Event-driven Body Motion Analysis For Real-time Gesture Recognition

Bernhard Kohn*, Ahmed Nabil Belbachir*, Thomas Hahn* and Hannes Kaufmann†
*AIT Austrian Institute of Technology GmbH 1220 Vienna, Austria, Email: bernhard.kohn@ait.ac.at
†Vienna University of Technology Interactive Media Systems Group 1040 Vienna, Austria kaufmann@ims.tuwien.ac.at

***Abstract*— This paper presents an evaluation of spatio-temporal data generated by a dynamic stereo vision sensor in a highdimensional space (3D volume and time) for motion analysis and gesture recognition. In contrast to traditional frame-based (synchronous) stereo cameras, dynamic stereo vision sensors asynchronously generates events upon scene dynamics. Motion activities are intrinsically (on-chip) segmented by the sensor, such that activity, gesture recognition and tracking can be intuitively and efficiently performed. In this work, we investigated the applicability of this sensor for gesture recognition. We developed a machine lerning method based on the Hidden Markow Model for training and automated classifications of gestures using the event data generated by the sensor. By training eight different activities (dance figures) with 15 persons we build up a library of 580 recorded activites. An average recognition rate of 97% has been reached.**

I. Introduction

Gesture recognition is a well investigated topic in the past years, motivated by providing new systems for human-machine interaction. Many robust technologies including 3D, time of flight or structured light cameras (e.g. Kinect) were established on which gesture recognition can be applied. Furthermore, many computer vision methods were investigated for robust and real-time recognition including Hidden-Markov-Model [1]–[3], artificial neural networks [4], [5], decision trees and support vector machines [6], [7].

All these computer vision methods deal with processing a synchronous sequence of intensity images taken at equidistant time intervals. The employed methods mainly make use of object detection in single frames and recovering the motion across the sequence of frames in a spatio-temporal representation of the pose and gesture.

In 2005, an event-driven dynamic vision sensor [8], [9] was developed for capturing scene dynamics and asynchronously generating events upon relative light intensity changes in the scene. Unlike conventional frame-based cameras, this sensor generates a continuous stream of events representing the motion path of the scene dynamic at high temporal resolution. Furthermore, the sensor outputs very low data volume compared to a frame-based sensor, such that the real-time application is possible. By offering a complete background subtraction on-chip, this sensor can be ideal for motion capturing and activity recognition of persons.

To our knowledge, there exists no refereed literature on gesture recognition estimation using this type of sensor. Several ideas on using this type of sensor for gesture recognition are mentioned on the Capo Caccia Cognitive Neuromorphic Engineering Workshops [10]–[12]. Therefore, we exploit this opportunity to evaluate the applicability of this sensor for real-time interaction taking into account its capability for real-time motion capturing in a 4D spatio-temporal space at lower data volume than standard 3D cameras.

This paper is structured as follows. To make the paper self-contained, we present in section II a short review on the event-driven stereo vision sensor. Section III introduces how gestures are represented by the dynamic stereo vision sensor. Furthermore we will briefly introduce the applied method for the evaluation and will show first results of gesture recognition with the dynamic stereo vision sensor. We finishes the paper with conclusions in the section IV.

II. Event-driven Stereo Vision Sensor

In this section, a general overview of the dynamic stereo vision sensor is given including figures illustrating the spatio-temporal data. The dynamic stereo vision sensor consists of a pair of arrays of 304x240 pixels each, built in a standard $0.18\mu m$ CMOS technology. The array elements (pixels) respond to relative light intensity changes by instantaneously sending their address, i.e. their position in the pixel matrix, asynchronously over a shared bus to a receiver using a request-acknowledge 2-phase handshake.

Such address-events (AEs) generated by the sensors arrive first at the multiplexer unit. Subsequently, they are forwarded to an FPGA, which attaches to each AE a timestamp at a resolution, which can be set between $12.5ns$ and $100\mu s$. The combined data (AEs and timestamps) are used as input stream for 3D map generation and subsequent processing.

Figure 1 depicts a spatio-temporal data representation of one dynamic vision detector, resulting from a two persons crossing the sensor field of view in a room-like environment. The events are represented in a 3D volume with the coordinates x (0:303), y (0:239) and t (last elapsed ms), the so-called space-time representation.

