Automatic Vergence Control Based on Hierarchical Segmentation of Stereo Pairs

R. Marfil, C. Urdiales, J. A. Rodríguez, F. Sandoval

Department of Tecnología Electrónica, ETSI Telecomunicación, Universidad de Málaga, Campus de Teatinos, 29071 Málaga, Spain

Received 18 July 2003; revised 31 October 2003

ABSTRACT: This article presents a video system to focus an object of interest in the field of view by means of a stereo robotic head. The system relies on a new hierarchical segmentation technique that split the stereo pair into regions in a combined way. After segmentation is achieved, resulting regions in both images are implicitly related by a set of links and, consequently, their disparity can be estimated. Then, the cameras vergence is changed according to the estimated disparity until the object of interest is focused. The system has proven to be robust against mild noise and illumination changes. Its main advantage is that calibration is not a critical issue for the system to work.

Key words: hierarchical segmentation; active binocular systems; depth perception; vergence control

I. INTRODUCTION

Surveillance is a very important field in computer vision applications, whose scenarios are very varied and, usually, complex. Images from these environments are either acquired with static cameras yielding wide-angle lenses or with cameras mounted on pan-tilt devices. The main advantage of the first approach is that all the information about the environment is available in every single image. Active camera-based systems may present a longer focal length instead and, consequently, better resolution. Particularly, active binocular systems have been proposed (Fairley et al., 1995) for recovery of 3D trajectories by means of distance calculation. Depth perception has been a key issue in computer vision for decades. Even though some approaches rely on applying constraints to work with a single camera (Bradshaw et al., 1997), most works are based on stereo vision instead. These approaches typically rely on establishing a relationship between two images captured by a binocular head. Unfortunately, it is not easy to establish such a correspondence because of differences between both cameras, errors in calibration, occlusions, etc. Basically, three different approaches are used to that purpose.

Correlation based methods rely on filtering the left and right images to cross-correlate them in order to determine the correct local matches. This approach may appear to be slow for medium-size images, but it is typically applied using a fixed window size so operations may be parallelized (Fauguères et al., 1995; Koschan and Rodehorst, 1995). The main drawback of these methods is that they are very sensitive to noise (Hansen and Sommer, 1996). Also, the computation of depth is prone to errors in surface discontinuities.

A second approach has been developed in the past decade (Sanger, 1988). It consists of estimating the local phase difference between the images in the pair because it is equivalent to the disparity in bandpass signals. This technique can be parallelized as well (Crespi et al., 1998) for the sake of speed. Gabor filters are typically used for the required bandpass filtering. Because these filters must be necessarily small to work in real time for average sized images, only small disparities are allowed, but coarse to fine working strategies have been proposed to solve this drawback (Westelius et al., 1994). Initially, phase-based approaches were reported to be very sensitive to orientation and scale (Fleet et al., 1991). Further works (Langley et al., 1990; Theimer and Mallot, 1994) achieved better performance by weighting the estimated disparity of different scales and orientations. The most important drawback of these strategies is that they do not operate correctly unless cameras are well calibrated.

Feature correspondence-based methods rely on matching a set of features, like lines, corners or edges, in a stereo pair (Lew et al., 1994). If those features are correctly located in both images, the method leads to an accurate resolution disparity map. However, these maps are usually very sparse and interpolation is not simple. Alternatively, regions may be matched instead of points. This approach leads to denser but less accurate disparity maps. The obvious problem in this case is how to obtain the same regions through segmentation of two different images that are probably subjected to different capture conditions because of the nature of the cameras and differences in the field of view. Because of this problem, the region matching stage is not obvious.

This article proposes a method to combinely segment a stereo pair so that correspondence between regions is implicitly achieved. Segmentation is performed in a hierarchical way by using linked pyramids as described in section I. Thus, the relationship between
regions existing in both images is preserved in the link structure and the disparity of those regions can be estimated as described in section III. This disparity is used to focus an object of interest in the field of view by means of a vergence control process as described in the same section. Section IV presents some experiments performed using a Biclops robotic head and different objects of interest. Finally, conclusions and future work are presented in section V.

