Live Demonstration: Gesture-Based remote control using stereo pair of dynamic vision sensors

Junhaeng Lee[1], T. Delbruck[2], Paul K. J. Park[1], Michael Pfeiffer[2], Chang-Woo Shin[1], Hyunsurk Ryu[1], and Byung Chang Kang[1]
[1] Samsung Advanced Inst. of Technology, Samsung Electronics Co. Ltd. Republic of Korea
[2] Institute of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland

***Abstract*— This demonstration shows a natural gesture interface for console entertainment devices using as input a stereo pair of dynamic vision sensors. The event-based processing of the sparse sensor output allows fluid interaction at a laptop processor load of less than 3%.**

Associated Track 11.1: Sensory Systems: Image and Vision Sensors

I. DEMONSTRATION

Natural gesture interfaces have been under intense recent development. The Kinect system provides excellent performance but its static power consumption of 12W is very high for always-on applications. Therefore it is interesting from an ecological standpoint to look for passive sensors that could do the same task at much lower power consumption, without giving up fluid interaction capability. Recent developments in activity-driven, event-based vision sensors have opened up a promising alternative to conventional frame-based vision [1]. By outputting a sparse and variable-rate stream of data originating asynchronously from pixels with local gain control, these sensors reduce processing cost and latency, while increasing dynamic range. In this demonstration, we show how a stereo pair of event-based vision sensors can be used in a gesture control system that allows intuitive and fluid control of a mock entertainment device. The operation of this demonstration is described in the accompanying paper.

II. DEMONSTRATION SETUP

The demonstration setup consists of a laptop and a stereo pair of event-based DVS128 cameras [2]. A large monitor or beamer shows the output. The user interface allows control of various modes of operation, such as game selection, volume, channel, menu selection, pausing and playing, and training.

III. VISITOR EXPERIENCE

Visitors see how the activity-driven event-based sensors enable fluid interaction at low computational and power cost. In particular they can see how stereo vision is used in this application to focus computational effort automatically on the nearest moving object and how the stereo disparity is used to automatically adjust detection thresholds. After about 2 minutes of training naïve users can usually operate the demonstration with high reliability.

IV. DEMONSTRATION READINESS

The demonstration has been fully functional for about 9 months and has already been shown at an internal meeting.

REFERENCES

[1] Activity-Driven, Event-Based Vision Sensors, T. Delbruck, B. Linares-Barranco, E. Culurciello, C. Posch, IEEE International Symposium on Circuits and Systems, 2010. ISCAS 2010, Paris, France, pp. 2426 - 2429.
[2] Dynamic Vision Sensor (DVS) - asynchronous temporal contrast silicon retina. Retrieved 10.10.2011. Available: siliconretina.ini.uzh.ch

Fig. 1. Gesture demonstration setup. a) shows the stereo DVS sensor output along with tracking markers. b) shows part of the user interface, including two games (baloon popping and 3D drumming), the video player controller, and the tutorial screen for the gesture set.

Gesture-Based remote control using stereo pair of dynamic vision sensors

Junhaeng Lee[1], T. Delbruck[2], Paul K. J. Park[1], Michael Pfeiffer[2], Chang-Woo Shin[1], Hyunsurk Ryu[1], and Byung Chang Kang[1]
[1] Samsung Advanced Inst. of Technology, Samsung Electronics Co. Ltd. Republic of Korea
[2] Institute of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland

***Abstract*—This paper describes a novel gesture interface based on a stereo pair of event-based vision sensors and neuromorphic event processing techniques. The motion trajectory of a moving hand is detected every 3 ms by spatiotemporally correlating the output events of the DVSs by using leaky integrate-and-fire (LIF) neurons after the stereo vergence fusion. The trajectory of each gesture is automatically spotted by setting the threshold of LIF neurons, and, subsequently, sixteen feature vectors are extracted from each spotted gesture trajectory. The thresholds of LIF neurons are adaptively adjusted based on the disparity obtained from the stereovision to achieve distance invariant performance of gesture spotting. Gesture patterns were classified by using hidden Markov model (HMM)-based gesture models. The implemented system was tested with 6 subjects (3 untrained subjects and 3 trained subjects) producing continuous hand gestures (22 trials of 9 successive gestures for each subject). Achieved recognition rates ranged from 91.9 % to 99.5% depending on subject.**