The bold colored dots represent the events generated in the recent $16ms$. The blue and red dots represent spike activity generated by a sensed light-intensity increase (ON-event) and decrease (OFF-event) resulting from the person motions, respectively. The small gray dots are the events generated in the elapsed $1.5s$ prior to the recent $16ms$. This event history highlights the dynamics path in the past $1.5s$ resulting from the moving persons, which is an ideal basis for continuous

978-1-4673-0219-7/12/$31.00 ©2012 IEEE

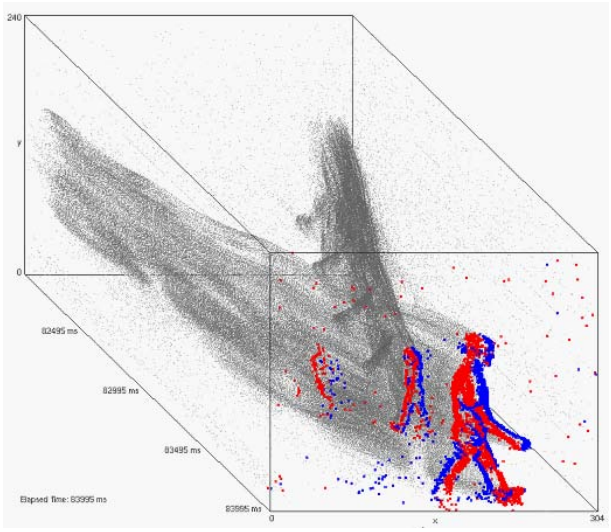

Fig. 1. Spatio-temporal representation of one dynamic vision detector capturing scene dynamics of 2 persons crossing the field of view. The events data are shown in a room-like representation with sensor mounted on the side wall.

Fig. 2. Still image of two persons from a conventional video camera (left); the corresponding events of one dynamic vision sensor (middle) and resulting event sparse depth map of two dynamic vision sensors (right) rendered in an image-like representation.

monitoring by simultaneous detection and tracking in space and time.

Figure 2 shows the the instant picture of the visual scene in figure 1 imaged by a conventional video camera (left) and its corresponding AEs using one dynamic vision detector (middle) rendered in an image-like representation. The white and black pixels represent spike activity generated by a sensed light-intensity increase (ON-event) and decrease (OFF-event) resulting from one persons motions, respectively. The gray background represents regions with no activity in the scene. The non-moving parts in the scene do not generate any data.

Using the stereo configuration, it is possible to asynchronously reconstruct the depth information of moving objects in real-time and in a very efficient way. The first working depth estimation of asynchronously generated data from an event-driven stereo sensor was provided by Schraml et al in [13], [14]. The processing unit of the dynamic stereo vision sensor embeds this event-based stereo vision algorithm, including the depth generation or the so-called sparse depth map. The resulting sparse color-coded depth map of the scene dynamics is provided at the right in Figure 2.

III. Evaluation of Dynamic Vision Sensor for Gesture Recognition

This section presents the applied method for gesture recognition using the dynamic stereo vision sensor data. The first results are also shown in this section.

A. Dynamic vision sensor in context to gesture recognition

In Figure 3, sample data of the dynamic stereo vision sensor are shown in a spatio-temporal form for a sequence of hand movements. As mentioned before the dynamic vision sensor only recognizes the moving parts of a person, so the background extraction is done by the sensor without additional cost, and the processing can be focused on the real interesting moving parts. Thus the sensor design shows a big advantage compared to traditional cameras and promises reasonable application for gesture recognition.

Gesture recognition is applied in several fields e.g. deaf sign language, navigation of virtual environments, gaming and many others. In this paper the gesture recognition will be applied to recognize certain dance figures (or activities) for use in a dance/fitness training game. A complete dance includes typically between 5 and 15 different activities. Our aim is to recognize, which activity a trainee performs at the moment. In Figure 4 eight activities are represented as time sequences. Such activities last typically few seconds. In this evaluation we try to recognize and differ between these eight activities.

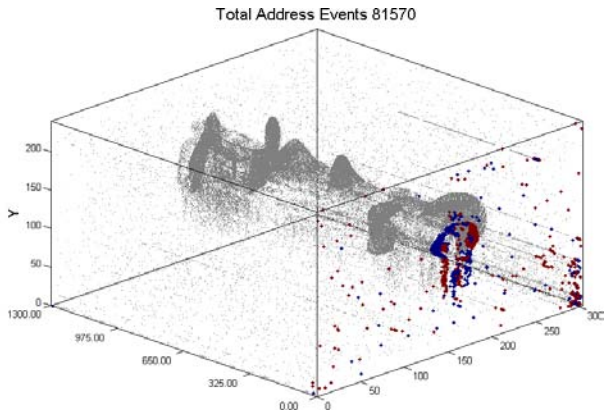

Fig. 3. Representation of the spatio-temporal generated events by one dynamic vision detector as a reaction to a person moving his arms.

