space (1m∼5m). Such long-range hand gesture interaction is an open research area which is not practically covered by current methods or devices such as Kinect. Practical solutions for the problem imply significant influence on future TV applications like a next-generation remote controller in large space. Conventional vision-based approaches including Kinect have fundamental weakness in their performance: Interaction speed and accuracy directly depends on camera's performance and cost. Typically, vision-based approaches for long-range hand gesture interaction require expensive cameras to capture visual information at high-speed and high–resolution such as Oblong's G-Speak system. To overcome the conventional limitations, we present a camera-free hand gesture interaction method based on spatiotemporally encoded illumination.

1.1 Contributions

We present an analysis of a photosensor-based interaction system that accurately and rapidly recognize 3D position of markers through spatiotemporally encoded

N. Streitz and C. Stephanidis (Eds.): DAPI/HCII 2013, LNCS 8028, pp. 22–31, 2013.

illumination. Our method proposes a practical solution for long-range hand gesture interaction with a cheap camera-free system. Included in this analysis are the following:

- performance of the hand gesture interaction method based on spatio-temporal encoding and photosensor markers,
- a unique design of a photosensor marker which consists of multiple photosensors to increase field of sensing and minimize occlusion errors,
- example applications for the hand gesture interaction, including 6 DOF manipulation of a graphic object, multi-user interaction for drawing operation, and manipulation of a deformable object.

1.2 Related Work

Vision-Based Interaction: Bare-hand interaction is earning high attention with the emergence of Kinect. Although such interaction provides the best user convenience obviating the need of wearing a device, it's been considered one of the most challenging interaction tasks. [1] presented a bare-hand interaction technique, called BiDi screen, based on a light field camera and a time-division multiplexed display. While it successfully demonstrated object manipulation by bare hands, the accuracy is not high enough to separately track each finger and the interaction speed is limited due to the time-division multiplexed operation over display frequency. Generally, bare-hand interaction techniques based on a camera vision system suffer from ambiguity problem among fingers or hands. In addition, the computational cost is high and in turn the interaction speed is slow. To overcome such limitations, [2] and [3] employed color markers and color gloves, respectively. Color information helps object ambiguity but still depth ambiguity remains in 3D interaction. So, multiple cameras, a camera-projector system, or a 3D camera have been adopted in 3D interaction domain. However, strong dependency between system cost and interaction performance appeared as an obstacle in developing practical applications. Besides, self-occlusion among fingers or hands has still remained a challenging task. To computationally tackle this matters, [4] presented an approach to jointly solve salient point association with hand pose estimation. An almost everywhere differentiable objective function was proposed to estimate hand gestures by simple local optimization taking edges, optical flow, salient points and collisions into account. [5] approached the full 3D pose estimation of a user hand with a unique wrist-worn device consisting of an IR camera-project setup and an IMU sensor. However, a fully flat or over-arching hand was still problematic in their method.

Photosensor-Based Interaction: Some researchers ([6], [7], [8]) presented interaction methods based on photosensors and an illumination device like a video projector demonstrating good augmented reality applications. While our method has a common ground in the usage of photosensors as markers, the core tracking method, spatio-temporal encoding, is completely different with their methods. The UNC's HiBall system[9] employed six photosensors and six lens to track a

user's pose (location and orientation) with ceiling-mounted light-emitting diodes (LEDs). SCAAT (Single-Constraint-At-A-Time) method, recursively estimating accurate pose information from a single inaccurate measurement, was applied to the system in order to improve tracking accuracy with less latency. However, the system had a drawback in cumbersome system installation which required for mounting huge number (approximately 3000) of LEDs onto a ceiling. Contrastingly, our system requires only two lighting devices consisting of 36 LEDs for 3D position tracking. Kang[10] introduced an indoor GPS metrology system with a unique probe unit, called 3D Probe, composed of three photosensors and a ball probe tip. 3D Probe captures an object's 3D surface via scanning and 3D coordinate of a certain object point is measured by analyzing light signals transmitted from multiple light sources at known location and orientation. While the method requires measuring light sources' geometry information at a calibration stage, our method is free of such information and more accurate with less number of light sources.

