IMAGE MATCHING ALGORITHMS IN STEREO VISION USING ADDRESS-EVENT-REPRESENTATION

A Theoretical Study and Evaluation of the Different Algorithms

M. Dominguez-Morales, E. Cerezuela-Escudero, A. Jimenez-Fernandez, R. Paz-Vicente
J. L. Font-Calvo, P. Iñigo-Blasco, A. Linares-Barranco and G. Jimenez-Moreno
Department of Architecture and Computer Technology, University of Seville, Seville, Spain

Keywords: Stereo vision, Epipolar restriction, Image matching, Address-event-representation, Spike, Retina, Area-based method, Features-based method.

Abstract: Image processing in digital computer systems usually considers the visual information as a sequence of frames. These frames are from cameras that capture reality for a short period of time. They are renewed and transmitted at a rate of 25-30 fps (typical real-time scenario). Digital video processing has to process each frame in order to obtain a filter result or detect a feature on the input. In stereo vision, existing algorithms use frames from two digital cameras and process them pixel by pixel until it is found a pattern match in a section of both stereo frames. Spike-based processing is a relatively new approach that implements the processing by manipulating spikes one by one at the time they are transmitted, like a human brain. The mammal nervous system is able to solve much more complex problems, such as visual recognition by manipulating neuron's spikes. The spike-based philosophy for visual information processing based on the neuro-inspired Address-Event- Representation (AER) is achieving nowadays very high performances. In this work we study the existing digital stereo matching algorithms and how do they work. After that, we propose an AER stereo matching algorithm using some of the principles shown in digital stereo methods.

1 INTRODUCTION

In recent years there have been numerous advances in the field of vision and image processing, because they can be applied for scientific and commercial purposes to numerous fields such as medicine, industry or entertainment.

As we can easily deduce, the images are two dimensional while the daily scene is three dimensional. This means that, between the passage from the scene (reality) to the image, it has lost what we call the third dimension.

Nowadays, society has experimented a great advance in these aspects: 2D vision has given way to 3D viewing. Industry and research groups have started to study this field in depth, obtaining some mechanisms for 3D representation using more than one camera. Trying to resemble the vision of human beings, researchers have experimented with two-camera-based systems inspired by human vision. Following this, it has been developed a new branch of research, focused on stereoscopic vision (S. T. Barnard, M. A. Fischler, 1982). In this branch, researchers try to obtain three-dimensional scenes using two digital cameras. Thus, we try to get some information that could not be obtained with a single camera, i.e. the distance that the objects are.

By using digital cameras, researchers have made a breakthrough in this field, going up to create systems able to achieve the above. However, digital systems have some problems that, even today, have not been solved. In any process of stereoscopic vision, image matching is the main problem that has consumed a large percentage of research resources in the field of stereoscopic vision, and it is still completely open to research. The problems related to evolving image matching are the computational cost needed to obtain appropriate results. There are lots of high-level algorithms used in digital stereo vision that solve the image matching problem, but this implies a computer intervention into the process and it is computationally expensive.

The required computational power and speed make it difficult to develop a real-time autonomous system. However, brains perform powerful and fast vision processing using millions of small and slow

Dominguez-Morales M., Cerezuela-Escudero E., Jimenez-Fernandez A., Paz-Vicente R., Font-Calvo J., Iñigo-Blasco P., Linares-Barranco A. and Jimenez-Moreno G..
IMAGE MATCHING ALGORITHMS IN STEREO VISION USING ADDRESS-EVENT-REPRESENTATION - A Theoretical Study and Evaluation of the Different Algorithms.
DOI: 10.5220/0003518500790084
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications* (SIGMAP-2011), pages 79-84
ISBN: 978-989-8425-72-0

cells working in parallel in a totally different way. Primate brains are structured in layers of neurons, where the neurons of a layer connect to a very large number (~104) of neurons in the following one (G. M. Shepherd, 1990). Most times the connectivity includes paths between non-consecutive layers, and even feedback connections are present.

Vision sensing and object recognition in brains are not processed frame by frame; they are processed in a continuous way, spike by spike, in the brain-cortex. The visual cortex is composed by a set of layers (G. M. Shepherd, 1990), starting from the retina. The processing starts when the retina captures the information. In recent years significant progress has been made in the study of the processing by the visual cortex. Many artificial systems that implement bio-inspired software models use biological-like processing that outperform more conventionally engineered machines (J. Lee, 1981; T. Crimmins, 1985; A. Linares-Barranco, 2010). However, these systems generally run at extremely low speeds because the models are implemented as software programs. For real-time solutions direct hardware implementations of these models are required. A growing number of research groups around the world are implementing these computational principles onto real-time spiking hardware through the development and exploitation of the so-called AER (Address Event Representation) technology.