B. Applied Method

By investigating the literature of gesture recognition, we noticed that the most promising method in term of the recognition rate of gestures seems to be the Hidden Markov Model (HMM). Further details on HMM can be found in the paper of Rabiner et al [1]. The best recognition rates reached with HMM reported so far are up to 92% - 93% [3], [15] for a small number of test cases (less than 100). It is frequently mentioned that Yamato et al [2] reached a recognition rate of 96%. This value is only reached, if the training data includes sequences of the test persons. Yamato et al mentioned that more realistic results will be given by using a leave one out cross validation,

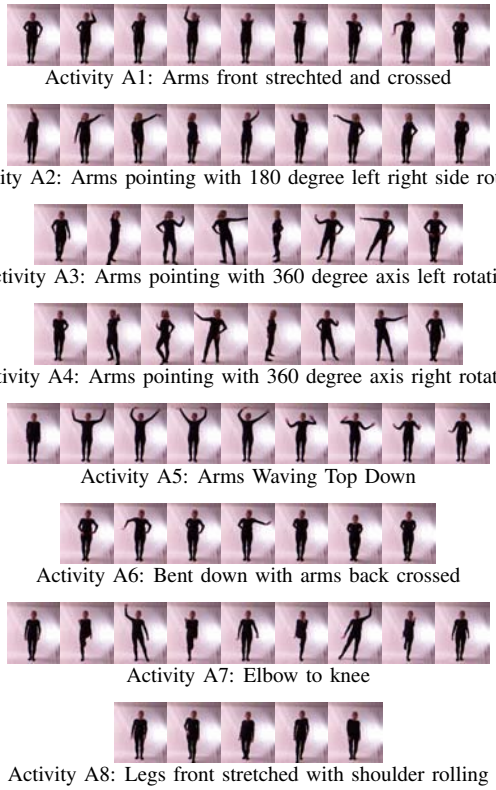

Fig. 4. Snapshots of eight different activities A1-A8 shown as a time sequence. The activities A3 and A4 are nearly identical, only the direction of rotation is different.

which means the test persons are not included in training data. In this case Yamato et al reaches a recognition rate of 71%.

In this paper a left-right continuous HMM with mixed Gaussian output probability is used for the learning phase [3]. We use a very similar feature vector to the one of Yamato et al. We simply lay a mesh with a 8x8 pixel size over the acquired data. For time slices of $40ms$ the accumulated AE count of active pixels is divided by the total number of pixels in a block. This is called the relative pixel count. The resulting vector is of one dimension and contains all calculated relative pixel counts. As the sensor is 304x240 pixels in size a total of 1140 8x8 pixel blocks has to be processed. As a second vector, the stereo data are used to calculate features referring to the depth information. For each 8x8 pixel block the maximum of depth is calculated and divided trough the maximum depth of the total sensor size. This feature type is called the relative depth. Ideally the feature vector used for training the HMMs should not be large. Therefore, we decided to compress the 1140 elements by using a discrete cosine transformation (DCT). In case of the mono or overlay data we use the first 16 coefficients of DCT for training the HMMs. In case of the stereo data we combine the first eight coefficients of the resulting DCT of the relative pixel count vector and of the relative depth vector to form a feature vector of 16 elements and use it for training the HMM.

C. Experimental Setup

For testing the algorithms several recordings of eight different activities with 15 persons has been made. In total

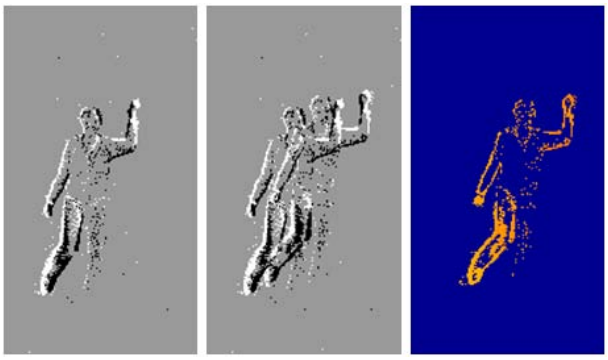

Fig. 5. Mono data refers only to one of detectors (left image), overlay data is generated by data from both detectors plotted into the same time slice (middle image) and color-coded stereo data (right image).

