A Real-time Pedestrian Classification Method for Event-based Dynamic Stereo Vision

S. Schraml, A.N. Belbachir, *Member IEEE*
Neuroinformatics, Safety & Security Department,
AIT Austrian Institute of Technology
Donau-City Strasse 1/5, A-1220, Vienna Austria.
{stephan.schraml ; nabil.belbachir}@ait.ac.at

N. Brändle
Dynamic Transportation Systems, Mobility Department,
AIT Austrian Institute of Technology
Giefinggasse 2, A-1210, Vienna Austria.
norbert.braendle@ait.ac.at

Abstract

This paper proposes a real-time implementation of a clustering and classification method using asynchronous events generated upon scene activities by an event-based dynamic stereo vision system. The inherent detection of moving objects offered by the dynamic stereo vision system comprising a pair of dynamic vision sensors allows event-based stereo vision in real-time and a 3D representation of moving objects. The clustering and classification method exploit the sparse spatio-temporal representation of sensor's events for real-time detection and separation between moving objects. The method makes use of density and distance metrics for clustering asynchronous events generated by scene dynamics (changes in the scene). It has been evaluated on clustering the events of moving persons across the sensor field of view. The method has been implemented on the Blackfin BF537 from analog device and tested on real scenarios with more than 100 persons. The results show that the resulting asynchronous events can be successfully clustered in real-time and that the classification rate of pedestrians is successful in more than 92% of the cases.

1. Introduction

Event-based stereo vision [2] aims to duplicate the human vision system in reacting to scene dynamics by generating events including the depth information, using a pair of vision sensor. An event-based 2D Dynamic Vision Sensor (DVS) was introduced in [9] including a set of autonomous self-spiking pixels reacting to relative light intensity changes. Its advantages include high temporal resolution, extremely wide dynamic range and complete redundancy suppression due to included on-chip preprocessing. It exploits very efficient asynchronous, event-driven information encoding, Address-Event Representation (AER) [4], for capturing scene dynamics (e.g. moving objects).

A preliminary result on the realization of event-based 3D vision in a stereo sensor, with a pair of DVSs and a stereo matching algorithm for calculating depth information, is reported in [2]. Such a system exploits the on-chip pre-processing offered by the DVS for efficient and real-time 3D vision and object classification with regards to two aspects: Firstly, the data volume is reduced as compared to conventional image frame-based stereo systems due to the efficient representation of scene dynamics using on-chip pre-processing of the visual information. Indeed, real-time stereo vision is computationally demanding, implying the allocation of large and costly processing and memory resources. The dynamic vision sensors inherently support on-chip edge detection with a low data volume by means of massively parallel focal plane processing, to allow real-time 3D representation. Secondly, the sensor sensitivity to the relative light-intensity changes allows robustness against illumination conditions. Furthermore, since it is not necessary to integrate light as in frame-based sensors, the sensor is also highly sensitive to scene dynamics in weak illuminations with high temporal response.

Spatio-temporal data processing has been introduced by Fahle [5] and Adelson [1] in the early 80's. However, methodologies for representing low-level spatio-temporal cues and high-level models suitable to explain spatio-temporal evidence are still scarce. The main reasons why joint spatio-temporal processing has not been addressed in detail originates from different factors: (i) digital computers operate using "atomistic" principles, where operations are broken down into sequence of steps and processing is performed independently for each step on discrete data; (ii) common vision sensors provide temporal data sequences in form of distinct images (frames) and (iii) the computational burden imposed by the large amount of data in the space-time volume has been a limitation for efficient operation.

The space-time processing approach is an appropriate strategy for the robust analysis of visual data encompassing dynamic processes such as motion, variable shape, and appearance, whereas traditional frame-based approaches require additional modeling tools (e.g. Markov chains) for dynamical processes. In the development of methodologies for the space-time domain over the last two decades, the research focus has mostly remained on the development of low-level cues, which have incrementally

become more descriptive (e. g. transition from simple motion cues to space-time shape).
Those efforts have been invested for automated extraction of relevant information (in space and in time) from image sequences using frame-based image sensors. Mainly due to the temporally (rate) and spatially (frame) discrete nature of digital image sequences provided by these standard sensing devices, a constant data volume is continuously produced. Such frame-based sensors are not well suited for space-time processing as (i) the data contain substantial temporally redundant information within each frame, and (ii) temporally discrete with coarse resolution (typically 25 frames per second), and (iii) increasing the temporal resolution (thus the amount of visual data) leads to prohibitive computational complexity.