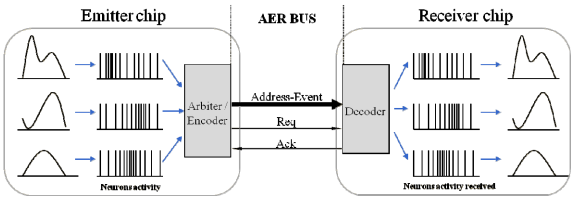

Figure 1: Rate-coded AER inter-chip communication scheme.

AER was proposed by the Mead lab in 1991 (M. Sivilotti, 1991) for communicating between neuromorphic chips with spikes. Every time a cell on a sender device generates a spike, it transmits a digital word representing a code or address for that pixel, using an external inter-chip digital bus (the AER bus, as shown in figure 1). In the receiver the spikes are directed to the pixels whose code or address was on the bus. Thus, cells with the same address in the emitter and receiver chips are virtually connected by streams of spikes. Arbitration circuits ensure that cells do not access the bus simultaneously. Usually, AER circuits are built with self-timed asynchronous logic.

Several works are already present in the literature regarding spike-based visual processing filters. Serrano et al. presented a chip-processor able to implement image convolution filters based on spikes that work at very high performance parameters (~3GOPS for 32x32 kernel size) compared to traditional digital frame-based convolution processors (B. Cope, 2006; B. Cope, 2005; A. Linares-Barranco, 2010).

There is a community of AER protocol users for bio-inspired applications in vision and audition systems, as evidenced by the success in the last years of the AER group at the Neuromorphic Engineering Workshop series. One of the goals of this community is to build large multi-chip and multi-layer hierarchically structured systems capable of performing complicated array data processing in real time. The power of these systems can be used in computer based systems under co-processing.

First, we describe digital stereo matching algorithms, their pro and cons, and their operation. Then we propose an AER theoretical algorithm, based on the digital ones, which can be developed in AER systems using a FPGA to process the information.

2 DIGITAL STEREO MATCHING ALGORITHMS

The two-camera model in stereo vision draws on the biological model itself, where the object distance can be determined thanks to the distance between the two eyes; this could be the considered as the third dimension. The eye separation produces displaced images of the same scene captured by two eyes (as shown in figure 2), i.e. the image of an eye is almost the same as other's but displaced a certain distance, what is inversely proportional to the distance between eyes and objects.

According to this, stereo vision process includes

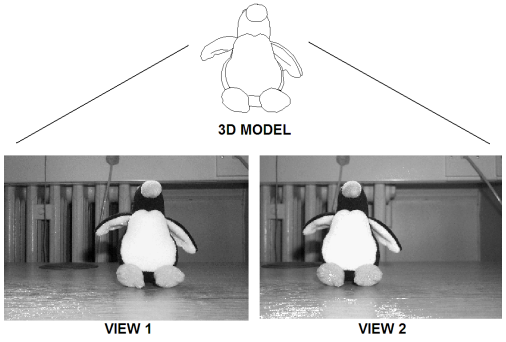

Figure 2: Stereo-vision system.

six main steps (S. T. Barnard, 1982): image acquisition, camera modeling (depending on the system geometry. D. Papadimitriou, 1996), features extraction from the scene, image matching according to these features, determining the objects distance, and interpolating, if necessary. In these steps, image matching is regarded as the most complex of all.

The projection for a three-dimensional-space point is determined for each image of the stereo pair during the image matching. The solution for the matching problem demands to impose restrictions on the geometric model of the cameras and the photometric model of the scene objects. Of course, this solution implies a high computational cost.

A common practice is trying to relate the pixel of an image with its counterpart on the other one. Some authors divide the matching methods depending on the restrictions that exploits. According to this, a high-level division could be as follows:

- Local methods: Methods that applies restrictions on a small number of pixels around the pixel under study. They are usually very efficient but sensitive to local ambiguities of the regions (i.e. regions of occlusion or regions with uniform texture). Within this group are: the area-based method, features-based method, as well as those based on gradient optimization (S.B. Pollard, 1985).
- Global methods: Methods that applies restrictions on lines of the image or the entire image itself. They are usually less sensitive to local peculiarities as add support to regions difficult to study locally. However, they tend to be computationally expensive. Within this group are the dynamic programming methods and nearest neighbour methods.

Each technique has its advantages and disadvantages and, depending on the system restrictions and the cameras geometry (G. Pajares, 2006).

