PERFORMANCE IMPROVEMENT OF DEEP LEARNING BASED GESTURE RECOGNITION USING SPATIOTEMPORAL DEMOSAICING TECHNIQUE

Paul K. J. Park, Baek Hwan Cho, Jin Man Park, Kyoobin Lee, Ha Young Kim, Hyo Ah Kang, Hyun Goo Lee, Jooyeon Woo, Yohan Roh, Won Jo Lee, Chang-Woo Shin, Qiang Wang, and Hyunsurk Ryu

Samsung Electronics, SAIT, Samsung-ro 130, Yeongtong-gu, Suwon-si, 443-803 Korea

ABSTRACT

We propose a novel method for the demosaicing of event-based images that offers substantial performance improvement of far-distance gesture recognition based on deep Convolutional Neural Network. Unlike the conventional demosaicing technique using the spatial color interpolation of Bayer patterns, our new approach utilizes spatiotemporal correlation between pixel arrays, whereby timestamps of high-resolution pixels are efficiently generated in real-time from the event data. In this paper, we describe this new method and evaluate its performance with a hand motion recognition task.

Index Terms— Dynamic Vision Sensor, motion, recognition, demosaicing, convolutional, neural network

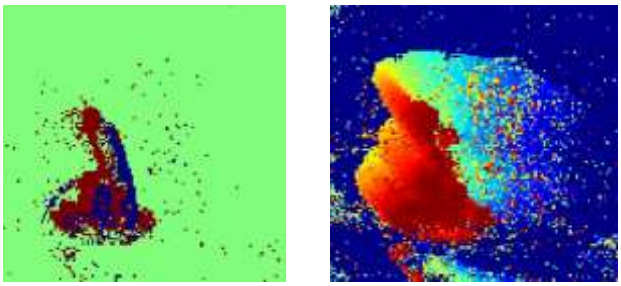
1. INTRODUCTION

The demand for next-generation image recognition systems for smart IT devices (such as TVs and electric vehicles), especially those based on innovative machine learning algorithms, has been rapidly increasing. For example, biometric technologies, as in face, fingerprint, and iris recognition, have been gaining significant traction in the market. In particular, motion recognition that accurately interprets sequential hand images is expected to play an important role in a variety of applications, especially in the form of an innovative motion-based user interface (UI) technology that transcends the limitations of current touch-based interactions as it provides a more natural user experience without the requirement for precise contact with the device. In this case, it is necessary to use distance-free gesture recognition for TV application because users want to control TVs in various distances. Thus, multi-scale inputs and size augmentation of training data can be used in deep Convolutional Neural Networks (CNNs) in order to achieve scale-invariant recognition [1]. However, it has been reported that the recognition performance would be deteriorated when the size of image is beyond the trained scale [2]. In this paper, we demonstrate that the demosaicing of far-distance images improves the recognition accuracy of deep CNN. In particular, we propose a novel technology for demosaicing the images of event-based sensor using spatiotemporal correlation. Our results show that, even though the far-distance small-size images were not trained in CNN, its demonstrated performance on super-resolved image data generated by using the proposed technique is comparable to that of the trained images. This is because the proposed technique restores the inherent properties (for example, optical flows of moving objects) of original event data efficiently and accurately, unlike using simple spatial interpolation [3]. Thus, this research suggests that the proposed demosaicing method holds promise for use in reliable recognition of even-based image sensor for future smart IT devices, not only enhancing the recognition performance of far-distance small-size images, but even enabling more precise motion analysis in sub-pixel resolution.

2. DYNAMIC VISION SENSOR

The neuromorphic vision sensor is a unique image sensor inspired by biological visual systems [4]. It operates as an activity-driven and event-based vision sensor known as the Dynamic Vision Sensor (DVS). The conventional CMOS Image Sensor (CIS) outputs the static and color image while the DVS output is the dynamic (temporal difference) and discrete binary image. In principle, the amount of information in a DVS image is much less than that in a CIS image, but its response time is faster. It should be noted that CIS is appropriate for use in precise and static tasks such as face and object recognition while DVS is more optimal for fast and dynamic motion (gesture) recognition. Thus, it has been recently shown that this sensor is as fast as used for real-time hand motion UI [5]-[6]. In principle, each DVS pixel produces a stream of asynchronous events just as the ganglion cells of biological retina do. Because it responds to the temporal contrast in a few microseconds, it is possible to quickly detect fast-moving objects. In addition, the DVS blocks all stationary background images since its pixels respond to the temporal contrast only. Thus, edges of a moving object can be simply detected without post image processing. By processing the information of local pixels

