Address-Event Video Streaming over Wireless Sensor Networks

Eugenio Culurciello, Joon Hyuk Park
E-Lab
Yale University
New Haven, CT 06520
eugenio.culurciello@yale.edu

Andreas Savvides
Embedded Networks and Applications Lab
Yale University
New Haven, CT 06520
andreas.savvides@yale.edu

***Abstract*— We explore an algorithm and methodology for real-time video compression and communication over sensor network. Video is encoded using the address-event representation (AER) and using custom image sensors capable of detecting intensity-differences (motion) information. The work focuses on keeping a constant low bit-rate over sensor network channels based on ZigBee radios, so that the motion information in an image can degrade gracefully when the resolution or the image content is varied over time. We show that a 320x240 video stream can be compressed up to 50 times while being intelligible for interpreting motion patterns. This compression requires zero computation when the image is encoded with AER. The work will be applied to remote monitoring of home-care patients and surveillance.**

I. Introduction

In recent years, Sensor Networks have proliferated significantly to a variety of applications and sensing modalities [1]. Image Sensor Network (ISN) can be applied to a variety of applications, such as surveillance, monitoring and security [2], where the transmission of video is paramount. ISN have recently acquired particular interest for their possible use to monitor the elderly aging at home [3]. In fact, processing video with sensor network will eliminate the need for elderly to remember to wear cumbersome instrumentation and sensors. A sensor network for home-care environments uses video transmission as sensory modality for identifying patients' behavior [3]. Video streams are processed to extract features that can be processed with sensory grammars such as BehaviorScope [4].

Because of the large data collected by image sensors, and the low-power and low-bandwidth nature of sensor networks, transmitting video over Sensor Networks is still a difficult endeavor despite the progress in video signal processing. The XBow MicaZ and the Intel iMote2 ZigBee-powered sensor network nodes can communicate data at approximately 250kbps and will thus be barely able to stream a 320x240 frame-difference video (for motion detection) at about 2fps. This frame rate makes it hard to understand the content of these videos. Even worse, all the communication capabilities of the node are used by streaming the video, compromising the networking share of data across nodes and the usefulness of the distributed sensor network deployment. Table I summarizes the capabilities of popular nodes for streaming video. The video can be compressed with standard DCT-based techniques before being communicated by the sensor network, but the consequences are: a low battery life, increased latency and the use of the node processor for compression only [5], [6], [7].

	Frame size [kb]	Mica2 [fps]	iMote [fps]
320x240 Edge	76.8	0.5	3.26
320x240 Differencing	153.6	0.25	1.63
320x240 Grayscale	614.4	0.06	0.41

TABLE I
VIDEO STREAMING CAPABILITIES OF POPULAR SENSOR NETWORK NODES.

In this paper, we report a zero-computation inexpensive methodology for compressing frame-difference video. This compression enables wireless sensor nodes to stream temporal frame-difference video over wireless networks at high rates. The technique can be applied to home-care ISNs to monitor patients un-intrusively. The use of frame-difference videos preserves the privacy of the users because intensity images are not stored or recorded by the ISN.

II. Frame-Differencing Motion Videos

Motion in a scene can be revealed by using temporal differences between consecutive frames. Frame-differencing is a form of temporal compression that examines redundancies between adjacent frames in a moving image sequence. A frame-difference movie is obtained by subtracting each pixel intensity values from the previous sampled frame from the corresponding pixel intensity values of the current frame. To display the frame-difference video on a standard display, if the difference between the pixel intensities are negative, the pixel is set to white. If the difference is positive, the pixel is set to black. If there is very little or no difference in pixel intensity between the two frames, the pixel is set to gray. An example of frame-differencing is shown in Figure 1. Frame-differencing is important to segment moving objects from the background and is usually the first processing step employed in artificial video processing [4]. There are advantages of streaming frame-difference video. First, due to the reduction of intensity information, the observer will not be able to determine the identity of the target. On the other hand, the

1-4244-0921-7/07 $25.00 © 2007 IEEE.

