Computer Vision and Image Understanding **83,** 97–117 (2001)
doi:10.1006/cviu.2001.0924, available online at http://www.idealibrary.com on IDEAL®

Disparity Estimation on Log-Polar Images and Vergence Control

R. Manzotti, A. Gasteratos, G. Metta, and G. Sandini

Laboratory for Integrated Advanced Robotics (LIRA-Lab), Department of Communications Computer and System Science, University of Genoa, Viale Causa 13, Genoa I-16145, Italy

Received September 27, 1999; accepted June 14, 2001

An important issue in the realization of an autonomous robot with stereoscopic vision is the control of vergence. Together with version, it determines uniquely the position of the fixation point in space. Vergence control is directly related to both depth perception and *binocular fusion*. Previous works in this field employed either a measure of correlation of stereo images or some kind of disparity-related estimate. In this paper, we present a new method of extracting a global disparity measure for vergence control, which does not require a priori segmentation of the object of interest. Our method uses images acquired by retina-like sensors and, therefore, the computation is performed in the log-polar plane. The technique we present here is: (i) global, in the sense that it is an integral measure over the whole image, (ii) computationally inexpensive, considering that the goal was to use it in the robot control loop rather than to accurately measure some 3D world features. Moreover, the proposed technique is robust and independent of the average illumination as well as of other features of the target such as size, shape, and direction of motion. It provides a precise and linear estimate of the vergence error, which is the only requirement from the control point of view. Several experimental results on a real robotic setup demonstrate the effectiveness of the proposed technique. © 2001 Academic Press

Key Words: active vision; log-polar; disparity; vergence.

1. INTRODUCTION

In building a robotic stereo head, an extremely important degree of freedom is represented by vergence. Given a stereoscopic vision system (Fig. 1), the vergence angle, together with version and tilt angles, describes uniquely the fixation point in space. The problem we are dealing with here is controlling the vergence angle only with the assumption that other subsystems maintain the object of interest close to the image center. For instance, we could imagine a tracking module, which deals with the problem of following the target in space

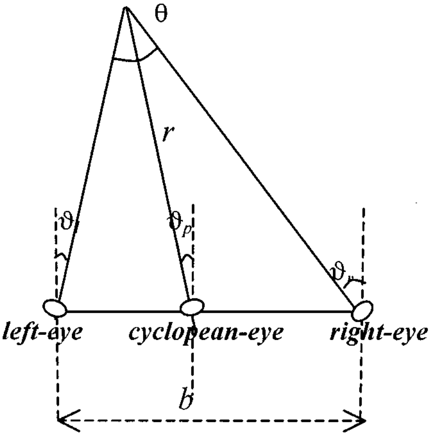

FIG. 1. Stereoscopic vision system. The version (ϑ_p) and vergence (ϑ_v) angles are shown.

(by controlling version and tilt) or a binocular saccade-like control to quickly foveate a possibly interesting target [1].

If vergence could be controlled effectively, several advantages would arise in subsequent image processing. These include easier fusion of stereo images in one “cyclopean image” and easier figure–ground segmentation (e.g., by means of zero disparity filters [2]). Of course, if a tracked object is stable in the retinas, further image processing is facilitated. Indeed, it is not just a coincidence that most of the biological stereoscopic vision systems in the higher branches of the evolutionary tree possess a developed and specialized vergence control system [3]. If we consider biological systems, several kinds of disparity exist (horizontal, vertical, and rotational) in order to implement horizontal and cyclo-vergence [4]. In this paper only the horizontal component is considered, which is the most relevant for the control of the vergence angle, in the hypothesis that the other d.o.f. are fixed. Maintaining a correct vergence angle should not be seen as a goal in itself but as a way of improving the performance and robustness of successive visual computation. Particularly relevant in this respect is the intrinsic limit introduced by a correct binocular fusion in the computation of binocular disparity and 3D feature extraction. Vergence angle, moreover, provides a measure of absolute distance, even if limited to a point in space, as well as a reference point in the environment [5]. Therefore, a general principle is that vergence control should be implemented, keeping in mind its further use by the whole system [6].

In the implementation presented here we consider “only” disparity measure, which, even if not the only source of information useful for vergence control [7], is possibly the most direct and relevant. Disparity estimation has been performed according to several techniques. These are based either on correlation [8–10], matching [11, 12], phase difference [13, 14], or Bayesian methods [15, 16]. Despite this big variety of disparity estimation methods, examples employing log-polar images are quite occasional [17–19]. The reason is the complex geometrical layout of log-polar images, which apparently is not well suited for disparity computation. Log-polar images, however, are ideal for vergence control tasks. They provide high resolution in the fovea, where the target should be located, and a wide field of view at the same time [20]. These features, along with a computationally simple mapping technique used for disparity estimation, allows real-time performance to be achieved with high accuracy.

The experimental results reported in order to show the feasibility of the approach have been carried out with one main goal in mind: to demonstrate that horizontal disparity alone allows an efficient and robust vergence control, assuming that vertical, torsional, and focus degrees of freedom (df) are fixed. The paper is organized as follows. The essentials of log-polar images are given in Section 2. Several techniques, which have been applied for vergence control, are discussed in Section 3. The proposed disparity estimate technique is extensively described in Section 4. In Section 5, the application of the proposed technique for vergence control is presented. Section 6 contains the experimental results and, finally, Section 7 contains some concluding remarks.

2. LOG-POLAR IMAGES

Studies on the primate visual pathways from the retina to the visual cortex have shown that the geometrical layout follows an almost regular topographic arrangement [21–23]. These results can be summarized as follows:

- The distribution of the photoreceptors in the retina is not uniform. They lay more densely in the central region called fovea, while they are sparser in the periphery. Consequently, the resolution also decreases, moving away from the fovea toward the periphery. It has a radial symmetry, which can be approximated by a polar distribution.
- The projection of the photoreceptors array to the primary visual cortex can be well approximated by a logarithmic-polar (log-polar) distribution mapped onto a rectangular-like surface (the cortex). Here the representation of the fovea is quite expanded; i.e., more neurons are devoted to it, and the periphery is represented using a coarser resolution.

From the mathematical point of view the log-polar mapping can be expressed as a transformation between the polar plane (ρ, θ) (retinal plane), the log-polar plane (ξ, η) (cortical plane), and the Cartesian plane (x, y) (image plane),

$$\begin{cases} \eta = q \cdot \theta \\ \xi = \ln_a \frac{\rho}{\rho_0} \end{cases} \tag{1.1}$$

where ρ_0 is the radius of the innermost circle, $1/q$ is the minimum angular resolution of the log-polar layout, and (ρ, θ) are the polar coordinates. These are related to the conventional Cartesian reference system by:

$$\begin{cases} x = \rho \cos \theta \\ y = \rho \sin \theta. \end{cases} \tag{1.2}$$

Figure 2 illustrates the log-polar layout as derived by Eqs. (1.1) and (1.2). In particular, the grid on the left represents a standard Cartesian image mapped according to Eq. (1.1). The plot on the right shows the corresponding cortical image.

3. VERGENCE CONTROL

Vergence control issues have been addressed by means of several different techniques. It is worth noting that all of them are somehow related to the estimation of 3D features. These techniques can be classified as follows:

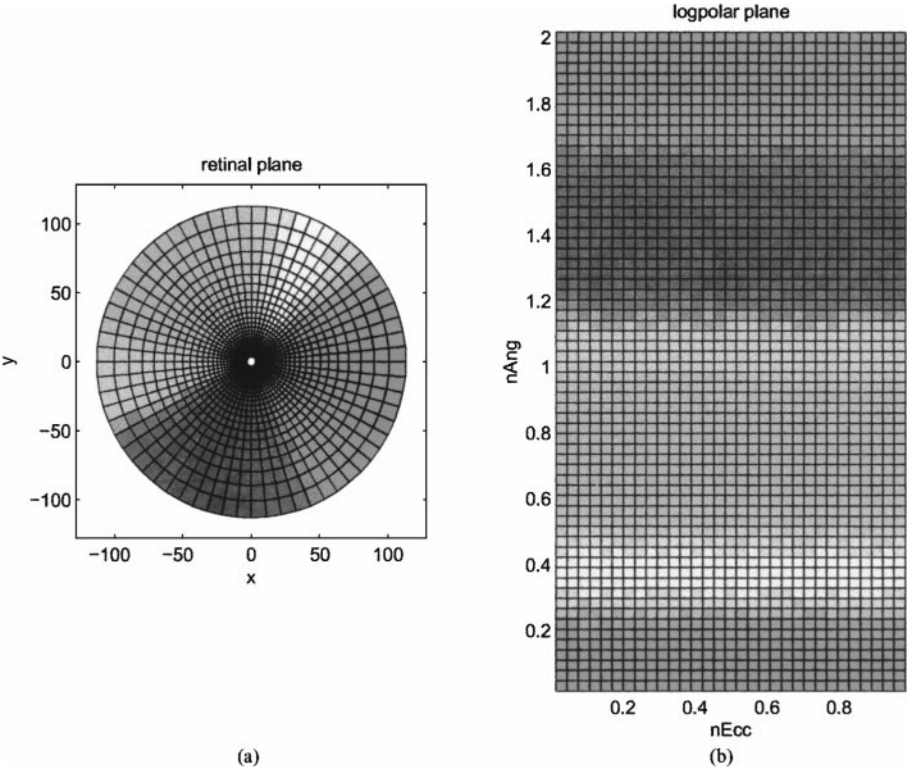

FIG. 2. The log-polar transformation. (a) The log-polar layout mapped into the Cartesian space and (b) the corresponding cortical image.