Instrumented Glove-Based Interaction: Hand interaction techniques based on various sensors such as mechanical and electrical sensors have been actively commercialized. Immersion Corporation's CyberGlove[11] and Fifth Dimension Technologies' Data Glove are good examples. Generally, such products are configured into an instrumented glove assembled with a delicate sensor system. CyberGlove measures hand posture through 18 electrical sensors in long and thin strip shape which are sewn into a glove fabric. Each sensor experiences change in resistance depending on bending amount. Hand gesture is estimated by the amount of deformation at each sensor position, which is measured by electric current change. User-dependent calibration is indispensable in most sensor-based approaches since different hand size or shape affects consistent bending measurement and gesture estimation.

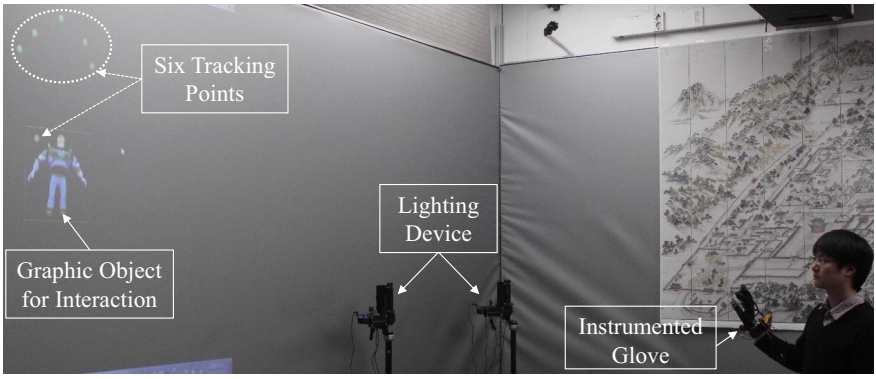


Fig. 1. A user interacts with a graphic object, Buzz, through natural hand gestures in a large space. Our interaction system consists of an instrumented glove with six photosensor markers and two lighting devices.

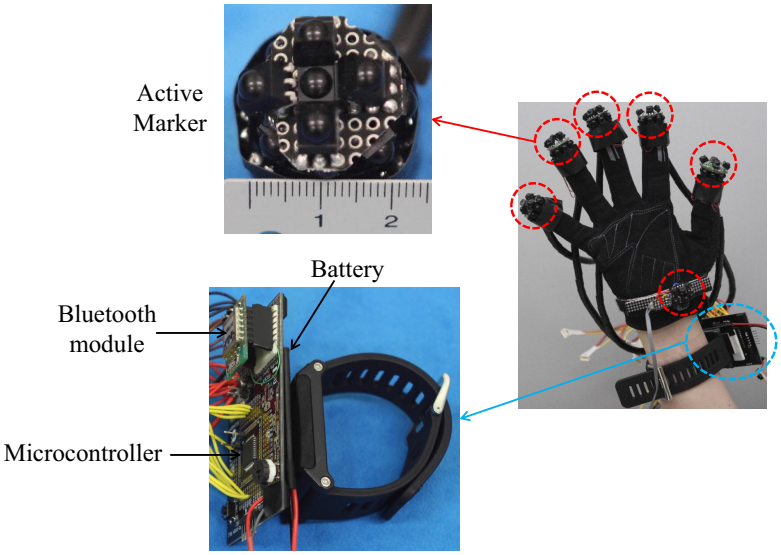
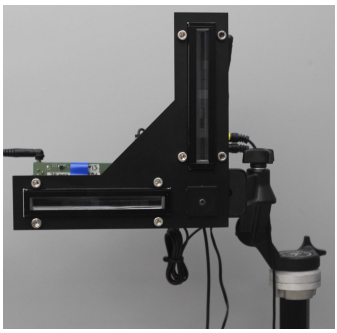


