ATTENTIONAL MANAGEMENT FOR MULTIPLE TARGET TRACKING BY A BINOCULAR VISION HEAD

Fábio de Freitas Caetano **Jacques Waldmann**

Instituto Tecnológico de Aeronáutica
Departamento de Sistemas e Controle - Divisão de Engenharia Eletrônica
CTA-ITA-IEES - 12228-900 - São José dos Campos – SP
jacques@ele.ita.cta.br

Abstract: The development of an attention strategy for multiple target tracking and its integration with a binocular vision head is presented. Monocular motion-based segmentation yields image regions that are further clustered into targets. The emulation of a fovea significantly reduces the computational workload of target segmentation. Attentional management is based on assigning an interest value to each target. Its evaluation weighs binocular disparity, the number of detected moving pixels in a target, its pixel density, velocity and duration along the image sequence. The vision head performs saccades, ballistic motions and smooth pursuit around its tilt and vergence axes with characteristics similar to those of the anthropomorphic visual apparatus. System performance is evaluated on a 166 Mhz PC host with unexpensive off-the-shelf hardware in two situations. In the first case, small targets translate independently over a textured background. In the second, a moving human subject yields multiple targets undergoing rotation, non-rigid motion and scale changes. A processing rate about 0.8 stereo pair per second is attained. The first case shows the tracking and image stabilization capabilities for small translating targets only. The second case reveals the impact of hardware limitations on the system sensitivity to distortions in gray level patterns. It is induced by interframe rotation and changes in the illumination and scale factor as the tracked human subject moves around in the scene. As a consequence, undesired attentional reorientations occur. However, the system succeeds in keeping the subject within the field of view in such conditions though not always stabilized within the image region of fixation.

Keywords: Active vision, visual attention, binocular tracking, fovea-aided multiple target segmentation, robotic vision head, real-time control, autonomous robotic systems.

1 INTRODUCTION

Since immemorial times the sense of vision supports human beings in their search for food, shelter, evasion from predators and overall survival. Everyday, humans engage countless times in complex visual tasks such as detecting and tracking objects that somehow draw their attention. Detection is here considered as the confirmation that an object is present on the image whereas tracking concerns maintaining a given object within the field of view. The latter is often accomplished by controlling eye position and velocity. The connection between tracking and attention, which is a critical assumption in this work, is made explicit here. It is hypothesized that the visual tracking behavior in humans is driven by a sort of attentional management. It seems reasonable to claim that some sort of decision is required whether an object raises the system attention for tracking purposes and whether another object entering its field of view becomes more interesting, thus producing an attentional reorientation.

The literature describes robotic vision systems that reproduce some of the features found in biological vision systems that successfully evolved in nature. The motivation behind such efforts can be traced to the need for automated industrial and information processes based on visual feedback. An effective visual tracking system can be useful in a wide range of applications such as automatic surveillance (Batista *et allii*, 1998), traffic monitoring (Dagless *et allii*, 1993) vision-aided autonomous vehicles (Balkenius and Kopp, 1996; Huber and Kortenkamp, 1995), multiple target detection, selection and tracking by an aircraft or missile (Bar-Shalom and Li, 1993) and novel man-machine interfaces such as guidance support to the visually impaired (Crisman *et allii*, 1998; Molton *et allii*, 1998). Research efforts aligned with the work presented here exist on the areas of visual tracking (Andersen, 1996; Murray and Basu, 1994; Uhlin, 1996), visual attention (Bandera *et allii*, 1996; Culhane and Tsotsos, 1992) and the development of anthropomorphic-inspired robotic heads for active vision (Andersen, 1996; Weiman and Vincze, 1996).

1.1 Visual Tracking

Model-driven tracking is based on high level information to describe objects on the image (Andersen, 1996). Object features are sought which are robust to scale changes, translation and rotation of the object projection on the image. Image analysis resorts often to object-centered representations. However, unknown objects can be incorrectly processed with detrimental effects on system performance. Therefore, model-driven tracking may not be suitable in an unstructured environment as the completeness of the model database and the robustness of the representation scheme become critical. Moreover, it seems reasonable to assume that the model-driven approach imposes a heavy computational workload when dealing with day-to-day situations as expected in this work.

