Real-time Classification of Pedestrians and Cyclists for Intelligent Counting of Non-Motorized Traffic

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

We propose a real-time method for counting pedestrians and bicyclists by classifying bulks of asynchronous events generated upon scene activities by an event-based 3D dynamic vision system. The inherent detection of moving objects offered by the 3D dynamic vision system comprising a pair of dynamic vision sensors allows event-based stereo vision in real-time and a 3D representation of moving objects. A clustering method exploits the sparse spatio-temporal representation of sensor's events for real-time detection and separation between moving objects. The method has been demonstrated for clustering the events and classification of pedestrian and cyclists moving across the sensor field of view based on their dimensions and passage duration. Tests on real scenarios with more than 100 cyclists and pedestrians yield a classification performance above 92%.

1. Introduction

Pedestrian and bicyclist counts are a key performance measure necessary to evaluate the impacts of infrastructure improvement, to develop estimates of pedestrian risk and to understand the environmental correlates of walking and cycling [7]. Statistics about pedestrian and cyclist frequencies allow municipalities to monitor urban mobility and trigger planning infrastructure design improvements.

Figure 1 shows an example of a poor design, resulting in pedestrians using a narrow ramp [6]. While a number of reliable bicycle counting technologies are already commercially available, at this time, most of the existing automated *pedestrian* counting technologies are not well-adapted to counting pedestrians in outdoor urban environment. We refer to [7] for an overview of commercially available people counting technologies. Vision-based systems have seen large progress in recent years, even for crowded scenes in oblique camera views [2] [3]. The large coverage area and the rich visual input of vision systems has the potential to distinguish between and counting multiple types of non-rigid objects, enabling detection and counting of multiple classes of traffic participants with a single counting system. If the vision system is based on processing stereo information and depth computation, harsh environmental conditions such as cast shadows or rain can be better fulfilled than with a mere 2D visual processing [8].

Figure 1: Pedestrians prefer the narrow ramp to the shallow stairs [6]

Event-based stereo vision [1] aims at duplicating 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).

The event-based 3D vision in a stereo sensor has been realized with a pair of DVSs and a stereo matching algorithm for calculating depth information and is reported in [1][12]. Such a system exploits the on-chip pre-processing offered by the DVS for efficient and real-time 3D vision and object classification with regard 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 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, thus allowing 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.

This paper presents the application of the system in surveillance scenarios for classifying non-motorized traffic (pedestrians, cyclists) in a spatio-temporal domain towards a compact remote stand-alone 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 **Fehler! Verweisquelle konnte nicht gefunden werden.** provides a brief review of the architecture of the event-based 3D vision system including the 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 pedestrians and cyclists. A summary is provided in Section 5 to conclude the paper.

2. Dynamic Stereo Vision Sensor

The existing dynamic stereo vision sensor [1] [12] is reported in this section including data examples generated by the system. The system, including the sensor board, DVS chip and DSP board, is depicted in Figure 2. It includes two DVSs as sensing elements [9], a buffer unit consisting of a multiplexer (MUX) and First-In First-Out (FIFO) memory, and a digital signal processor (DSP) as processing unit.

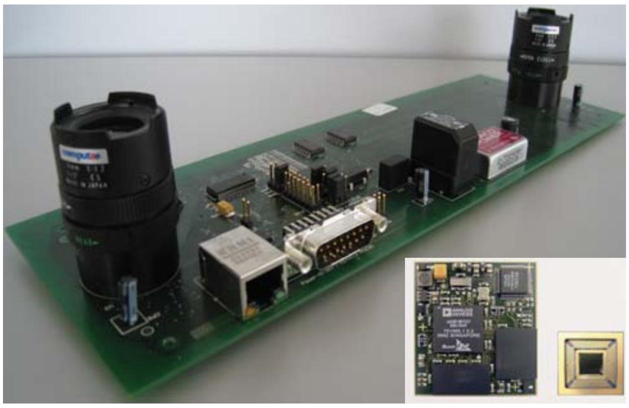

Figure 2: Photo of the stereo sensor. In the lower right corner the DSP Bf537 and the DVS sensor chip are shown. The DSP is mounted on the back of the board

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 3 depicts a space-time representation of one DVS’ data, resulting from a one person crossing the sensor field of view from far left to near right. 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 motion, respectively. The small gray dots are the events generated in the elapsed 2 seconds prior to the recent 10ms. Thes generated in the past 1.41 sec represent the path of the moving person, which is an ideal basis for clustering and tracking in space and time.

