Stereo Matching: From the Basis to Neuromorphic Engineering

M. Domínguez-Morales, A. Jiménez-Fernández,
R. Paz-Vicente, A. Linares-Barranco and G. Jiménez-Moreno

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/45901

1. Introduction

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 between 25 and 30 frames per second (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. Classical machine vision started using a single camera (A. Rosenfeld, 1969) as a system sensor in order to perform a treatment for each of the frames obtained by that camera. This method provided a controlled environment but it lacks certain aspects from human vision, such as 3D vision, distance calculation, trajectories, etc.

Nowadays, humankind has experienced a breakthrough in the field of computer vision. This advancement is related to the introduction of a greater number of cameras in the scene (C. Dyer, 2001). Trying to mimic human vision, researchers usually work with a two-camera system, called stereo vision system. In stereo vision, existing algorithms use frames from two digital cameras and process them. Video processing in stereo vision covers many stages during its journey: from the pre-calibration of the cameras (J. Weng et al, 1992; Q. Memon & S. Khan, 2001) to the final outcome, such as distance measurements or 3D reconstruction (R. Tsai, 1987; J. Douret & R. Benosman, 2004). Each step works with frames, processing them pixel by pixel until the pattern that it is looked for is found, or until the treatment that the system if focused on is done. It is important in these systems to calibrate the camera timing to obtain synchronized frames from both cameras. Stereo vision has a wide range of potential application areas including three dimensional map building, data visualisation and robot pick and place.

This chapter will focus on the most difficult step in stereo vision if it is taken into account the computational cost. This step is the stereo vision matching. Throughout this section, a basic knowledge of the common approaches used by stereo matching algorithms is assumed. Also all the steps in the stereovision process will be shown to a lesser extent to see the interaction of each one with the matching process. The purpose of this chapter is to analyse the significant pieces of work produced in the area of stereo vision. In order to do this, a categorisation will be introduced before a global introduction to the stereo vision.

After this introduction to a classical stereo vision system and all the steps that are part of the stereo vision process, this work will focus on a relatively new approach to a digital system implementation: this work will introduce the reader to the world of Neuromorphic Engineering as a new paradigm for codifying, process and transmit data.

Finally, the aim of this work is to show a first approach of a stereo vision system using the principles of Neuromorphic Engineering and applying them to solve one important problem in a stereo vision system: the matching process.

2. Classic machine vision

The goal of Computer vision is to process images acquired with cameras in order to produce a representation of objects in the world (A. Roselfeld et al, 1982). There already exists a number of working systems that perform parts of this task in specialized domains. For example, a map of a city or a mountain range can be produced semi-automatically from a set of aerial images. A robot can use the several image frames per second produced by one or two video cameras to produce a map of its surroundings for path planning and obstacle avoidance. A printed circuit inspection system may take one picture per board on a conveyer belt and produce a binary image flagging possible faulty soldering points on the board.

However, the generic "Vision Problem" is far from being solved. No existing system can come close to emulate the capabilities of a human. Systems such as the ones described above are fundamentally brittle: As soon as the input deviates ever so slightly from the intended format, the output becomes almost invariably meaningless.

There are different models to work with in machine vision. At first, researches looked for industrial applications using a single-view system with only one camera. These systems have lots of limitations due to have only one point of view of the scene. An important breakthrough was to implement systems with multiple points of view: it can be used multiple cameras or a camera in movement. With this modification, the industrial applications experienced a huge improvement in its efficiency: with multiple cameras a three-dimension scenario can be reconstructed and the previous errors produced by using a single camera (no depth knowledge) can be solved. However, researchers top goal in this area are trying to mimic human vision behaviour and functionality. That is why in the area of computer vision there is a big amount of researchers working with stereo vision systems, where a two-camera model is used (S. T. Barnard & M. A. Fischler, 1982).

In machine vision, the two-camera model draws on the biological model of stereovision itself (R. Benosman & J. Devars, 1998), where thanks to the distance between the eyes, the depth can be estimated. This corresponds to the third dimension. The fact of the distance between the two eyes produce a disparity between the visions obtained from each eye (see Figure 1): there is an offset between the information of each eye. In short, the two eyes see the scene in a similar way but with some displacement and, this displacement is inversely proportional to the distance between the eyes and the object itself.

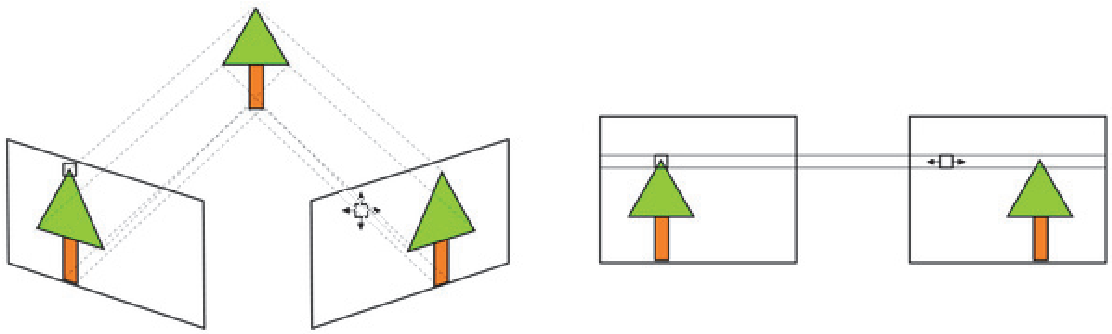

Figure 1. Stereo vision disparity.

Another inherent aspect of stereoscopic vision systems is their geometry. It can be chosen depending on the optical axes geometry: parallel or converging. The human visual system works with converging axes, so the eyes are focused on the objects of interest. When the object is next there is an axes convergence over that object. On the other case, if the object is situated at a certain distance there is almost no eyes convergence. In this case it is common to suppose that the optical axes are in parallel way.

As the reader has probably guessed from the previous introduction, when a stereoscopic vision system is used, two of the common steps in the video process are the image acquisition and the camera system modelling. A greater detailed decomposition of the stereoscopic vision process can be seen in Figure 2.

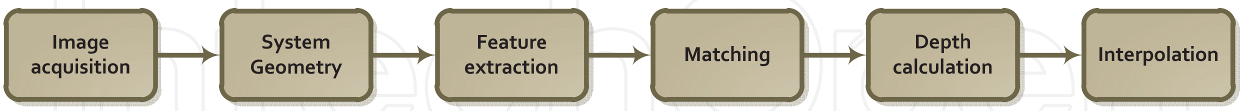

Figure 2. Steps in a stereo vision process.