having relative intensity changes instead of entire images at fixed frame rates, the computational requirements can be reduced. Thanks to these sparsities and asynchronicities, the recognition task can be also achieved with low computational cost and latency [7]-[8]. Figure 1 (a) shows the event outputs (time duration = 20 ms) measured when the hand was moving before DVS. Two types of events (for example ON and OFF events) are generated in DVS. DVS outputs each event packet including x and y positions, event type, and firing time. Using this information, we can reconstruct the timestamp image whose pixel values represent the last firing time as shown in Fig. 1 (b). Thus, the trajectory, direction, and optical flow of moving objects can be easily analyzed by using this timestamp image [9]. In addition, in this paper, we propose a novel demosaicing technique using spatiotemporal correlation between timestamps.



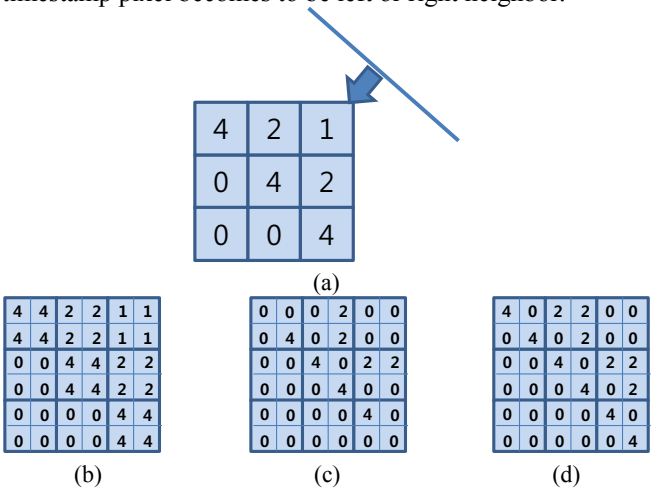
(a) (b)

Fig. 1. (a) Event output and (b) timestamp image of DVS

The DVS used in the experiment has 640 × 480 pixels and its pixel size is 9 μm × 9 μm. Even though this resolution is not sufficient to recognize the precise shape of objects in far distance, it is high enough for motion recognition. High speed USB 3.0 interfaces were used to deliver the asynchronous events with 10 microsecond resolution to a Personal Computer (PC) for motion gesture recognition.

3. SPATIOTEMPORAL DEMOSAICING

Figure 2 shows the operating principle of the proposed demosaicing technique. In this case, we assumed that the object was moved to the direction of left bottom corner from the right top corner. Thus, the first timestamp is generated in the pixel of right top corner then next events becomes to be fired to the direction of left bottom corner consecutively as shown in Fig. 2 (a). Here, we consider three kinds of demosaicing techniques to achieve x4 resolution. Firstly, the original timestamp of one pixel can be simply replicated to the corresponding four pixels of x4 image as shown in Fig. 2 (b). In addition, Figure 2 (c) shows that the timestamp of new super-resolved pixel can be generated by using the temporal interpolation between timestamps of three neighboring pixels at one corner. Thus, if one timestamp value of three neighboring pixels is close to that of the center pixel within 20 ms, the original timestamp can be delivered. Otherwise, new super-resolved pixel cannot be generated. Lastly, super-resolved timestamps can be generated by using spatiotemporal correlations among eight neighboring pixels at all corners as shown in Fig. 2 (d). For example, if the last timestamp among eight neighboring pixels is located at the one of four corners, then the timestamp of original center pixel can be delivered to two super-resolved pixels that lie along the diagonal line. In addition, if the last timestamp pixel is top or bottom, two pixels along the vertical line are generated. In this case, the location of vertical line can be decided by comparing the timestamps between left and right pixels. In the similar way, the horizontal pixels can be generated when the last timestamp pixel becomes to be left or right neighbor.