(1) segmentation techniques [11, 12, 19, 24, 25];
(2) fusion index [17, 26];
(3) direct disparity estimation [9, 27].

3.1. Segmentation Techniques

This group requires some sort of heuristics to identify the object of interest (segmentation). The main problem, in this case, is their lack of flexibility. We would like to stress the fact that the segmentation of the object from its background is not an easy task itself. Many of the proposed systems do not use any direct control of the vergence angle, but they rather control each eye separately. These approaches may, for instance, fail in the presence of a false matching; i.e., the robot might try to follow two different targets. Moreover, in biological systems, vergence control is a relatively low-level functionality, which does not require an actual segmentation or recognition of the target object. On the contrary, these techniques act at a higher level of abstraction, i.e., the object needs to be segmented and/or recognized.

3.2. Fusion Index

This second group exploits the fact that if an object is correctly verged, the stereo images should be very similar, at least around the fixation point. Of course, this does not hold exactly as the two images are never the same. However, under standard conditions, such as typical fixation distance, optical parameters, and kinematics of robot heads, the difference between the images is rather small. In that case, the images are said to be binocularly fused. The goal of a vergence control system is to minimize this difference. The global index of binocular fusion is an example of such a difference estimate. It can be computed using the normalized correlation technique [28],

$$C(I_l, I_r) = 1 - \frac{\sum_{\eta,\xi}(I_r(\eta,\xi) - \mu_r)\cdot(I_l(\eta,\xi) - \mu_l)}{\sqrt{\sum_{\eta,\xi}(I_r(\eta,\xi) - \mu_r)^2 \cdot \sum_{\eta,\xi}(I_l(\eta,\xi) - \mu_l)^2}}, \tag{1.3}$$

where I_r and I_l are the right and left images, respectively, and μ_r and μ_l represent their mean values, respectively. $C(t)$ is almost invariant to changes of illumination. It is normalized in the range [0, 1]. The normalized correlation measure can be employed in a standard proportional control law,

$$\dot{\theta} = -K \cdot \dot{C}, \tag{1.4}$$

where K is a constant gain and $\dot{\theta}$ and $\dot{C}$ are the first derivatives of the vergence angle and the fusion index respectively [17, 26].

However, this approach has several drawbacks:

(1) $C(t)$ is not a linear estimation of the angular error.
(2) $C(t)$ is constant, if the eyes are still and the object is not moving ($\dot{C} = 0$).
(3) $C(t)$ has minima whose values are variable with image characteristics in a nonlinear and unpredictable fashion.

For these reasons, though feasible, the use of $C(t)$ in vergence control is limited.

3.3. Direct Disparity Estimation

The last group concerns the use of direct estimates of the binocular disparity. It is, in our view, the most promising one, because it uses disparity directly. The latter can be easily related to the vergence control error. In fact, it can be shown that binocular disparity is related to depth, which is in turn related to the vergence angle. Consider again the situation depicted in Fig. 1; using only the sine law, it is easy to derive

$$r = b\frac{\cos(\vartheta_r)\cos(\vartheta_l)}{\cos(\vartheta_v)\cos(\vartheta_p)}, \tag{1.5}$$

where r is the distance of the fixation point from the baseline, b the baseline length, ϑ_r, ϑ_l the eye angles, $\theta = \vartheta_l - \vartheta_r$ represents the vergence angle, and $\vartheta_p = (\vartheta_l - \vartheta_r)/2$ the version (or gaze) angle. In principle, Eq. (1.5) can be used to find out the vergence angle required to move the fixation point from one location to another.

On the other hand, disparity is related to depth z (relative to the fixation point reference frame) by the approximate equation,

$$z \cong \frac{K(\theta, b)}{\alpha\tan(\theta)}(x_r - x_l), \tag{1.6}$$

where $K = b/\sin(\theta)$, b is the baseline length, α is the focal length of the cameras, and $x_r - x_l$ the binocular disparity [29]. Note that Eq. (1.6) is valid only in a neighborhood of the fixation point. Using Eqs. (1.5) and (1.6) together, it is thus possible to link binocular disparity to vergence error directly. Moreover, we have actually converted the problem of estimating the vergence error in that of estimating disparity. (That is, starting from disparity, it is possible to compute depth by using Eq. (1.6), and from depth, which in this case represents the required motion, recover the vergence angle movement by employing Eq. (1.5). The goal of the controller is indeed that of zeroing depth.)

Unfortunately, the direct disparity estimation approach has been undermined by some practical difficulties. In fact, in order to reduce the computational burden, an attention region should be selected in advance. This means that the object has to be segmented from the background; this has been proved to be a hard task in itself. On the other hand, this is not really an issue if we utilize log-polar images and we constrain ourselves to a global estimate of the object disparity. Of course, there is nothing magic about the use of retina-like sensors. However, assuming that the object of interest is close to the center of the image, its relevance (e.g., number of "pixels") becomes higher than that of more peripheral objects (its projection in the log-polar plane is effectively magnified) [30].

4. DISPARITY

In order to compute disparity it is necessary to solve a correspondence problem. That is, we have to establish which pixels on the left and right image planes map to the same point in space. Formally, considering a standard Cartesian images case, this can be written as

$$d(x, y) = \arg\max_{d}\{\Delta(I_l(x, y), I_r(x + d, y))\}, \tag{1.7}$$

where, Δ is a similarity measure for each possible shift $d \in [-d_{\max}, d_{\max}]$. The similarity measure can be either a sum of squared difference (SSD) or other criteria such as the normalized correlation.

Assuming symmetric vergence and a simple pinhole camera model, the image and motor coordinates systems are related by the equation,

$$d = 2f \cdot \sin\left(\frac{\Delta\vartheta_v}{2}\right), \tag{1.8}$$

where $\Delta\theta$ is the difference between the actual and the correct vergence angle (the vergence angle with null disparity), f the camera focal length, and d the measured disparity. It is worth noting that Eq. (1.8) is monotonic in the required domain (typically $\Delta\theta \in [-\pi/2, \pi/2]$). Roughly speaking because the measure is well formed in Lyapunov sense ([31]) the closed loop system will be stable even if a simple PD controller is used. Furthermore, the shift of image pixels can be represented by a disparity operator $disp_{cart} : \Re^3 \to \Re^2$, defined as

$$disp_{cart}(x, y, d) \cong (x + d, y), \tag{1.9}$$

where d represents the disparity and (x, y) a point in the image plane.

By applying Eq. (1.9) to an image point (x, y), we can generate the corresponding matching point at disparity d. The rationale of defining $disp_{cart}$ will be clear when we will deal with nonuniform mapping such as the log-polar one. Combining Eqs. (1.7) and (1.9) yields

$$d(x, y) = \arg\max_{d}\{\Delta(I_l(x, y), I_r(disp_{cart}(x, y, d)))\}, \tag{1.10}$$

In a general case, in order to apply Eq. (1.10) and recover the disparity, we need to provide

(i) the disparity function *disp*, which embeds the description of the image geometry;
(ii) the similarity function Δ, which defines a distance measure between pixel blocks.