V. INTRODUCTION

Natural and intuitive user interfaces (**UI**) based on body gestures, like Microsoft's Kinect, seems to be a promising solution for next generation UIs for multimedia devices like TVs, computers, game consoles, and mobile devices. The availability of the Kinect and of 3D time-of-flight cameras has generated a great deal of recent progress in gesture UIs [1]. However, these active sensors have several common drawbacks like constant high power dissipation, sensitivity to bright ambient lighting, and possible interference between multiple devices. Event based sensors like "dynamic vision sensors (**DVS**s)" [2] are promising candidates to address these problems. Each of these vision sensor pixels outputs a low-latency sparse stream of asynchronous events representing only changes in scene reflectance. Thanks to the sparseness of the event-stream outputs from the sensor, fast and efficient post-processing is possible. Using these sensors also allows us to efficiently solve the more restricted stereo fusion problem of tracking moving hands.

This paper describes a real-time hand gesture UI with 3ms response time using a stereo pair of DVSs, combined with a novel neuromorphic event processing algorithm for clustering moving hands [3]. The paper describes technical details on the gesture recognition process, consisting of clustering and tracking of a moving hand, gesture spotting, and gesture pattern classification. The paper concludes with experimental results and discussion.

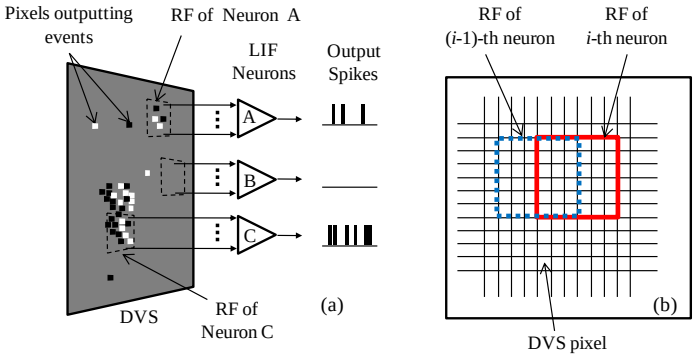


Fig. 1. (a) Illustration of correlating events using LIF neurons. (b) Receptive field (RF) of LIF neurons on DVS pixels. A pair of neighboring neurons is set to have overlapping region of their RF to create spatial correlation between them [3].

VI. ALGORITHMS

The DVS provides a simple but powerful way to detect moving objects. Since it outputs events only from the pixels which sense temporal dynamics of the scene, all stationary background images (i.e. non-moving objects) are blocked out. Detecting and tracking of moving objects is done by a novel method for spatiotemporally correlating the output events of the DVSs. We achieved this by using leaky integrate-and-fire neurons (**LIF** neurons) as shown in Fig. 1. The LIF neuron is a commonly used mathematical model of biological neurons. Its internal state (i.e. membrane potential) increases as it receives spikes (i.e. events) from presynaptic neurons, and fires a spike when the membrane potential exceeds the predefined threshold. After firing a spike event, its membrane potential jumps down by a certain amount or to a fixed level. The spatial correlation (i.e. intra-neuronal spatial correlation) is considered by defining the receptive field of the LIF neuron as shown in Fig. 1(a). The receptive field of a neuron is a region of space in which the pixels of the DVS make synaptic connections with the neuron. If there is strong enough spatiotemporal correlation between input events (i.e. if the pixels in the receptive field fire spikes at nearly the same time),