Artigo Submetido em 02/02/2000
1a. Revisão em 31/07/2000
Aceito sob recomendação do Ed. Consultor Prof. Dr. Jacques Szczupak

Hence, the data-driven approach is selected to circumvent the complexity inherent to the model matching problem when dealing with unstructured environments and to keep the computational burden bearable to unexpensive PC-based hardware. The latter approach utilizes pixel-based information such as brightness, motion, texture and contrast and sets of image points result as possible targets (Culhane and Tsotsos, 1992).

Binocular tracking requires control of both fields of view to project the image of any object raising the system interest on the region of fixation of both cameras. The human eye carries out this task by means of saccades and smooth pursuit (Araújo *et allii*, 1996; Batista *et allii*, 1997; Cowie and Taylor, 1997; Rotstein and Rivlin, 1996; Uhlin, 1996). Saccades are fast and usually have a large amplitude. They are associated with the allocation of attention to another region of the 3D space. Rapid minute corrections of the eyes' positions during the fixation of a static object occur and are called microsaccades. The high speed of a saccade precludes the processing of visual information due to motion blur. Smooth pursuit consists of a slow eye motion during tracking. It aims at maintaining the viewed object projection on the fovea where the retina resolution peaks. A mismatch between sensor velocity and object retinal velocity causes the object projection to slip away from the fovea. It is usually remedied by means of a microsaccade and smooth pursuit then resumes. Whenever attention is shifted to a different portion of the scene a saccade to that region occurs followed by smooth pursuit and microsaccades. In the present work, the binocular vision head is controlled according to the following guidelines:

1- Ballistic motions. These correspond to what is described as saccades in the literature. They employ high speed movements to control the camera position. It is employed to coarsely aim at an object that has just attracted the system attention. The rapid motion prevents the processing of visual information. The camera must halt to resume image processing. Otherwise, overshoot and oscillations result in blurred images.

2 - Saccades. Here differently from the literature, these are corrections to the camera angular velocity. They are engaged whenever a velocity mismatch between the camera and the object projection occurs causing significant slip error relative to the region of fixation. The camera velocity is then updated to reduce this error. Image processing continues during this type of motion.

3- Smooth pursuit occurs when the slip error is within acceptable bounds and the tracked object projects on the region of fixation thus obviating any corrections in camera velocity. Ideally, image processing benefits from sharper images as the target projection is stabilized on the region of fixation.

1.2 Visual Attention

Early visual processing in humans is often characterized as having two functionally distinct modes. The first, preattentive, in which information is processed in a spatially parallel manner that circumvents the need for attentional resources, and the second, which involves the allocation of attentional resources to specific locations or objects for more complex analysis. The latter mode, also known as the "where to look next" problem, is a key issue when dealing with various objects which compete for the system attention. Research concerning the allocation of attentional resources in human beings has yielded evidence for another functional dichotomy: attentional orienting can be voluntary, controlled by strategic goals (endogenous attentional control), or involuntary, driven by particular stimulus events (exogenous attentional control), as in Folk *et allii*, 1992. They claimed the existence of psychophysical evidence supporting the "contingent involuntary orienting hypothesis", according to which involuntary shifts of attention to a stimulus event depended on whether the event shared a feature property that was critical to the performance of the desired task. Involuntary shifts of attention were systematically dependent on the relationship between the stimulus properties of the distractor cue and the properties required to locate and process the target. Inspired in such observations, it is proposed here that the design of a machine capable of visually tracking multiple targets should result in a system behavior consistent with the desired goal while reducing undesired attentional reorientations. On the other hand, the management of surroundings awareness by the system should provide flexibility to consider changing environment conditions and their impact on system goals in order to adequately modify the attentional control settings according to the need for adaptation and autonomous behavior.