Considering the log-polar case, we just need to replace the *disp* operator with a suitable one for that particular topology. However, the mapping is not simple using log-polar images. A simple horizontal shift in Cartesian coordinates is mapped to a complex curve in log-polar coordinates. Given a (ξ, η) pair the disparity operator $disp_{\log}\colon \Re^3 \rightarrow \Re^2$ is defined as

$$disp_{\log}(\xi, \eta, d) = (\xi_n, \eta_n). \tag{1.11}$$

The transformation is now given by

$$\begin{pmatrix} \xi_n \\ \eta_n \end{pmatrix} \cong \begin{pmatrix} \frac{1}{q} \cdot \arctan\left[\frac{\rho_0 \cdot a^{\xi} \cdot \cos(\eta/q)}{\rho_0 \cdot a^{\xi} \cdot \sin(\eta/q) + d}\right] \\ \sqrt{(\rho_0 \cdot a^{\xi} \cdot \sin(\eta/q) + d)^2 + \rho_0^2 \cdot a^{2\xi} \cdot \cos^2(\eta/q)} \end{pmatrix}. \tag{1.12}$$

By applying the *disp* operator we can roughly simulate a horizontal shift of (I_l, I_r) without actually moving the cameras. In practice this is only approximately true because of two main reasons:

(1) Space-variant images introduce a corresponding space variant distortion. In fact, we cannot recover the missing information belonging to the periphery, where the resolution is coarser.

(2) A real camera motion also distorts the image points along the vertical axis. However, the *disp* operator, as it has been defined in Eq. (1.11), does not take this into account.

Concerning the space-variant resolution we are limited by the fact that the information loss on the periphery cannot be avoided. With regard to the contribution of the vertical disparity, it is a rather small effect, and we can neglect it as a first approximation. Figure 3 shows how a regular grid in the left image would be mapped in the right one for increasing values of disparity (in the range zero to nine pixels). These are exactly the graphical representations of Eqs. (1.11) and (1.12).

Furthermore, as our objective was to measure a global disparity index, Eq. (1.10) can be further simplified by including in the aforementioned “neighborhood” of the current pixel all the image pixels. In this case the disparity index d is no longer dependent on the position of a single pixel. A final comment regards the computational load associated with Eq. (1.12). We can note that the equation depends only on the log-polar geometry. In fact it is not dependent on the actual images. Therefore, all calculations can be performed in advance for all possible disparities and stored into a fixed connection map (CM). A CM is basically a look-up table (LUT), which implements a mapping according to Eq. (1.12). From a more general point of view, we can see the CM as a network, where the CM values represent network nodes, and the connections the log-polar geometry itself. This suits very well our conceptual biological bias. It is easy to imagine how these correspondence maps might be implemented in parallel using several layers of neurons.

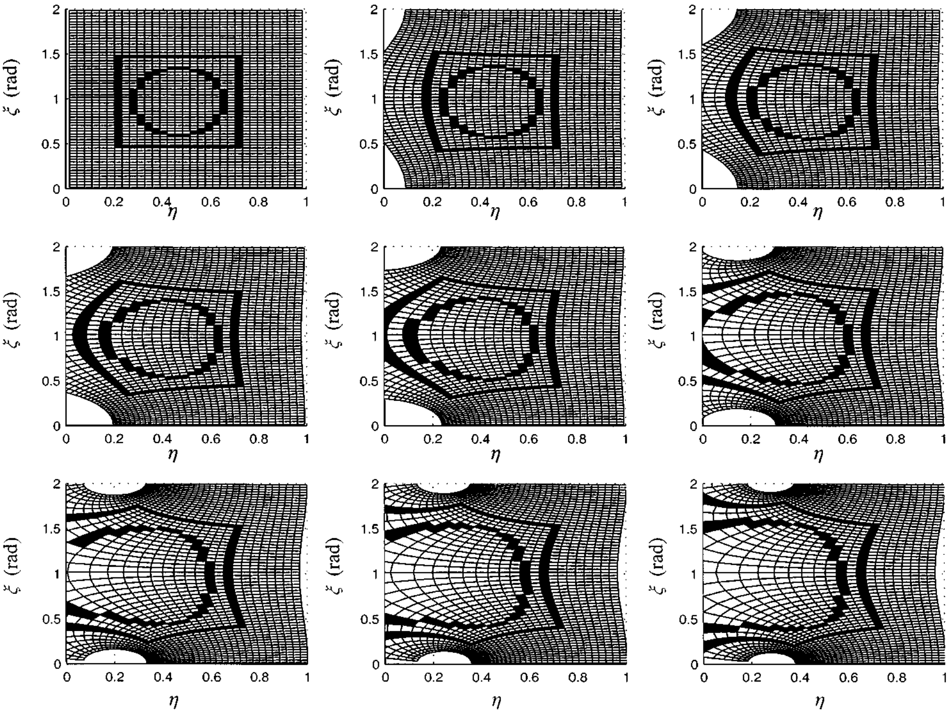

FIG. 3. The geometrical transformation of the cortical mesh. The disparity values increase from zero and nine pixels (from the left to the right and form the top to the bottom, respectively). In each graph the horizontal axis represents the eccentricity, while the vertical axis the angular position. The small circle and the box show the effects of the transformation.

4.1. Disparity Computation

As is stated above, in order to compute the disparity, we need to store the geometrical transforms. The corresponding set of CMs can be defined as

$$CM\_SET = \{CM_d, d = 0 \ldots N\} \tag{1.13}$$

$$CM_d = \{(\xi_n, \eta_n)_{\xi,\eta} : \xi = 0 \ldots \xi_{\max}, \eta = 0 \ldots \eta_{\max}\}, \tag{1.14}$$

where $\xi_{\max}\ \eta_{\max}$ is the size of the log-polar images in cortical coordinates, N the total number of CMs, and

$$(\xi_n, \eta_n)_{\xi,\eta,d} = disp_{\log}\left(\xi, \eta, \left(\frac{d - \frac{N}{2}}{N}\right) \cdot \xi_{\max}/2\right). \tag{1.15}$$

Using this notation, an image transform can be rewritten as

$$I_r = disp_{\log}(I_l, d) \tag{1.16}$$

Implicit in this notation is the fact that we are actually mapping all the pixels of the left image I_l into those of the right one I_r. That is,

$$I_r(\xi, \eta) = I_l((\xi_n, \eta_n)_{\xi,\eta,d}) \tag{1.17}$$

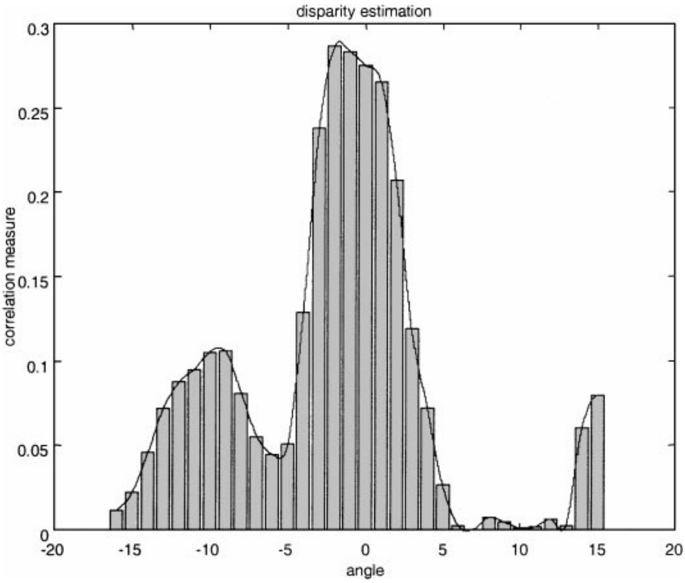

FIG. 4. A measured disparity function obtained with an object almost in vergence. It is possible to notice the maximum of the function near the center of the curve.

It is worth noting that the set of CMs has a finite size. Consequently, in order to evaluate the arg max function over d we need to evaluate only a finite number of possible disparities. In other words, at each time instant t the result is the discrete function (which is function of the disparity d),

$$\chi(I_l, I_r, d) = C(I_r, disp_{\log}(I_l, d)), \tag{1.18}$$

where $C(I_r, I_l)$ is the normalized correlation function of Eq. (1.3); i.e., we employed the normalized correlation as similarity criterion Δ.

Figure 4 shows a real-case plot obtained using the procedure described above. Without loosing the generality we can define the disparity function, for a given stereo pair (I_l, I_r), as

$$\chi(d) = \chi(I_l, I_r, d). \tag{1.19}$$

The global disparity index is simply

$$d_t = \arg\max_x(\chi(x)). \tag{1.20}$$

It is worth stating that:

(1) Given the fact that the disparity function $\chi(d)$ is an integral correlation measure over the whole image it is extremely robust and reliable.

(2) $\chi(d)$ represents the global image disparity because its values are the result of the correlation of the entire image with its corresponding image, shifted by the geometrical transformation due to d.