Local methods will be discussed later in more detail. We won't go further into global methods because they are rarely used due to their computational cost.

2.1 Area-based Matching Algorithms

We calculate the correlation between the distribution of disparity for each pixel in an image using a window centered at this pixel, and a window of the same size centered on the pixel to be analyzed in the other image (as shown in figure 3). The problem is to find the point to be adjusted properly at first. The effectiveness of these methods depends largely on the width of the taken window. Thus, we can assume that the larger the window, the better the outcome. However, the computing power requirements these methods. We have to find a window size large enough to ensure finding a correspondence (S.B. Pollard, 1985) between two images in most of the cases, but the window width should not be overwhelmed as it would cause a huge latency in our system. Also, if the window size is close to the total size, it would be deriving to the global methods, which we were dismissed because of their computational inefficiency.

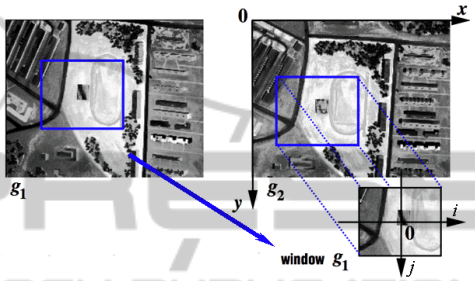

Figure 3: Window correlation.

The main advantage of these correlation mechanisms has been previously named in multiple times, and it is their computational efficiency (T. Tuytelaars, 2000). This characteristic is crucial if we want our system to perform fairly well in real time. On the other hand, the main drawbacks in digital systems primarily focus on results:

- Working directly with each pixel: we can observe a high sensitivity to distortions due to the change of point of view, as well as contrast and illumination changes.
- The presence of edges in the windows of correlation leads to false matches, since the surfaces are intermittent or in a hidden image has an edge over another.
- Are closely tied to the epipolar constraints (D. Papadimitriou, 1996).

Therefore, area-based stereo vision techniques look for cross correlation intensity patterns in the local vicinity or neighbourhood of a pixel in an image, with intensity patterns in the same neighbourhood for a pixel of another image. Thus, area-based techniques use the intensity of the pixels as an essential characteristic.

2.2 Features-based Matching Algorithms

As opposed to area-based techniques, the features-

based techniques need an image pre-processing before the image matching process. This pre-processing consists of a feature extraction stage from both images, resulting in the identification of features of each image. In turn, some attributes have to be extracted to be used in the matching process. Thus, this step is closely linked to the matching stage in those matching algorithms based on features because without this step, we would not have enough information to make inferences and obtain the image correlation.

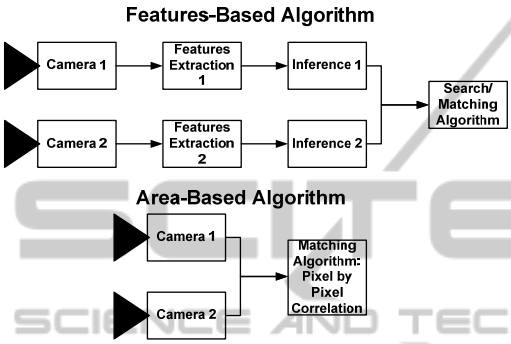

Figure 4: Area-based and features-based algorithms.

For features-based stereo vision, symbolic representations are taken from the intensity images rather unless of directly using the intensities. The most widely used features are: breakpoints isolated chains of edge points or regions defined by borders. The three above features make use of the edge points. It follows that the end points used as primitives are very important in any stereo-vision process and, consequently, it is common to extract the edge points of images. Once the relevant points of edge have been extracted (as shown in figure 5), some methods use arrays of edge points to represent straight segments, not straight segments, closed geometric structures which form geometric structures defined or unknown.

Aside from the edges, the regions are another primitive that can be used in the stereo-vision process. A region is an image area that is typically associated with a given surface in the 3D scene and is bounded by edges.

Figure 5: Edge detections in a features-based algorithm.

With the amount of features and depending on the matching method that will be used, an additional segmentation step may be necessary. In this step, additional information would be extracted from the known features. This information is calculated based on inferences from the known characteristics. Thus, the matching algorithm that receives the inferred data possess much more information than the algorithm that works directly on the pixel intensity.
Once the algorithm has both vectors with the inferred features from the two images, it searches in the vectors looking for similar features. The matching algorithm is limited to a search algorithm on two features sets. We can say that the bulk of computation corresponds to the feature extraction algorithm and the inference process.