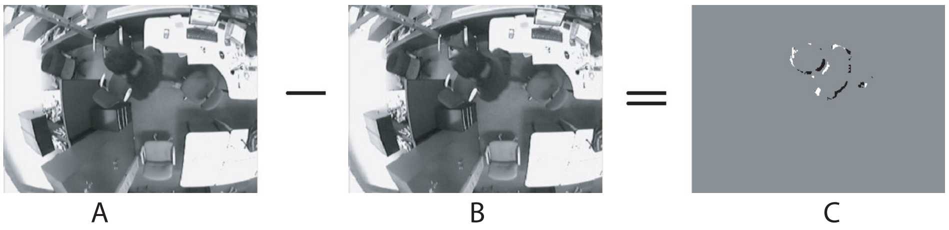
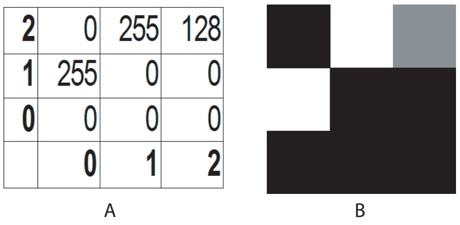


Fig. 1. Frame-differencing: (A) current frame, (B) previous frame, (C) result of subtracting the corresponding pixel intensity values of (B) from (A).

observer will still be able to understand the actions of the subject. This will protect the privacy of the subject being observed. Second, the processed videos require less bandwidth than a grayscale video. Each pixel value in a grayscale video has a size of 8 bits. Instead, a frame-difference images uses only 2 bits per pixel, and AER requires a single bit per event, as explained in the next section.



2	0	255	128
1	255	0	0
0	0	0	0
	0	1	2

Fig. 2. A simple 3x3 frame-difference image: (A) the pixel values of the image, (B) corresponding image.

III. Address-Event Image Sensors

Frame-difference videos can be further compressed by using the Address-Event Representation (AER). AER is a biologically-inspired asynchronous protocol for encoding and communicating sensory data between a transmitting sensor and a receiving processing unit [8]. In an frame-difference AER image sensor [9], events are signaled when changes in pixel intensity reach a threshold voltage. An event has the simple form of the address of the transmitting element (hence the origin of the term address-event). The advantage of AE image sensors is that they do not need to be queried for information. Instead, they push information to the receiver once they have gathered it. This is a important feature in data-driven sensor networks, since image sensors can detect features of interest in the environment by itself and provide hardware triggers that do not need to be polled for information [9], [8]. Also AER sensors automatically provide a rank encoding of data, based on importance, an important feature that can be leveraged for compression. In a frame-difference AER image sensor the pixels that experience large intensity changes will generate events first and more frequently than others, so the data from these pixels will become available immediately to a receiver [8]. The frequency that the event is generated is proportional to the value of the pixel. So for a simple frame-difference 3x3 image shown in Figure 2, the event train of the output would look as in Figure 3. The node receives these events from the image sensors and is able to reconstruct the image.

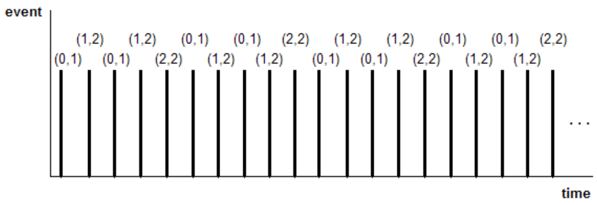


Fig. 3. Train of events generated by an AER image sensor given the image in Figure 2.

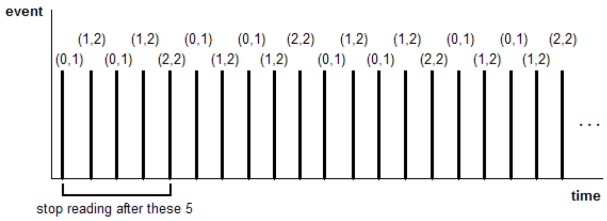


Fig. 4. Compressing the image in Figure 2 using AER only requires reading less events, and no computation whatsoever.