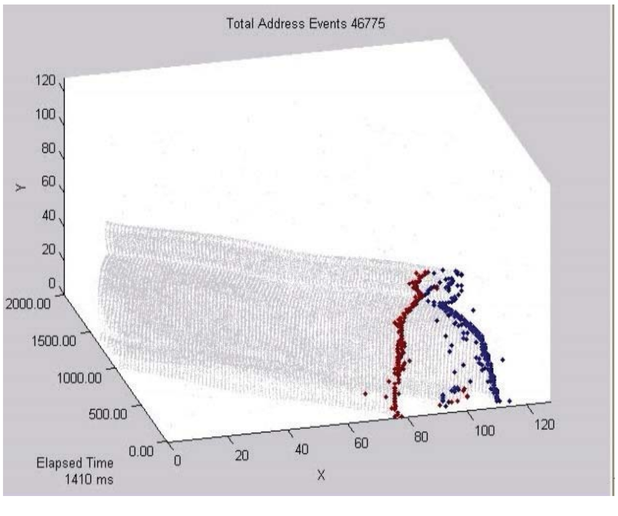

Figure 3: Event representation of scene dynamics (one person crossing the field of view from far left to near right) in a space-time domain using one DVS

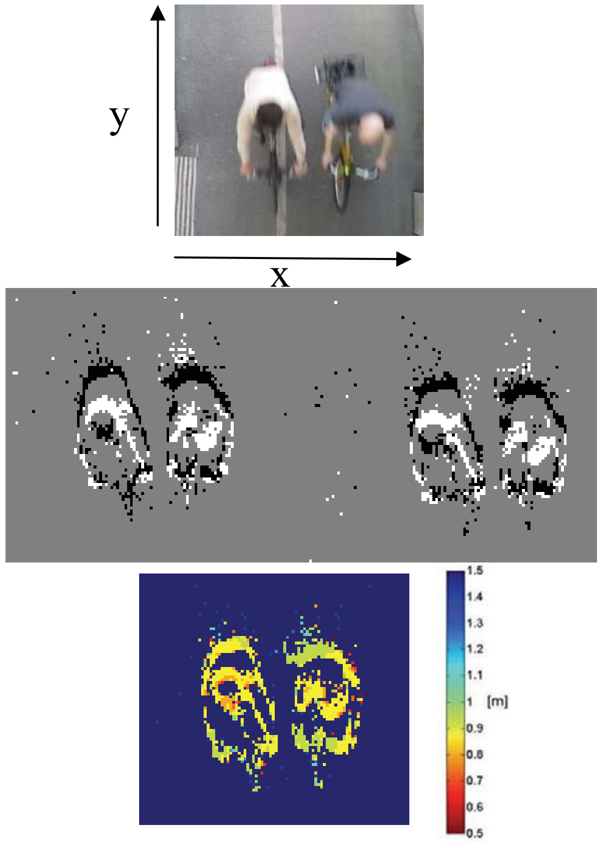

Figure 4: Still image of two cyclists from a conventional video camera (top); the corresponding AE a pair of dynamic vision sensors (middle); resulting event “sparse” depth map (bottom)

Figure 4 shows an example of a visual scene imaged by a conventional video camera (top) and its corresponding AEs using a pair of DVSs (middle) represented 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 depicted in Figure 4(top) is provided at the bottom in Figure 4. The algorithm used for real-time depth estimation is described in [1][12] in details.

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 (pedestrians, cyclists). 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 clustering method is described in detail in [11]. It combines density-based [10] and distance based clustering for robustness of the clustering. 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, where 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 asynchronously generated and having (x,y) coordinates, their polarity p (OFF or ON), the timestamp t, the reconstructed depth z. The data stream is neither stored for interactive processing nor grouped in frames. For each AE a cluster assignment is evaluated; afterwards, the AE is discarded. The polarity