The main advantages of these techniques are:

- Better stability in contrast and illumination changes.
- Allow comparisons between attributes or properties of the features.
- Faster than area-based methods since there are fewer points (features) to consider, although require pre-processing time.
- More accurate correspondence since the edges can be located with greater accuracy.
- Less sensitive to photometric variations as they represent geometric properties of the scene.
- Focus their interest on the scene that has most of the information.

Despite these advantages, features-based techniques have two main drawbacks, which are easily deduced from the characteristics described above. The first drawback is the high degree of dependence on the chosen primitives of these techniques. This can lead to low quality or unreliable results if the chosen primitives are not successful or are not appropriate for these types of images. For example, in a scene with few and poorly-defined edges, delimiters would not be advisable to select regions as primitive.

Another drawback is derived from the characteristics of the preprocessing stage. We described this step as a feature extraction mechanism of the two images and the inference or properties of the highest level in each of them. As stated above, there is a high computational cost associated to this pre-processing stage, to the point that using digital cameras with existing high-level algorithms running on powerful machines cannot match the real time processing.

3 REAL-TIME SPIKES-BASED MATCHING ALGORITHM

We have seen two major lines of research about image matching algorithms for digital systems.

- Area-based matching algorithms: very focused on speed, without using pre-processing and with low efficient results.
- Features-based matching algorithms: more focused on getting good results at the expense of a pre-processing and inference phases, resulting in a higher latency.

As a first approximation, we could consider making an adaptation of the features-based algorithms to obtain a consistent algorithm with good results. However, in this case we have a lot of problems mainly derived from the early stages of pre-processing and inference. In order to define an algorithm that is feasible in our system we have to take into account its features and the goals we want to achieve.

In the introduction we mentioned AER systems, motivations, current development and research lines related to them. Our main goal is to design and build an autonomous and independent system that works in real time basis, with no need to use a computer to run high-level algorithms. The efficiency of our system is not as important as real-time processing. Although we sacrifice quality in the results, we cannot afford to perform pre-processing and inference stages, which slow down our system making the real-time processing impossible. Moreover, due to the independence requirement, a computer cannot be used to run high level algorithms.

The information in an AER system is a continuous flow that cannot be stopped: the information can only be processed or discarded. Each spike is transmitted by a number of communication lines, and contains information from a single pixel. Moreover, the intensity of a pixel dimension is encoded in the spike frequency received from that particular pixel. The AER retinas used by research groups are up to 128x128 tmpdiff resolution, which means that measure brightness changes over time. Thus, taking a load of 10% in the intensity of the pixels, we would be in the range of more than four hundred thousand pulses to describe the current state of the scene with a single retina. This is too much information to be pre-processed.

We want our system to be independent and based on an FPGA connected to the outputs of two AER retinas. The FPGA processes the information using the proposed algorithm and transmits the resulting information using a parallel AER bus to an USBAERmini2 PCB (R. Berner, 2007). This is responsible for monitoring AER traffic received and transmitted by USB from and to a computer. Should be noted that we just use the computer to verify the algorithm running on the FPGA works as required, the computer itself is not used to process any information.

Taking into account the digital algorithms, the second option is to use a variant of the area-based matching algorithms (T. Tuytelaars, 2000). In this case, the topic to consider would be the results because, as discussed above, these algorithms do not require preprocessing but not ensure result reliability.

Among the problems related to the area-based matching algorithms, AER retinas include failures caused by variations in the brightness and contrast. This involved the properties of AER retinas we use, which do not show us all the visual information it covers, but the information they send is the spatial derivative in time. This means that we appreciate the information only for moving objects, while the rest of non-mobile environment is not "seen". In addition, these retinas have very peculiar characteristics related to information processing. These characteristics make them immune to variations in brightness and contrast (lightness and darkness does not interfere with transmitted information). With this retina property managed to avoid the major drawback of area-based algorithms. Our proposed algorithm is linked to the information received from the AER bus. It also inherits similar properties from area-based algorithms, but adapted to the received information. We propose an algorithm able to run in a standalone environment in a FPGA, which receives traffic from two AER retinas through a parallel bus.

We will count the received spikes for each pixel of the image and store this information in a table. So that, we have pixel intensity measures derived from the moving objects detected by each retina at every moment. The algorithm will mainly find correspondences between the two tables of spike counters. Given that, there are two problems involving the algorithm and the properties of AER retinas.

A major issue about the retinas is their unique properties (AER output frequency, firing intensity threshold of the pixels, etc) of each retina and its BIAS setting. Thus, the information of the same pixel in both retinas is not sent at the same time and there could be a frequency variation between both of them. To prevent this, we propose a fuzzy matching algorithm which will not seek exactly the same levels of intensity, but would admit an error in the

range of 5 and 8% of these levels (in our case, with 256 intensity levels, we would admit a 12-20 levels error. G.Pajares, 2006).