IV. AER Compression

In a traditional scanned image sensor, every pixel of the image frame is read sequentially and then processing is performed on the image frame to compress the data. This imposes a computational burden on the microcontroller of a sensor node. Instead, compression of an AER stream requires no computation whatsoever. All that is required is to read less events outputted by the address-event image sensor, as portrayed in figure 4. This is due to the fact that events are automatically randomized by the sensor itself.

It is important to note that AER image sensors will not always be beneficial in bandwidth usage. Only *sparse images* can be significantly compressed. As an example, for representing a 320x240 frame-difference image, a scanning image sensor will need three colors, white, black and gray, which requires 2 bits per pixel. Therefore, for each differential frame, 320 x 240 pixels x 2 bits/pixel = 153600 bits need to be transmitted by the network. Due to the nature of AER, if the image to be transmitted is not sparse, an AE image sensor could end up using more than 153.6 kbits of bandwidth. For a 320x240 image, 9 bits are needed for the value of row address and 8 bits are needed for the value of column address for a total of 17 bits per event. So the bandwidth requirement of an AE image sensor is a function of the number of events (N_{ev}) per frame: N_{ev} x 17 bits/event = size of each frame. As shown in Figure 5 for 320 x 240 pixels images, after about 9000 events, encoding with AER is less efficient than using a scanning image sensor. However, our application is able to utilize a compressed AE image that requires much less bandwidth than a scanning image sensor, as shown in later sections. We have experimented with contour-based AER-encoded images as well, but have found that we could not achieve compression ratio of more than 2 (for 320 x 240 pixels images), due to the larger number of active pixels in the contour image (the image is *less sparse*). This can be explained by Figure 5: a 320 x 240 pixels motion image generates less than 1000events/frame, while a contour one more than 4000events/frame, compromising compression levels.

V. Test Methodology

In order to determine how much a video can be compressed, a psychophysical test was performed on human subjects. The subjects ranged in ages of 18 to 43 and had varying backgrounds. The test video was captured at 320x240 from a wide-lens camera looking straight down at a laboratory. The camera was located 10 feet above the ground. The test video mimics the actions of a person being observed in a home-care environment. In the video, a person gets up from his desk, walks over to another chair and sits down, then stretches out with his arms crossed, then gets up and sits back at his desk. Figure 6 shows some of the frames from the video. The test video was converted to a frame-difference video at various compression levels using an AER emulator from ENALAB [3]. A non-AER frame-difference video was also created from the same video. To determine if people would be able to understand the uncompressed, non-AER frame-difference video, this video was shown to 10 subjects. All the people that viewed the uncompressed non-AER video were able to determine the actions of the person in the video. The compressed AER video was shown to 10 different subjects at a very high compression level, and the compression ratio was gradually decreased until the test subject was able to give a clear description of what was happening in the test video. Figure 7 shows the gradual decrease of compression level and the corresponding image reconstructed from the AER stream.

VI. Results

Ten subjects were tested in the method given in the previous section. Table II reports the number events required before the subjects were able to give a clear description of the person's actions in the test video. The average number of events required for the understanding of the video is 180 events per frame with a standard deviation of 33. Since this result could have been biased from the prior shows of lower-quality version of the same video, we verified that this average is not affected by the bias. Therefore, another test was run where a single 180 events/frame movie was shown to a different set of subjects instead of gradually increasing the quality of the video. Eight subjects who were tested in this method were all able to give a clear description of the video. If only 180 events per frame is required for human understanding, we can calculate the compression ratio and the bandwidth requirement for the video. For a 320x240 image, 17 bits are required to represent the pixel addresses (9 for column and 8 for row). This means that each frame in the video requires 180 events x 17 bits/event = 3.06kbits Comparing this value to the bandwidth requirement of a non AER imager 320 x 240 pixels x 2 bits/pixel = 153.6kbits the compressed address-event frame-differencing video requires 50 times less bandwidth than a scanning image sensor. The Mica2 would be able to stream this video at 12fps and the MicaZ / Intel iMote2 at 83fps.

