Gaze control of an active vision system in dynamic scenes

Yann Ducrocq, Shahram Bahrami, Luc Duvieubourg, François Cabestaing

Gaze control of an active vision system in dynamic scenes

Ducrocq Yann [1], Bahrami Shahram [1],
Duvieubourg Luc [2] and Cabestaing François [2],
yann.ducrocq@eipc.fr, shahram.bahrami@eipc.fr,
Luc.Duvieubourg@univ-lille1.fr, francois.cabestaing@univ-lille1.fr

[1] Ecole d'Ingénieurs du Pas-de-Calais, Département Automatique, Campus de la Malassise, 62967 Longuenesse Cedex - FRANCE
[2] Laboratoire LAGIS - UMR CNRS 8146, Université des Sciences et Technologies de Lille, Cité Scientifique - Bâtiment P2, 59655 Villeneuve dAscq - FRANCE

Abstract. In this paper, we present an active vision system that imitates the human viewing strategy. We know that gaze control is a major activity in scene observation. In fact, a person instinctively directs his visual fixation point toward regions of interest containing salient information. This information is then acquired by the region of the retina near the center of the field of view, i.e. the fovea. In this region, the acuity and the color sensitivity are higher. In order to mimic this human behavior, we have developed a novel active vision system based on a particular stereovision rig. It is composed with one camera, one prism and a set of mirrors. To direct the fixation point, the prism is rotated about its axis by a motorized stage. The system is designed for accurate dynamical adjustments of gaze and for operating at video rate. To validate our approach to active perception, we have developed a simple observation strategy for a real scene with dynamic objects. We present an experiment on real videos, which show its efficiency.

1 Introduction

Cameras have become widely popular sensors in mobile robotics because of their closeness to human perception. Since vision is our most valuable sense, researchers are inclined to incorporate such capabilities in the mobile machines they build. The traditional vision systems are generally based on the concepts defined in the Marr paradigm [1]. The latter is widely used, but unfortunately does not lead to successful vision applications performing recognition and navigation tasks. In fact, most researchers consider either a static or mobile sensor, but never a sensor with adjustable properties. That is why Aloimonos [2], Bajcsy [3], and Ballard [4] have suggested a radical change of the Marr paradigm by proposing the concept of **active vision**. For example, with a moving acquisition system we can greatly increase the field of observation of the sensor. A second example is an active stereovision system with two mobile cameras that can control the vergence, that is, the points of interest that are observed with

no disparity. Since the theoretical framework described by Aloimonos, Bajcsy and Ballard, many projects have been developed that mimic as much as possible the human visual system. They mainly relate to vision heads for mobile robot navigation [5, 6].

Others works are focused on the development of humanoid robots imitating human behavior as accurately as possible. Kozima [7] has developed an infant-like humanoid robot, fitted out with a foveated stereo vision head, to investigate the underlying mechanisms of social intelligence and communicate with human beings. In [8], Aryananda has presented an active-vision humanoid robot head able to interact with people using simple visual and verbal actions during long periods of time.

An other domain where active vision systems are widely used is driver assistance. Clady [9] developed a vision system for tracking a specific moving object, i.e. the most dangerous car, in a sequence of road images. In [10], Pellkofer and Dickmanns described an approach to optimal gaze control for autonomous vehicles equipped with a vision system. Their goal was to determine the direction to observe and to predict the future situation. In [11], they proposed a vision sensors composed of four cameras with different focal lengths. Their system was mounted on a pan-tilt camera head and was designed in order to perform smooth pursuit and saccadic movements.