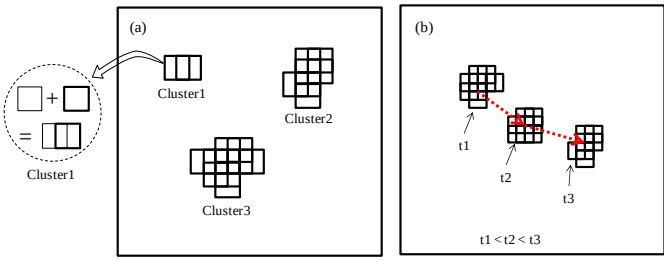
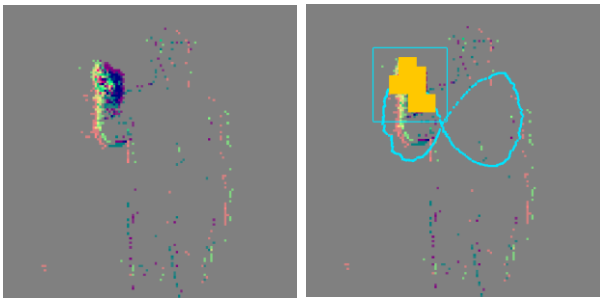


Fig. 2. (a) Clustering and (b) tracking of moving object by detecting active LIF neurons.



(a) (b)

Fig. 3. Screen captures of moving hand clustering and tracking. a) Stereo DVS input over 20ms; grey represents no events, red pixels have events from right DVS, and green from left DVS. b) Yellow filled rectangles are RFs of linked active neurons, and the sky blue hollow rectangle represents the cluster defined by these neurons. Sky blue dots are the trajectory of the hand.

the LIF neuron will fire a spike. Thus, by detecting the firing LIF neurons, we can easily find the active regions which contain highly spatiotemporally correlated events. Fig. 1(a) illustrates this process in which LIF neurons correlate output events from the DVS and detect the active regions. We set the LIF neuron to overlap half of its receptive field with each of its neighbors in all 4 directions to create spatial correlation between LIF neurons (i.e. inter-neuronal spatial correlation) as illustrated by the solid and dotted squares in Fig. 1(b). Clustering of events from the moving hand is achieved directly by grouping the firing LIF neurons 'linked' with overlapping receptive fields as shown in Fig. 2(a). Any two adjacent LIF neurons are defined as 'linked' if they have common receptive field and fire events within a certain time window. The location of the cluster is calculated by averaging the locations of all firing neurons in the group weighted by their average firing rate. The trajectory of the moving gesture is obtained by tracking the clusters as shown in Fig. 2(b). Although the event processing in LIF neurons is done asynchronously by the sensor input events, the trajectory information is sampled at fixed intervals (here every 3ms) during processing to be manipulated for gesture recognition. Fig. 3 shows an example of clustering and tracking of a moving hand based on LIF neurons. A large number of events (~50k per second) are generated by the DVS pixels seeing the moving hand, which stimulated the LIF neurons with the RF on the pixels to fire. On the other hand, only few events are observed from the pixel area of the stationary body, which is not enough to make LIF neurons fire.

Gesture spotting is a process to detect the start and the end points of a motion gesture. Spotting gestures is hard because the hand makes continuous trajectories even when it does not produce any meaningful gesture patterns. Previously, it has been shown that the threshold model or garbage model based on hidden Markov models (HMM) were effective to detect the spatiotemporal characteristics of gesture trajectories [4]. However, these models can be computationally expensive due to their large number of states with an ergodic transition matrix and frame-based calculation. In this paper, we utilize the properties of the spatiotemporal correlators (i.e. the LIF neurons) to spot the gesture trajectories, which practically adds no additional computational cost. By appropriately setting the threshold of LIF neurons, we can naturally segment the trajectory of hand movement whenever the hand moves slowly, stops, or abruptly changes its moving direction. The DVS generated much less events in such conditions for a moment (thus, LIF neurons did not fire), and we can easily and accurately detect this moment.