These six steps are performed in a sequenced order. From all these stages, the most complex known of all, and which determinates the final results obtained, is the fourth one: image matching process. Next, all these steps will be shown quickly before getting into the matching process itself.

2.1. Image acquisition

This step can be done in many different ways. The images, of frames, can be taken simultaneously in time or using a fixed time interval between images. The most important factor in the image acquisition is the kind of application they are going to be used to. It is not the same to consider a cartographic application or a self-controlled vehicle application because there are different needs in each case.

2.2. System geometry

The camera model is a representation of the most important physical and geometrical attributes of the camera. This model has a relative component because it relates the coordinate system of one camera from the other one. In this work, it has been used a geometric model where both cameras are separated a certain distance from each other, but their optical axes are not in a parallel way, so they collide at a determinate distance. More detail of the geometric model will be explained when the full system is presented.

2.3. Feature extraction

In this step, the identification elements of the image are extracted. From those elements, in a second pass of this step, high-level attributes will be extracted. They will be used in the matching step. So this process is closely linked to the next one and, in many aspects, the election of a matching method or another depends on the feature extraction method (or in the absence of it).

2.4. Matching

The correspondence problem consists of finding a unique mapping between the points belonging to two images of the same scene. If the camera geometry is known, the images can be rectified, and the problem reduces to the stereo correspondence problem, where points in one image can correspond only to points along the same scanline in the other image. This step, because of its complexity and its repercussions on the final results, is the most important in the stereo vision process; and that is why the correspondence problem will be deepened in the next epigraph.

2.5. Depth calculation

After the matching process, the system has the correspondences between the elements that appear in one of the projection with the elements of the other one. With this problem solved, depth calculation is a relative easy problem, which consists only in a simple triangulation. However, in some occasions, the execution of this process reveals some non-correlations obtained from the previous step results. These mistakes are due to a lack of precision or to unreliability results.

Thanks to the epipolar restrictions (that would be presented in epigraph 3.1) the projections of a third-dimensional object into both cameras are well-known if the system geometry has been defined properly in the second step. Considering a geometric relation with triangles similarities, if two concrete projections reflected in each camera are related to the same third-dimensional point (solved in the matching process), the coordinates of this object in the space can be calculated and, with them, the third coordinate (Z) is know so the depth too. After this process, it is obtained a depth map of the scene (see Figure 3).

2.6. Interpolation

This step is not always applied, it depends on the mechanisms used in the rest of the steps and the application problem that the system is trying to solve, because in some cases the results obtained at the end of depth calculation process are enough (dense depth map). In other cases, the results show a big amount of three-dimensional points with its correspondence in both cameras but to do an interpolation process these points are not enough.

One of the easiest methods used to solve the interpolation problem is the interpretation of the disparity map obtained from previous steps (see Figure 3). After that the system would obtain a continuous function to obtain the depth of any point in the space given for the projections on both cameras.

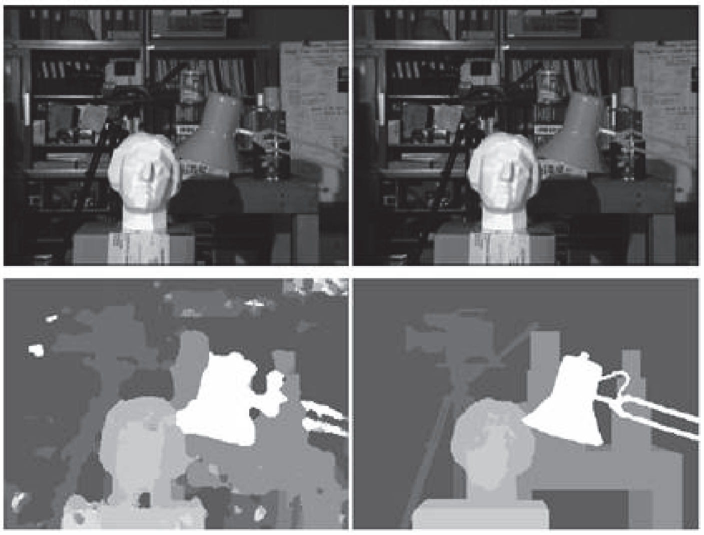

Figure 3. Disparity map and Depth map for a concrete stereo scene.

At this point, all the steps in the stereo vision process have been detailed. Next, the stereo matching process will be exposed in depth.

3. Stereo matching problem

The image matching process has the duty of determinate, for a concrete three-dimensional point, which is its projection on each of the two-dimensional space of both cameras. At the beginning of this step, the results from the other steps are available and can be used to facilitate the matching. First, a local matching has to be done and, to check the results consistency, has to be done a global matching process, which obtain the final results of the whole process (M. Dominguez-Morales et al, 2011). Both matching process use properties from the physic reality to determinate their success. These properties are applied like restriction to the system and are detailed below (see Figure 4 for mostly common used restrictions):

a. **Similarity:** the similarity restriction is much related to the results obtained in the previous step (features extraction). Both projections of the same three-dimensional entity should have similar properties or attributes; like shapes, colours, sizes, vertex number, etc.
b. **Uniqueness:** this restriction applies the condition that one feature in the projection of one of the cameras has one, and only one, feature related to it on the projection of the other camera. However, there are some cases where this restriction may cause more problems than solutions, i.e. the system geometry can produce that one feature does not have a correspondence because of the occlusion of the visual space in the other camera.
c. **Positional order:** given two features in a concrete projection of the scene, this restriction applies the condition that on the other projection both features have to appear in the same order. In most cases this restriction has no problem at all, however, in some cases where both features are very close this restriction may not work correctly.
d. **Disparity continuity:** this property assumes that changes in the image disparity are generally smooth, i.e., if a disparity map is considered it is presented in a continuous way except for a few discontinuities. This principle also appears in different forms and, sometimes, with some small variation, as the case of Minimum Differential Disparity (G. Medioni & R. Nevatia, 1985).
e. **Structural relations:** this principle supposes that objects are made of edges, vertices or surfaces with a certain structure and a geometric arrangement between these elements. In fact, with this restriction the system is looking for geometrical features between the features extracted in the previous step of the whole stereo vision process. Good results can be obtained if the scene has well-defined geometrical objects but, on the other hand, the application of this restriction can get the system worse results if there is not an optimal environment.
f. **Epipolar restriction:** this restriction allows the system to reduce the searching space for the matching process between pixels because of the system geometry. This restriction is very important and very used in the stereo vision system and, to understand it, some introduction to projective geometry has to be done. That is why this restriction is extended in epigraph 3.1.