II. HIERARCHICAL SEGMENTATION

Pyramid-based image processing is not new. These structures have been widely used for low-pass filtering, data reduction, and, particularly, hierarchical image segmentation. Basically, a pyramid is a graph \((G(V,E))\) consisting of a set of vertices \(V\) linked by a set of edges \(E\). We refer to the vertices as nodes and to the edges as links. The base of the pyramid is designated as level 0. Each node \(n\) in a pyramid is identified by \((l,i,j)\), where \(l\) represents the level and \((i,j)\) are the \((x,y)\) coordinates within the level. In the classic pyramid, each node at level \(l\) is linked to the \(2 \times 2\) nodes underneath at level \(l - 1\). In order to achieve an efficient segmentation by using a pyramid, Burt et al. (1981) proposed the adaptive linking principle, consisting of rearranging the links between nodes at different levels of the pyramid in a controlled way. When a pyramid is adaptively stabilized, each node of the structure is linked to an irregular region of cells, yielding a homogeneous grey level at the base, which is the original image. Thus, if the pyramid is cut at a given level \(L\) (working level) yielding \(N\) nodes, the image is implicitly segmented into \(N\) homogeneous regions. In this article, we propose a new method to combinedly segment a stereo pair by using the aforementioned pyramids. The main novelty of our proposal is that, instead of working with a single pyramid, we use two pyramids at the same time. These pyramids are interlinked so that relationships between regions of pixels can be established in a combined hierarchical way (Fig. 1).

A. Structure Generation. In order to initialize the required structures, a pyramidal structure is built over the left and right images by means of the following steps:

1. Let level \(l = 0\), being level 0 the original frame.
2. For each set of \(2 \times 2\) nodes (sons) at level \(l\), generate a single node (parent) at level \(l + 1\), whose grey level is equal to the average of the grey level of its sons.
3. Create a link from each son to its parent.
4. Let \(l = l + 1\). Repeat step 2 until the pyramid is built.

It can be noted that every node of the structure is linked to a nonhomogeneous square area of cells at the base of the pyramid whose area depends on the node level.

B. Structure Stabilization. Stabilization consists of rearranging the links of the two pyramids built over the stereo pair, left and right, so that a node at each level of the left pyramid may be linked to an irregular homogeneous region at the base level of the same left pyramid, but also to a homogeneous region at the base of right pyramid. Equally, each node of the right pyramid is linked to an irregular homogeneous region at the base level of the same right pyramid and to a homogeneous region at the base of left pyramid. After this process is accomplished, segmentation provides a matching criterion between corresponding regions from the two images (Fig. 1). To stabilize two pyramids, left and right, in a combined way, the following steps are required:

1. Let \(l = 0\).
2. This step is performed for both pyramids left and right:
   - For each node \(n(i, j, l, \text{left/right})\), find the most similar parent at level \(l + 1\) in a \(3 \times 3\) vicinity above that node at pyramid left/right and link them [Fig. 2(a)].
   - For each node \(n(i, j, l, \text{left/right})\), find the most similar parent at level \(l + 1\) in a \(1 \times 3\) vicinity above node \(n(i, j + \Delta j, l, \text{left/right})\) at pyramid right/left and link them. \(\Delta j\) represents the disparity of the node between the left and right images. It is equal to the difference between the position of the centroids of the regions at the bases linked to such a node in left and right images [Fig. 2(b)]. If no disparity estimation is available, it is initially set to 0.
3. Regenerate level \( l + 1 \) of pyramid left/right. The grey value of each father cell is recomputed as the average of the sons linked to it at level \( l \) in pyramids left and right. It must be noted that a parent may yield from 0 to 36 sons at its own pyramid, and a maximum of 12 sons at the opposite pyramid. If level \( l + 1 \) does not suffer any change, proceed to step 4. Otherwise, return to step 2.