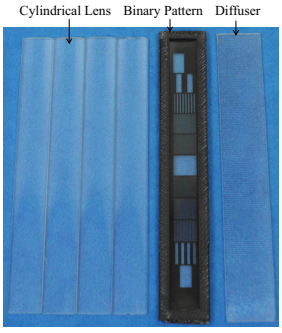
Fig. 2. Our unique glove system consisting of markers, a control unit and a battery achieves long-range hand gesture interaction at high-speed (250 Hz for 3D) and fair accuracy (5mm at 3m)

2 System Configuration

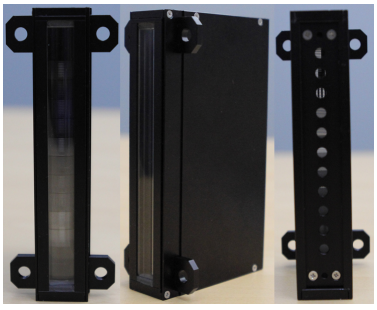
Our prototype camera-free interaction system shown in Fig.1 is composed of an instrumented glove (Fig.2) and a pair of lighting device (Fig.3). The glove is integrated with six photosensor markers, a microcontroller, a Bluetooth module, and a battery. The six markers are located at the five fingertips and the end of palm as shown in Fig.2 right. Each marker consists of multiple photosensors facing different views to cover a wide sensing range. For the photosensor, we used Vishay TSOP7000 with 455 kHz PCM (Pulse-code modulation) frequency which provides a high-speed sensing rate (1 kHz for six markers' 1D tracking). Our system's tracking speed is inversely proportional to tracking DOF (Degrees of Freedom) due to the temporal encoding nature of our method: 500 Hz and 333 Hz for 2D and 3D tracking, respectively. For 2D tracking, the lighting device (Fig.3 (a)) consisting of a pair of 1D lighting unit (Fig.3 (c)) are used to project spatiotemporally encoded light in X and Y plane. Similarly, 3D tracking requires projecting 1D lighting unit along X, Y, and Z axis, which allows 333 Hz tracking speed. However, practically such projection demands a large space so we alternatively present a stereo combination method with a pair of 2D lighting device as shown in Fig.1, which achieves 3D tracking at 250 Hz. A microcontroller (Microchip PIC18F45K20) controls all electronic devices including photosensor



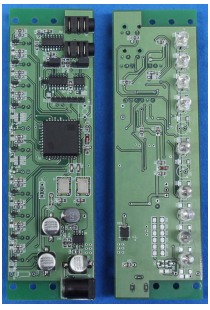
(a) A lighting device



(b) Optical components



(c) Assembly mount



(d) LED unit

Fig. 3. A lighting device to project spatiotemporally encoded light. The 2D lighting device (a) is assembled with a pair of 1D lighting unit (c) positioned perpendicular to each other.

markers and a Bluetooth module in the glove. A Bluetooth module (FB755AC) transmits markers' position value to a remote server and receives event signals for haptic feedback to a user. All electronic devices in the glove are powered by a thin lithium polymer battery (3.7V/1000mA).

The 1D lighting units (Fig.3 (c)), consisting of nine IR LEDs (Vishay TSFF5210 with 180mW/sr), spatiotemporally encode user's interaction space by binary infrared light signals. Eight LEDs encodes space with 8 bit binary signals and one LED at the center of the unit sends a start signal to synchronize photosensor markers. Fig.3 (b) shows optical components including a binary pattern film, four cylindrical lens, and a diffuser. Two cylindrical lenses and a diffuser scatter LED light over a large interaction space. Other two cylindrical lenses focus LED light on the interaction space. Fig.3 (d) shows an electronic circuit board to control the nine LEDs.

