On the AER Stereo-Vision Processing: A Spike Approach to Epipolar Matching

Manuel Jesus Domínguez-Morales[1], Elena Cerezuela-Escudero[1],
Fernando Perez-Peña[2], Angel Jimenez-Fernandez[1],
Alejandro Linares-Barranco[1], and Gabriel Jimenez-Moreno[1]

[1] Robotic and Technology of Computers Lab, University of Seville, Spain
{mdominguez,ecerezuela,ajimenez,alinares,gaji}@atc.us.es
[2]Applied Robotics Research Lab, University of Cadiz, Spain
fernandoperez.pena@uca.es

Abstract. Image processing in digital computer systems usually considers 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 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 finds a pattern match in a section of both stereo frames. To process stereo vision information, an image matching process is essential, but it needs very high computational cost. Moreover, as more information is processed, the more time spent by the matching algorithm, the more inefficient it is. Spike-based processing is a relatively new approach that implements 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. The aim of this work is to study the viability of a matching mechanism in a stereo-vision system, using AER codification. This kind of mechanism has not been done before to an AER system. To do that, epipolar geometry basis applied to AER system are studied, and several tests are run, using recorded data and a computer. The results and an average error are shown (error less than 2 pixels per point); and the viability is proved.

Keywords: Address-Event-Representation, spike, neuromorphic engineering, stereo, epipolar geometry, vision, dynamic vision sensors, retina.

1 Introduction

In recent years there have been numerous advances in the field of vision and image processing, because these matters can be applied for scientific and commercial purposes to numerous fields such as medicine, industry or entertainment. As it can be deduced, the images are two dimensional while the daily scene is three dimensional.

This means that, between the passage from the scene (reality) and the image, there is a loss of what we call the third dimension. Nowadays, society has experienced a great advance in these aspects: 2D vision has given way to 3D viewing. Industry and research teams 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, a new branch of research has been developed, focused on stereoscopic vision [1]. 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. distance estimation.

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 completely. In any process of stereoscopic vision, image matching is the main problem that has consumed a large percentage of research resources in this field, and it is still completely open to research. Matching is the process performed in every stereo system to find the pixel within a camera matrix which corresponds to a particular one of the opposite camera. This process is critical, because it allows obtaining high-level results like distance calculation [2-3] or shape description. The main problem related to image matching is the computational cost needed to obtain appropriate results. There are lots of high-level algorithms used in digital stereo vision that can solve this problem, but they involve a high computational cost. Nowadays, mathematicians are trying to use several techniques in order to reduce the number of possible matches in the second camera. Calibration mechanisms and Epipolar Geometry are applied to pre-configure these systems and to obtain better and more optimal results.

In parallel to all these computational vision evolution, Neuromorphic Engineering arises, whose operation principles are based on the biological models themselves. 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, where the neurons of a layer connect to a very large number (~104) of neurons in the following one [4]. Connectivity mostly 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 of a set of layers [4], 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 [5-7]. However, these systems generally run at extremely low speeds because the models are implemented as software programs. For real-time solutions direct hardware implementations are required. A growing number of research groups around the world are implementing these principles onto real-time spiking hardware through the development and exploitation of the so-called AER (Address Event Representation) technology.

AER was proposed by the Mead lab in 1991 [8] 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 Fig. 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 [9-10]. There is a community of AER protocol users for bio-inspired applications in vision and audition systems. 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.

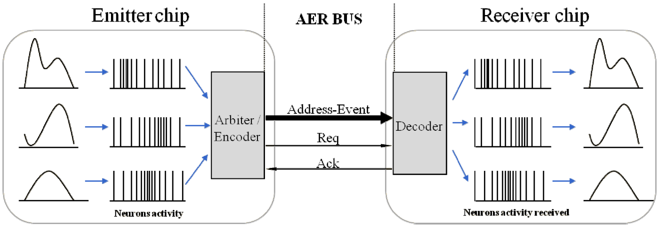

Fig. 1. Rate-coded AER inter-chip communication scheme

The optical sensor used by these research groups is the DVS128 AER retina [11]. This sensor works in such a way that each pixel only detects the derivate in time of the luminosity. That means that this retina only perceives luminosity variations or, as it is very common, objects in movement. This fact simplifies the amount of information transmitted, and it helps the researcher to focus on the important information of the scene. First, the epipolar geometry principles used in stereo vision image matching algorithms will be described. After that, these principles will be applied to a stereo vision AER system with the information obtained by the DVS-retina. Finally, the results and error measurements will be shown.