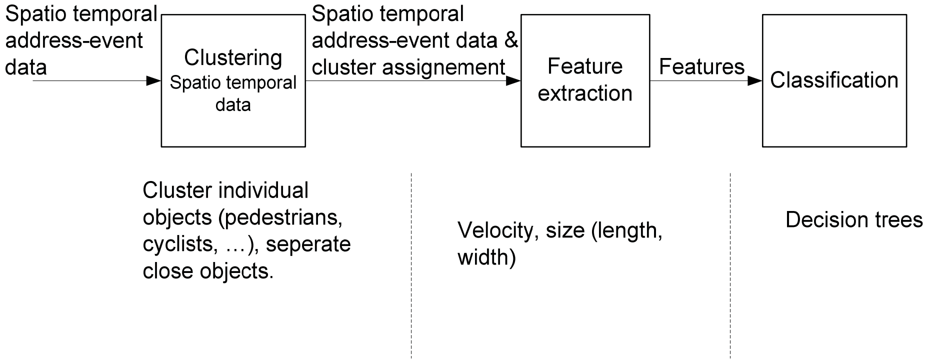
Spatio temporal address-event data
Clustering
Spatio temporal data
Spatio temporal address-event data & cluster assignement
Feature extraction
Features
Classification
Cluster individual objects (pedestrians, cyclists, …), seperate close objects.
Velocity, size (length, width)
Decision trees

Figure 5: Classification steps

information is not used for the clustering.
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. When the maximum density is at the object center, the cluster reaches high.

- ***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 a local density resulting from generated events exceeds a threshold and when these AEs cannot be assigned to an existing cluster because they lie outside the maximal size of the existing clusters.

- ***Suppression 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 online, meaning that AE are clustered in one step such that individual AE are assigned to a cluster at once. There is no reassignment or rearranging of AE or clusters. This clustering is 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. Real-time 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.

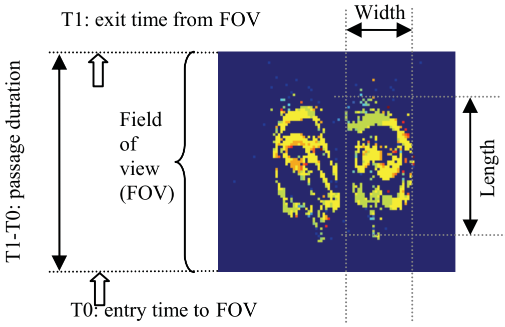

Figure 6: Graphical illustration of the features used for classification

4. Experimental Results

The event-based stereo vision sensor has been overhead mounted over a road with two lanes for pedestrians and cyclists. Images from the test environment and the sensor mounting position are shown in Figure 7. The classification performance has been therefore, evaluated for live scenarios of non-motorized traffic.

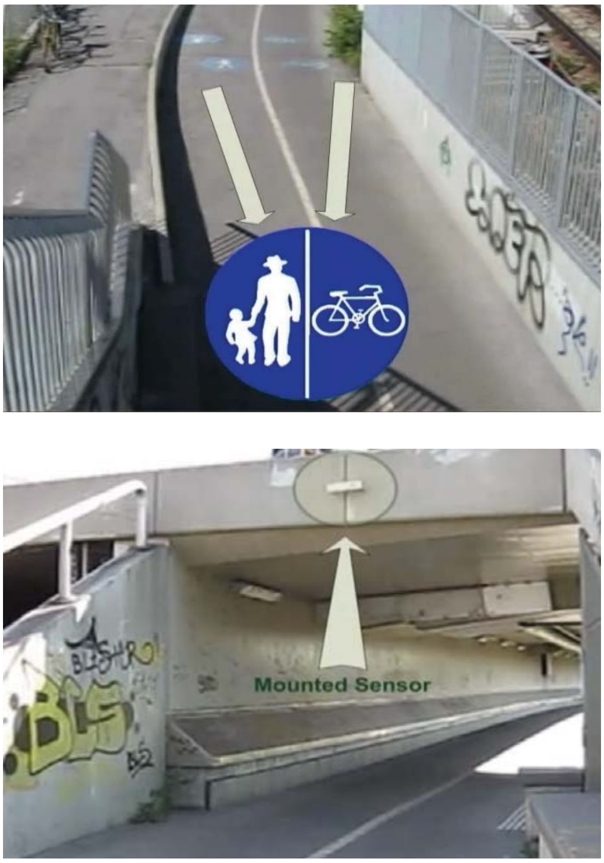

Figure 7: Picture of the test environment (top) and sensor mounting position (bottom)

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 8 shows selected test cases.

Figure 9 shows, generated events from two cyclists (also in Figure 4 (bottom left) 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 depth information (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.