Stereoscopic restrictions previously described can be applied in different orders depending on the application they are used in. Moreover, there are restrictions that can be used or not. In a typical scenario, the most used ones are: epipolar restriction, similarity, uniqueness and continuity (related to the disparity). Some authors may name these restrictions with different names and fuse some of them into one restriction, but at the end all authors applied similar methods and combinations between restrictions. So, changes on the order of application of these steps may produce two typical alternatives: in both of them the epipolar restriction and the similarity one are very important, as well as uniqueness restriction and continuity (see Figure 4).

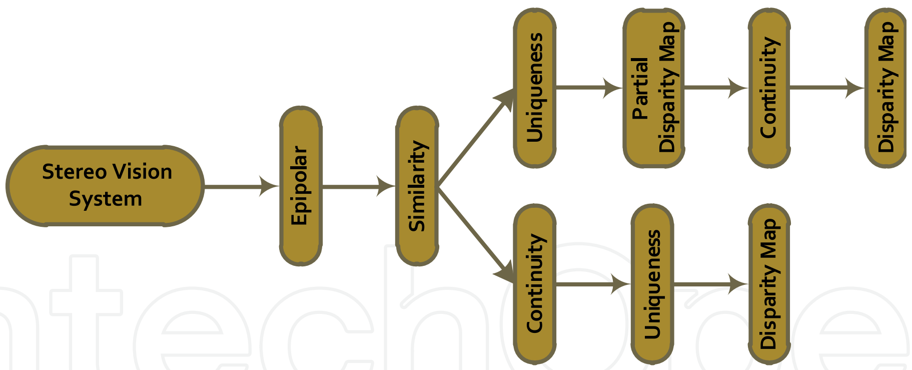

Figure 4. Restrictions application order in the matching process.

3.1. Epipolar restriction

The epipolar geometry is the intrinsic projective geometry between two views. The application of projective geometry techniques in computer vision is most notable in the Stereo Vision problem which is very closely related to Structure-from-Motion. Unlike general motion, stereo vision assumes that there are only two shots of the scene. In principle, then, one could apply stereo vision algorithms to a structure from motion task.

Applying projective geometry to stereo vision is not new and can be traced back from 19th century photogrammetry to work in the late sixties by Thompson (E. Thompson, 1968). However, interest in the subject was recently rekindled in the computer vision community thanks to important works in projective invariants and reconstruction by Faugeras (O. Faugeras, 1992) and Hartley (R. Hartley, R. Gupta, & T. Chang, 1992).

Epipolar restriction is independent of scene structure, and only depends on the cameras' internal parameters and relative pose. The epipolar geometry between two views is essentially the geometry of the intersection of the image planes with the pencil of planes having the baseline as axis (the baseline is the line joining the camera centres). This geometry is usually motivated by considering the search for corresponding points in stereo matching, and this explanation will start from that objective here.

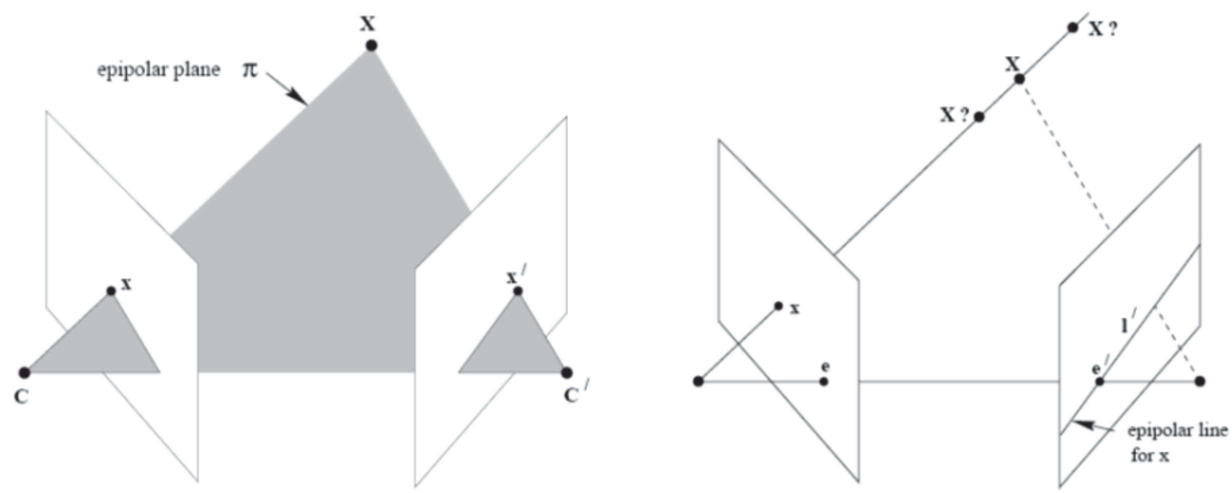

Figure 5. Epipolar restriction.

Suppose that a point X in a third-dimensional space is imaged in two views (see Figure 5), at x in the first, and x′ in the second one. The relation between both points is inherent to the scene and can be seen in Figure 5. As shown in the image: points x and x′, space point X, and camera centres are coplanar (denote this plane as ◎). Clearly, the rays back-projected from x and x′ intersect at X, and the rays are coplanar, lying in ◎. This latter property is the one that is of most significance in searching for a correspondence.

Suppose now that x is the only known point, it can be determined how the corresponding point x′ is constrained. The plane ◎ is determined by the baseline and the ray determined by x. From above it is known that the ray corresponding to the (unknown) point x′ lies in ◎, hence the point x′ lies on the line of intersection l′ of ◎ with the second image plane. This line l′ is the image in the second view of the ray back-projected from x. In terms of a stereo correspondence algorithm the benefit is that the search for the point corresponding to x need not cover the entire image plane but can be restricted to the line l′. These lines are known as epipolar lines. So the matching problem is reduced to seek for the corresponding point; not in the whole image, but only in those points lying on the epipolar line of the other camera.