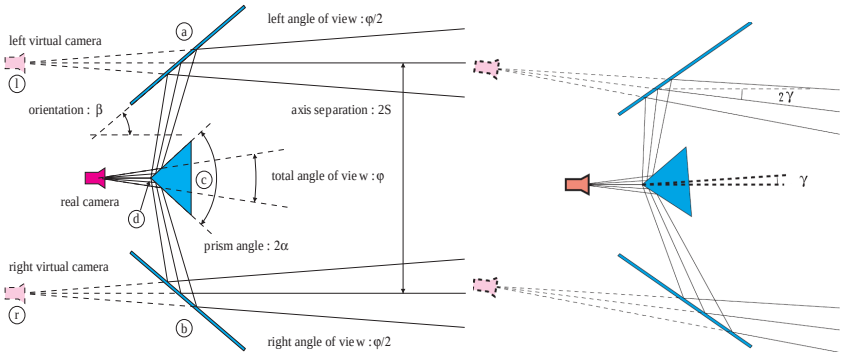
But, all the active vision systems mentioned before are mechanically complex, big and heavy. In this case, inertia does not permit a fast and precise movement of the whole system. In order to overcome this drawback, we have developed an active stereo vision system able to rapidly change its gaze, thanks to a very compact and light motorized part. This article is organized as follows. Section 2 presents an overview of the stereoscopic rig with an analysis of gaze orientation. Section 3 describes a simple strategy for controlling the gaze of an active vision system. In section 4, the applicability of this approach to our active stereoscopic rig is demonstrated with experiments on a real image sequence. Finally, conclusion and outlooks are given in section 5.

2 Active Stereoscopic Rig

2.1 Sensor description

A schematic top view of the stereoscopic sensor is presented in figure 1(a). The sensor is composed of a single camera, of two lateral plane mirrors, marked as ⓐ and ⓑ in figure 1, and of a central prism with two planar reflective surfaces, marked as ⓒ.

This setup is similar to the system described in [12], except that the prism can rotate about the axis defined by the edge at the intersection of the reflective surfaces, marked as ⓓ. The mirrors project the left and right images of the stereo pair onto both halves of the imaging surface of the camera, yielding the equivalent of two virtual cameras with parallel axes. Without orientation of the prism and for a specific pose of the lateral mirrors, it can be shown that



(a) without rotation (b) rotation of angle γ

Fig. 1. Stereoscopic setup without and with rotation of the prism

the optical axis of a virtual camera is parallel to the optical axis of the real camera [13].

The optical axis of the camera intersects the edge of the prism, which is projected as a vertical straight line through the center of the image. This straight line, which does not move when the prism is rotated, splits the single image into two half images that would be separately formed by the virtual cameras marked as $\textcircled{l}$ and $\textcircled{r}$ in figure 1(a). Rotating the prism changes the orientation of both optical axes while keeping them parallel [14]. Figure 1(b) allows one to understand the modifications of the optical properties of the setup when the prism is rotated about its vertical edge. For a rotation γ of the prism the orientation of the field of view is twice the angle γ, but in the opposite direction. In fact this orientation of the stereoscopic setup is similar to what happens when we change the direction of our gaze without rotating the head.

2.2 Active vision servoing setup

Our active vision host system is a personal computer, equipped with a picolo frame grabber. The stepper motor (200 steps per revolution), which is coupled with a reducer, is controlled by a GCD 93-70 Phytron module (*cf.* figure 2). The maximum angular speed is near $70^o/s$ and the accuracy 0.001^o. Compared to human capacities, whose eyes movements angular velocity can reach $900^o/s$ [15], the dynamical performance of our sensor is poor. However, in this particular application to driver assistance, it is able to shift the field of view quickly enough and track vehicles precisely.

3 Gaze Control Strategy

3.1 Minsky's frame system theory

People, in order to identify objects, make an extensive use of their characteristics. Then, they can decide if it is important or not to pay attention to a given object.