4. Let \( l = l + 1 \). Return to step 2 until the whole structure is stabilized.

After steps 1–4 are accomplished, any node at any stabilized level of the pyramids is linked to an homogeneous region of pixels at its base. Hence, given any level of the structures, the images at the base of pyramids left and right are segmented into a number of regions equal to the number of nodes of such a level. These regions present homogeneous grey levels and disparity values. It must be noted that the relationship between the resulting regions is implicitly provided by the link structure. A node at the working level of the left pyramid is linked to a homogeneous region of pixels at the left image, but also to an homogeneous region of pixels at the right image. If the process is working correctly, both regions are the same. Figure 3 shows the stabilization results of the proposed method for a simple real stereo pair. In Figure 3(a) both images are stabilized alone according to the classic adaptive stabilization strategy. In Figure 3(b) the images are combinedly stabilized by using the proposed technique. It can be appreciated in the pyramid built over the right frame of the pair in Figure 3(b) that the combined stabilization process affects its upper levels to reflect the organization of the left frame. The most important difference between both structures is that nodes in different frames in Figure 3(a) are unrelated, whereas nodes in different frames in Figure 3(b) are interlinked.

The main disadvantage of the proposed algorithm is that the number of classes of any segmented image is always equal to \( N \), being \( N \) the number of nodes of the last stabilized level. Naturally, \( N \) is not necessarily equal to the number of regions in the scene. If \( N \) is larger than the adequate number of regions in the stereo pair, classes do not correspond to a single entity, and they tend to be arranged randomly depending on the layout of the environment in each image. If \( N \) is too low, wrong class fusions occur and resulting classes might be nonhomogeneous. We propose to work with an adaptive number of classes to avoid undesired class fusions. We apply an unsupervised merging process \textit{a posteriori} originally presented in Rodriguez et al. (2001) to achieve stable regions. Basically, the process consists of analysing regions presenting similar grey values and overlapped bounding boxes. If any of the contour pixels of such regions are in contact, both regions are merged into a single one. The main advantage of this approach is that the number of resulting classes is no longer equal to the number of nodes at a given level and therefore the segmentation algorithm works in a more flexible way.

C. Main Features of the Proposed Technique. The proposed technique presents several novelties with respect to previous works with pyramids:

- Instead of stabilizing a single structure, two pyramids are interlaced and stabilized in a combined way. Figure 1 shows the proposed link structure.
- Links among nodes of the same pyramid are arranged as proposed in Burt et al., 1981 [Fig. 2(a)].
- During the stabilization of pyramids, both pyramids are regenerated.
- The parent of node \( n(i,j,l,\text{left/right}) \) for the same pyramid should be among the \( 3 \times 3 \) nodes immediately above in the same pyramid. Nevertheless, its parent for the pyramid related to the other camera may not occupy the same position. Hence, its parent at pyramid is searched among the \( 1 \times 3 \) nodes immediately above \( n(i,j+\Delta j,l,t) \) at the opposite pyramid, being \( \Delta j \) the disparity for node \( n(i,j,l,t-1) \) [Fig. 2(b)]. It must be noted that, for a stereo pair of calibrated cameras, disparity has no vertical component.
- Any level \( l + 1 \) in pyramid left/right is regenerated from two levels: level \( l \) in pyramid left/right and level \( l \) in pyramid right/left. In this case, the following can happen:
  - If the disparity \( \Delta j \) for a given node \( n(i,j,l,\text{left/right}) \) is correctly estimated, the node yields the same grey level that node \( n(i,j+\Delta j,l,\text{right/left}) \). Hence, both nodes may be correctly linked to a similar parent among the \( 1 \times 3 \) ones.

Figure 3. Levels \( 128 \times 128, 64 \times 64, 32 \times 32, \) and \( 16 \times 16 \) of (a) pyramids independently stabilized built over a stereo pair; (b) pyramids combinedly stabilized built over a stereo pair.
immediately above at level $l + 1$ in pyramid right/left. In this case, the parent grey level should not be very affected by having sons in both pyramids. Therefore the process converges. If previous levels are correctly stabilized, nodes $n(i, j, l, left/right)$ and $n(i, j + \Delta j, l, right/left)$ are linked to homogeneous regions at images left and right. Hence, the parent node is now linked to homogeneous regions at both images.