The linear epipolar geometry formulation also exhibits sensitivity to noise (i.e. in the 2D image measurements) when compared to nonlinear modelling approaches. One reason is that each point can be corresponded to any point along the epipolar line in the other image. Thus, the noise properties in the image are not isotropic with noise along the epipolar line remaining completely unpenalized. Thus, solutions tend to produce high residual errors along the epipolar lines and poor reconstruction. Experimental verification of this can be found in the references (A.J. Azarbayejani, 1997).

After the epipolar restriction has been detailed, this work will continue with the general matching problem. Next, before the discussion of the matching process problems and the application to Neuromorphic Engineering, a global classification of the matching algorithms will be shown.

3.2. Matching algorithms classification

From the previous explanations about matching process, it can be resumed that 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 some 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:

a. **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).

b. **Global methods:** Methods that applies restrictions on the entire image itself. They are usually less sensitive to local peculiarities and they add support to some regions that are difficult to study in a local way. However, they tend to be computationally expensive. Within this group are the dynamic programming methods and nearest neighbour methods (M. Bleyer & M. Gelautz, 2004).

Each technique has its advantages and disadvantages and these ones depends on the system restrictions and the cameras geometry (G. Pajares et al, 2006), as said before. The best matching method would be one that applies the advantages of each of the methods explained before; this is a method that processes the given information using local and global methods and, after it, compares both results and combines them to obtain better results than both of them separately. This fact is very difficult to obtain because the system would need huge computational resources and would not work in a real time system.

Local methods will be discussed later in more detail because they are the most used ones. This work won't go further into global methods because they are rarely used due to their high computational cost.

3.2.1. Area-based matching algorithms

Area-based techniques to solve matching problems in a typical stereo vision system use intensity patterns in the neighbourhood of a concrete pixel to determinate its correlation. It is calculated the correlation between the distribution of disparity for each pixel in an image using a window centred at this pixel, and a window of the same size centred on the pixel to be analysed in the other image (see Figure 6). 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, it can be assumed that the larger the window, the better the outcome. However, the computing power requirements increase in these methods as the window becomes bigger. The biggest problem in these methods is 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 were not taken into account because of their computational inefficiency.

The main advantage of these correlation mechanisms has been previously named in multiple times and it is very easy to deduce it: the computational efficiency (T. Tuytelaars et al., 2000). This characteristic is crucial if the resulting system is wanted to be performed fairly well in real time. On the other hand, the main drawbacks in digital systems primarily focus on results:

- Working directly with each pixel: it can be observed 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 & T. Dennis, 1996).

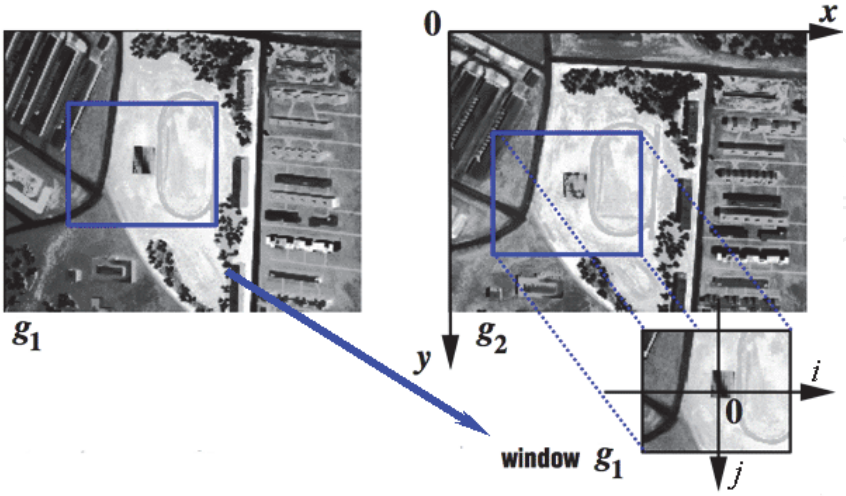

Figure 6. Window correlation.

Therefore, area-based stereo vision techniques look for cross correlation intensity patterns in the local vicinity or neighbourhood of a pixel in an image (L. Tang, C. Wu & Z. Chen, 2002; B. McKinnon & J. Baltes, 2004), 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.

3.2.2. Features-based matching algorithms

As opposed to area-based techniques, the features-based techniques need an image pre-processing before the image matching process (see Figure 7). 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, the algorithm would not be able to have enough information to make inferences and obtain the image correlation.

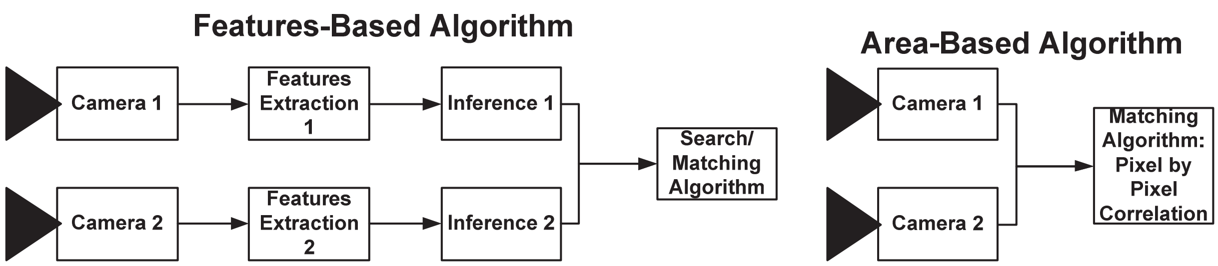

Figure 7. Area-based and features-based algorithms

For features-based stereo vision, symbolic representations are taken from the intensity images instead 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 (see Figure 8), some methods use arrays of edge points to represent straight segments, not straight segments, closed geometric structures which form geometric structures defined or unknown.