(3) The significant element in $\chi(d)$ is its maximum, which corresponds to the most common disparity value among the image points.

(4) Most of the noise is rejected, simply because it is located out of the global maximum.

The robustness of the proposed method is a consequence of the fact that, by choosing the disparity value corresponding to the maximum correlation, the contribution of the not relevant part of the image is implicitly rejected. Therefore, even if the side lobes are large in comparison to the maximum of the correlation function, they do not modify the value of $\chi(d)$.

As a matter of fact, a precise measurement of disparity requires a great resolution. In fact, even a small disparity (in the subpixel range) might correspond to a large variation in the object distance (more than 10 cm). To prevent such a loss of accuracy it is possible to increase the image resolution. However, this is, in general, a resource consuming approach. Consequently, we devised a few techniques to increase the estimated accuracy without varying the image resolution itself. They are presented in detail in the next paragraphs.

4.1.1. Disparity nonuniform sampling. One of the main drawbacks of the previously described correlation technique is the lacking of accuracy. The fact that we have used a set of precomputed CMs restricts disparity to one of these CMs. Therefore, the accuracy is limited by the disparity difference, which corresponds to the minimum distance between CMs. In other words, the maximum accuracy is

$$d^1_{\min} = \frac{\xi_{\max}}{N}, \tag{1.21}$$

where N is the number of CMs, and $\xi_{\max}$ the log-polar image radius.

A possible solution would be to increase the numbers of CMs (increasing N) but this would increase the computational cost accordingly. A second solution would be to reduce the radius of the log-polar image $\xi_{\max}$, reducing, consequently, the field of view. An alternative approach exploits the space variant techniques used by using the same number of CMs distributed in a nonlinear fashion. In this way resolution for small disparities is improved without increasing the computational load. This solution does not affect the overall system performance, since for high-disparity values there is no need for precise disparity estimation because the higher the value of disparity the higher is the distance of the world point from fixation. On the other hand, when disparity is small, high precision is required, since the control should minimize even small errors.

We can therefore replace Eq. (1.13) with

$$CM\_SET_{\mathrm{var}\,iant} = \left\{CM_d^{\mathrm{var}\,iant}, d = 0 \ldots N\right\} \tag{1.22}$$

$$CM_d^{\mathrm{var}\,iant} = \{(\xi_n, \eta_n)_{\xi,\eta} \colon \xi = 0 \ldots \xi_{\max}, \eta = 0 \ldots \eta_{\max}\}, \tag{1.23}$$

where $\xi_{\max}, \eta_{\max}$ is the log-polar image size in cortical coordinates, N is the total number of CMs, and

$$(\xi_n, \eta_n)_{\xi,\eta,d} = disp_{\log}(\xi, \eta, \mathrm{sign}(i - N/2)(\mathrm{abs}(\arctan((1 - N/2)/N)))^{\lambda} \cdot \xi_{\max}/2), \tag{1.24}$$

where λ is positive number. λ acts as a steep enhancer in the nonlinear transformation and it is typically $\lambda = 3$.

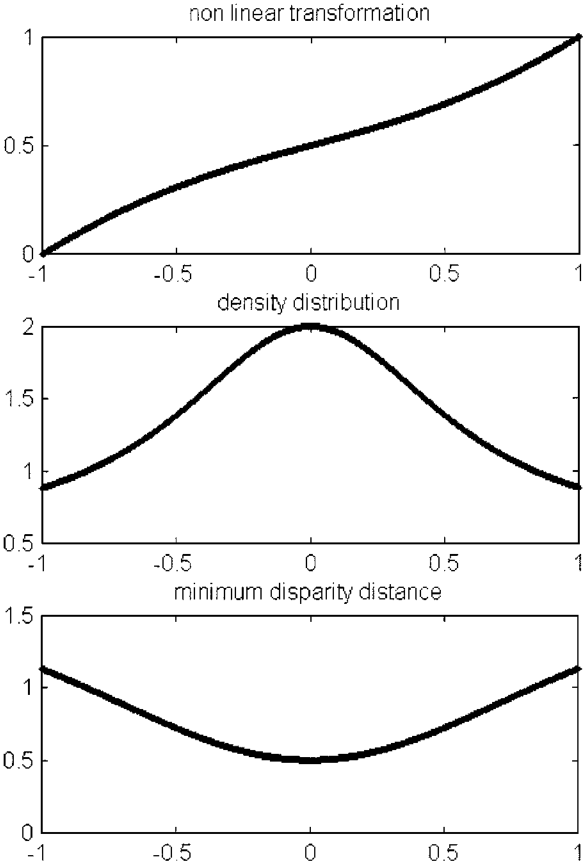

FIG. 5. The nonlinear distribution of CMs. (Top) The transformation function. (Middle) The density of the distribution. (Bottom) The maximum detectable precision of disparity.

Figure 5 is indeed the plot of equation 1.24 for $\lambda = 3$. The minimum resolution is now equal to

$$d^1_{\min} = \frac{\xi_{\max}}{2} \cdot \arctan(1/N) < d^1_{\min}. \tag{1.25}$$

By observing Eq. (1.24) we might think that the accuracy can be increased just by modifying the nonlinear function, i.e., increasing λ. However, by reducing the disparity step, we reduce also the offset applied to the original log-polar mesh. For small disparity values, the offset would be less than one foveal pixel. This means that the log-polar mesh would remain roughly the same. In other words, if $d_{\min}$ is too small,

$$I_r = disp_{\log}(I_l, d_{\min}) = I_r. \tag{1.26}$$

Hence, as it is intuitively obvious, it is not possible to compute directly a value of disparity smaller than one foveal pixel.

4.1.2. Quadratic interpolation. It is clear that $\chi(d)$ is actually a sampled version of an underlying continuous function (i.e., we might imagine to define an infinite disparity set

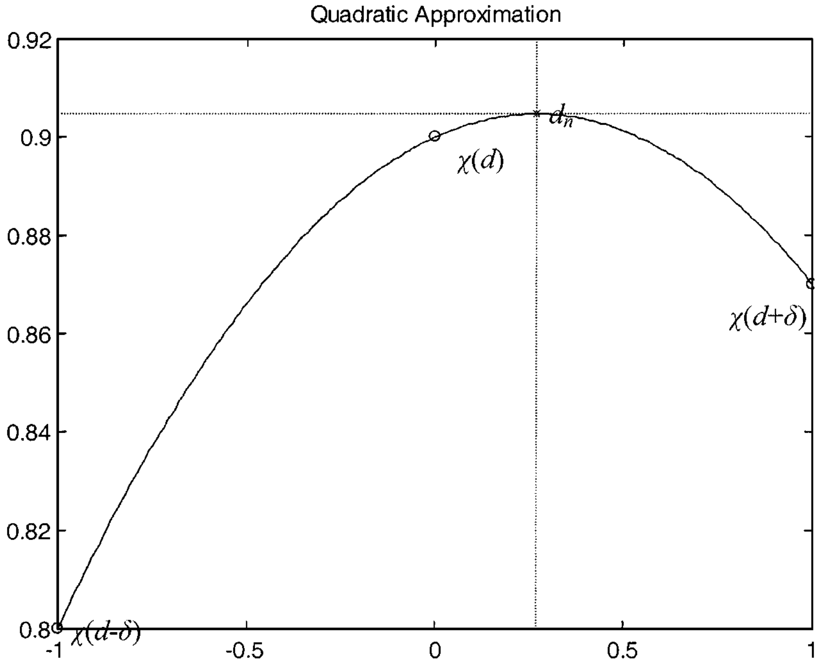

FIG. 6. An example of interpolation by using the quadratic approach (Eq. (1.10)). The accuracy improved to about 0.3 pixels.

instead of the finite one), which has the same formulation as $\chi(d)$ but it has no constraints on its argument: d $\in \Re$. It is thus possible to apply an interpolation technique on the samples $\chi(d)$. A straightforward solution is to apply a quadratic interpolation technique as sketched in Fig. 6. In equation form,

$$d_n = \frac{\chi(d-\delta) - \chi(d+\delta)}{2 \cdot (\chi(d-\delta) - \chi(d) + \chi(d+\delta))}, \tag{1.27}$$

where d is the disparity value computed without interpolation and δ is the minimum disparity value ($\delta = d_{\min}$) or a suitable multiple.

Given the heuristic of this last technique there are no particular reasons to prefer an interpolation technique from another. Note also that the interpolation is not performed on the final data (the estimated disparity) but on the raw data of the disparity function. Of course, this means that it is not just a mathematical smoothing but it represents a real improvement over the noninterpolated counterpart.