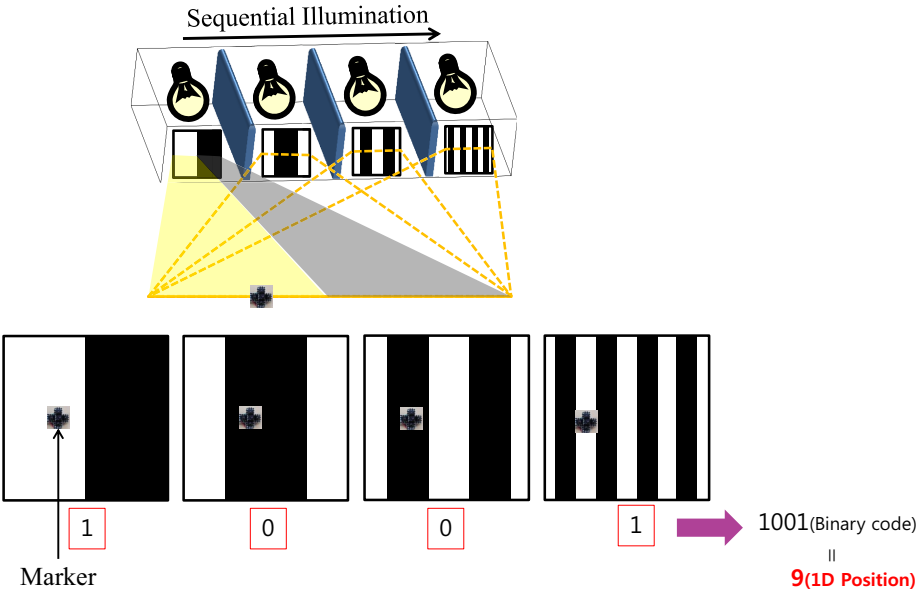


Fig. 4. When IR LEDs sequentially project binary patterns once at a time, a maker receives a unique binary code which is transformed to an 1D position value

3 Tracking Method

Our tracking method is based on spatiotemporal encoding technique [12] which temporally illuminates 8 bit binary light signals into the interaction space via the lighting device shown in Fig.3. Photosensor marker's position is simply obtained by decoding the 8 bit light signal sensed by the marker's photosensors. For example, while the lighting device sequentially illuminates 4 bit binary patterns in Fig.4, the marker receives either 0 or 1 depending on black or white light stripe region. In the figure, the marker sequentially receives 1001 which is a unique code for the position. In such a manner, 1D position is encoded by binary light codes and Cartesian coordinate position is obtained by converting a marker's received binary signal to a real, decimal value. 2D tracking is simply the two dimensionally extended case of 1D tracking with a pair of 1D lighting device which are orthogonally positioned along X and Y axis. 3D tracking can be done in the same manner with three lighting devices positioned along X, Y and Z axis. Tracking with these configurations from 1D to 3D doesn't require any calibration since a position value is directly obtained by the received light signal. However, in the case of stereo combination with a pair of 2D lighting device as shown in Fig.1, a calibration step is required to calculate 3D position from two 2D position values. We applied Miaw's calibration method ([13]) which models the relation between a 3D real position value and a 1D sensed value with seven unknown parameters.

Fig. 5. A demo video verifying our system's resolution (5mm at 3m working distance)

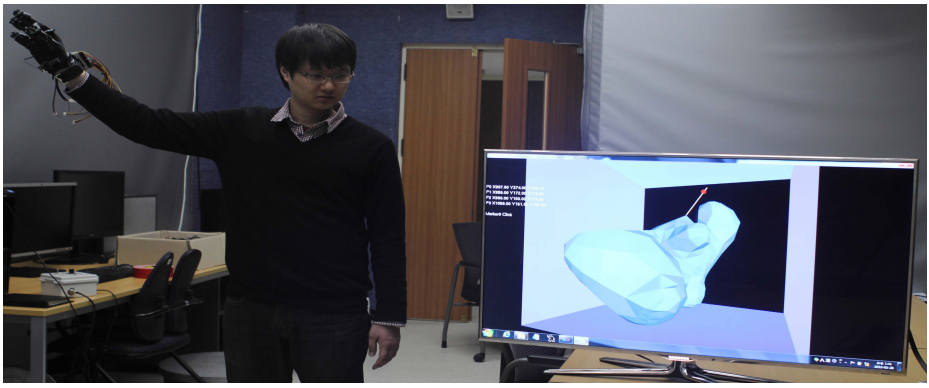
The tracking speed and the sensing accuracy of our system depends on photosensor's response time and lighting device's encoding resolution, which is given by the resolution of printed binary patterns in Fig.3 (c). Hence, improving the both tracking speed and accuracy is loosely related with cost factor, which is one of distinct benefits in our method comparing to vision-based approaches.