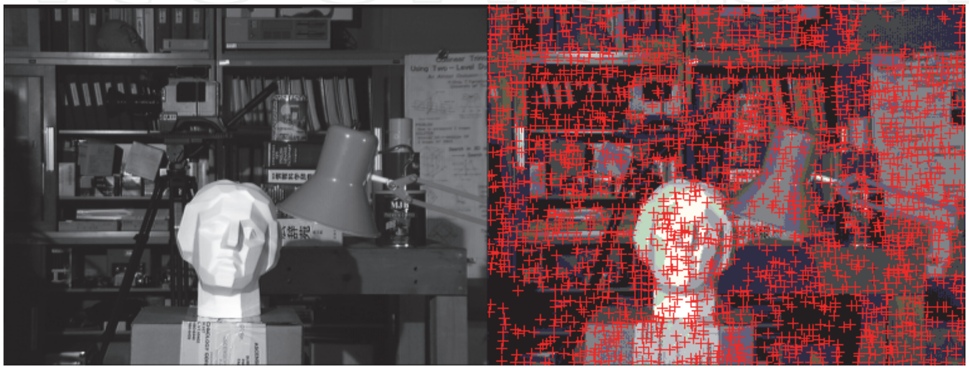

Figure 8. Edge detections in a features-based algorithm.

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.

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 possesses 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. So, it is understandable to say that the bulk of computation corresponds to the feature extraction algorithm and the inference process. This fact will affect to the system it is going to be located in (in a real-time system with a low power consumption it is difficult to use this kind of algorithm). 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 pre-processing stage. Previously, this step was described 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.

However, in general purpose equipment, this technique is the most commonly used because of its results. In classic machine vision, this research branch has been the most deepened in (P. J. Herrera et al, 2009; D. Scaramuzza et al, 2008, P. Premaratne & F. Safaei, 2008).

With these explanations, a global perspective to matching algorithms has been presented as well as classified in different types. All of them have been exposed and evaluated with their advantages and drawbacks. Next, this work introduce the reader to the concept of Neuromorphic Engineering and, after that, a stereo matching approximation to a neuromorphic system is shown.

4. Neuromorphic engineering

Throughout history, many times engineers have achieved solutions to very difficult problems inspired by nature behavior to solve them. This has been applied in many diverse fields, so it is very common to find these bio-inspired systems in the near environment. This is the origin of Neuromorphic Engineering.

However, there are too much unsolved problems in nature that, maybe, could be solved using this kind of mechanisms applied directly to the problems themselves. In neuromorphic engineering, researchers look for the human being "controller" or, what is the same, the nervous system; trying to mimic it, using inverse engineering (V. Chan et al, 2007; Shih-Chii Liu et al, 2010). These systems obtained after looking for answers in the nervous system are called neuro-inspired systems (M. Domínguez-Morales et al, 2011). They are a subset of the bio-inspired systems that try to solve common engineer problems using systems based on the manner that nervous system codifies and processes the information. This is a continuous evolving research branch thanks to the work of many neuromorphic engineers.

Focusing on the vision problems, digital vision systems process sequences of frames from conventional frame-based video sources, like cameras (as was shown in previous epigraphs). For performing complex object recognition algorithms, sequences of computational operations must be performed for each frame (this is like the processing chain in stereo vision that was shown previously). The required computational power and speed required make it difficult to develop a real-time autonomous system. However brains perform powerful and fast vision processing using millions of small and slow cells working in parallel in a totally different way. Primate brains are structured in layers of neurons, in which the neurons in a layer connect to a very large number (~104) of neurons in the following layer (G. M. Shepherd, 1990). Many 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 (a spike is like an electronic pulse produced in the brain by neurons), 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.

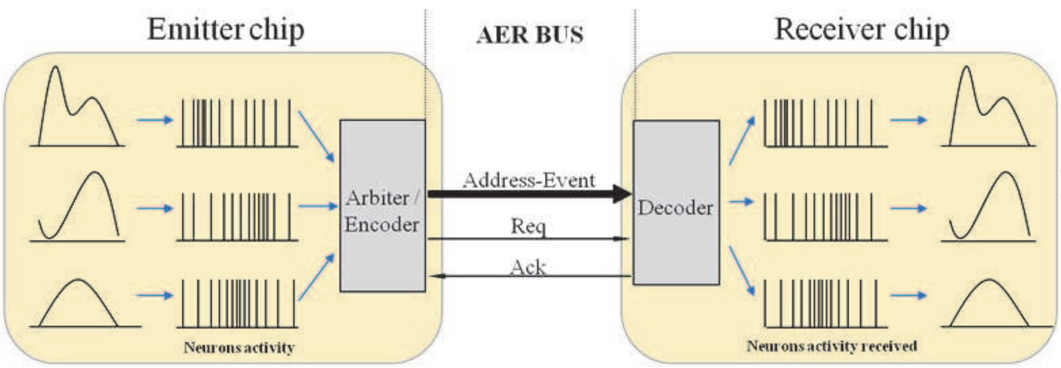

Figure 9. Rate-coded AER inter-chip communication scheme.

AER was proposed by the Mead lab in 1991 (M. Sivilotti, 1991) for achieving a communication between neuromorphic chips with spikes (see Figure 9). 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.

5. Stereo matching in AER system

Hitherto, the reader has had the possibility of getting into the state of the art in stereo vision systems, as well as learning about bio-inspired systems. In this epigraph, a stereo matching algorithm for an AER system will be explained.

First, it is very important to know what type of bio-inspired camera (retina) is going to be used. In this work, and many others done in the same research group, a couple of DVS128 retinas are used (P. Lichtsteiner, C. Posh & T. Delbruck, 2008). This kind of retina has a resolution of 128 rows plus 128 columns, so it has 16384 pixels. The importance of this retina is not the resolution itself, but the work behaviour. These retinas implement the brightness derivative in time, so they only see changes in luminosity or, after a simplification, objects in movement. The mechanism of transmitting the information is centred on a couple of arbitrators (one for the row and the other for the column) and sent via a parallel bus using seven lines for the row ($2^7 = 128$) and another seven for the column.

However, there is no transmission about intensity of the pixel itself. This information is codified in time using the pulses frequency: this is the pulse frequency modulation (PFM). So there are two different possibilities when trying to emulate classical machine vision algorithms behaviour using these retinas: first one is implementing some kind of spiking algebra (A. Jimenez-Fernandez, 2010; A. Jimenez-Fernandez et al, 2012) to attach the problem and solve it in a different way, this option is an important branch of research currently in development and some excellent results have been obtained using it; the second one is trying to adapt classical algorithms to the new paradigm in some way. The final step of this work evaluates similarities between classical stereo vision matching algorithms and AER retinas obtained data, to obtain a first-approach matching algorithm in an AER stereo vision system, full-working on programmable hardware (VHDL over a FPGA).