Stereo vision is a standard way to detect the distance between the sensors and an object. It is also useful for filtering out background noise by using the measured distance. In this paper, background noise is suppressed by using the stereo vergence fusion technique which combines events from the stereovision DVSs as shown in Fig. 4. The disparity (i.e., the positional difference between two images of a given point seen by the left and right eyes) is measured to achieve stereo vergence fusion by cross-correlating 1D fading event histograms of the left and the right eyes, where columns are summed into histogram bins, and finding the peak shift, which represents the dominant disparity. When we make a hand gesture in order to give a command, the events from the DVSs are mainly caused by the moving hand. Thus, stereo vergence based on the disparity significantly enhances the evidence of the moving hand by concentrating the events for this object. On the other hand, the events from the background movement do not overlap precisely by stereo vergence fusion. As a result, the stereo vergence significantly improves the performance of clustering in the environment of background movement noise. The clustering and tracking of the moving hand in the stereovision is achieved by using the LIF neurons after the stereo vergence as shown in Fig. 4(c). In this stage the background movement noise is further suppressed by using a stereovision association algorithm. We classified the events from the left and right DVSs based on their origin (i.e., left or right DVS) and polarity (i.e., ON or OFF event) when they were input to the LIF neurons for clustering. The LIF neurons adaptively adjust the synaptic weight for the left DVS based on the event rate for the same polarity from the right DVS inside its receptive field and vice versa. For example, if there is no event with ON polarity coming from the right DVS, the synaptic weight for ON events between the LIF neuron and the left DVS becomes zero and, as a result, the events from the left DVS no longer contribute to the membrane potential of the LIF neuron. This way, correlating events contribute multiplicatively to tracking rather than just by summing. The disparity of stereovision is also used for adaptive control of the threshold values of LIF neurons. The threshold values of

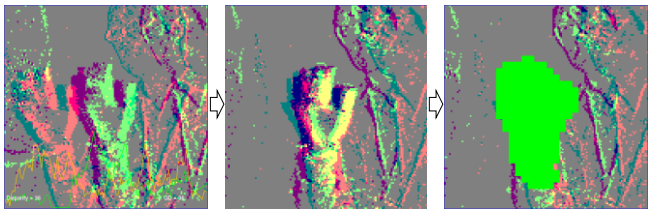


Fig. 4. Procedure of stereovision processing consisting of (a) disparity detection (b) stereo vergence fusion, and (c) stereovision association and clustering [3].

the LIF neurons are used for automatic gesture spotting. The performance of the gesture spotting depends on the threshold values of the LIF neurons. Thus, it is important to adaptively adjust the threshold value by considering the distance to achieve distance-invariant performance of gesture spotting.

Fig. 5 shows the block diagram of the proposed gesture recognition technique. After tracking the moving hand and spotting the gesture trajectory, a sequence of 16 feature values is extracted from each spotted trajectory (Fig. 3). We defined each feature as one out of 16 quantized directions of movements. We achieve size and space-invariant gesture recognition by cutting every spotted section into a fixed number of subsections (i.e. 16 sections in this paper) with equal length, and obtaining a fixed number of features for every trajectory regardless of its length. The sequence of features is sent to the pattern recognition module. The pattern of hand gesture is classified by using HMM-based gesture models. Since the trajectory of a hand gesture can be spotted during tracking, spatial pattern classification is sufficient for gesture recognition. The HMM-based gesture models are trained by using the Baum-Welch algorithm [5].

The gesture UI system for the live demonstration has two operating modes, graphical user interface (GUI) mode and gesture command mode. In the GUI mode, the position of a cursor on the screen is controlled based on the position of the moving hand just like a mouse interface in PC. On the other hand, in the gesture command mode, gesture patterns are used as a set of commands like buttons on a remote controller for TV. The major difference between the two modes is the capability of gesture spotting and continuous feedback to the users. In the gesture command mode, gesture spotting is essential for gesture pattern classification while it is not necessary in the GUI mode in which smooth and endless tracking of a moving hand is required. In the GUI mode, it is often difficult to track a slowly moving hand since it produces events at a rate that is too low to stimulate LIF neurons to fire. Thus, in this case, there is no firing neuron to detect for clustering and tracking. We solved this problem by a subthreshold tracking technique, in which a cluster is tracked based on membrane potentials of LIF neurons instead of their firing rate. If there is no linked neuron group detected to update a cluster during tracking (due to slow hand movement), the next position of the cluster is obtained from a virtual group of neurons, which virtually represents a group of firing neurons. The virtual group is created based on the current position and size of the cluster. All neurons in the area of the