A spatio-temporal clustering method for asynchronously generated events has been presented in [11]. This paper presents the real-time evaluation of the clustering method in [11] for the DVS' events, represented in a spatio-temporal domain and its implementation on the Blackfin BF 537 from Analog Device for real-time object classification in real surveillance scenarios towards a compact remote stand-alone 3D vision system. The system consists of a stereo sensor and a processing unit including event-based clustering and classification algorithms. The paper is structured as follows: Section 2 provides a brief review of the architecture of the event-based 3D vision system including core algorithms. The clustering and classification method using the sensor data is presented in Section 3. Section 4 describes evaluation results on real-world recordings of surveillance scenarios. A summary is provided in Section 5 to conclude the paper.

2. Dynamic Stereo Vision Sensor

The architecture of the dynamic stereo vision system [12] is depicted in Figure 1 including two DVSs as sensing elements [9], a buffer unit consisting of a multiplexer (MUX) and First-In First-Out (FIFO) memory, and a Blackfin BF537 digital signal processor (DSP) from Analog Device as processing unit.

The DVS consists of an array of 128x128 pixels, built in a standard 0.35µm 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 15 bit 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 the DSP over a FIFO. The DSP attaches to each AE a timestamp at a resolution of 1ms. The combined data (AEs and timestamps) are used as input stream for 3D map generation and subsequent processing.

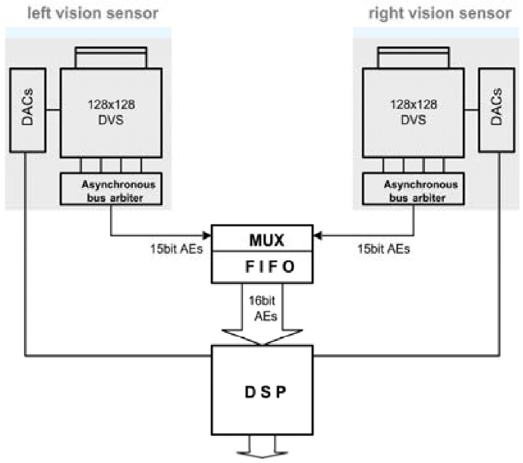

Figure 1: Hardware architecture and signal flow of the Stereo DVS

Figure 2 shows the realized stereo vision system where the vision chips are on the front face (left image) and the BF537 core module is on the back face (right image). Other modules for the sensor control and data interface can also be seen on both sides.

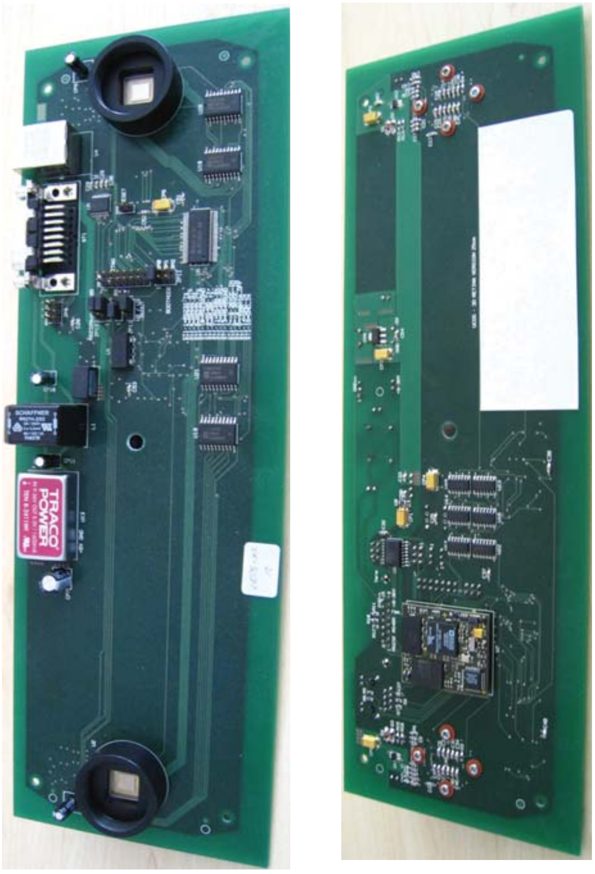

Figure 2: Image of the realized event-based stereo vision system: front face (left) and back face (right)