As a first approximation, it could be considered making an adaptation of the features-based algorithms to obtain a consistent algorithm with good results (see epigraph 3.2.2). However, in this case there is lots of efficiency problems mainly derived from the early stages of pre-processing and inference used to obtain the full set of features from each frame and the second-level features obtained by inference. In order to define an algorithm that is feasible in an AER stereo vision system it has to be taken into account its properties and the goals that the system is wanted to achieve.

At epigraph four it was mentioned an introduction about neuromorphic engineering and, deeper, a first look to AER systems, its motivations, current development and research lines related to them. The main goal in this work and in everyone related to bio-inspired systems is to design and build an autonomous and independent system that works in a real-time ambit, with no need to use a computer to run high-level algorithms. The efficiency of the system does not have to be as important as real-time processing; due to the nature of AER systems, it is not important to make some error in calculations, because its processing is applied in a continuous way and it will be automatically self-corrected over time. Although quality is sacrificed in the results, it cannot be afforded to perform a pre-processing and inference stages, which slow down the full system making impossible to obtain a real-time processing. Moreover, due to the independence requirement derived from the AER system, sending information to a computer via a typical serial port can alter timing constraints of these systems and make it difficult to correlate pixels from both retinas, unless some kind of timestamp is transmitted with each spike. This fact will increase the bandwidth used and can make the computer to lose information. Another major setback to consider in this case is that information transmitted by AER system is closely linked to time and to the number of spikes received and, in serial communication, information is sent in packets and it may have a large time span without receiving any spike.

Resuming, 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 a 128x128 resolution (nowadays some groups work with a 320x240 resolution retina), which means that measure brightness changes over time. Thus, taking a load of 10% in the intensity of the pixels, it 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. Furthermore, the stereo system has two retinas (double data rate), so the information transmission is a critical point to be taken into account.

This system is required to be independent and based on an FPGA connected to the outputs of two AER retinas (see Figure 10). The FPGA will process the information using the concrete algorithm and transmits the resulting information using a parallel AER bus to an USBAERmini2 PCB (R. Berner et al, 2007), which is used like a monitor between the output of the system and the computer. This is responsible for monitoring AER traffic received and transmitted by USB from and to a computer. Should be noted that the computer is only used to verify that the concrete 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 pre-processing 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 used, which do not show all the visual information it covers, but the information they send is the spatial derivative in time. This means that it is only appreciated the information of 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. The next 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. It is proposed an algorithm able to run in a standalone environment in a FPGA, which receives traffic from two AER retinas through a parallel bus.

The algorithm itself work as this: at a first step, it counts the received spikes for each pixel of the image and store this information in a table. So that, this table has 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. If the reader has paid attention to what explained in the rest of the chapter, it may be noticed that in practice what is being done is a process of digitalization of the bio-inspired system. In fact, at this point the algorithm is performing integration over time of the information received from each retina. So, the stored tables are really two frames, each one corresponding to one different retina. 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 pre-configuration setting. Thus, the information of the same pixel in both retinas does not have to be sent at the same time and there could be a frequency variation between both of them; furthermore, both retinas are not presented in a parallel way, so they are put in a concrete angle. To prevent this, it is proposed a fuzzy matching algorithm based on spikes frequency. First, the hardware system will be shown.

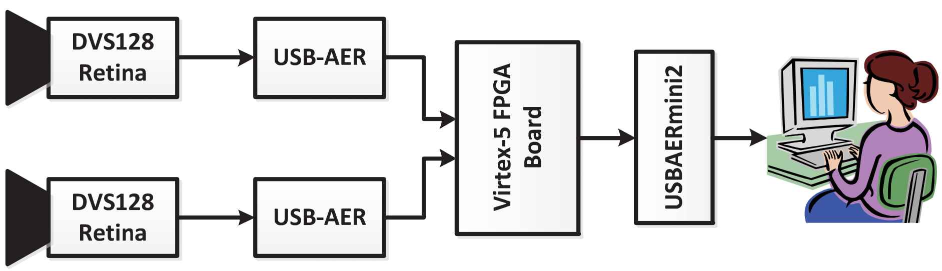

Figure 10. Complete system with all the elements used

All the elements that compose the system are these (from left to right, see Figure 10): two DVS128 retinas (P. Lichtsteiner, C. Posh & T. Delbruck, 2008), two USB-AER, a Virtex-5 FPGA board (AVNET reference), an USBAERmini2 (R. Berner et al, 2007) and a computer to watch the results with jAER software (jAER reference). Next, the non-explained components in this system will be talked about them.

USB-AER (see Figure 11, left) board was developed in the Robotic and Computer Technology Lab during the CAVIAR project (R. Serrano-Gotarredona et al, 2009), and it is based on a Spartan II FPGA with two megabytes of external RAM and a cygnal 8051 microcontroller. To communicate with the external world, it has two parallel AER ports (IDE connector). One of them is used as input, and the other is the output of this board. In the whole system two USB-AER boards have been used, one for each retina. In these boards it has been synthetized in VHDL a filter called Background-Activity-Filter, which allows the system to eliminate noise from the stream of spikes produced by each retina. This noise (or spurious) is due to the nature of analog chips and since anything can be done to avoid it in the retina, it has been filtered in some way. So, at the output of the USB-AER boards, there is the same information given by the retinas but filtered (better quality information).

The other board used is a Xilinx Virtex-5 board, developed by AVNET (AVNET reference). This board is based on a Virtex-5 FPGA and mainly has a big port composed of more than eighty GPIOs (General Purpose Inputs/Outputs ports). Using this port, it has been connected an expansion/testing board, which has standard pins, and it has been used to connect two AER inputs and one output to it.

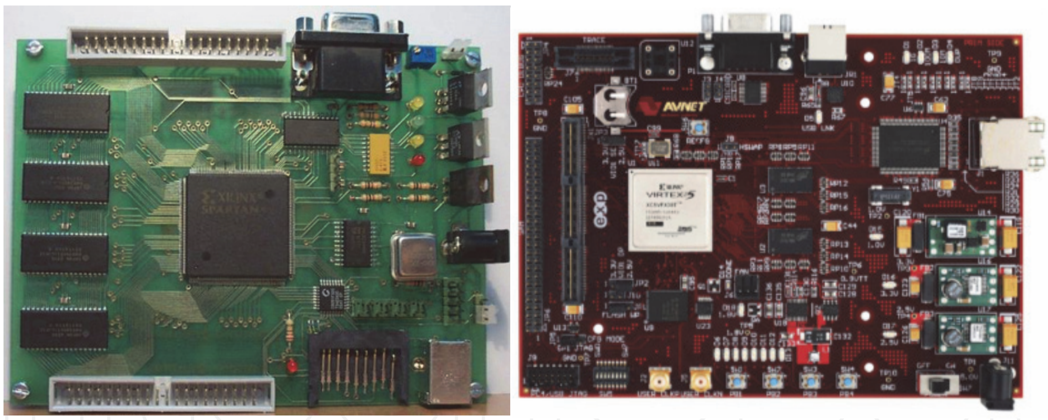

Figure 11. Left, USB-AER board; right, Virtex-5 FPGA board.