- If $\Delta j$ is not correctly estimated for $n(i, j, l, left/right)$, then nodes $n(i, j, l, left/right)$ and $n(i, j + \Delta j, l, right/left)$ may yield very different grey levels. Nevertheless, $n(i, j, l, left/right)$ is forced to choose a parent among the $1 \times 3$ ones immediately above $n(i, j + \Delta j, l, right/left)$. In this case, the chosen parent grey level changes to accommodate the new son. Eventually, it will reach a stable compromise value, but the regions the parent is linked to at the base may not be homogeneous.

### III. DISPARITY CALCULATION AND VERGENCY CONTROL

After the two images of a given stereo pair have been segmented, the resulting regions in the left and right images are implicitly related by the pyramids link structure. Then, their disparity can be estimated by calculating the displacement of the centroid of each pair of regions. Because our method works with regions, it is less noisy than differential methods relying on matching windowed areas of pixels. Figure 4(a) shows the combined stereo pair in Figure 3. The disparities for both frames are presented in Figure 4(b). Each region is colored according to the distance in pixels between its centroids at the right and left image in the stereo pair. Consequently, regions presenting the same shape and color are correctly matched in both images. Bright regions present a high disparity, while dark regions present a low disparity. Figure 4(c) presents the depth map extracted from the disparities in Figure 4(b). Objects closer to the camera are presented in a brighter color.

Figure 5(a,b) present the left and right images of a more complex stereo pair, respectively. The image in Figure 5(d) presents the disparity map estimated for the stereo-pair. It can be noted that the hands of the first dummy present the higher disparity because they are closer to the camera and unfocused, as can be seen in Figure 5(c). The farther the dummy is from the camera, the lower its disparity.

In order to focus a given object in the scene, it is necessary to move the cameras so that its disparity becomes minimum. The disparity is minimum when the object is at the same position in both cameras. Hence, camera movements to be performed basically depend on the relative position of the object to be focused with respect to the cameras. To move the cameras, a Biclops robotic head is used. This head yields pan, tilt, and vergence movements.

The vergency control system consists of three different modules (Fig. 6). The whole system is parallelized by means of a distributed control architecture so that all modules can work at the same time. This architecture allows data exchanging between different software and hardware modules working on the same or different machines in a client/server way.

Initially, the capture module captures two images at the same time. These images are fed to the disparity calculation module, which calculates the disparity of the regions in the field of view according to the previously described method. In order to focus a

![Figure 4](image4.png)  
**Figure 4.** Depth map calculation: (a) combined stereo pair; (b) disparities of the right and left images of a stereo pair; (c) depth map of the stereo pair.

![Figure 5](image5.png)  
**Figure 5.** Disparity calculation for a stereo pair: (a) left image; (b) right image; (c) left and right images overlapped; (d) estimated disparity map.
Given region, the hardware control module controls the movements of the Biclops head via the serial port. As previously commented, the vergence angle of the cameras is changed according to the region disparity. The region is focused when the disparity value returned by the disparity computation module is minimum. It must be observed that it is necessary to establish some criteria to define an area as a region of interest so that the system tries to focus it. Basically, these criteria depend on the application and environment conditions.

Finally, it must also be noted that each time a vergence movement is performed, a new stereo pair is captured and segmented. The previous disparity values are then used to stabilize the pyramids for the new stereo pair. It must be noted that, at first, no disparity estimation is available. Hence, the initial segmentation results might not be adequate. Nevertheless, the iterative nature of the adaptive stabilization process compensates this factor and, after segmenting a few stereo pairs, results quickly converge.