5. VERGENCE CONTROL FROM DISPARITY COMPUTATION

Vergence control has been implemented by applying the following control law:

$$\dot{\theta} = -K \cdot d_t, \tag{1.28}$$

where K is a constant gain, $\dot{\theta}$ is the first derivative of the vergence angle, and d_t is the disparity index (see Eq. (1.20)). The advantages of this approach can be summarized as follows:

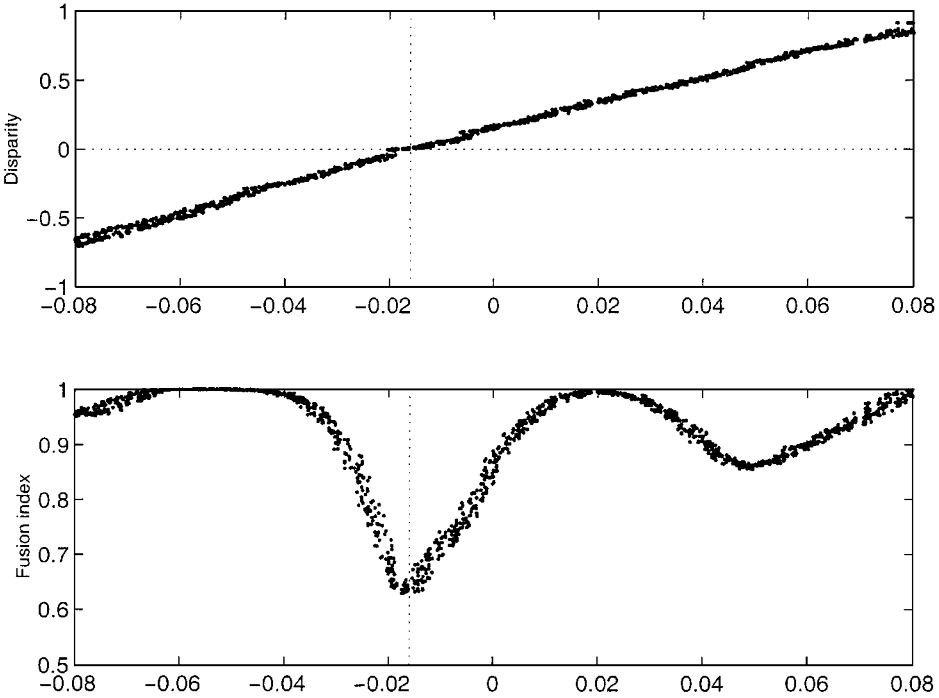

FIG. 7. The disparity function. These two plots were obtained by changing the vergence angle at a constant speed (total span was 0.16 radians). For each angular position we measured the disparity (top) and the fusion index *C(t)* (bottom), respectively.

(1) d_t is a linear estimation of the angular error, as shown in Fig. 7.

(2) d_t provides the same information irrespective of the state of the system, because it does not matter if the cameras are moving or if they are still.

(3) d_t is robust to noise, environmental modifications, changes in lighting conditions, and object properties.

A few words are needed to explain the previous points. The properties of d_t are graphically illustrated in Figs. 7 and 10, in which it is possible to see that d_t is linear in a range of approximately 0.2 radians. This range is compatible with the goal of keeping the right angle of vergence, a task that is usually achieved for small values of disparity. Regarding the second point it is important to stress that one of the major drawbacks of other implementations of vergence control was the inability of detecting the correct angle without moving the cameras (e.g., [32]). In other words, for each pair of images we have a meaningful value for the controller without having to acquire a new pair in order to compute a gradient. Last, the robustness of d_t is partially derived from the robustness of the correlation function $C(\,)$ and partially from the property of Eqs. (1.18), (1.19), and (1.20).

6. CARTESIAN IMAGES VERSUS LOG-POLAR IMAGES

The main motives in using log-polar images derive from their computational advantages, their geometrical properties, their wide field of view, their high resolution in the fovea, and their implicit selection of a target in the central part of the image. All these properties are exploited in this implementation of disparity estimation. Of course, a "conventional" Cartesian system may present each of these properties but not all of them simultaneously. For example, it is easy to build a Cartesian system with the same resolution and field of view as a log-polar one but at a much greater computational expense. Similarly, it is possible to build a system with the same number of pixel but with a limited field of view or, alternatively, a decreased resolution.

With respect to the disparity estimation and in order to make a comparison with a Cartesian system we assume the following: Given a log-polar image of size $\xi_{\max} \times \eta_{\max}$, there is no Cartesian equivalent image. Two of the following three parameters should be fixed: the pixel size, the number of the pixels, and amplitude of the field of view. Therefore, three different comparative approaches can be constructed and studied:

(1) a Cartesian image having the same number of pixels and the same field of view with the log-polar one;

(2) a Cartesian image with the same number of pixels corresponding to a square area smaller than the whole image but with the same number of pixels with the log-polar one;

(3) a Cartesian image having the same amplitude of the field of view and the same resolution of the fovea, but with a much larger number of pixels.

Let us examine, one by one, these cases. It is assumed that $\eta_{\max} = 2\xi_{\max}$, as it is in the usual case (e.g., [20, 26, 32]), and that the Cartesian equivalent image is square. Besides, it is reasonable to assume that there would be no oversampling of pixels in the fovea and, thus, the following equations must hold:

$$\rho_0 = \eta_{\max} \cdot \frac{rf_{\min}}{2\pi} \qquad a = \frac{\rho_0 + rf_{\min}}{\rho_0}, \tag{1.27}$$

where $rf_{\min}$ is fixed and represents the diameter of the receptive field of minimum size in the fovea.

Same field of view, same number of pixels. Given a log-polar image of size $\xi_{\max} \times \eta_{\max}$, it is possible to use a Cartesian image of dimension $x_{\max} = \sqrt{\xi_{\max} \cdot \eta_{\max}} = \sqrt{2} \cdot \xi_{\max}$. Obviously, having to cover the same field of view, this image has to correspond to the same area as the log-polar image: this entails a corresponding resolution (in the fovea) of

$$\frac{\rho_{\max} \cdot rf_{\min}}{x_{\max}} = \sqrt{2} \cdot rf_{\min} \cdot \frac{\rho_0}{\xi_{\max}} \cdot a^{\xi_{\max}}. \tag{1.28}$$

Using Eqs. (1.27), the previous can be rewritten as

$$\frac{\rho_{\max} \cdot rf_{\min}}{x_{\max}} = \sqrt{2}\pi \left(1 + \xi_{\max}^{-1}\right)^{\xi_{\max}}. \tag{1.29}$$

This means that given the constrains we have assumed there is a great reduction of resolution in the Cartesian counterpart. The disparity suffers the same loss in resolution.

Same resolution in the center, same number of pixels. A Cartesian image of size $x_{\max} = \sqrt{\xi_{\max} \cdot \eta_{\max}} = \sqrt{2} \cdot \xi_{\max}$ is assumed. It covers only the central part of the image and comprises the same resolution as the log-polar image in the fovea ($rf_{\min}$). In this case

TABLE 1
A Case Study between Log-Polar Technique and Cartesian Counterparts

Methods	Computational load	Resolution in fovea	Field of view	Drawbacks
Log-polar disparity	$\xi_{\max} \cdot \eta_{\max} = 2\xi_{\max}^2$	1	1	
Same amplitude	$\xi_{\max} \cdot \eta_{\max} = 2\xi_{\max}^2$	$\sqrt{2}\pi(1+\xi_{\max}^{-1})^{\xi_{\max}}$	1	Low resolution; no implicit selection of target; integral estimation of disparity $C(\ )$ fails at small targets
Same resolution	$\xi_{\max} \cdot \eta_{\max} = 2\xi_{\max}^2$	1	$\frac{\sqrt{2}}{2\pi}(1+\xi_{\max}^{-1})^{-\xi_{\max}}$	Small field of view; the target gets lost easily
Same amplitude and same resolution	$\frac{1}{2\pi^2} \cdot \xi_{\max}^2 \cdot (1+\pi\xi_{\max}^{-1})^{2\xi_{\max}}$	1	1	Increased computational load; no implicit selection of target; integral estimation of disparity $C(\ )$ fails at small targets

the loss in the amplitude of the field of view is equal to (using Eq. (1.27))

$$\frac{x_{\max}}{\rho_{\max} \cdot r f_{\min}} = \frac{\sqrt{2}}{2} \cdot \frac{\xi_{\max}}{\rho_0 \cdot a^{\xi_{\max}} \cdot r f_{\min}} = \frac{\sqrt{2}}{2\pi} \left(1 + \xi_{\max}^{-1}\right)^{-\xi_{\max}}. \tag{1.30}$$

Therefore, the probability of loosing the target increases accordingly.