The Virtex-5 (see Figure 11, right) implements the whole processing program, which works with the spikes coming from each retina, processes them and obtains the final results. The system behaviour and its functionality are shown in the following sections.

The fuzzy algorithm proposed works as explained next. It will not seek exactly the same levels of intensity in both virtual frames, but would admit an error in the range of 5 and 8% of these levels (in the proposed case, with 256 intensity levels, that means that it is admitted a range of 12-20 levels error). The exact parameters used for the final results depend on the environment where the matching process is applied: if it is an open space it needs a bigger error range, but if it is used a laboratory environment, there is no need for a big error range. This parameter can be adjusted modifying the VHDL code.

The first problem is solved, but it is almost very inefficient algorithm if a pixel by pixel correlation is used. To solve it, the principles of area-based matching algorithms are used in this algorithm. The system is not limited to one pixel correspondence search, neither a global search of the pixel; each correspondence is looked for over a window of a concrete width in the other virtual frame. This window is centred in the position given by the original pixel of the first retina. The size of the window used by the algorithm depends on the environment too, but also in the system calibration. In the system proposed, both retinas are allocated at a distance of 13’5 centimetres between them and in an angle of 86’14 degrees to obtain a focal collision of one meter. 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). In fact, the system divides the correspondence map into regions and each process looks for a correspondence over its region alone. This full process can only be done because of the VHDL implementation over an FPGA, which allows a massive parallel multiprocessing execution (depending on the number of logic cells allocated in the FPGA) (A. Jimenez-Fernandez et al, 2010-2).

The system functionality has been presented, but there is an important element that has not been talk about yet. This element is in charge of controlling the virtual frames timing. In fact, in a classical machine vision system is easy to understand that a video is sent using frames and, each frame, contains the information of the entire scene in a concrete instant. Furthermore, these frames are transmitted using a concrete period (called fps or frames per second). Returning to the system described in this work (see Figure 12), artificial frames have been created but the information stored in them in the integration of the information during all the time, there is no timestamp that divides one virtual frame from the next. That is why the system needs a daemon that preserves this harmony. This daemon will be considered a process in charge of reset all the spikes counters time to time to avoid counters overflow.

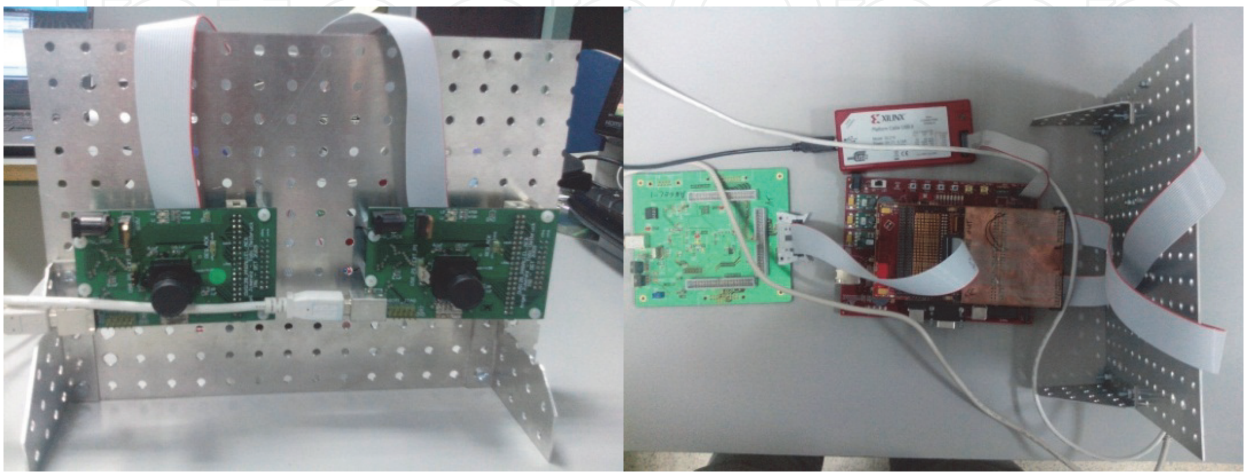

Figure 12. Virtex-5 with AER retinas and USBAERmini2.

There are many adjustable parameters on this system depending on the environment where the stereo vision matching is applied and on the area it is going to be used. The whole system is under testing stages: first it was submitted to simulation tests over Matlab software and, at the present, the system is under testing directly on the FPGA.

6. Conclusions

Humanity has experienced great changes in the field of vision, but they have always been aimed to get the results of the vision system of a human being.

In this work, an introduction to stereo vision systems has been shown, as well as an explanation about all the typical steps used in a common stereo vision system. It has entered deeply into the most important stage: the matching process, which has been theoretically analysed in depth, showing and explaining most typical algorithms that solve this problem in classic machine vision.

Next, a relatively new processing and encoding paradigm has been explained with its advantages and drawbacks. It has been discussed the existing relations between classical stereo matching process and the stereo matching process related to this new paradigm (AER stereo matching process).

Finally, an Address-Event-Representation stereo matching algorithm has been detailed using classical stereo vision concepts and adapting them to the bio-inspired system. As well, the AER stereo system has been shown and all the elements that compose it have been exposed.

Author details

M. Domínguez-Morales, A. Jiménez-Fernández,
R. Paz-Vicente, A. Linares-Barranco and G. Jiménez-Moreno
Robotic and Computer Technology Group - University of Seville, Spain

Acknowledgement

This work has been supported by the Spanish Science and Education Ministry Research Projects TEC2009-10639-C04-02 (VULCANO).