4 Experimental Results

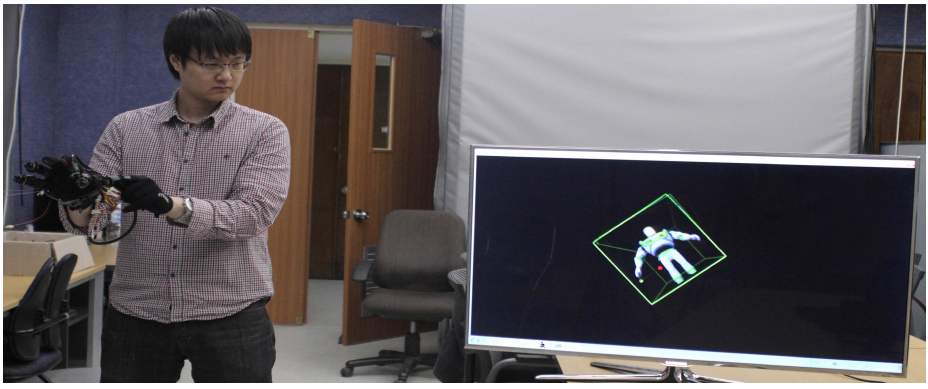
Demo video Fig.5 proves that our system's interaction resolution is 5mm at 3m distance between lighting devices and a marker. The marker's graphic representation, a white ball, moves according to actual marker's back and forth movement in 5mm interval in the video. In the supplementary videos of which still cuts are shown in Fig.6, we demonstrate hand gesture interaction where hand gestures are recognized as input commands to interact with digital information mimicking natural hand gestures for handling real objects. Fig.6 (a) shows an interaction demo with a deformable graphic object. With our system, a user can grasp the object at any point, elongate, press, and release it freely. Fig.6 (b) demonstrates interaction with a rigid object, Buzz. A user can grasp it at any point, freely manipulate it in translational and rotational movements, and change scaling. Fig.6 (c) shows multi-user interaction for a drawing operation. Two users performs drawing jobs in parallel with our system.

5 Conclusion

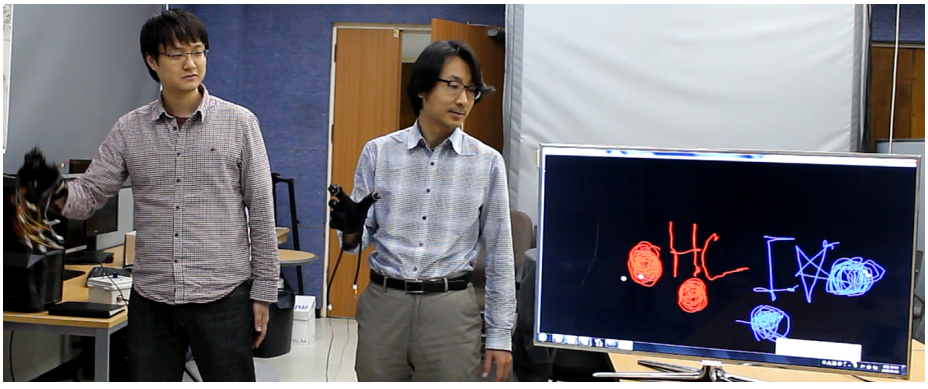
We presented a novel hand gesture interaction technique with a large working space (1m∼5m) based on spatio-temporal encoding method. Our unique tracking system equipped with photosensor markers is optimized for hand gesture tracking with covering large sensing range and minimizing finger's self-occlusion. Demo videos verify high possibility in practical applications demonstrating natural and free hand gesture interaction at high speed and good accuracy.