Same field of view, same resolution, and bigger number of pixels. A Cartesian image of size $x_{\max} = \frac{2}{r f_{\min}} \rho_0 \cdot a^{\xi_{\max}}$ is used, in order to achieve the same *field of view*. Using Eq. (1.27), $x_{\max}$ is related to $\xi_{\max}$ as

$$x_{\max} = \frac{\sqrt{2}}{2\pi} \cdot \xi_{\max} \cdot \left(1 + \pi \xi_{\max}^{-1}\right)^{\xi_{\max}}. \tag{1.31}$$

Besides, in each of the previous cases, there is no implicit selection of the target. This means that, by using a global correlation function like $C(\ )$ (Eq. (1.3)), a target occupying a large portion of the image is needed. If the image is log-polar, a target in the center of the image could correspond to a larger number of pixels, which improves significantly the estimation of disparity. A synopsis of this case study is provided in Table 1.

7. EXPERIMENTS

In order to demonstrate the feasibility of the approach we tested the algorithm on a binocular robotic setup. This consists of five degrees of freedom robot head (Fig. 8a). The head kinematics allows independent vergence (both cameras can move independently around a vertical axis) and a common tilt motion. Furthermore, the neck is capable of a tilt and pan independent motion. However, in the following experiments, only two *df* (i.e., the

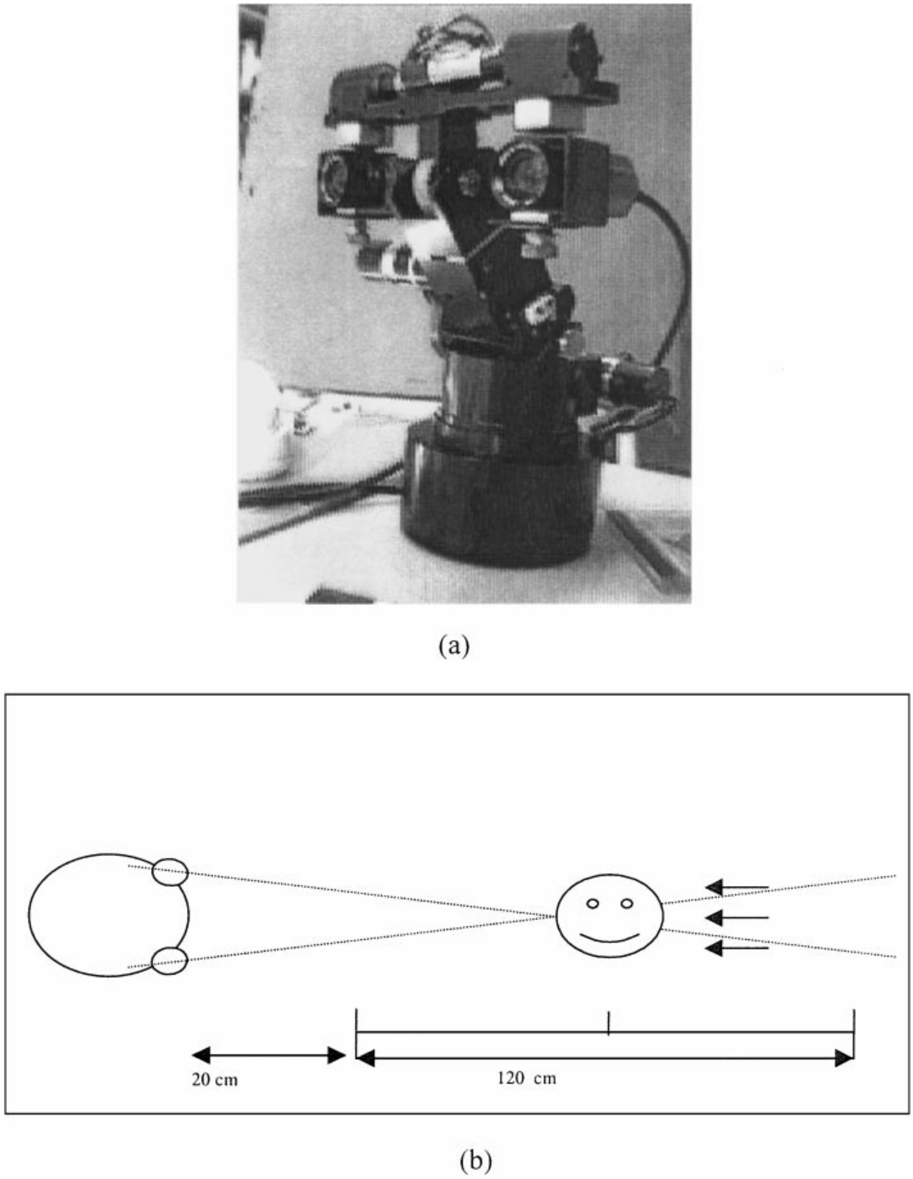

FIG. 8. The experimental setup. (a) The 5 *df* robot head. (b) A clown head is moving back and forth in front of the robot head at different speeds. The amplitude of the movement was 120 cm; the nearest point was at 20 cm from the cameras.

camera pans) were used to test the algorithm. The robotic setup is equipped with two space-variant color cameras (20). Experiments were carried out on a standard Windows NT-based Pentium II 400-MHz machine. The log-polar images were 32 × 64 pixels and they were processed at a video rate (25 frames/s). Actually, the overall computation time for control cycle was only 10 ms. The limiting factor in this case was the image acquisition process. The proposed technique was tested under two different experimental conditions under both controlled and uncontrolled stimulation. An object (the clown head of Fig. 9) was fixed on a programmable moving slider (Compumotor 3000 motor/drive system) capable of moving a

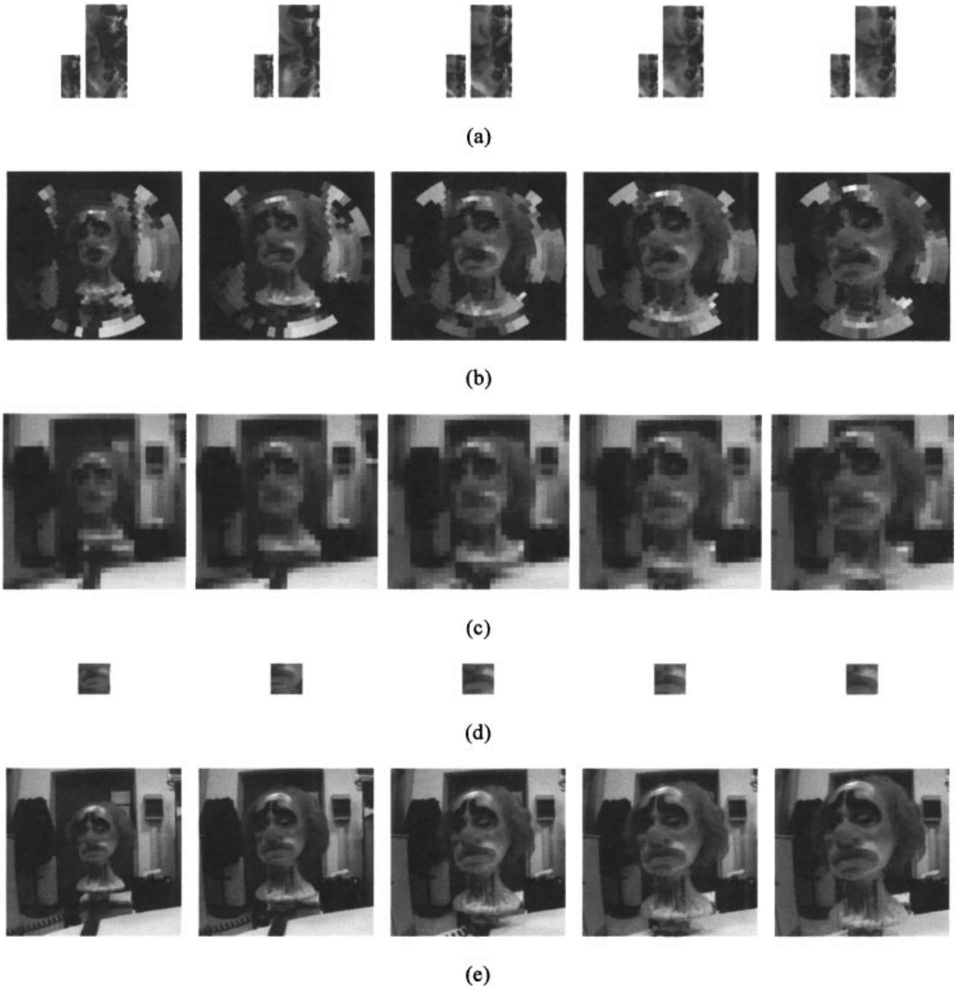

FIG. 9. An image sequence of five thumbnails (left image) of the subject, acquired during the experiment, showing the tracking while the target approaches the system: (a) in cortical plane (original on the left, zoomed-in on the right), (b) Cartesian reconstructed from the cortical ones, (c) Cartesian format with the same number of pixels and the same field of view as the ones in (a), (d) Cartesian format with the same number of pixels and the smaller field of view to the ones in (a), and (e) Cartesian format with bigger number of pixels and the same field of view as the ones in (a).