Figure 3 depicts a space-time representation of one DVS' data, resulting from a two persons crossing the sensor field of view. The events are represented in a 3 D volume with the coordinates x (1:128), y (1:128) and t (last elapsed ms), the so-called space-time representation. The bold colored dots represents the events generated in the recent 10 ms. 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 1.9 seconds prior to the recent 10ms. The four dashed lines are added in the figure to highlight the event path in the past 1.9 sec for the moving persons, which is an ideal basis for clustering and tracking in space and time.

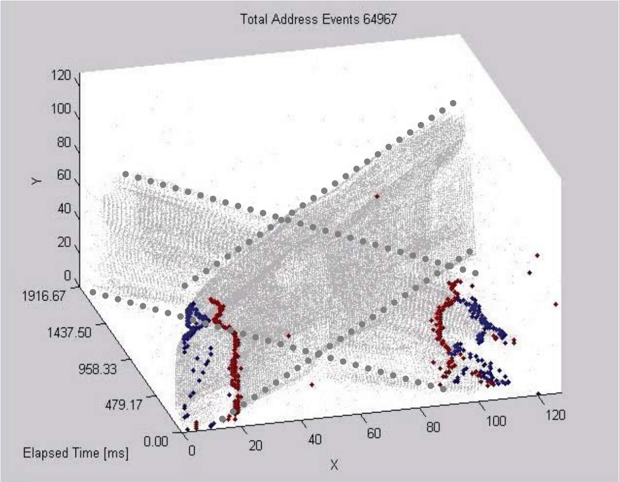

Figure 3: Event representation of scene dynamics (2 persons crossing the field of view) in a space-time domain using 1 DVS

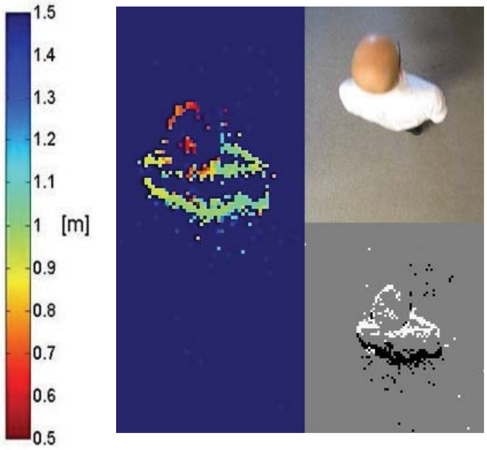

Figure 4: Still image of a person from a conventional video camera (top right); the corresponding AE from one dynamic vision sensors (bottom right); resulting "sparse" event depth map color coded (left)

Figure 4 shows an example of a visual scene imaged by a conventional video camera and its corresponding AEs using one DVS 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 two cyclists 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. The processing unit (DSP) embeds event-based stereo vision algorithms, including the depth generation or the so-called *sparse depth map*. The resulting sparse color-coded depth map of the scene is depicted in Figure 4 (left). The algorithm for real-time depth estimation has been described in [2][12] in detail.

3. Real-time Clustering and Classification

The 3D DVS continuously and asynchronously generates events as reaction to moving objects crossing the sensor field of view. The objective of this clustering method is to group together (asynchronous) events belonging to the same moving object e.g. the same person. The objective of classification is to recognize the clustered objects' events and separate them into pedestrians and cyclists. Figure 5 provides an overview of the processing steps which are described in the following.

3.1. Real-time Clustering

The used clustering method is described in [11] in detail. It combines density-based [9] and distance based clustering for robustness. Similarity between AEs is given by a distance function f(Cluster,AE) calculating the distance of the AE to the cluster center and expressed in the assignment of the AEs to the same cluster. The distance used is the sum of Manhattan distance in space-time (x,y,t) between the pixel coordinates of the AE and the cluster center is used. The cluster center is defined as the moving average of (x,y) coordinates of the assigned AE's. The clustering input data is a stream consisting of the temporal sequence of AEs having (x,y) coordinates, their polarity p (OFF or ON), the timestamp t and the reconstructed depth z. The data stream is neither stored for interactive processing nor grouped in frames. For each AE, a cluster assignment will be evaluated; afterwards, the AE will be discarded.