(a) Hand gesture interaction with a deformable object



(b) 6 DOF manipulation demo for a graphic object



(c) Multiple user interaction demo for a drawing operation

Fig. 6. Our method provides natural hand gesture applications with six markers' 3D tracking at high speed (250 Hz) and fair resolution (5mm at 3m working distance)

Our system has strengths in accuracy, speed, low price, and robustness to noise comparing with conventional long-range interaction techniques. Robust marker identification, less erroneous performance for finger's self-occlusion and little cost-performance dependency are additional advantages of our photosensor marker-based system.

Ackowledgements. This work was supported by the Global Frontier R&D Program on ⟨Human-centered Interaction for Coexistence⟩ funded by the National Research Foundation of Korea grant funded by the Korean Government(MSIP) (2010-0029752). We appreciate Maciej's help on the advices regarding interesting applications in physics simulation and especially the provision of source code for Soft Body 3.0 ([14]) which was used for our demonstration in Fig.6 (a).

References

1. Hirsch, M., Lanman, D., Holtzman, H., Raskar, R.: BiDi screen: a thin, depth-sensing LCD for 3D interaction using light fields. ACM Transactions on Graphics 28 (2009)
2. Mistry, P., Maes, P., Chang, L.: WUW - Wear Ur World - A Wearable Gestural Interface. In: The CHI 2009 Extended Abstracts on Human Factors in Computing Systems (2009)
3. Wang, R., Popovic, J.: Real-time hand-tracking with a color glove. ACM SIGGRAPH (2009)
4. Ballan, L., Taneja, A., Gall, J., Van Gool, L., Pollefeys, M.: Motion capture of hands in action using discriminative salient points. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 640–653. Springer, Heidelberg (2012)
5. Kim, D., Hilliges, O., Izadi, S., Butler, A.: Digits: freehand 3D interactions anywhere using a wrist-worn gloveless sensor. In: Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology (2012)
6. Nii, H., Sugimoto, M., Inami, M.: Smart Light Ultra High Speed Projector for Spatial Multiplexing Optical Transmission. In: Procams Workshop (held with IEEE CVPR) (2005)
7. Raskar, R., Beardsley, P., Van Baar, J., Wang, Y., Dietz, P., Lee, J., Leigh, D., Willwacher, T.: RFIG Lamps: Interacting with a Self-describing World via Photosensing Wireless Tags and Projectors. ACM Transactions on Graphics (SIGGRAPH) 23 (2004)
8. Lee, J.C., Hudson, S.E., Summet, J.W., Dietz, P.H.: Moveable Interactive Projected Displays using Projector Based Tracking. In: ACM Symposium on User Interface Software and Technology (UIST), pp. 63–72 (2005)
9. Welch, G., Bishop, G.: SCAAT: Incremental Tracking with Incomplete Information. In: Proceedings of SIGGRAPH 1997, Computer Graphics Proceedings. Annual Conference Series (1997)
10. Kang, S., Tesar, D.: Indoor GPS Metrology System with 3D Probe for Precision Applications. In: Proceedings of ASME IMECE 2004 International Mechanical Engineering Congress and RD and D Expo (2004)
11. Kessler, D., Hodges, L., Walker, N.: Evaluation of the CyberGlove as a Whole-Hand Input Device. ACM Tran. on Computer-Human Interactions 2(4), 263–283 (1995)

12. Raskar, R., Nii, H., Dedecker, B., Hashimoto, Y., Summet, J., Moore, D., Zhao, Y., Westhues, J., Dietz, P., Barnwell, J., Nayar, S., Inami, M., Bekaert, P., Noland, M., Branzoi, V., Bruns, E.: Prakash: lighting aware motion capture using photo-sensing markers and multiplexed illuminators. ACM Transactions on Graphics 26, 36 (2007)
13. Miaw, D.: Second Skin: motion capture with actuated feedback for motor learning. MIT Thesis (2010)
14. Matyka, M., Ollila, M.: A pressure model for soft body simulation. In: Proc. of Sigrad (2003)