580 different recordings has been acquired. Every activity is performed up to 5 times. Iterations with bad performance are not used for the gesture recognition.

To simulate a training session in a TV room, the dynamic stereo vision sensor has been mounted in an elevation of $1.3m$ (typical to a sensor placed over a TV set). The distance to the performing person was about $2m$.

D. Results

To investigate the influence of the depth information, the gesture recognition algorithms are applied to three different data types, which are named as mono, overlay and stereo data (see figure 5). In the case of mono, only data from one of the two detectors of the stereo system is used for training the HMMs (left image of figure 5). In the case of the overlay data, both detectors are used. The data of both sensors are placed in the same time slice, so that the shift (disparity) of both detectors is seen (middle image figure 5). In these both cases no depth information is available, so the feature vector is based only on the relative pixel count. The stereo data will include the depth information delivered by the dynamic stereo vision sensor (right image figure 5). The used feature vector is based on a combination of the relative pixel count and the relative depth as explained before.

For applying the HMM-based gesture recognition algorithms, the Murphy Toolbox [16] on a MATLAB r2010b system was used. For all three types of data (mono, overlay and stereo) "leave one out" cross validations have been calculated with varying combinations of the Gaussian mixture numbers and states. "Leave one out" means to train the HMMs without samples of person N and then do the testing with the samples of person N. This is done for all persons, which results in 15 evaluation matrices (for each person one matrix). The sum of all evaluation matrices is called the confusion matrix. A variation of the Gaussian mixtures (M) has been made from 4 to 8, the states (Q) are iterated from 8 to 14. The best resulting confusion matrices are shown in table I. Gesture recognition based on a single detector yields an average recognition rate of 91%, the overlay data reaches the highest recognition rate of 97%, whereas the stereo datas recognition rates amounts to 96%. It can be noticed that typical wrong recognitions results

	A1	A2	A3	A4	A5	A6	A7	A8	G	E	R[%]
A1	64	8	0	0	0	0	0	5	64	13	83,1
A2	0	76	1	0	0	0	0	4	76	5	93,8
A3	0	1	65	0	0	0	0	6	65	7	90,3
A4	0	0	0	75	0	0	1	4	75	5	93,8
A5	0	0	0	0	27	0	1	1	27	2	93,1
A6	1	1	0	0	0	85	0	3	85	5	94,4
A7	0	0	0	0	0	0	82	1	82	1	98,8
A8	0	1	6	0	7	0	1	53	53	15	77,9
Sum									527	53	**90,7**

	A1	A2	A3	A4	A5	A6	A7	A8	G	E	R[%]
A1	77	0	0	0	0	0	0	0	77	0	100
A2	0	80	0	0	0	0	0	1	80	1	98,8
A3	0	2	66	0	0	0	0	4	66	6	91,7
A4	0	0	0	78	0	0	0	2	78	2	97,5
A5	0	0	0	0	29	0	0	0	29	0	100
A6	0	1	1	0	0	86	2	0	86	4	95,6
A7	0	0	0	0	1	0	79	3	79	4	95,2
A8	0	0	0	0	0	0	0	68	68	0	100
Sum									563	17	**97,4**

	A1	A2	A3	A4	A5	A6	A7	A8	G	E	R[%]
A1	77	0	0	0	0	0	0	0	77	0	100
A2	0	76	0	0	0	0	1	4	76	5	93,8
A3	0	2	66	0	0	0	0	4	66	6	91,7
A4	0	1	0	73	0	0	2	4	73	7	91,3
A5	0	1	0	0	28	0	0	0	28	1	96,6
A6	0	1	0	0	0	86	0	3	86	4	95,6
A7	0	0	0	0	1	0	80	2	80	3	96,4
A8	0	0	0	0	0	0	0	68	68	0	100
Sum									554	26	**95,7**

TABLE I

CONFUSION MATRIX FOR THE MONO (TOP, $Q = 12$, $M = 8$), OVERLAY (MIDDLE, $Q = 10$, $M = 6$) AND STEREO (BOTTOM, $Q = 10$, $M = 6$) DATA. INDICES A1 - A8 - EIGHT ACTIVITIES, G - GOOD RECOGNITION, E - ERROR RECOGNITION, R - RECOGNITION RATE IN PERCENT.

in the activity A8.