2 Epipolar Geometry

Epipolar geometry [12] is the intrinsic projective geometry between two views. It is independent of scene structure, and it 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 centers). This geometry is motivated by considering the search for corresponding points in stereo matching.

Suppose a point *X* in space is imaged in two views (one from each camera), at *x* in the first, and *x'* in the second. In this case, the relation between the corresponding image points *x* and *x'*, the spatial point *X* and the camera centers is that all of them are coplanar (located in the same plane, *π*). Clearly, the rays back-projected from *x* and *x'* intersect at *X*, and the rays are coplanar, lying in *π*. This latter property is the most significant in searching for a correspondence (see Fig. 2).

Supposing now that we know only *x*, we may ask how the corresponding point *x'* is constrained. The plane *π* is determined by the baseline and the ray defined by *x*. From above we know that the ray corresponding to the point *x'* lies in *π*, hence the point *x'* lies on the line of intersection *l'* of *π* 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* does not need to cover the entire image plane but it can be restricted to the line *l'*. The importance of this fact is that, if a transformation mechanism between *x* and line *l'* can be obtained using a pre-calibration step, a simple AER spiking system implemented on programmable hardware (FPGA, etc.) will be able to discriminate the possible matches calculating the opposite epipolar line. To do that, the spikes building blocks can be used to operate with simple elements between spikes [13]. This mechanism needs the Fundamental Matrix to do that transformation.

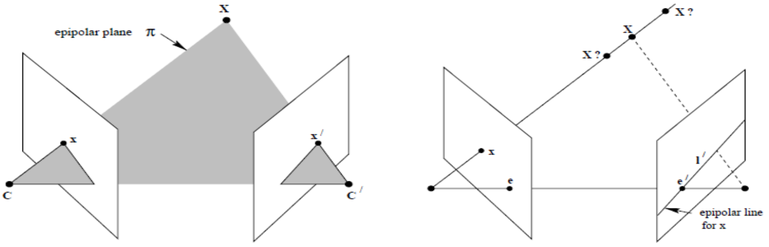

Fig. 2. Epipolar geometry explanation

Next, a quick introduction to the calibration mechanism will be shown, as well as the projection matrixes calculation.

3 Pre-calibration Summary

Camera calibration is, mainly, finding the internal quantities of the camera that affect the imaging process, like the position of the image center, the focal length, the lens distortion, etc. These parameters are joined to obtain the 'camera matrix'. This is an important process for rebuilding a world model. Several calibration mechanisms exist in classical computer vision. Some of them use a two-step process to obtain the camera matrix: first approach and optimization. Pin-Hole is the most used camera model (the way in which the camera system interacts with the world). This fact, determines the mechanism used to calibrate the system.

In this work, a Pin-Hole model [14] and a Tsai calibration mechanism [15] have been used to calibrate the system. Also, as Tsai used in his experiments, a calibration grid was built to allow the retina to see the calibration points. In this case, the way in which the retina works (section 1) obstructs the calibration process because, in order to obtain good calibration results, the objects used in the calibration process cannot be in movement (remember that the retina only perceives the variation of the luminosity). Hence, the calibration grid built is different from the one used before. It is composed of a matrix of *8 by 8* LED lights and has been connected to a microcontroller that switches on and off each one of the LEDs. The calibration system is shown in Fig.3.

By capturing the information from the LED grid and processing it in a computer, the *64* projection points are obtained on each retina (after the application of some image filters). Using the Tsai mechanism [15], both camera matrixes (one for each retina) are calculated (first step). These matrixes are used to determine the 2D points in each retina, related to the 3D points in the space. Also, they have the internal information about each retina (internal parameters), so they describe the physical and the environmental properties of the cameras. Next step is testing and error measurement: using several space points and testing the conversion from 3D to 2D points, the final result is an average error of less than 1%. Also, these matrixes are optimized (second step) using the Faugeras mechanism [16]. After testing, new matrixes work with an error even smaller than the one obtained before (around 0.8%).