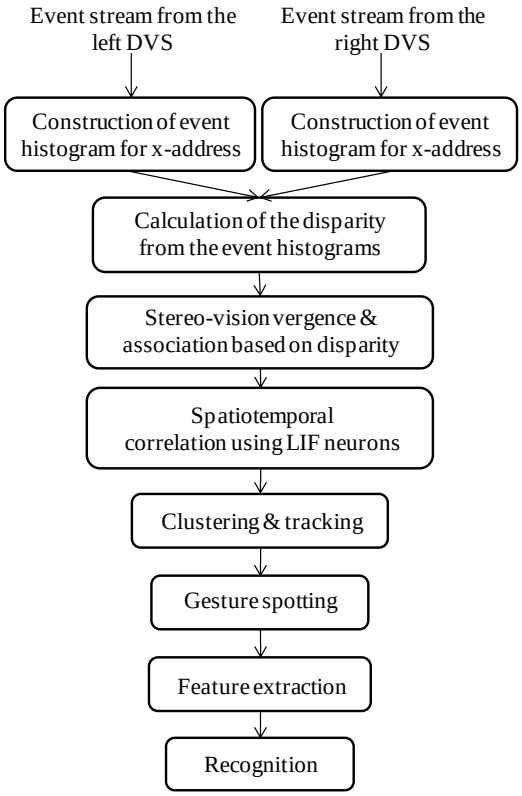


Fig. 5. Functional block diagram of overall gesture recognition procedure

cluster are assigned as members of the virtual group. The location of the virtual neuron group is calculated by averaging the locations of all member neurons in the group weighted by their membrane potentials.

VII. EXPERIMENTAL SETUP

Fig. 6(a) shows the experimental setup we used to evaluate the proposed gesture UI system. Two DVSs were fixed horizontally above a computer monitor to create a stereovision, pair and were connected to a computer via USB cables. The DVS chip used here consumes 23mW and has 128×128 pixels with a pixel size of 40μm×40μm [2]. The focal length and the maximum aperture ratio of the lenses used for the DVSs were 12mm and 1:1.4, respectively. Each HMM model was trained with 30 training gestures. Subsequently, we evaluated the gesture system for six subjects using the same models. Two of them (Subject-1 and Subject-2) had experienced the gesture systems several times before participating in the evaluation process. Another person (Subject-3) was exposed to the gesture system for around 20 minutes before the evaluation. The other three subjects (Subject-4, Subject-5, and Subject-6) were not allowed to experience the gesture system at all. The gesture sequence used in the test is shown in Fig. 6(b). The subjects were asked to produce these gestures continuously in order with hand motion as clear and natural as possible. No feedback about the classification result or the trace of hand motion was given to the subjects during gesture recording. The sequence of gestures in Fig. 6(b) was tried 22 times by each subject. Thus, 198 gestures were tested for each subject.

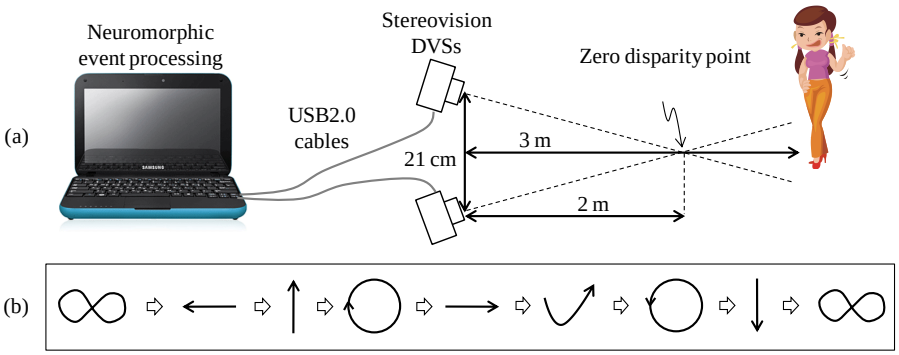


Fig. 6. (a) Experimental setup. (b) Sequence of test gestures [3].