(a) (b) (c) (d)

Fig. 2. Timestamp maps of (a) original and x4 resolution generated by using (b) copy (pixel replication), (c) temporal interpolation, and (d) spatiotemporal correlation

Using these demosaicing techniques, we obtained super-resolved (x4) DVS images as shown in Fig. 3. The super-resolved image based on timestamp copy has exactly same shape as the original image. In this case, the number of events is four times larger than the original image. In addition, the quality of the super-resolved image based on temporal interpolation seems to be more deteriorated than the case using spatiotemporal correlation as shown in Fig. 3 (c) and (d).

(a) (b) (c) (d)

Fig. 3. (a) DVS images (time duration = 5 ms) of (a) original resolution, (b) x4 resolution based on timestamp copy (pixel replication), (c) x4 resolution based on temporal interpolation, and (d) x4 resolution based on spatiotemporal correlation

Figure 4 shows the optical flows measured when the user waved his hand. Basically, the optical flow of event-based image sensor can be derived by using the timestamps of neighboring pixels [10]. Because the number of super-resolved events is larger than that of the original image, the measured optical flows of super-resolved images are higher than the original case. However, it should be noted that the optical flow of super-resolved image based on the spatiotemporal correlation is nearly identical to the original one. This is because the inherent temporal properties of event based image sensor are maintained well even though new super-resolved pixels are generated by using the proposed spatiotemporal correlation technique.

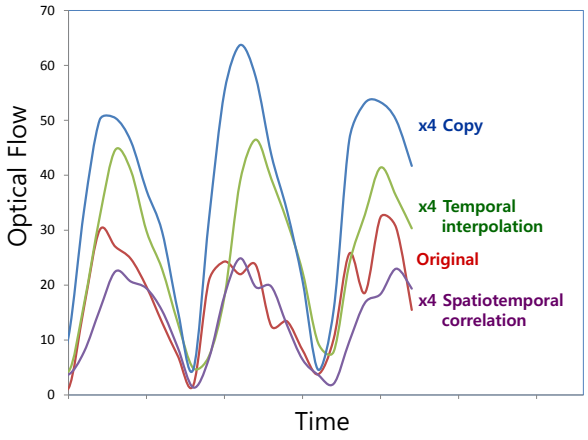


Fig. 4. Optical flows measured when the user waved his hand

4. GESTURE RECOGNITION

To evaluate the performance of the proposed demosaicing technique with a hand motion recognition task for TV application, we defined 6 hand gestures as shown in Table 1.

TV function	Gesture		Description
Gesture on		Hand Wave	Fast waving with open hand facing the screen.
Controller activation		Close	Change hand position from open to close. Hand closed with all fingers touching the thumb.
Command selection		Open	Change hand position from close to open.
Mode change		Flip	Fast change hand position from open to lateral, and back to open.
Pointer on		Finger wave	Fast waving with index finger and other fingers folded.
Pointer selection		Pinch	Fast moving with index finger from open to pinch, and back to open.

Table 1. List of gestures for TV control

To detect hand motions, we utilized structured output Deep Belief Network (s-DBN). It had been reported that this detection network is superior to Random Forest (RF) in terms of performance and speed [11]. The gesture of this detected hand was classified by using the frame-fusion CNN. In this case, in order to classify moving hand gestures, we trained 7 gestures (6 gestures + 1 others) in the model. Figure 5 shows the operating principle of frame-fusion technique. Three different events frames are merged into one 3ch image. For example, the first channel data is composed of the events generated from t-20ms to t and the second channel data is made by using the events from t-60ms to t-40ms, and the 3rd channel data corresponds to t-100ms to t-80ms, respectively.