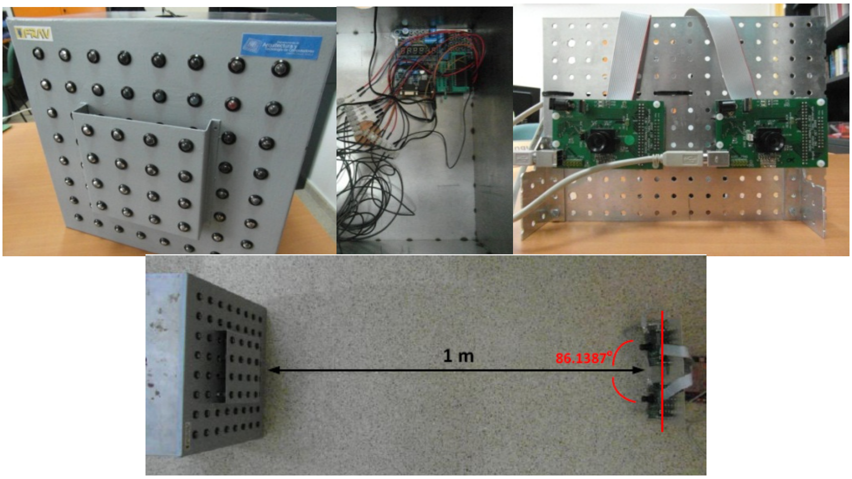

Fig. 3. Calibration system description

The main question is: having the 2D point of one retina, how can it be determined which point in the other retina corresponds to the same 3D point? This is necessary in order to obtain the coordinates of the 3D point and this process is known as the 'matching' step. To do that, in this work, epipolar geometry is applied to obtain the epipolar lines related to these points. Next, the mechanism used to calculate the epipolar lines from both projection matrixes will be shown.

4 Epipolar Lines Applied to a Calibrated AER Stereo-Vision System

The aim of this work, given a 2D point coordinates from one retina, is to discriminate the possible matches on the second retina. To do this, the epipolar line within the matching point will be calculated. Summarizing the classical machine vision principles, given a camera matrix *P* and a 3D point coordinates (X, Y, Z); the 2D point represented on the retina (U, V) is obtained using equation 1.

$$\begin{pmatrix} U \\ V \\ t \end{pmatrix} = P \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}, \qquad P = \begin{pmatrix} q_{11} & q_{12} & q_{13} & q_{14} \\ q_{21} & q_{22} & q_{23} & q_{24} \\ q_{31} & q_{32} & q_{33} & q_{34} \end{pmatrix}, \qquad t = scale\ factor \tag{1}$$

To obtain the epipolar line, the inverse process will be done: given a 2D point from one of the retinae, the corresponding 3D point should be determined and, with its information and using the camera matrix of the second retina, the 2D coordinates of the second one will be obtained. However, a transformation from 3D to 2D entails a loss of information, so the distance information (coordinate *Z*) is impossible to obtain given only one 2D point (inverse process). So, how can the matching point (2D point in the other retina) be obtained? As said before, the final result will give the line within which the matching point is, not the exact point itself. The triangulation process is necessary to find the intersection of the rays projected from both retinae. Using the Pin-Hole model and both camera matrixes, the equation system can be modified to obtain the fundamental matrix (combining both retinae cameras), resulting on equation 2.

$$F \cdot M = 0, \qquad where\ M = (X_i \quad Y_i \quad Z_i \quad \lambda)^t$$

$$F = \begin{pmatrix} P_{L1,1} - u_i \cdot P_{L3,1} & P_{L1,2} - u_i \cdot P_{L3,2} & P_{L1,3} - u_i \cdot P_{L3,3} & P_{L1,4} - u_i \cdot P_{L3,4} \\ P_{L2,1} - v_i \cdot P_{L3,1} & P_{L2,2} - v_i \cdot P_{L3,2} & P_{L2,3} - v_i \cdot P_{L3,3} & P_{L2,4} - v_i \cdot P_{L3,4} \\ P_{R1,1} - u'_i \cdot P_{R3,1} & P_{R1,2} - u'_i \cdot P_{R3,2} & P_{R1,3} - u'_i \cdot P_{R3,3} & P_{R1,4} - u'_i \cdot P_{R3,4} \\ P_{R2,1} - v'_i \cdot P_{R3,1} & P_{R2,2} - v'_i \cdot P_{R3,2} & P_{R2,3} - v'_i \cdot P_{R3,3} & P_{R2,4} - v'_i \cdot P_{R3,4} \end{pmatrix}, \tag{2}$$

where *M* is the spatial point, *(u_i, v_i)* are the coordinates of the 2D points projected on the left retina, *(u'_i, v'_i)* are the coordinates of the 2D point projected on the right retina, P_L is the left camera matrix, P_R is the right camera matrix and *F* is the fundamental matrix (to obtain a 3D point given both 2D points). This system is solved using Singular-Value Decomposition (SVD), obtaining the resulting points in the last column of the V matrix: $(V_{1,4}\ \ V_{2,4}\ \ V_{3,4}\ \ V_{4,4})^t$. Finally, dividing by the scale factor: $(V_{1,4}/V_{4,4}\ \ V_{2,4}/V_{4,4}\ \ V_{3,4}/V_{4,4}\ \ V_{4,4}/V_{4,4})^t$;which is: $(X_i\ \ Y_i\ \ Z_i\ \ 1)^t$. The system *F*, can be obtained step by step, starting on the projection matrixes. Demonstration:

$$q_{11}X + q_{12}Y + q_{13}Z + q_{14} = u = Ut, \qquad q_{21}X + q_{22}Y + q_{23}Z + q_{24} = v = Vt$$
$$q_{31}X + q_{32}Y + q_{33}Z + q_{34} = t \tag{3}$$