IV. EXPERIMENTS AND RESULTS

In order to probe the feasibility of the proposed approach several experiments have been performed. The experimental setup is shown in Figure 7. Basically, it is composed by the aforementioned Biclops head mounted with two ordinary videoconference cameras to acquire the stereo pair images and a programmable moving slider to generate controlled vergence stimuli in a realistic environment. Experiments have been carried out on two standard Windows-based Pentium II 400 MHz machines. One computer is used to capture and preprocess the stereo images, and the second one is employed to control the head movements and run the disparity estimation algorithm. The images are digitized to either 128 × 128 pixels or 256 × 256 pixels and 256 grey levels. Because the cameras have its own automatic gain control, the average grey level of the images has to be normalized so they do not present significant differences. Then, the disparity of the stereo pair is calculated. It must be noted that images are acquired in real time and, consequently, subjected to illumination changes and noise. Hence, disparity may change for different captures even when the cameras do not move. In order to achieve resistance against these fluctuations, disparity is calculated three times for each vergence angle and then averaged. Anyway, the whole algorithm can process two stereo pairs per second with 128 × 128 pixels images.

To illustrate the system performance in different conditions three experiments are described. In the first experiment we compare the disparity maps generated by different methods. The second and third experiments are based on the works of Capurro et al. (1997). Thus, the second experiment shows the behavior of the proposed system in response to controlled changes of depth (vergence tracking experiment), and the third one is used to show the behavior of the vergence controller in response to sudden appearance of objects at different depths (saccadic vergence experiment). Anyway, it must be remarked that the proposed approach must be refined to correctly work in dynamic environments.

A. Disparity Map Estimation Experiment

In this section, several stereo algorithms are compared. Particularly, the tested algorithms are the proposed algorithm, the dynamic programming (DP) method (Bobick and Intille, 1999), the scanline optimization (SO) method (Scharstein and Szeliski, 2002), the cooperative algorithm (CA) (Zitnick and Kanade, 2000), and the traditional shiftable window sum of squared differences (SSD). The DP works by computing the minimum-cost path through each x-d slice in the disparity space image (Bobick and Intille, 1999). Every point in this slice can adopt one of three different states: match, left-visible only, or right-visible only. If we assume that the ordering constraint is being enforced, a valid path can take at most three directions at a

![Figure 6. Modular description of the proposed vergence control system.](image)

![Figure 7. The experimental setup: (a) the Biclops robotic head and the programmable moving slider; and (b) different objects are moving back and forth in front of the robotic head at different speeds.](image)
point associated with different deterministic state change. Dynamic programming is used to accumulate the minimum cost of all paths at this point. Points in match state charge is the matching cost at this point in the disparity space image. Points in the other two states charge is a fixed occlusion cost. The SO is a global optimization approach designed to evaluate different smoothness terms (Scharstein and Szeliski, 2002). It also works on individual x-d slices of the disparity space image, and it optimizes one scanline at a time. However, the method is not symmetric and does not use ordering constraints. Thus, the algorithm does not need an occlusion cost parameter. The cooperative algorithm is a stereo algorithm for obtaining disparity maps with occlusion explicitly detected (Zitnick and Kanade, 2000). Basically, it builds a 3D array of match values in the disparity space; each element of this array corresponds to a pixel in the reference image and a disparity value relative to the other image. An update function of match values is also built for use with real images. The update function generates continuous and unique values by diffusing support among neighboring match values and by inhibiting values along similar lines of sight. Initial math values are used to retain details during each iteration. After the match values have converged, occluded areas are explicitly identi-

Table I. Comparative performance of five stereo algorithms using $B_{NO}$ and running times.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>$B_{NO}$</th>
<th>Running Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSD</td>
<td>5.67</td>
<td>0.150</td>
</tr>
<tr>
<td>SO</td>
<td>5.08</td>
<td>0.150</td>
</tr>
<tr>
<td>DP</td>
<td>4.56</td>
<td>0.120</td>
</tr>
<tr>
<td>Cooperative algorithm</td>
<td>3.49</td>
<td>7.420</td>
</tr>
<tr>
<td>Proposed algorithm</td>
<td>3.67</td>
<td>0.350</td>
</tr>
</tbody>
</table>

Figure 8. (a) Left image of the Tsukuba stereo pair; (b) ground truth; and disparity maps obtained by: (c) dynamic programming (DP); (d) scanline optimization (SO); (e) cooperative algorithm (CA); (f) SSD; and (g) proposed method.
Finally, the SSD is a traditional stereo algorithm where the matching cost is the squared difference of intensity values at a given disparity. In order to smooth the obtained depth map, an aggregation procedure is done by summing matching cost over square windows with constant disparity. Final disparities are computed by selecting the minimal aggregated value at each image pixel.