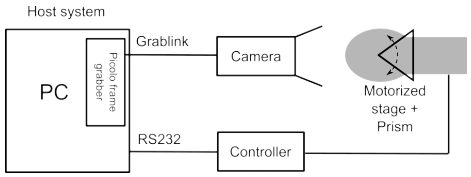


Fig. 2. Active vision servoing set

For many years, researchers have tried to reproduce artificially this cognitive perception behavior. Minsky [16] has recently proposed a representation scheme of human knowledge, which organizes basic knowledge into chunks called frames. Frames are supposed to capture the essence of concepts or typical situations. This includes information about how to use the frame, about expectations, about what to do if expectations are not confirmed and so on. Therefore, this model assumes that, when we meet a new situation, we select the appropriate frame from our memory.

According to this principle, there are two types of frames in the human memory: *general* frames, acquired by learning or by generalization of lived experiments, and *instanced* frames, which are more precise and testify of the memory of known objects or lived situations. Considering these different frames, we can extract information from an observed scene in order to select the elements which are the most important.

For example, for the analysis of a visual scene, frames describe the scene observed from different points of view. In this case, transformations between one frame and another represent the effects of motion. We distinguish two types of frames:

- Object frames that describe the shape of a real object, but also the evolution of this shape when it is observed from various angles.
- Location frames that are aggregations of object frames used to represent the global scene.

Spatial relations between several object frames represent the positions of objects with respect to each other. The temporal relations between object frames aim at characterizing object motion. Location frames describe the scene by the instanced particular frame objects.

Minsky's goal was to supply a theoretical model allowing robots to perceive, make reasonings and react in imperfectly known environments. Actually, he did not propose a computer implementation of his theory. However, the concept of frame that Minsky describes, is very similar to the concept of class and class instances, i.e. objects, that is used in object-oriented programming.

In a computer implementation of Minsky's approach, frames are represented by class instances that are constantly updated in order to represent accurately objects of the scene. Class variables represent object characteristics, such as: type of object, spatial position, speed, date of perception, etc [17]. These data are updated according to various criteria : time since the last perception, number

of times when the object was perceived, object importance for the task to be performed, dynamic properties, etc.

3.2 Our Strategy of Perception

When driving, human strategy of perception depends on the road context. In fact, the situations on a motorway or in a city are different. In a city, we meet several types of users, roads and rights of way. On a motorway, not all users are allowed and the traffic must respect different specific rules. The strategy of perception depends also on the goal selected by the driver : to follow a car or to reach a destination by following successive roads. A relevant perception strategy must take into account the selected driving context.

In this study, we focus on a simple situation of driving on a one way road in presence of other vehicles. So, we need to orient the sensor gaze ahead toward objects detected on the road. After looking ahead a time depending on our velocity, we shift our gaze on the object which seems to be the most relevant. We smooth pursue the latter, either after a time or until an other object becomes more relevant. Then, we shift back to the ahead gaze. Basing on the Minsky's theory explained before, we developed a strategy of perception which is computed and updated for each frame. We call 'Observation Index' (OI) our basic strategy for scene perception. It allows for selecting the object that should be observed among all the objects identified in the scene. If multiple candidates for fixation are present in the list, the one with the maximum OI is selected. Candidate objects are determined using stereo images and several associated features are computed such as position, speed, size.... For each candidate object an OI is computed, based on a combination of these features.

We build a table including OIs to represent the activity of the scene, i.e. its actual content and variations of this content. Finally, the sequence fixation points to be observed is derived from this table. The date and duration of observation of each candidate object, which are also derived from its OI, are stored in the same table. For every image, the table is updated for each candidate object, even if the latter is not in the field of view of the system.

4 Experiment

To test the validity of the described approach, experiments were carried out in which the prism orientation is controlled to change the gaze. An example of stereoscopic images (two half images) is shown in figure 3.