Using *t*:
$$\begin{gathered} q_{11}X + q_{12}Y + q_{13}Z + q_{14} = U(q_{31}X + q_{32}Y + q_{33}Z + q_{34}) \\ q_{21}X + q_{22}Y + q_{23}Z + q_{24} = V(q_{31}X + q_{32}Y + q_{33}Z + q_{34}) \end{gathered} \tag{4}$$

Grouping the coefficients of the 3D coordinates:

$$\begin{gathered} (q_{11} - Uq_{31})X + (q_{12} - Uq_{32})Y + (q_{13} - Uq_{33})Z + (q_{14} - Uq_{34}) = 0 \\ (q_{21} - Vq_{31})X + (q_{22} - Vq_{32})Y + (q_{23} - Vq_{33})Z + (q_{24} - Vq_{34}) = 0 \end{gathered} \tag{5}$$

Changing the name of the previous expressions for a_1, b_1, c_1, d_1, a_2, b_2, c_2 and d_2:

$$a_1X + b_1Y + c_1Z + d_1 = 0, \qquad a_2X + b_2Y + c_2Z + d_2 = 0$$
$$X = \frac{Z(b_1c_2 - b_2c_1) + (b_1d_2 - b_2d_1)}{(a_1b_2 - a_2b_1)}, \qquad Y = \frac{Z(a_2c_1 - a_1c_2) + (a_2d_1 - a_1d_2)}{(a_1b_2 - a_2b_1)} \tag{6}$$

With the given values of one vision sensor, two equations with three unknown terms can be obtained. Summarizing, given the 2D point of one retina, the line in space, where the 3D point is, can be calculated (line *l*). Using two 3D random points situated on *l* and calculating its projections on the second retina, two 2D points are obtained. The line obtained by linking these 2D points is known as the 'epipolar line' (placed on the second retina): the matching point has to be situated over it. Using two random *Z* coordinates (i.e. *-10* and *10*), two 3D points are calculated. With these points, their projections over the second retina are calculated using its camera matrix. After that, the projected line on the second retina can be shown using the inclination calculated between both 2D points. If there is no error, the matching point must be situated on this line. Next, results and error measurements will be shown.

5 Matching Results and Error Measurement

With the results, epipolar lines between both retinas have been calculated. To see all the data, Fig.4 shows the epipolar lines obtained from the opposite retina and the 'target' points of this retina. In one case, it shows the epipolar lines calculated from the right retina with the points projected on the left retina. The other case is the opposite. To appreciate them better, examples have been run with *64* and *8* points.