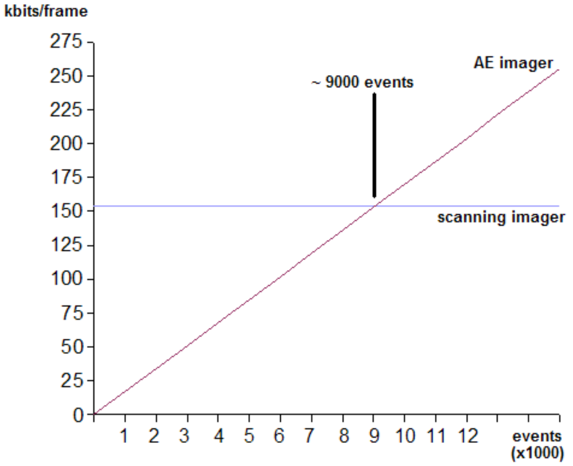


Fig. 5. Bandwidth usage comparison of traditional scanning imagers and AER imagers for a 320x240 frame-difference image.

s1	s2	s3	s4	s5	s6	s7	s8	s9	s10
150	200	150	150	200	150	200	250	200	150

TABLE II

Results from showing a random group of people a gradually improving frame-difference movie. Top: subject, bottom: events needed for interpreting video.

VII. Conclusion

The communication capabilities of current wireless sensor nodes do not enable the network to stream high-detail videos. To reduce the network load while streaming videos, processed videos such as edge or frame-difference videos can be used. There are many applications, as Home-Care networks, where streaming frame-difference videos is necessary. However, even for such videos, the bandwidth requirements are still too high

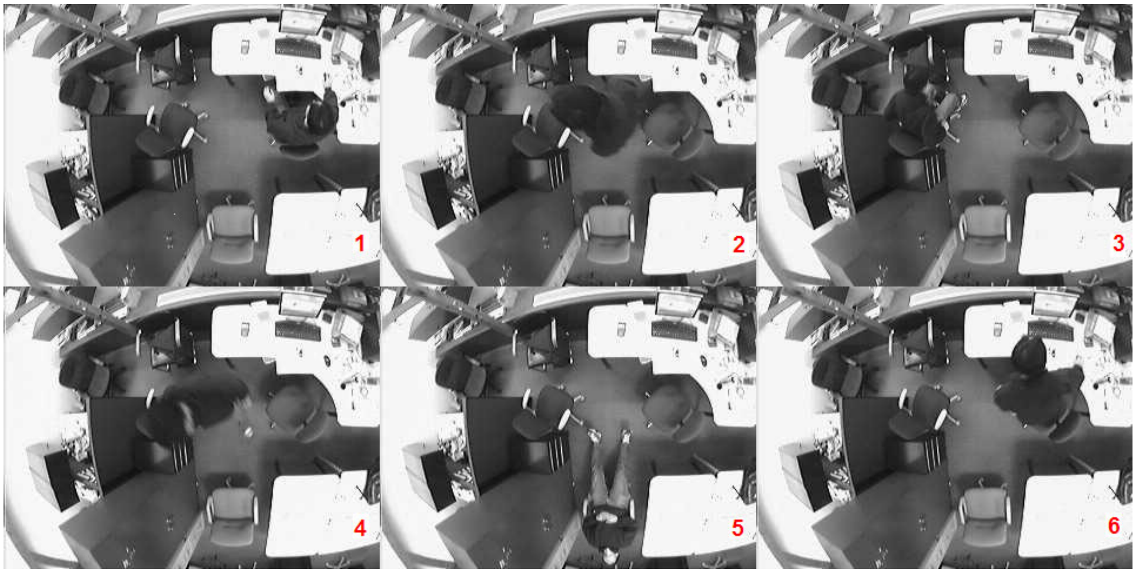


Fig. 6. Clips from the experimental video. 1: shows a person sitting at a desk. 2: shows the person walking to another chair. 3: shows the person sitting down. 4: shows the person walking to a different chair. 5: shows the person stretched out in the chair. 6: shows the person sitting back at the desk.