small object with speed ranging from 0 to 5 m/s. We then generated different back and forth motion profiles. In this situation only vergence control was required (no version) to actually track the moving stimulus. A sequence of log-polar images was recorded while the system was tracking. This is presented in Fig. 9a (cortical images). It should be noted that the actual size of the images (32×64) is the one on the left, while the one on the right is an enlarged image that we have added for the reader. In Fig. 9b, the Cartesian images reconstructed from the cortical sequence are presented. The same sequence is presented in Figs. 9c, 9d, and 9e in Cartesian format. Each of the sequences presented in these figures (9c to 9e) corresponds to one of the three cases described in the previous section. More specifically Fig. 9c presents Cartesian images with the same number of pixels (44×44) and the same field of view as the ones in Fig. 9a. Figure 9d illustrates a Cartesian image sequence with the same number

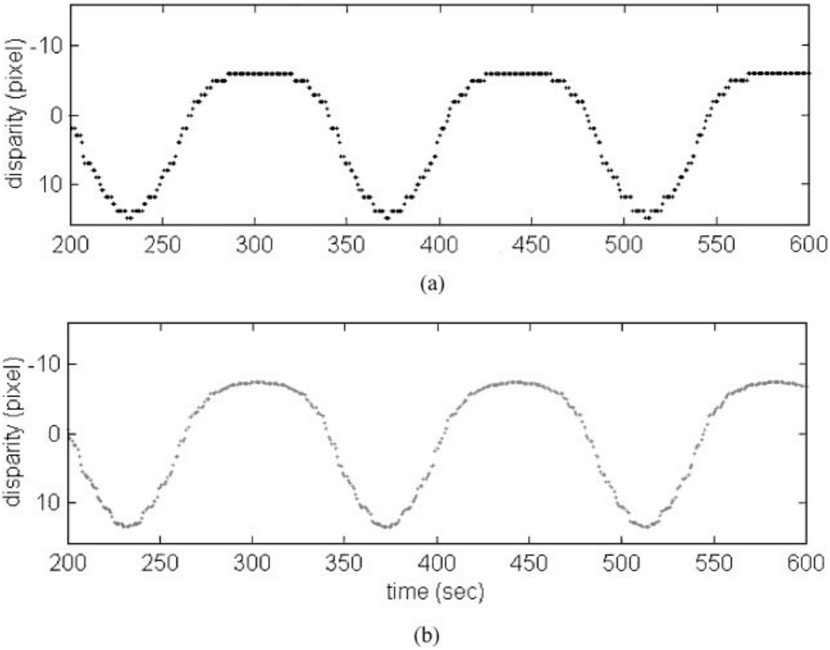

FIG. 10. Effect of the quadratic interpolation. The target, in this case, was moving along a sinusoidal trajectory with fixed frequency. The cameras were still, i.e., the vergence control was not applied. (a) The estimated disparity without interpolation and (b) the improved final result, after the interpolation.

of pixels (44 × 44) as the ones in Fig. 9a, but corresponding to a square area smaller than the whole image. Finally Fig. 9e shows Cartesian images with same field of view and the same resolution of the fovea, but with a much larger number of pixels (128 × 128). Observing Figs. 9a and 9e, it becomes obvious that the amount of clutter the system is rejecting, only by using space variant images. The object was able to move along either a triangular or a sinusoidal trajectory. Motion amplitude was 1.20 m. The nearest point was at 20 cm in front of the cameras (Fig. 8b). This corresponded to an angular difference (in term of vergence angle) of 45°.

Concerning the closed-loop case, Fig. 10 depicts the effect of the quadratic interpolation. Figure 10a is the disparity estimation without interpolation; Fig. 10b is the same trace using the described quadratic interpolation. It is possible to notice the difference in the smoothness of the estimate (the upper plot is step-wise). The data were obtained using the setup described by Fig. 8. In this case the cameras were still, and the object was moving along a sinusoidal trajectory (at a fixed frequency).

Figure 11 presents the closed loop condition where the estimated global disparity is used to drive vergence control. In this case the control was activated and the trajectory had a triangular profile. The four plots represent the estimated disparity (Fig. 11a), the fusion index $C(t)$ (Fig. 11b), the motor command (speed) (Fig. 11c), and the vergence angle (Fig. 11d). An interesting effect is visible by comparing the disparity plot with the vergence angle: the biggest error (in module) corresponds to the minimum distance of the slider, whilst the minimum estimated error corresponds to the maximum distance. Besides, when the change in direction occurs at the minimum distance there is a relatively big overshoot in the disparity estimate. On the other hand, when it corresponds to the maximum distance there

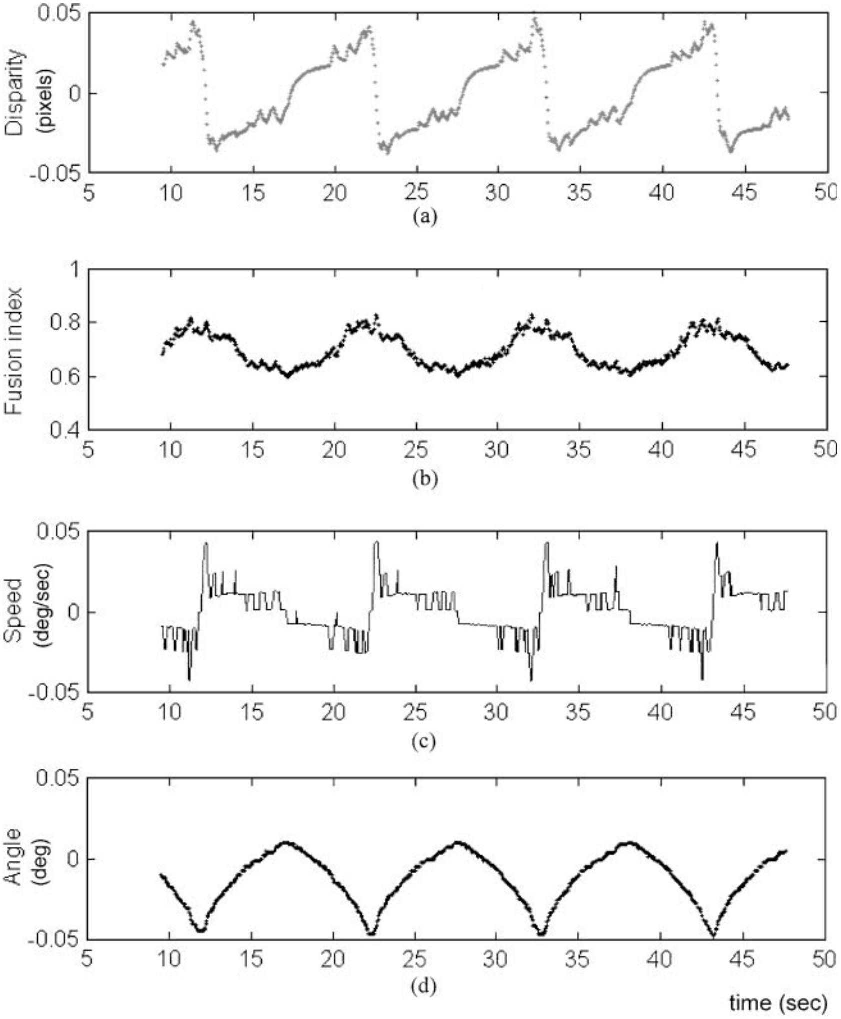

FIG. 11. Vergence control. The object moved along a triangular trajectory at a constant frequency. The vergence control was activated. (a) The estimated disparity, (b) the fusion index *C(t)*, (c) the motor command (speed), and (d) the vergence angle.

is almost no overshoot at all. Nevertheless, by observing the absolute scale of the computed error we can see that the maximum error was only 0.05 log-polar pixels.