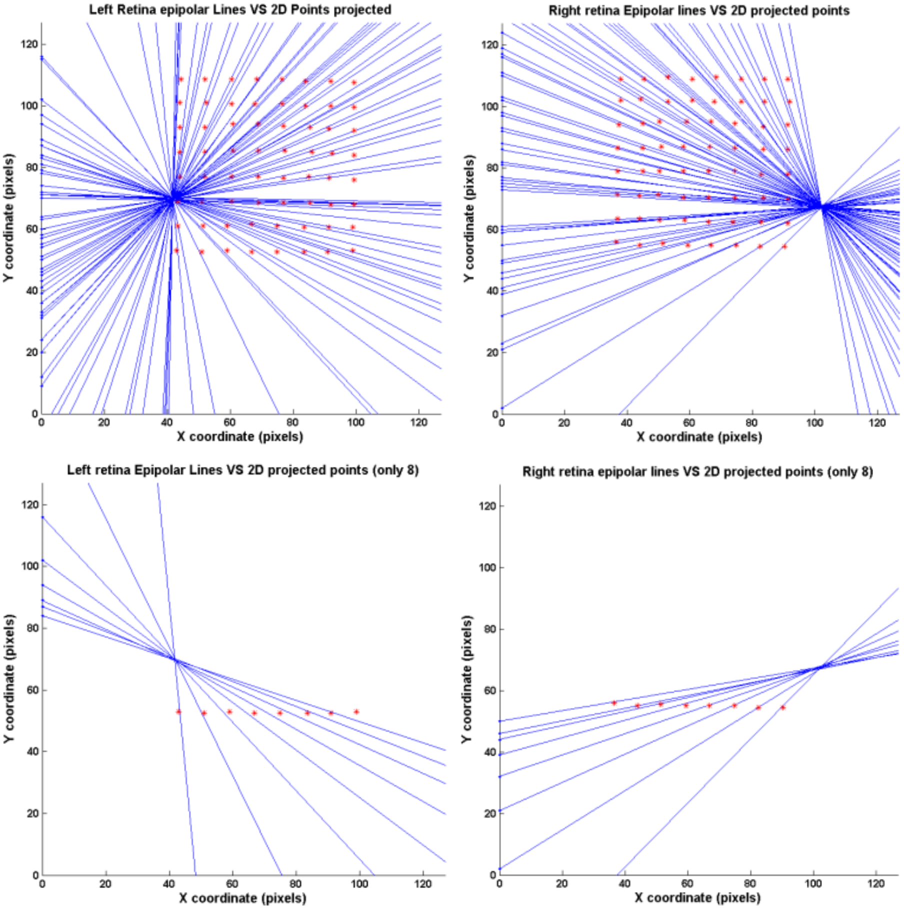

Fig. 4. Epipolar lines VS 2D projected points

It is easier to understand the results by watching the 8-point tests. For example: in the bottom-left image (left retina), the first point is the one in the left. A red asterisk can be seen, which is the point captured by the left retina, and a blue line passing next to the point, which is the epipolar line obtained by the right retina and projected on the left retina. As it can be seen, the point is situated over the line (almost no error).

To calculate the error, the distance between each epipolar line to the point itself has been calculated. Average error for the 64-point test is *2.2299* pixels per point for the left retina and *1.3957* for the right one. For the 8–point test the average error is *1.4240* for the left retina and *1.2239* for the right one. The obtained error is tolerable.

6 Conclusions

In this work, a matching algorithm based on epipolar geometry and a calibrating step has been tested under a neuromorphic AER-DVS stereo vision system. Mathematical principles have been shown and demonstrated. The implemented system has been detailed, tested using sixty-four 3D points coordinates and its error has been measured. All the results, errors, graphs and formulas have been presented.

The proposed system obtains a tolerable error (less than 2 pixels per point) to work under a spiking system with two DVS128 retinae [11]. It has been proved that this mechanism works with spiking data and under the restrictions of an AER system.

As further research, the next step is to implement the full mechanism into programmable hardware, like FPGAs, in order to obtain an autonomous system without the computer intervention.

References

1. Barnard, S.T., Fischler, M.A.: Computational Stereo. Journal ACM CSUR 14(4) (1982)
2. Dominguez-Morales, M., et al.: An approach to distance estimation with stereo vision using Address-Event-Representation. In: International Conference on Neural Information Processing, ICONIP (2011)
3. Dominguez-Morales, M., et al.: Live Demonstration: on the Distance Estimation of Moving Targets with a Stereo-Vision Aer System. In: ISCASS (2012)
4. Shepherd, G.M.: The Synaptic Organization of the Brain, 3rd edn. Oxford University Press (1990)
5. Lee, J.: A Simple Speckle Smoothing Algorithm for Synthetic Aperture Radar Images. Man and Cybernetics SMC-13 (1981)
6. Crimmins, T.: Geometric Filter for Speckle Reduction. Applied Optics 24, 1438–1443 (1985)
7. Linares-Barranco, A., et al.: AER Convolution Processors for FPGA. In: ISCASS (2010)
8. Sivilotti, M.: Wiring Considerations in analog VLSI Systems with Application to Field-Programmable Networks. Ph.D. Thesis, Caltech (1991)
9. Cope, B.: Implementation of 2D Convolution on FPGA, GPU and CPU. I.C. Report (2006)
10. Cope, B., et al.: Have GPUs made FPGAs redundant in the field of video processing? FPT (2005)
11. Lichtsteiner, P., Posh, C., Delbruck, T.: A 128×128 120dB 15 us Asynchronous Temporal Contrast Vision Sensor. IEEE Journal on Solid-State Circuits 43(2), 566–576 (2008)
12. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press (2004)
13. Jimenez-Fernandez, A., et al.: Building Blocks for Spike-based Signal Processing. In: IEEE International Joint Conference on Neural Networks, IJCNN (2010)
14. Rosenfeld, A.: First Textbook in Computer Vision: Picture Processing by Computer. Academic Press, New York (1969)
15. Tsai, R.Y.: A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. IEEE Int. Journal Robotics and Automation. 3(4), 323–344 (1987)
16. Faugeras, O.: Three-dimensional computer vision: a geometric viewpoint. MIT Press (1993)