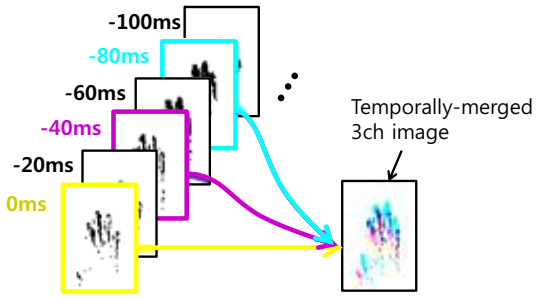


Fig. 5. Temporal merging of hand images

Using this temporal-fusion data, the data augmentation process was performed and finally we obtained 3.6 million early frame-fusion images for training. In our experiments, we utilized the minor-changed GoogLeNet (i.e., 5x5 convolution layer was replaced with double 3x3 layers in the inception module) as a classification module [12]. To investigate the performance of gesture recognition, we recorded test DB including various distances between DVS and a user. The measured distance is ranging from 1m to 5m with a 1-m step. Each data set includes 100 gesture videos and each video has 6 consecutive gestures. Figure 6 shows the F values (defined by using recall and precision accuracies) while varying the distances between DVS and a user. The results show that the F values measured from 2 to 4 meters are around 90%. In this case, we compared the performance of two post-processing methods (voting vs. LSTM). LSTM (Long Short Term Memory) is a recurrent neural net which has additional weight matrixes to remember or forget the temporal patterns [13]. Our results show that LSTM outperforms the temporal voting. This is mainly because LSTM predicts the gestures dynamics which are basically the sequences of high level visual features. The LSTM layer used in the experiment has 1,024 hidden dimensions and 1 core. However, it should be pointed out that the recognition accuracies of 1-m and 5-m distances are significantly reduced because the trained images were in the range of from 2m to 3m.

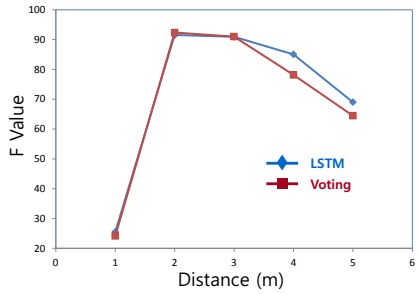


Fig. 6. Recognition performance (F values) measured while varying the distances

To improve the performance of far distance, we applied the proposed demosaicing technique to our gesture recognition system. Figure 7 shows the recognition recall accuracies measured at 5-m distance according to various gesture types. As we expected, the demosaicing technique based on the proposed spatiotemporal correlation has the best performance. The accuracy of original images measured to be 77% at 5-m distance. This accuracy was dramatically improved to be up to 93%, which was comparable to 3-m case (93.7%). However, the measured performance of the temporal interpolation technique was rather degraded due to the incomplete timestamp reconstruction as shown in Fig. 3 (c). In particular, it should be noted that, the performance improvement of short-time gestures (HC, HO, HF) is larger than in case of long-time gestures (HW, FW). Our results confirm that the proposed spatiotemporal demosaicing technique restores the inherent properties of original event data more efficiently and accurately than the spatial interpolation.

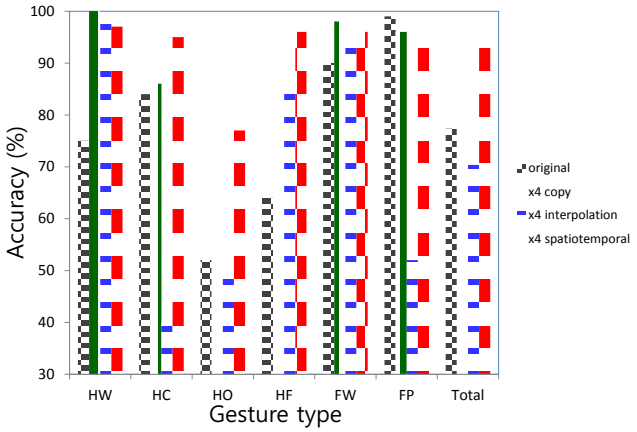


Fig. 7. Recognition accuracies measured at 5 m distance according to various gesture types (HW: hand wave, HC: hand close, HO: hand open, HF: hand flip, FW: finger wave, and FP: finger pinch)

In addition, for the near-distance recognition, we utilized the sub-sampled DVS images as shown in Fig. 8. Fig. 8 (a) shows the DVS hand image detected at 1-m distance. Only one finger is recognized as a hand while the other fingers as body. In this case, we can detect full hand by using the sub-sampled DVS images as shown in the Fig. 8 (b). By using these sub-sampled images, we can detect, track, and recognize the hand gestures in the near distance (~1 m).