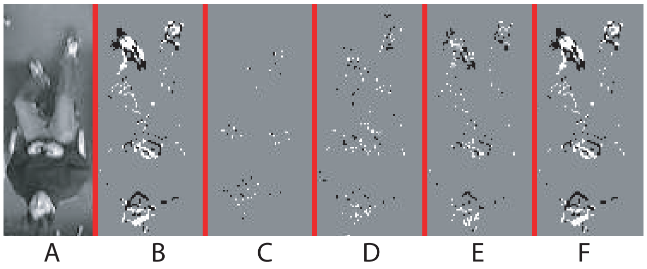


Fig. 7. (A) The original image. (B) non AER frame-difference. (C) AER frame-difference image composed from 50 events. (D) AER frame difference image composed from 150 events. (E) AER frame-difference image composed from 250 events. (F) AER frame-difference image composed from 10000 events.

for the network. To address this issue, compression can be used, but there is too much overhead for such a process to run on a node. In this paper we have demonstrated that address-event image sensors can alleviate the compression of videos employing zero resources from the sensor network node's microcontroller. We demonstrated with a psychophysical experiment that compression of a 320x240-pixel frame-difference images using address-event resulted in 50 times less bandwidth usage than non-AER scanned video. This enables sensor network nodes to stream frame-difference video at very high frame rates for use in home-care environments.

VIII. Future Work

We are instrumenting sensor network nodes with AER custom image sensors [9], [8] for a demonstration of the video streaming capabilities and feature extraction in a home-care environment. Up-to-date information on this effort can be obtained from our websites at http://www.eng.yale.edu/enalab and http://www.eng.yale.edu/elab, and will be presented live at ISCAS 2007.

Acknowledgment

This work was partly funded by NSF award 0622133 and CISE/CNS award 0448082. The authors would also like Thiago Teixeira for the fruitful discussions.

References

[1] Various Authors, *Proceedings of IEEE Special Issue on Sensor Networks*. IEEE, August 2003.

[2] T. He, P. Vicaire, T. Yan, Q. Cao, G. Zhou, L. Gu, L. Luo, R. Stoleru, J. Stankovic, and T.Abdelzaher, "Achieving long-Term Surveillance in VigilNet," in *Infocom 2006*, April 2006.

[3] T. Teixeira, E. Culurciello, E. Park, D. Lymberopoulos, A. Barton-Sweeney, and A. Savvides, "Address-event imagers for sensor networks: Evaluation and modeling," in *SPOTS '06*, November 2006.

[4] D. Lymberopoulos, A. Ogale, A. Savvides, and Y. Aloimonos, "A sensory grammar for inferring behaviors in sensor networks," in *Information Processing in Sensor Networks (IPSN)*, April 2006.

[5] E. M. C.F. Chiasserini, "Energy-efficient coding and error control for wireless video-surveillance networks," in *Telecommunication Systems*, vol. 26, JUN-AUG 2004, pp. 369–387.

[6] M. Rahimi, D. Estrin, R. Baer, H. Uyeno, and J. Warrior, "Cyclops: image sensing and interpretation in wireless networks," in *Second ACM Conference on Embedded Networked Sensor Systems, SenSys*, Baltimore,MD, November 2004.

[7] D. Lymberopoulos and A. Savvides, "XYZ: A motion-enabled, power aware sensor node platform for distributed sensor network applications," in *Information Processing in Sensor Networks (IPSN)*, April 2005.

[8] E. Culurciello and A. Andreou, "CMOS image sensors for sensor networks," *Analog Integrated Circuits and Signal Processing*, vol. 49, no. 1, pp. 39 – 51, 2006.

[9] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128128 120dB 30mW asynchronous vision sensor that responds to relative intensity change," in *IEEE International Solid State Circuits Conference (ISSCC)*, vol. 4, February 2006, pp. 508 – 509.