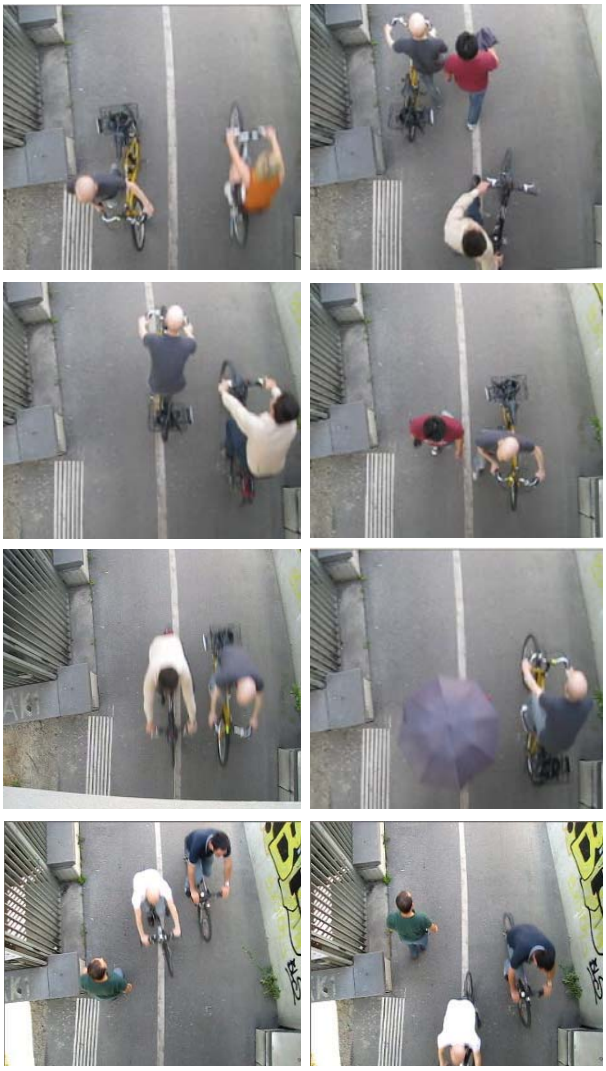

Figure 8: Selected surveillance test scenarios of non-motorized traffic

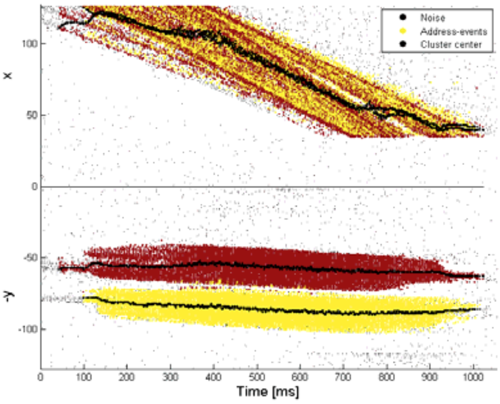

Figure 9: Illustration of tracking two riding cyclists during their passage across the sensor FOV (x-coordinate; y-coordinate; time

Figure 8 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.

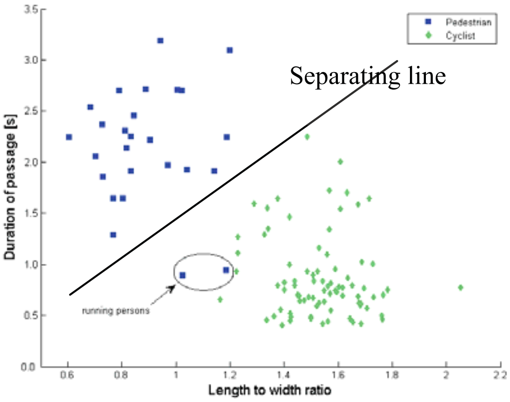

Figure 10: Classification results of riding cyclists and pedestrian using the 2D size vs. passage duration

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. A full evaluation of the classification, including false positives, still 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 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 the application of the stand-alone event-based 3D vision system for real-time surveillance in the classification of non-motorized traffic like pedestrians and cyclists. The system includes a spatio-temporal clustering method and a decision tree for the real-time classification. The preliminary results on real-scenarios 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 for 128 test cases. A validation on a larger test set will be performed. This system shows potential for non-motorized traffic surveillance applications.

6. References

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

[2] M.D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-meier and L. Van Gool, “Markovian Tracking-by-Detection from a Single, Uncalibrated Camera”, in Proc. PETS2009.

[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] 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

[6] J. Gehl. “Life between Buildings: Using Public Space”, The Danish Architectural Press,1996

[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.