The first problem is solved, but we will have a very inefficient algorithm if we use pixel by pixel matching. To solve it, we use the principles of area-based matching algorithms. So we do not limit ourselves to one pixel correspondence search, but search within a window of variable width, which will conduct the correspondence of each pixel in the window. To emphasize, the correspondence between each pair of pixels from both retinas will be done separately using the fuzzy matching algorithm explained above (G.Pajares, 2006).

Another feature that improves the algorithm efficiency is what we name the "re-dog" (reset-dog). It is a process that resets the internal counters of the pixel intensity measures from time to time to avoid counters overflow.

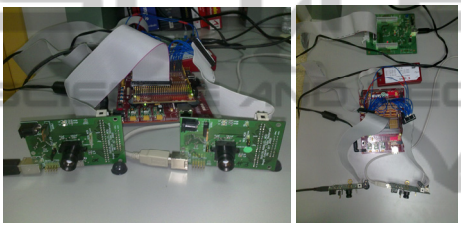

Figure 6: Virtex-5 with AER retinas and USBAERmini2.

To improve the efficiency of the search for correspondences, we can divide the image into quadrants, so our program has different threads to do correlation within each quadrant, but also has a problem related to the objects being placed between various quadrants. At present we have implemented the matching algorithm in a Virtex-5 FPGA model. We have connected to this board the two inputs from the AER retinas and it has an output to an USBAERmini2 PCB (R. Berner, 2007), which is responsible for monitoring traffic and send it to the computer using the USB bus (as shown in figure 6). We are now testing the algorithm.

This work is within the Spanish national project VULCANO, and will be used for calculating the distance to the object and, on this basis, controls a robotic arm that interacts with those objects.

4 CONCLUSIONS

In this paper we have discussed the current trends in the literature on computer vision, how it has evolved, step by step to the multi-camera view due to existing needs. We have analyzed the image matching algorithms for digital systems, their advantages and disadvantages.

After this first analysis, we have seen the problems evolving real-time systems with more than one camera, especially if the information from two cameras has to be merged. Because of this, we have mentioned AER technology as an approach to neuro-inspired computing. Thus we have named the most important details of this technology, and its adaptation to computer vision.

Finally, after identifying the problem and the means to resolve it, we have proposed an algorithm that is under test, allowing us to obtain information from two AER retinas and process information from both of them, so we can get an image matching algorithm as close as possible to real-time processing with an acceptable efficiency.

ACKNOWLEDGEMENTS

This work has been supported by Spanish government grant VULCANO (TEC2009-10639-C04-02) and the European project CARDIAC (FP7-248582).

REFERENCES

G. M. Shepherd, 1990. *The Synaptic Organization of the Brain*. Oxford University Press, 3rd Edition.

J. Lee, 1981. *A Simple Speckle Smoothing Algorithm for Synthetic Aperture Radar Images*. Man and Cybernetics, vol. SMC-13.

T. Crimmins, 1985. *Geometric Filter for Speckle Reduction*. Applied Optics, vol. 24, pp. 1438-1443.

D. Papadimitriou, T. Dennis, 1996. *Epipolar line estimation and rectification for stereo image pairs*. IEEE Trans. Image Processing, 5(4):672-676.

S. B. Pollard, et al, 1985. *PMF: a stereo correspondence algorithm using a disparity gradient limit*. Perception, 14:449-470.

S. T. Barnard, M. A. Fischler, 1982. *Computational Stereo*. Journal ACM CSUR. Volume 14 Issue 4.

G. Pajares et al, 2006. *Fuzzy Cognitive Maps for stereovision matching*. Pattern Recognition, 39, 2101-2114.

R. Berner et al. In ISCAS'07, *A 5 Meps $100 USB2.0 Address-Event Monitor-Sequencer Interface*.

A. Linares-Barranco et al. In ISCAS'10, *AER Convolution Processors for FPGA*.

B. Cope et al, 2006. *Implementation of 2D Convolution on FPGA, GPU and CPU*. Imperial College Report.

B. Cope et al. In FPT'05, *Have GPUs made FPGAs redundant in the field of video processing?*

M. Sivilotti, 1991. *Wiring Considerations in analog VLSI Systems with Application to Field-Programmable Networks*. Ph.D. Thesis, Caltech.

T. Tuytelaars et al. In BMVC'00 *Wide baseline stereo matching based on local, affinely invariant regions*.