Assignment: AEs, with local density above a dedicated threshold, imply the calculation of their radial distance to every cluster. The strength (influence) of the AE on each cluster is evaluated. The AE is assigned to the cluster most influencing. The evaluation function depends on the AE distance, the radial dilation of the cluster, and the weight of the cluster. This latter is calculated from the sum of all assigned AES (number of AEs in a cluster). AEs with

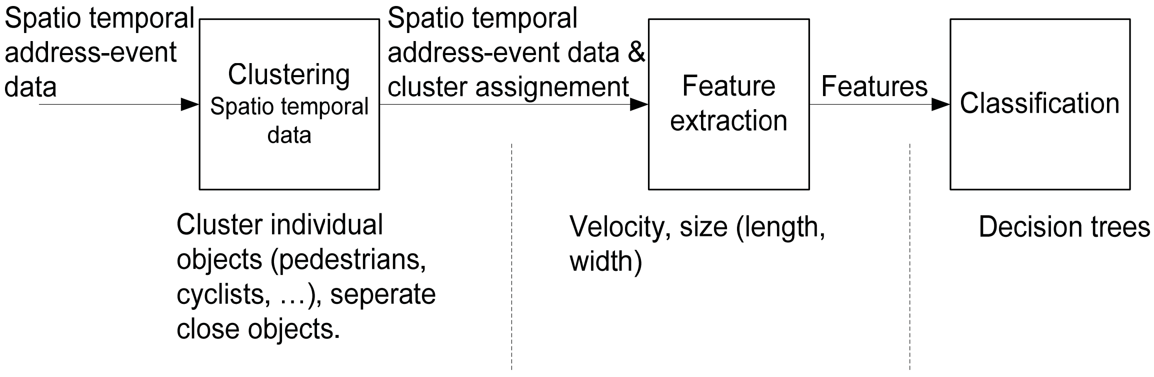
Spatio temporal address-event data
Clustering
Spatio temporal data
Spatio temporal address-event data & cluster assignement
Feature extraction
Features
Classification
Cluster individual objects (pedestrians, cyclists, …), seperate close objects.
Velocity, size (length, width)
Decision trees

Figure 5: Classification steps

density lower than a threshold are considered as outliers and are suppressed. An AE is not assigned to a cluster, when the radial distance to cluster is greater than the maximum object dilation. A new cluster can be created when AEs do not fit exiting clusters and the local density of these AEs exceeds a dedicated threshold.

Cluster properties: the properties of cluster comprise the radial dilation, the cluster weight, the coordinates-related dilation, cluster center, cluster size and passage duration. These properties are updated for each assigned AE.
The individual steps of the clustering method (cluster creation and AE assignment) are further detailed as follows:

- ***Cluster:*** the cluster represents a bulk of frequent AEs, which have density-based interrelationship around a center. If the density has its maximum at the object center then the object reaches a high stability.

- ***Interrelationship between AEs***: it depends on the cluster strength within the AEs locations. The interrelationship is not explicitly calculated but derived from the assignment to a common cluster.

- ***Creation of new clusters:*** a cluster is created when the generated AE cannot be assigned to an existing cluster because it lies outside the maximal size of all existing clusters and when a local density resulting from generated events exceeds a threshold.

- ***Termination of a cluster:*** A cluster can be removed whenever it is not timely actual and no new AEs are assigned to it for a dedicated time period.

- ***Temporal continuity:*** the continuity of a cluster is ensured when continuously actual AEs are assigned to it that is the case of moving objects.

- ***Stability:*** a cluster is stable whenever the maximum density of the assigned AEs lies at the center of the object. This is the case of AEs generated from moving pedestrians and cyclists, but not valid in case of umbrellas (in our application case).

- ***Parameters:*** there are four parameters used for the clustering. Two thresholds for the clustering creation and noise suppression and two parameters for the dilation in x and y axes. The dilation parameters define the size of the cluster outside its center. These latter have to be chosen with respect to the expected size of the observed objects (like pedestrian and cyclists in our case) and should not be greater than twice the size of the smallest object. The cluster creation threshold has to be chosen to allow clustering of objects with a low AEs density.