Two objects, i.e. candidates for gaze fixation, are moving in the scene. Object 1 is twice faster than object 2. At the beginning of the sequence, object 1 is close to the sensor, then the two objects are at the same distance, and finally object 1 gets further. The OI for this experiment is the inverse of the distance between the object and the sensor. Therefore, the observed object is the closer one. For fixation we bring the object in the middle of the dominant half image, here the right one. Gaze is controlled with three steps: 1) template matching

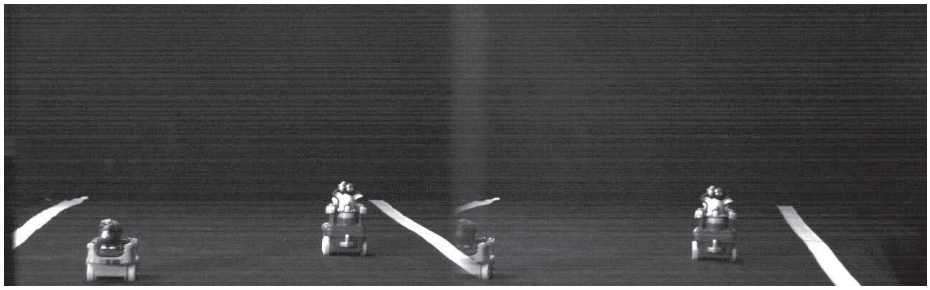


Fig. 3. Stereoscopic images

yields the center of gravity of the detected objects, as well as speed and 3D measurements; 2) retinal errors, i.e. distance between image center and object center, are computed; 3) the error signal is used by a PID to control the orientation of the prism.

In the sequence of images of figure 4 either object 1 (on the left) or object 2 (on the right) is in the middle of the image, thanks to the orientation of the field of view. Table 1 gives the object distances and the computed prism angle. In image 24 of the sequence, object 2 becomes closer than object 1. The active vision shifts the gaze to bring object 2 in the middle of the image.

Table 1. Table

Image number	1	11	13	18	23	24	29	34	39
Object number	2	1-2	1	1-2	1	1	1-2	1-2	1-2
Prism rotation (^{o})	0	0	+4.1	0	+3.6	0	-0.2	0	-0.3
$Z_1(m)$	-	2.8	3	3.2	3.4	3.4	3.6	3.8	4
$Z_2(m)$	2.9	3.1	-	3.3	-	-	3.5	3.6	3.7

5 Conclusion

This paper presents a basic approach to gaze control with a particular active vision system. In the first part of the paper we have described our active vision system. It is built to be very fast and precise because of the low inertia of the mobile prism. In the second part, we have established a basic strategy for the perception of a real scene. Experiments seem to validate our approach. In the future, we will characterize more precisely the process in order to derive a more robust control law for increasing the gaze control speed.