Table 1. Recognition accuracy.

	Group of trained subjects			Group of untrained subjects		
	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6
# of test gesture sequences (# of gestures)	22 (198)	22 (198)	22 (198)	22 (198)	22 (198)	22 (198)
# of errors	1	2	4	3	16	11
Recognition rate (%)	99.49	98.99	97.98	98.48	91.92	94.44

VIII. Results and Discussion

The result of the evaluation for the 6 subjects is summarized in Table 1. The recognition rate was higher than 97% for the group of trained subjects (Subject-1,2,3). The recognition rate ranged from 91.92 % to 98.48 % for the group of untrained subjects (Subject-4, 5, 6). Except for Subject 5, we obtained good hit rate (better than 94 %) for the untrained subjects even though they were not exposed to the proposed gesture system at all. For Subject-5, we observed large distortions in motion trajectories of its hand. We found that this was mainly due to the black and white stripe patterns on the shirt of Subject-5. Significant artifacts were observed, caused by these stripe patterns. We also evaluated the performance of the proposed gesture system as a function of the distance between the DVSs and a subject. Distance invariant performance was achieved by automatically adjusting the LIF threshold based on the disparity, so that for subjects who are further away, who create smaller visual motion and hence fewer events, have a smaller threshold to cross to activate the gesture pattern. The recognition rates of the proposed technique were robust to scene illuminance conditions. Two decades of dynamic range from 10 to 1000 lux could be achieved by using only two threshold settings of LIF neurons. (10 lux represents dim indoor lighting and 1000 lux was the limit of our setup.) Scene illuminance invariant performance can be easily achieved by detecting and using the light intensity for adjusting LIF neurons in the future version of DVS. More than 50% of the computing power in the gesture recognition part was consumed by the clustering algorithm based on LIF neurons. Peak processor load on the PC was 3%. Such bio-inspired systems could in the future be efficiently implemented in dedicated neuromorphic hardware chips [6], which could be readily integrated with the DVS.

IX. Conclusion

This paper describes a gesture system based on an event-based vision sensor, the DVS, and an event-driven processing technique based on LIF neurons. The result shows that the recognition rate was better than 91 % regardless of user experience. Extracted dominant stereo disparity is used to verge the stereo inputs in order to concentrate activity onto LIF integrator neurons. Distance invariant performance and operation over a large dynamic range of 2 decades of illumination were achieved by automatically controlling the LIF neurons' threshold based on measured disparity.

References

[1] Kinect Overview [Online]. Available: http://www.xbox.com/en-GB/kinect ?xr=shellnav. [Accessed 13 Oct. 2011].

[2] P. Lichtsteiner, C. Posch, and T. Delbruck, "An 128x128 120dB 15us-latency temporal contrast vision sensor," IEEE J. Solid State Circuits, vol. 43, pp. 566-576, Feb. 2007.

[3] -, "Real-time gesture interface based on event-driven processing from stereo silicon ratinas," IEEE Trans. SMBC, under review.

[4] H.-K. Lee and J. H. Kim, "An HMM-based threshold model approach for gesture recognition," IEEE Trans. PAMI, vol. 21, no. 10, pp. 961-973, Oct. 1999.

[5] J. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and Hidden Markov Models," Tech. Rep., International Computer Science Institute, ICSI-TR 97-021, 1997.

[6] G. Indiveri, E. Chicca, R. J. Douglas, "Artificial cognitive systems: From VLSI networks of spiking neurons to neuromorphic cognition," Cogn. Comput., vol. 1, pp. 119-127, Jan. 2009.