- ***Parameter sensitivity:*** The clustering algorithm is not sensitive to the two parameters related to the noise suppression and the cluster creation. However, the dilation parameters have to be adequately chosen with regard to the object size, which is also depend from the sensor mounting position and the distance between the sensor and the objects. Objects with small density may not trigger creation of clusters.

The AE clusters are computed in a single pass, meaning that AE are clustered in one step such that individual AE are directly assigned to a cluster. There is no reassignment or rearranging of AE or clusters. This clustering approach runs in real time and the complexity is proportional to the number of events “O(n)”, such that each event is processed only once. Furthermore, this method ensures fast calculation and assignment of events to clusters to be suitable for large data sets and for embedded systems.

3.2. Feature Extraction & Classification

After having built clusters from events through moving objects, descriptive cluster features are used to separate between pedestrians and cyclists with the help of a decision tree. A set of values for pedestrian, cyclists has been defined as a basis for classification.

We have investigated parameters like object dimensions (length, width, and height), temporal information (velocity, passage duration) and density (number of events per object). In a first attempt, we use three features (length, width and passage duration) for the classification as illustrated in Figure 6. The object height and density did not improve the classification robustness. Further investigations are still needed.

For the decision tree, thresholds on length, width and passage duration are set in order to distinguish between the multiple objects, leading to the classification results shown in the next section.

4. Experimental Analysis

In order to evaluate the embedded computer vision and scene interpretation performance using the event-based stereo vision system, a quantitative estimation of the generated AEs from the system for different objects size has been made as well as their related processing effort on the Blackfin BF-537 processor from Analog devices with 32 MB RAM and 600 MHz.

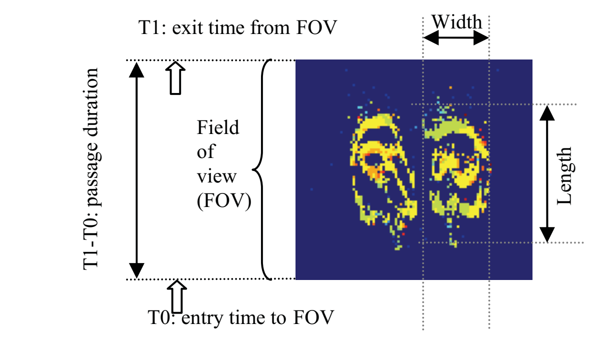

Figure 6: Graphical illustration of the features used for classification

We have collected real-world data for the evaluation of the event-based 3D system and the classification method. Test scenarios have been collected with a total of 128 passages (82 riding cyclists; 26 pedestrians, 13 walking cyclists and 7 pedestrians with umbrellas). Figure 7 shows selected test cases from an overhead mounting of the dynamic stereo vision sensor. The sensor monitor a road with two lanes one for pedestrian and the other for cyclists.