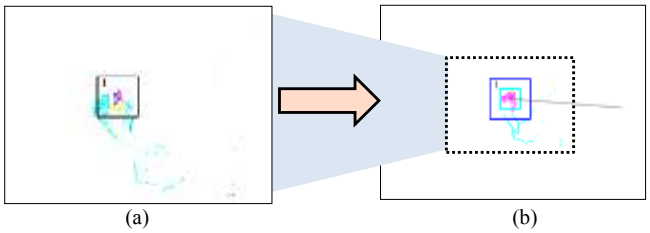


Fig. 8. (a) DVS hand image detected at 1-m distance and (b) sub-sampled DVS image

Using this technique, we confirmed that the recall accuracy was significantly improved as shown in Fig. 9 (15% → 81%). Thus, the results show that, even though our gesture recognition system was trained to maximize the accuracy of 3-m distance, the performance degradation of different-scale data can be compensated sufficiently by using the multi-scale network structure. In this case, the proposed demosaicing technique based on spatiotemporal correlation can be utilized to scale up the event-based image very efficiently. Moreover, we envision that as the event-based image sensors become even more widespread, for example with the electrical vehicle in addition to the TV application, the proposed technique will also play a critical role in the development of motion recognition and dynamic signal processing at sub-pixel resolution.

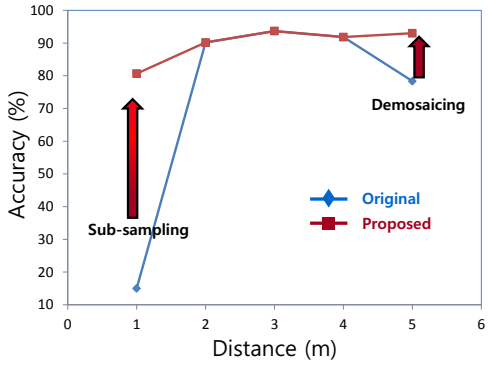


Fig. 9. Performance improvement of gesture recognition using multi-scale inputs

5. REFERENCES

[1] D. Yoo, S. Park, J.-Y. Lee, and I. S. Kweon, "Multi-scale Pyramid Pooling for Deep Convolutional Representation," *CVPR*, pp. 71-80, 2015.

[2] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *ICLR*, 2014.

[3] D. Paliy, R. Bilcu, V. Katkovnik, and M. Vahviläinen, "Color filter array interpolation based on spatial adaptivity," *Electronic Imaging*, 2007.

[4] P. Lichtsteiner, C. Posch, and T. Delbruck, "An 128x128 120dB 15us-latency temporal contrast vision sensor," *IEEE J. Solid State Circuits*, pp. 566-576, 2007.
[5] J. H. Lee, P. K. J. Park, C.-H. Shin, H. Ryu, B. C. Kang, and T. Delbruck, "Touchless hand gesture UI with instantaneous responses," *IEEE ICIP*, pp. 1957-1960, 2012.
[6] P. K. J. Park, J. H. Lee, C.-H. Shin, H. Ryu, B. C. Kang, G. A. Carpenter, and S. Grossberg, "Gesture recognition system based on Adaptive Resonance Theory," *IEEE ICPR*, pp. 3819-3822, 2012.
[7] P. K. J. Park, K. Lee, J. H. Lee, B. K. Kang, C.-H. Shin, J. Woo, J.-S. Kim, Y. Suh, S. H. Kim, S. Moradi, O. Gurel, and H. Ryu, "Computationally efficient, real-time motion recognition based on bio-inspired visual and cognitive processing," *IEEE ICIP*, pp. 932-935, 2015.
[8] J. H. Lee, T. Delbruck, M. Pfeiffer, P. K. J. Park, C.-H. Shin, H. Ryu, and B. C. Kang, "Real-Time Gesture Interface Based on Event-Driven Processing From Stereo Silicon Retinas," *IEEE Trans. Neural Netw. Learning Syst.*, pp. 2250-2263, 2014.
[9] J. H. Lee, K. Lee, H. Ryu, P. K. J. Park, C.-H. Shin, J. Woo, and J.-S. Kim, "Real-time motion estimation based on event-based vision sensor," *IEEE ICIP*, pp. 204-208, 2014.
[10] K. Lee, H. Ryu, S. Park, J. H. Lee, P. K. J. Park, C.-H. Shin, J. Woo, T.-C. Kim, and B. C. Kang, "Four DoF Gesture Recognition with an Event-based Image Sensor," *IEEE GCCE*, pp. 293-294, 2012.
[11] W. Mao, Q. Wang, X. Wang, P. Guo, S. Wang, G. Shao, K. Lee, and P. K. J. Park, "Real-time human body parts localization from dynamic vision sensor," *IEEE ICIP*, pp. 4783-4787, 2015.
[12] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," *ARXIV:*1502.03167, 2015.
[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, pp. 1735-1780, 1997.