As far as the performance in an unconstrained environment is concerned the behavior of the robotic head was tested for several prolonged experimental sessions. Its performances have been reliable, robust, and consistent. The frequency response of the system was practically flat over the tested range (10^{-1}–2 Hz). Nevertheless, these last experiments were conducted in a qualitative fashion and, therefore, their significance is limited. What we can say is that the head proved to be robust to noise in the periphery of its visual field as well as to noise derived by abrupt modification in light intensity or target object unexpected movement. While the robustness with respect to peripheral noise was mainly due to log-polar images, the capability of responding quickly to object modifications both in its dynamic (speed, directions, trajectory) and in its appearance (shape, position, color, reflectance, shape, and shades) factors is due to the robustness of the processing itself.

8. CONCLUSIONS

In this paper we presented a global disparity estimation method and its use in controlling the vergence angle of a stereoscopic artificial vision system. We were able to extract a measure of the vergence error by using space variant stereo images. This measurement was proved to be linear, fast, and robust to environmental changes. Consequently, it has been used to implement a closed loop vergence control that, through the use of both nonlinear distribution of CMs and space variant images, achieves accuracy much smaller than one pixel (typically around 10^{-1} pixels relatively to image size). Last but not least, we were able to extract the disparity information at frame rate using a standard hardware platform. Our claim is that it is possible to efficiently estimate disparity using log-polar images, which apparently seems not well suited for such computation. Furthermore, in this paper we showed (i) that disparity contains all useful information in order to control the vergence angle and (ii) that there is no need to use other cues to estimate the error. Besides, we claim that log-polar images permit an implicit selection of the central part of the overall scene that justifies the use of a global estimator.

Experimental results substantiate the above claim. The robotic head control proved to be robust to noise in the periphery of its visual field as well as from noise derived by abrupt modification in light intensity (as showed in Section 3) or target object unexpected movement (as showed in Section 5). Investigation of the integration of the proposed method with tracking control and saccade-like movements is underway.

ACKNOWLEDGMENTS

This work has been supported by the ESPRIT Projects NARVAL-30185 and ROBVISION-28867 and by funds from the Italian MURST. Dr. Gasteratos is supported by a fellowship from the TMR project-VIRGO.

REFERENCES

1. F. Panerai, G. Metta, and G. Sandini, Visuo-inertial Stabilization in Space-variant Binocular Systems, *Robotics Autonomous Systems* **30**(1-2), 2000, 195–214.
2. R. Manzotti, R. Tiso, E. Grosso, and G. Sandini, *Primary Ocular Movements Revisited*, Technical Report, 7/94, LIRA-Lab, Genova, 1994.
3. M. V. Srinivasan and S. Venkatesh, *From Living Eyes to Seeing Machines*, Oxford University Press, London, 1997.
4. J. P. Howard and B. J. Rogers, *Binocular Vision and Stereopsis*, Clarendon Press, Oxford, 1995.
5. D. H. Ballard and C. M. Brown, Principles of animate vision, *Comput. Vision Graphics Image Process*. **56**(1), 1992, 3–21.
6. J. Aloimonos, I. Weiss, and A. Bandyopadhyay, Active vision, *Internat. J. Comput. Vision* **1**(4), 1988, 333–356.
7. F. A. Miles and C. Busettini, Ocular compensation for self motion. Visual mechanisms, *Ann. NY Acad. Sci.* **656**, 1992, 220–232.
8. K. Pahlavan, T. Uhlin, and J.-O. Eklundh, Integrating primary ocular processes, in *Proceedings, ECCV92—European Conference of Computer Vision, Santa Margherita Ligure, Italy*, 1992.
9. J. Taylor, T. Olson, and W. N. Martin, Accurate vergence control in complex scenes, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Seattle, WA, 1994.
10. T. Kanade and M. Okutomi, A stereo matching algorithm with an adaptive window: Theory and experiment, *IEEE Trans. Pattern Anal. Mach. Intelligence* **16**(9), 1994, 920–932.

11. M. R. M. Jenkin, A. D. Jepson, and J. K. Tsotsos, Techniques for disparity measurement, *Comput. Vision Graphics Image Process. Image Understand.* **53**(1), 1991, 14–30.

12. D. De Vleeschauwer, An intensity-based coarse-to-fine approach to reliably measure binocular disparity, *Comput. Vision Graphics Image Process. Image Understand.* **57**(2), 1993, 204–218.

13. C. Tieh-Yuh and A. C. Bovik, Stereo Disparity from multiscale processing of local image phase, in *International Symposium on Computer Vision*, 1995.

14. S. Rougeaux, *Real-time active vision for versatile interaction*, Ph.D., Universite d'Evry, Courcouronnes, France, 1999.

15. L. Matthies, Stereo vision for planetary rovers: stochastic modeling to near real-time implementation, *Internat. J. Comput. Vision* **8**, 1992, 71–91.

16. C. V. Stewart, R. Y. Flatland, and K. Bubna, Geometric constraints and stereo disparity computation, *Internat. J. Comput. Vision* **20**(3), 1996, 143–168.

17. A. Bernardino and J. Santos-Victor, Vergence control for robotic heads using log-polar images, in *IROS'96*, Osaka, Japan, 1996.

18. A. Bernardino and J. Santos-Victor, Binocular visual tracking: Integration of perception and control, *IEEE Trans. Robotics Automation* **15**(6), 1999, 1080–1094.

19. N. Oshiro, N. Maru, N. Atsushi, and M. Fumio, Binocular tracking using log polar mapping, in *Proceedings, IROS'96*, Osaka, Japan, 1996.

20. G. Sandini, A. Alaerts, B. Dierickx, F. Ferrari, L. Hermans, A. Mannucci, B. Parmentier, P. Questa, G. Meynants, and D. Sheffer, The Project SVAVISCA: A space-variant color CMOS sensor, in *AFPAEC'98*, Zurich, 1998.

21. M. Daniel and D. Whitteridge, The representation of the visual field on the cerebral cortex in monkeys, *J. Physiol.* **159**, 1961, 203–221.

22. U. Schwarz and F. A. Miles, Ocular responses to translation and their dependence on viewing distance. I. Motion of the observer, *J. Neurophysiol*. **66**, 1991, 851–864.

23. E. L. Schwartz, A Quantitative model of the functional architecture of human striate cortex with application to visual illusion and cortical texture analysis, *Biol. Cybernetics* **37**, 1980, 63–76.

24. M. M. Marefat, L. Wu, and C. C. Yang, Gaze stabilization in active vision. I. Vergence error extraction, *Pattern Recognition* **30**(11), 1997, 1829–1842.

25. M. M. Marefat, L. Wu, and C. C. Yang, Gaze stabilization in active vision. II. Multi-rate vergence control, *Pattern Recognition* **30**(11), 1997, 1843–1853.

26. E. Grosso, R. Manzotti, R. Tiso, and G. Sandini, A space-variant approach to oculomotor control, in *IEEE Internat. Symposium on Computer Vision*, Coral Gables, FL, 1995.

27. I. Horswill and M. Yamamoto, A $1000 active stereo vision system, in *Proceedings, IEEE Symposium on Visual Languages*, 1994.

28. J. Nielsen and G. Sandini, Learning mobile robot navigation: A behavior-based approach, in *IEEE International Conference on Systems*, Man, and Cybernetics, San Antonio, Texas, 1994.

29. E. Grosso, G. Metta, A. Oddera, and G. Sandini, Robust visual servoing in 3D reaching tasks, *IEEE Trans. Robotics Automation* **12**(8), 1996, 732–742.

30. C. Capurro, F. Panerai, E. Grosso, and G. Sandini, A binocular active vision system using space variant sensors: Exploiting autonomous behaviors for space application, in *International Conference on Digital Signal Processing*, Nicosia, Cyprus, 1993.

31. G. Casalino, M. Aicardi, A. Bicchi, and A. Balestrino, Closed loop steering for unicycle like vehicles: A simple Lyapunov like approach, *IEEE Robotics Automation Magazine* **2**(1), 1995, 27–35.

32. C. Capurro, F. Panerai, and G. Sandini, Dynamic vergence using log-polar images, *Internat. J. Comput. Vision* **24**(1), 1997, 79–94.