E. Discussion

The HMM based gesture recognition algorithms were successfully applied to the data of the dynamic stereo vision sensor. It is a clear increase of recognition rate observed by using overlay and stereo data. By the use of the dynamic stereo vision sensor the up to now highest reported recognition rate of 96% is reached. Mendoza et al [3] employed a single video based system, whereas Wang et al [15] used a video based stereo system.

It should be mentioned, that these experiments was a first try, with very simple feature vectors. There is a high potential, that using a more sophisticated feature vector even a higher and stable recognition rate can be reached. E.g. it is very striking, that several activities are rated by the HMM as activity A8. This needs to be clarified in further investigations. Also it irritates, that the recognition rate of stereo data is slightly less than using the overlay data. This could be due to the "lost" AE rate by the matching stereoscopic calculation.

IV. CONCLUSIONS

In this work, it was shown that the combination of Hidden Markov Model and dynamic stereo vision sensor data in a 4D space has a large potential for efficient and robust gesture recognition. Recognition rates of up to 97% have been reached. Dynamic stereo vision sensors provides a continuous and asynchronous stream of data upon scene dynamics, thus offering an on-chip background subtraction and generating motion data in a 4D representation (x,y,t,z). The evaluation of the gesture recognition on the different data types (mono, overlay and stereo) demonstrate that using the stereo system increases the recognition rate from 91% up to 97%. In the next step, larger test data with more than eighty persons performing the activities will be acquired to estimate the minimum amount of persons needed for learning different activities.

ACKNOWLEDGMENT

This work is supported by the project grant SilverGame "aal-2009-2-113" running under the Ambient Assisted Living - European Commission joint program. The authors would like to thank Reha-Zentrum Lübben for providing the dance choreography of the eight activities.

REFERENCES

[1] L. R. Rabiner and B. H. Juang, "An introduction to hidden markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 4–16, 1986.

[2] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov model," *Proc. CVPR 92 IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, pp. 379–385, 1992.

[3] M. A. Mendoza and N. P. de la Blance, "Hmm-based action recognition using contour histograms," *Lecture Notes in Computer Sciences*, pp. 394–401, 2007.

[4] H. Meng, N. Pears, and C. Bailey, "A human action recognition system for embedded computer vision application," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–6, 2007.

[5] K. Murakami and H. Taguchi, "Gesture recognition using recurrent neural networks," *Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology, CHI91*, pp. 237–242, 1991.

[6] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: a loval svm approach," *ICPR Proceedings of the 17th International Conference on Pattern Recognition*, vol. 3, pp. 32–36, 2004.

[7] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Transactions on Systems, Man and Cybernetics, Part C: applications and Reviews*, pp. 311–324, 2007.

[8] P. Lichtsteiner, C. Posch, and T. Delbrück, "A 128x128 120db 15μs latency asynchronous temporal contrast vision sensor," *IEEE Journal of Solid State Circuits*, vol. 43, pp. 566 – 576, 2008.

[9] C. Posch, D. Matolin, and R. Wohlgenannt, "A qvga 143db dynamic range frame-free pwm image sensor with lossless pixel-level video compression and time-domain cds," *IEEE Journal of Solid State Circuits*, vol. 46, pp. 259 – 275, 2011.

[10] F. Stefanini and E. Neftci, "A gestures recognition system using neuromorphic chips," http://capocaccia.ethz.ch/capo/wiki/2010/gestures10, May 2010.

[11] T. Delbruck, R. Benosman, and S. Ieng, "Dvs machine vision," http://capocaccia.ethz.ch/capo/wiki/2010/dvsvision10, May 2010.

[12] T. Delbruck, S.-C. Liu, C. Posch, R. Benosman, J. Conradt, and M. Cook, "Event based sensory processing," http://capocaccia.ethz.ch/capo/wiki/2011/ebp11, May 2011.

[13] A. Belbachir, Ed., *Smart Cameras*. Springer New York, 2009.

[14] S. Schraml, A. Belbachir, N. Milosevic, and P. Schoen, "Dynamic stereo vision for real-time tracking," *Proc. of IEEE ISCAS*, June 2010.

[15] Y. Wang, T. Yu, L. Shi, and Z. Li, "Using human body gestures as inputs for gaming via depth analysis," *IEEE International Conference on Multimedia and Expo*, pp. 993–996, 2008.

[16] K. Murphy, "Hidden markov model (hmm) toolbox for matlab," link http://www.mathworks.com/help/toolbox/stats/f8368.html, June 2005, accessed 11 May 2011.