7. References

A. Rosenfeld (1969). First Textbook in Computer Vision: *Picture Processing by Computer*, Academic Press, New York.

C. Dyer. (2001). *Volumetric scene reconstruction from multiple views*, In L.S. Davis, editor, Foundations of image analysis. Kluwer, Boston.

J.Weng, P. Cohen, & M. Herniou (1992). *Camera Calibration with Distortion Models and Accuracy Evaluation*, IEEE Trans. Patt. Anal. Machine Intell., vol. 14, no. 10, pp. 965–980.

Qurban Memon & Sohaib Khan (2001). *Camera Calibration and Three-Dimensional World Reconstruction of Stereo-Vision Using Neural Networks*, International Journal of Systems Science.

R. Tsai (1987). *A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses*, IEEE Transactions on Robotics and Automation.

J. Douret & R. Benosman (2004). *A multi-cameras 3D volumetric method for outdoor scenes: a road traffic monitoring application*, International Conference on Pattern Recognition (ICPR).

A. Rosenfeld & A. C. Kaк (1982). *Digital Picture Processing*, Academic Press, New York, 1976; 2nd ed. (2 vols.).

S. T. Barnard & M. A. Fischler (1982). *Computational Stereo*, Journal ACM Computing Surveys (CSUR), vol. 14 Issue 4.

R. Benosman & J. Devars (1998). *Panoramic stereo vision sensor*, International Conference on Pattern Recognition.

M. Dominguez-Morales et al (2011). *Image Matching Algorithms using Address-Event-Representation*, International Conference on Signal Processing and Multimedia Applications (SIGMAP).

G. Medioni & R. Nevatia (1985). *Segment Based Stereo Matching*, Computer Vision, Graphics and Image Processing, 31, 2-18.

E. Thompson (1968). *The projective theory of relative orientation*, Photogrammetria, no. 23(1): 67-75.

O. Faugeras (1992). *What can be seen in three dimensions from an uncalibrated stereo rig?*, Proceedings of the 2nd European Conference on Computer Vision, pages 563-578, Santa Margherita Ligure, Springer-Verlag.

R. Hartley, R. Gupta, & T. Chang (1992). *Stereo from uncalibrated cameras*, Proceedings of the Conference on Computer Vision and Pattern Recognition, pages 761-764, Urbana-Champaign.

A. J. Azarbayejani (1997). *Nonlinear Probabilistic Estimation of 3-D Geometry from Images*, PhD thesis, Massachusetts Institute of Technology.

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

M. Bleyer & M. Gelautz (2004). *A Layered Stereo Algorithm Using Image Segmentation and Global Visibility Constraints*, ICIP, pp. 2997-3000.

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

T. Tuytelaars et al (2000). *Wide baseline stereo matching based on local, affinely invariant regions*, British Machine Vision Conference (BMVC).

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

L. Tang, C. Wu & Z. Chen (2002). *Image dense matching based on region growth with adaptive window*, Pattern Recognition Letters, 23, 1169–1178.

B. McKinnon & J. Baltes (2004). *Practical region-based matching for stereo vision*, proceedings of 10th International Workshop on Combinatinal Image Analysis (IWCIA), Springer, pp. 726–738, LNCS 3322.

P. J. Herrera et al. (2009). *A Featured-Based Strategy for Stereovision Matching in Sensors with Fish-Eye Lenses for Forest Environments*, Sensors, 9, 9468-9492

D. Scaramuzza, N. Criblez, A. Martinelli, R. Siegwart (2008). *Robust feature extraction and matching for omnidirectional images*, Field and Service Robotics, Springer, vol. 42, pp. 71–81.

P. Premaratne & F. Safaei (2008). *Feature based Stereo correspondence using Moment Invariant*, proceedings of the 4th International Conference on Information and Automation for Sustainability (ICIAFS), pp. 104–108.

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

Vincent Chan, Shih-Chii Liu & A. van Shaik (2007). AER EAR : *A Matched Silicon Cochlea Pair with Address Event Representation Interface*, IEEE Transactions on Circuits and Systems, 54, 48-59.

Shih-Chii Liu et al (2010). *Event-based 64-channel binaural silicon cochlea with Q enhancement mechanisms*, proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), pp. 2027-2030.

M. Dominguez-Morales et al. (2011). *An approach to distance estimation with stereo vision using Address-Event-Representation*, International Conference on Neural Information Processing (ICONIP).

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

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

A. Linares-Barranco et al. (2010). *AER Convolution Processors for FPGA*, International Symposium of Circuits And Systems (ISCAS).

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

B. Cope et al. (2005). *Have GPUs made FPGAs redundant in the field of video processing?*, International Conference on Field-Programmable Technology (FPT).

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

P. Lichtsteiner, C. Posh & T. Delbruck (2008). *A 128×128 120dB 15 us Asynchronous Temporal Contrast Vision Sensor*, IEEE Journal on Solid-State Circuits, vol. 43, no 2, pp. 566-576.

A. Jiménez-Fernandez (2010). *Diseño y evaluación de sistemas de control y procesamiento de señales basados en modelos neuronales pulsantes*, Ph.D. Thesis, University of Seville (Spain).

A. Jiménez-Fernandez et al. (2012). *A Neuro-Inspired Spike-Based PID Motor Controller for Multi-Motor Robot with Low Cost FPGA*, Sensors, vol. 12, no 4, pp. 3831-3856.

R. Berner, T. Delbruck, A. Civit-Balcells & A. Linares-Barranco (2007). *A 5 Meps $100 USB2.0 Address-Event Monitor-Sequencer Interface*, International Symposium of Circuits And Systems (ISCAS).

R. Serrano-Gotarredona et al (2009). *CAVIAR: A 45k-Neuron, 5M-Synapse, 12G-connects/sec AER Hardware Sensory-Processing-Learning-Actuating System for High Speed Visual Object Recognition and Tracking*, IEEE Transactions on Neural Networks, vol. 20, no. 9, pp. 1417-1438.

T. Tuytelaars et al. (2000). *Wide baseline stereo matching based on local, affinely invariant regions*, British Machine Vision Conference (BMVC).

A. Jiménez-Fernandez et al (2010-2). *Building Blocks for Spike-based Signal Processing*, IEEE International Joint Conference on Neural Networks (IJCNN).

AVNET Virtex-5 FPGA board: http://www.em.avnet.com/drc

jAER software: http://sourceforge.net/apps/trac/jaer/wiki