To quantitatively evaluate these algorithms, we require data sets that either have a ground truth disparity map. In this experiment, we use the well-known stereo pair “head and lamp” from the University of Tsukuba (Nakamura et al., 1996). Figure 8 shows the Tsukuba stereo pair and the obtained results by applying the previously described methods. It can be noted that the methods make quite different errors. Thus, the disparity map obtained by the SSD method presents an important loss of detail that is due to the large window size (17 × 17) required for the method to work well. The disparity maps computed by the scanline-based algorithms (DP and SO) show a lot of detail but they present significant quantitative errors due to the lack of inter-scanline consistency. The results obtained by the cooperative algorithm and the proposed method are very similar, although the cooperative algorithm performs the best. In order to compare the different algorithms, we use the percentage of bad

Figure 9. (a) Object motion sequence, vergence angle and the real and estimated object position respect to the cameras; and (b) several stereo frames of the vergence tracking experiment.

Figure 10. Experiments with homogeneous backgrounds: initial and final stereo pairs for (a) a simple homogeneous object; (b) a textured object; (c) two grouped objects.
matching pixels in the nonoccluded regions, $B_{NO}$ (Scharstein and Szeliski, 2002). Table I shows the values of $B_{NO}$ for the different methods. The cooperative algorithm is the winner in this comparison. Finally, to compare the efficiency of the different methods, Table I also shows the running times for each of the five algorithms. In this case, the cooperative method is the slowest. Although the local and scanline-based methods are quite fast, the proposed method is only three times slower than them and it can be used to calculate the disparity in our experimental setup.

**B. Vergence Tracking Experiment.** In this experiment, an object is fixed on the programmable moving slider in front of the two cameras and then different back and forth motion profiles are generated. In this situation, only vergence control is required to track the moving object. It must be noted that at the beginning the algorithm computes the disparity of the whole image in order to select the region of interest. In these experiments, we consider an object to be of interest when it is closer to the camera than the rest of the items in the environment. Basically, this object is the main disparity source of the field of view. Consequently, if the global disparity of the stereo pair is minimized, the object is focused. The global disparity of the images is equal to the average disparity of all regions of the images. According to this criterion, the system does not work properly if at the beginning there are several nongrouped objects in the scene at similar distances from the camera or if the given region of interest is blended with the background. In order to solve this problem, it would be necessary to establish more complex criteria to choose beforehand which features should present a region to be considered a region of interest for each particular application. Thus, the system could minimize only the disparity of that region. In this experiment, after selecting the region of interest, the algorithm can track it easily because its depth varies smoothly. We consider that the experiment begins when the object of interest has been correctly selected.

Figure 9 illustrates one of the vergence tracking experiments. In this case, the motion sequence [Fig. 9(a)] can be divided in five phases: forth-stop-back-stop-back. The speed of the moving object is practically 1 cm/s in the forth-and-back movements (there are noise peaks in the speed value when the start and stop commands are sent to the moving slider). Figure 9(a) also shows the state of the vergence angle and the real and estimated distance between the object and the robotic head during the execution of the experiment. Figure 9(b) shows the acquired stereo pair for different frames of the sequence. It can be appreciated that the region of interest (the little rectangle) is practically centred in all images.

**C. Saccadic Vergence Experiment.** In the saccadic vergence experiment, the response of the system to sudden depth changes in the visual environment is tested. In a first set of tests, a simple background has been used so that it can be easily appreciated how the objects in the scene are focused. It must be appreciated that, in
this case, there is not any previous estimation of the object depth. Figure 10 presents the results for three different objects that are introduced in the field of view at different instants of time. The first one [Fig. 10(a)] is a common black diskette. It can be observed how the initial position of the cameras is corrected until the diskette is correctly focused and, hence, centred in both images. The second object is a textured one [Fig. 10(b)], but the system works correctly as well. In the third case [Fig. 10(c)], two objects have been put together in the field of view. It can be appreciated how the system tries to focus both objects so that they are centered in the two images of the stereo pair. In all cases, the algorithm takes an average time of 2 s to obtain the correct vergence angle.