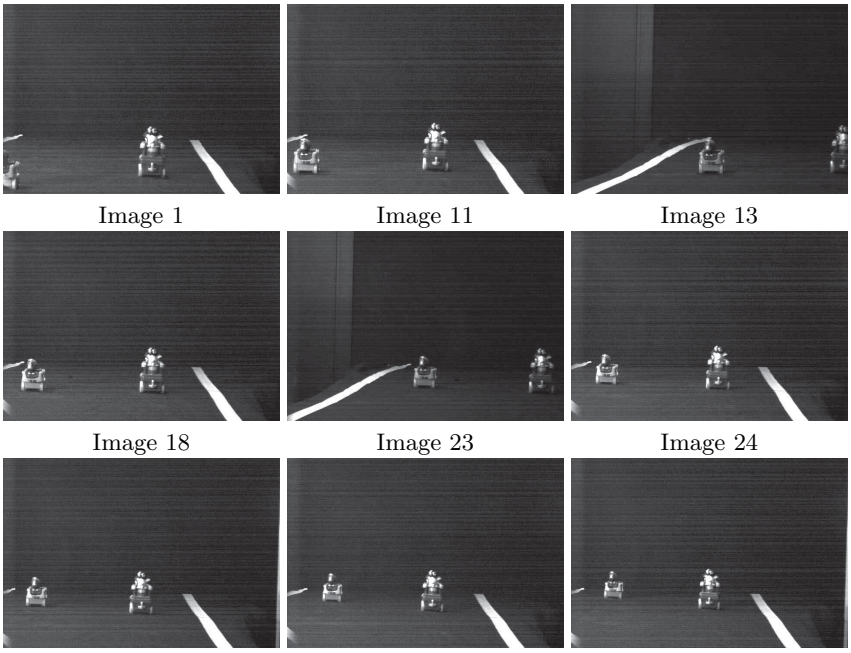


Fig. 4. Sequence of Images

Acknowledgements

The authors want to acknowledge the financial support of the French Nord-Pas de Calais Regional Council and EU FEDER under grant OBJ2-2005/3-4.1-253-7820.

References

1. Marr, D.: Vision : A Computational Investigation Into the Human Representation and Processing of Visual Information. W.H. Freeman and Company (1982)
2. Aloimonos, J.Y., Weiss, I., Bandopadhay, A.: Active vision. International Journal of Computer Vision **1** (1988) 333–356
3. Bajcsy, R.: Active perception. Proceedings of the IEEE, Special issue on Computer Vision **76** (1988) 996–1005
4. Ballard, D.H.: Animate vision. Artificial Intelligence **48** (1991) 57–86
5. Samson, E., Laurendeau, D., Parizeau, M., Comtois, S., Allan, J.F., Gosselin, C.: The agile stereo pair for active vision. Machine Vision and Applications **17** (2006) 32–50
6. Kuhnlenz, K., Bachmayer, M., Buss, M.: A multi-focal high-performance vision system. In: Proceedings of the 2006 IEEE International Conference on Robotics and Automation, ICRA 2006, May 15-19, Orlando, Florida, USA. (2006) 150–155
7. Kozima, H., Yano, H.: A robot that learns to communicate with human caregivers. In: International Workshop on Epigenetic Robotics, Lund, Sweden (2001)

8. Aryananda, L., Weber, J.: MERTZ: a quest for a robust and scalable active vision humanoid head robot. In: 4th IEEE/RAS International Conference on Humanoid Robots. (2004)
9. Clady, X., Collange, F., Jurie, F., Martinet, P.: Object tracking with a pan tilt zoom camera : application to car driving assistance. In: Proceedings of the IEEE International Conference on Robotics and Automation, ICRA'01, Seoul, Korea, May 23-25th. Volume 2. (2001) 1653–1658
10. Pellkofer, M., Dickmanns, E.D.: EMS-vision: gaze control in autonomous vehicles, Proceedings of the IEEE Intelligent Vehicles Symposium (IV'2000 (2000) 296–301
11. Pellkofer, M., Lutzeler, M., Dickmanns, E.D.: Vertebrate-type perception and gaze control for road vehicles. In: Springer Tracts in Advanced Robotics. Volume 6. (2003) 271–288
12. Mathieu, H., Devernay, F.: Système de miroirs pour la stéréoscopie. Rapport technique 172, INRIA, Sophia Antipolis (1995) Projet Robotvis.
13. Duvieubourg, L., Cabestaing, F., Ambellouis, S., Bonnet, P.: Long distance vision sensor for driver assistance. In: Proceedings of 6th IFAC Symposium on Intelligent Autonomous Vehicles (IAV'07), Toulouse, France (2007)
14. Duvieubourg, L., Ambellouis, S., Cabestaing, F.: Single-camera stereovision setup with orientable optical axes. Computational Imaging and Vision (2005) 173–178
15. Kandel, E.R., Schwartz, J.H., Jessel, T.M.: Principles of neural science. Elsevier, New York (1991)
16. Minsky, M.: Computer Science and the Representation of Knowledge. MIT Press (1979)
17. Herviou, D.: La perception visuelle des entités autonomes en réalité virtuelle : application à la simulation de trafic routier. PhD thesis, Université de Bretagne Occidentale (2006)