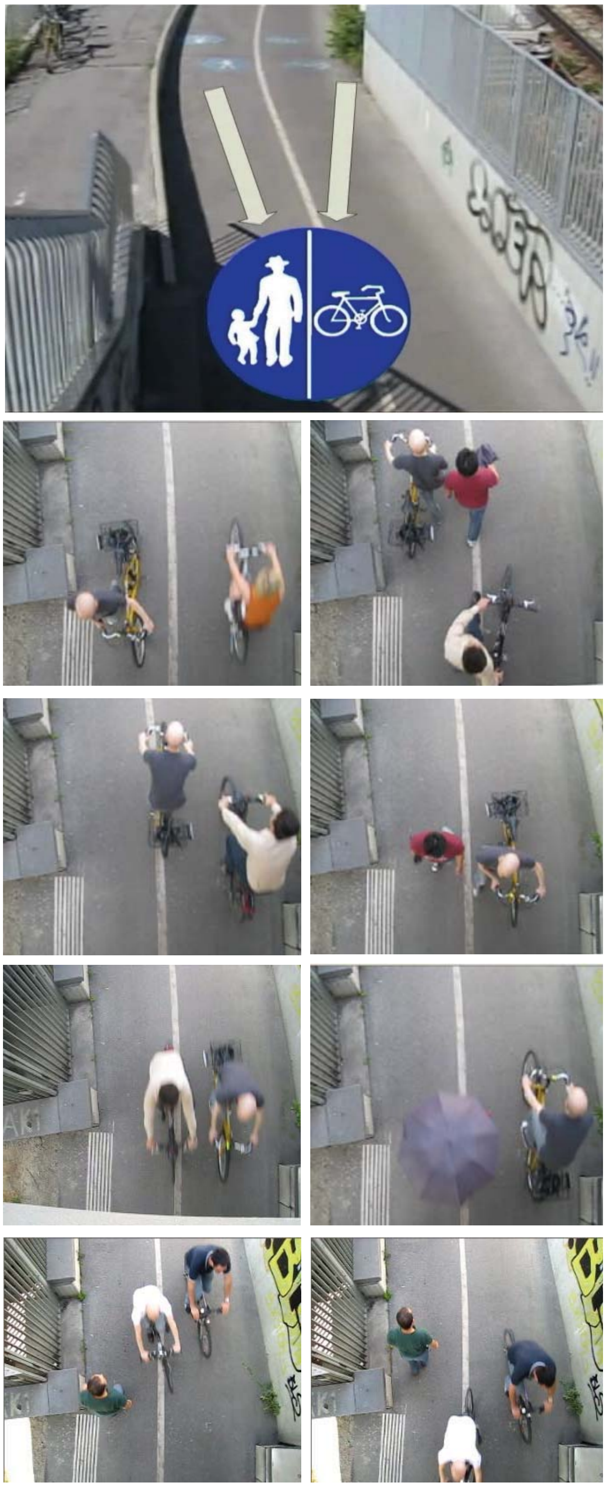

Figure 7: Overview of the test area (top) and selected test scenarios in pedestrian surveillance