Figure 11 shows the disparity between the two images of the stereo pairs for different vergence angles. The algorithm was run five times for each example so that it can be appreciated that the system is stable. Cameras are initially parallel in all cases because the position of the objects has not been previously estimated. It can be noted that the results are not exactly the same in all five tests with the same object. This occurs because all tests have been performed on-line and, consequently, capture conditions and illumination changed. It can also be observed that disparity changes in a semi-random way when vergence angles are very small, because the object of interest is unfocused and the differences between the images of the pair basically depend on the global structure of the scene. Nevertheless, it can be appreciated that all results tend to converge as soon as the object starts to be focused. It is also interesting to note that in order to achieve a minimum disparity for all tests and objects the vergence angle is close to 10°. This occurs because all objects in Figure 10 are positioned at the same distance from the camera, ~55 cm. Consequently, the distance to the object of interest can be estimated once the minimum disparity angle is obtained. Figure 12 shows the relationship between the vergence angle and distance for different runs of the algorithm. Data has been empirically obtained and it is coherent for different object to camera distances. A second set of tests have been performed for nonuniform backgrounds to prove that the system also works in more complex scenes. Figure 13 shows the results for two different objects under different conditions.

**Figure 13.** Experiments with nonhomogeneous backgrounds: initial and final stereo pairs for (a) an homogeneous object; (b) a nonhomogeneous object.

**Figure 14.** Disparity vs. vergence in a nonhomogeneous background for a) an homogeneous object; b) a nonhomogeneous object.
different lightning conditions. It can also be clearly observed that the cameras are not calibrated. In both cases, the initial and final stereo pairs are presented. It can be appreciated that in the final stereo pairs the objects of interest are centered in the images.

Finally, Figure 14 presents the disparity values for both tests in Figure 13 regarding different vergence values. In the first case, the system achieves a minimum disparity for an angle equal to 11°, the maximum for the Biclops head, whereas in the second case, this angle is equal to 9°. Even though distances can not be appreciated in an accurate way in the pictures, it can be noted that the battery in Figure 13(a) is closer to the camera than the box in Figure 13(b). As usual, it can also be appreciated that disparity values are not reliable when the cameras are parallel.

V. CONCLUSIONS AND FUTURE WORK

This article has presented a vergence control technique based on a hierarchical disparity calculation algorithm to focus an object of interest in a given scene. Disparity is estimated by combinedly segmenting a stereo pair so that regions defined at both images tend to be the same and are implicitly related. Vergence is changed according to this value until disparity is minimal. The proposed system has proven to work correctly for both simple and complex backgrounds as long as there is a single object to focus in the scene and it is significantly closer to the camera than the background. The algorithm can deal with noise and mild illumination changes, but its main advantage is that cameras do not need to be precisely calibrated to obtain good results. The main disadvantages of this algorithm is that it works only for a single object or a set of grouped objects and that the object of interest need to be significantly closer to the camera than the background if such a background is not homogeneous. Further work will focus on solving these problems by using different segmentation strategies.

ACKNOWLEDGEMENTS

This work has been partially supported by the Spanish Ministerio de Ciencia y Tecnología (MCYT) and FEDER funds, project TIC2001-1758. The authors thank Dr. Y. Ohta and Dr. Y. Nakamura for supplying the ground truth data from the University of Tsukuba. The authors also thank Dr. L. Zitnick and Dr. T. Kanade, whose source code of the cooperative algorithm has been obtained from http://www-2.cs.cmu.edu/~clz/stereo.html. Finally, the authors would like to thank to Dr. D. Scharstein and Dr. R. Szeliski for their Web page http://www.middlebury.edu/stereo. The clear description of the scanline optimization and dynamic programming methods presented in this Web page has permitted us to correctly implement these methods.

REFERENCES


H. Hansen and G. Sommer. Real time vergence control using local phase differences, Mach Graph Vis 5 (1996), 51–63.