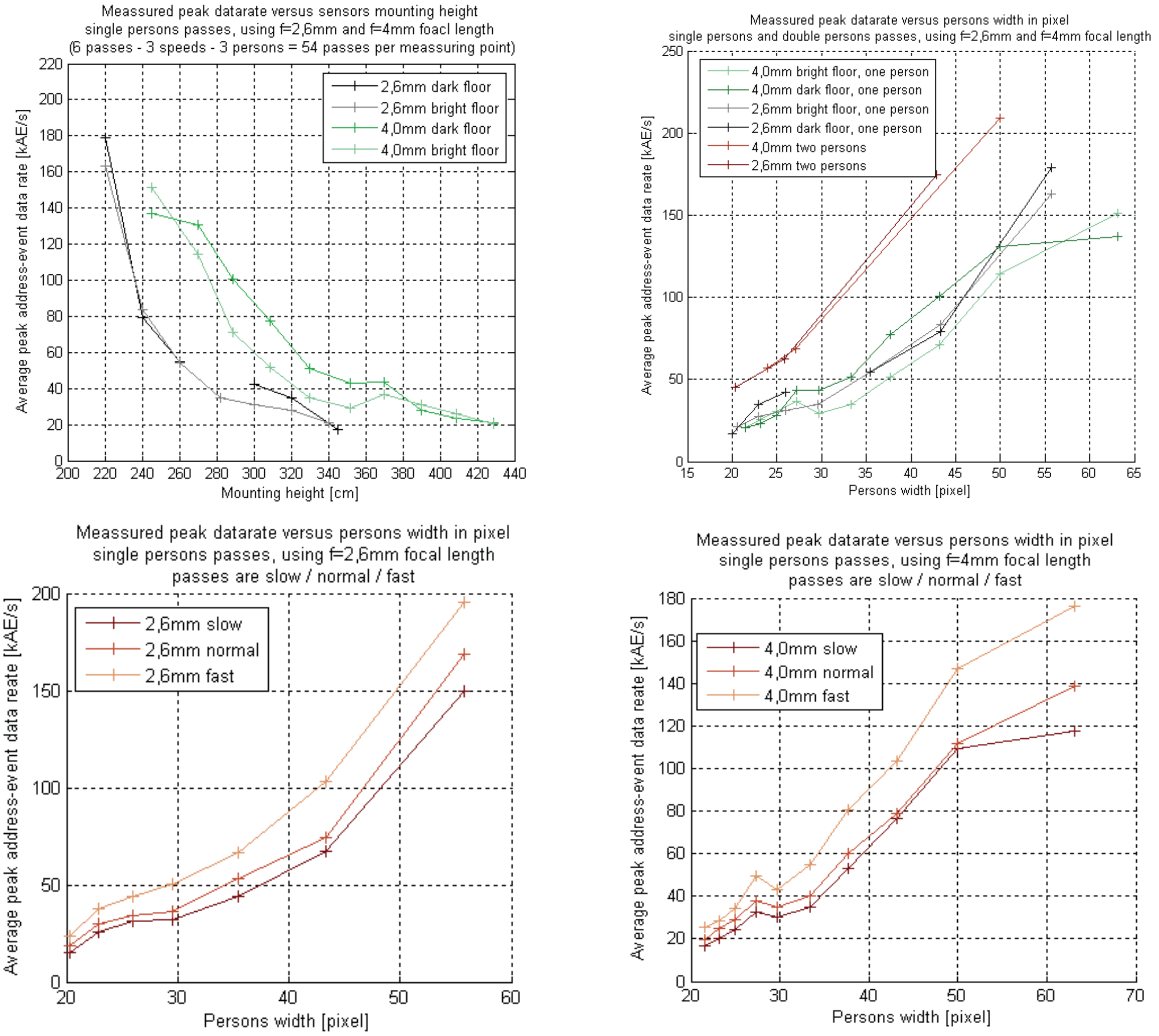

Figure 8: Illustration of resulting data rates for persons crossing the sensor field of view and its dependability on mounting positions & person size

Figure 8 shows an analysis of peak data rates received by the embedded system and its dependability of mounting positions and persons width. For a lower mounting position, the data rate is high due to the increasing size of the person on the sensor field of view. Tests showed that the Blackfin BF537 can process up to 350 kAE/s.

Figure 9 shows, generated events from two cyclists (also in Figure 6 crossing the sensor field of view. The top image shows AEs represented according to their x-coordinate in function of time and the representation in function to the y-coordinate is given in the bottom image. The z-coordinate computed from the stereo vision is not used in this classification. It was mainly used to remove outliers and cast shadow of the object. From Figure 9, it can be noticed that both object are separated and tracked along their passage duration.

Figure 10 shows classification results of riding cyclists and pedestrians for multiple scenarios using three criteria (length to width ratio in the x-axis and passage duration in the y axis). The separating line represents the thresholds used in the decision tree for the classification. The two objects classes are almost linearly, separable. However, running persons can coincide with slowly riding cyclists.

Table 1 and Table 2 present classification results for 2+1 classes (pedestrian and riding cyclist) and 4+1 (pedestrian, riding cyclist, walking cyclist and pedestrian with umbrella), respectively. In these tables only the true positive classification (correctly classified) is represented as a first step. Still a full classification evaluation needs to be performed. It can be noticed that riding cyclists are best distinguishable together with pedestrian and walking cyclist while pedestrians with umbrella are not efficiently classified. One reason for the bad classification of

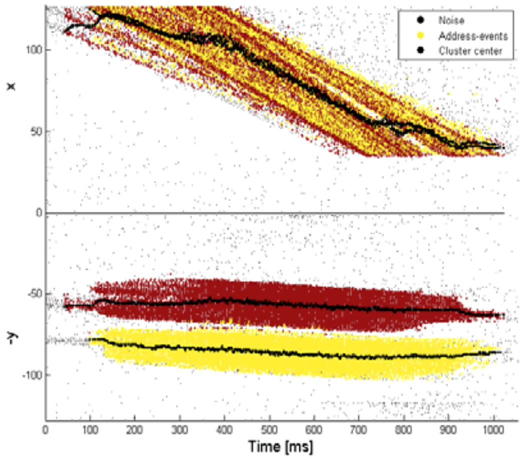

Figure 9: Illustration of tracking two riding cyclists during their passage across the sensor FOV (x-coordinate; y-coordinate; time)

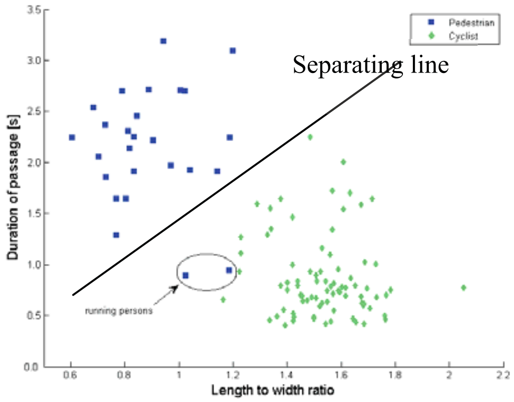

Figure 10: Classification results of riding cyclists and pedestrian using the size vs. passage duration

umbrellas might be the low density of the AEs and the difficulty to recognize them as one cluster. The other reason is probably the low number of test examples for this classification. This object (umbrella) still needs further investigation with more test data for robust analysis.

Type	Nb. cases	Correctly classified (true positive only)	Classification rate (%)
Riding cyclist	82	82	100
Pedestrian	26	24	92

Table 1: Classification results in 2+1 classes

Type	Nb. cases	Correctly classified (true positive only)	Classification rate (%)
Riding cyclist	82	79	96
Pedestrian	26	24	92
Walking cyclist	13	12	92
umbrella	7	3	43

Table 2: Classification results in 4+1 classes

5. Conclusions and Outlook

This paper presents a stand-alone and compact event-based 3D vision system including a spatio-temporal clustering method for real-time classification of pedestrians and cyclists. The preliminary results on real-scenarios, with the algorithm implemented in the BF537 Blackfin processor, have shown that the system can distinguish in real-time between riding cyclists, pedestrians, and walking cyclists in more than 92% of the cases using three criteria: length, width and time. Further investigations in clustering and criteria selection are needed to distinguish pedestrians with an umbrella. This evaluation is still preliminary as it was performed with a data set of 128 test cases. A validation on a larger scenarios set will be performed.

6. References

[1] E.H. Adelson and J.R. Bergen, “Spatiotemporal Energy Models for the Perception of Motion,” Journal of the Optical Society of America A,2, pp. 284-299, 1984.

[2] A.N. Belbachir, “Smart Cameras”, Springer, 2009.

[3] V. Chan, C. Jin and A. van Schaik, “An Address-Event Vision Sensor for Multiple Transient Object Detection,” in IEEE Transactions on Biomedical Circuits and Systems, vol. 1, issue 4, pp. 278 – 288, Dec. 2007.

[4] E. Culurciello, R. Etienne-Cummings, K. Boahen, “Arbitrated address event representation digital image sensor,” IEEE Elect. Letters, vol. 37, pp. 1443–1445, 2001.

[5] M. Fahle and T. Poggio, “Visual Hyperacuity: Spatio-temporal Interpolation in Human Vision,” Proceedings of the Royal Society of London B, 213, pp.451-477, 1981

[6] A. Fusiello, E. Trucco and A. Verri, “Rectification with Unconstrained Stereo Geometry”, in Proc. of the British Machine Vision Conf., pp. 400-409, BMVA Press, 1997

[7] R.Greene-Roesel, M.C. Diógenes, D.R Ragland and L.A. Lindau, “Effectiveness of a Commercially Available Automated Pedestrian Counting Device in Urban Environments: Comparison with Manual Counts”, Transport Research Board Annual 2008 Meeting, 2008

[8] G. Grubb, A. Zelinsky, L. Nilsson and M. Rilbe, “3D Vision Sensing for Improved Pedestrian Safety,” in Proceeding of the IEEE IVS, pp. 19-24, 2004

[9] P. Lichtsteiner, C. Posch and T. Delbrück, “A 128×128 120dB 15us Latency Asynchronous Temporal Contrast Vision Sensor,” IEEE JSSC, vol. 43, pp. 566 - 576, 2008.

[10] J. Sander et al., “ Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications,” Journal of Data Mining & Knowledge Discovery, Springer, vol. 2, pp. 169-194, 1998

[11] S. Schraml, A.N. Belbachir, “A Spatio-temporal Clustering Method Using Real-time Motion Analysis on Event-based 3D Vision,” in Proc. of the CVPR2010 Workshop on Three Dimensional Information Extraction for Video Analysis and Mining, San Francisco, 2010.

[12] S. Schraml, A.N. Belbachir, N. Milosevic and P. Schoen, “Dynamic Stereo Vision for Real-time Tracking,” in Proc. of IEEE ISCAS, June 2010.