Real-time visual tracking, either in biological systems or in their machine counterparts, requires dedicated hardware with a high processing capacity. Even with the evolution of dedicated hardware to meet the needs of parallel computing along with improved actuators and controllers for machine vision systems, often one is limited by the available computational resources. Inspired in nature's solution to this question, an attentional management strategy should attempt to carry out an efficient usage of limited system resources. Spatial compression of visual data occurs at the retina with a varying resolution that peaks at the fovea. The eyes scan the surroundings and position the projection of an area of interest onto the fovea to acquire high resolution data. Peripheral vision also plays a role at the level of shifting the system attention to other interesting objects within the field of view. Overall, the computational burden is kept bearable.

It is assumed that tracking is neither activated nor terminated without a purpose. Such purpose induces the attention of the visual system to focus on a specific region or to shift focus to another one. Interest on a given object may cease as the system goal is attained such as the completion of its recognition or the estimation of its trajectory parameters, for instance. A quantitative determination of the interest raised by multiple moving objects in a scene is considered here in order to establish priorities for tracking. An attention function based on image features is proposed to evaluate which part of the image contents should raise the system interest. The selection of task-relevant reliable image features, called here attentional features, and of an adequate structure for the attention function is critical since they directly influence the system behavior (Andersen, 1996; Araújo *et allii*, 1996a,1996b; Batista *et allii*, 1997; Culhane and Tsotsos, 1992; Hager and Belhumeur, 1996; Huber and Kortenkamp, 1995; Murray and Basu, 1994; Shi and Tomasi, 1994; Uhlin, 1996). Among the features that could raise the attention of the system are object brightness, shape, velocity, distance to the observer and the duration of its occurrence within the field of view. In yet another direction which involves the application of estimation theory, Kalman filtering and α-β-γ filters have been utilized to predict the positions of simultaneous objects in consecutive images but without addressing the aspect of establishing priorities for tracking (Andersen, 1996; Bar-Shalom and Li, 1993).

This paper presents and evaluates a binocular tracking system with its computational resources allocated according to an attentional management capability. It is expected to select and track a target that raises the system attention among other independently moving objects. System behavior is analyzed in terms of target competition for the system attention, target selection, attentional reorientation and the system ability to maintain the selected target within the field of view by controlling camera position and velocity. The term "target" is here employed to indicate a moving region on the image that raises attention. The analysis encompasses the effect upon the overall system behavior of a variety of issues such as motion detection and target segmentation with moving cameras in an unstructured environment, the real-time processing requirements for adequate operation and the adequacy of the attention function, i.e., its structure, composing attentional features and weight selection.

2 SYSTEM DESCRIPTION

Visual processing of images acquired by the vision head is depicted in figure 1. The dominant camera is the left one of the stereo assembly. It is assumed that anything that moves in a scene is a potential target. Motion detection employs image subtraction (Bispo and Waldmann, 1998; Murray and Basu, 1994). A search for moving regions proceeds with the head in static mode (static search). Monocular localization and clustering of moving regions into targets on the dominant image follow. Target matching, either across stereo images or along a sequence, employs the estimated centroid position and its surrounding gray-level pattern yielding estimates of target velocity on the dominant image and of depth-related stereo disparity. These features enter the attention function along with the number of moving pixels within the target area, its density and a time measurement which indicates for how long each target has been within the field of view. The evaluation of the attention function for each target is stored in a dynamic list which records the relative interest raised by the targets and thus supports the selection of the target to track. Position and velocity estimates of its centroid are used to compute angular increments which are the command signals for controlling the motion of the dominant camera. Additionally, the estimated disparity is the error signal for controlling the asymmetric vergence of the stereo assembly which leads to binocular fixation. This process continues as targets evolve within the field of view. Static search for target motion resumes whenever no motion is detected in the scene after eight images.

In the following, the various vision algorithms and their integration in a vision machine are described in more detail. Images are monochromatic with a resolution of 160 pixels x 120 pixels and 256 gray levels unless otherwise stated. It has been so selected as a compromise between sufficiently accurate information and keeping the computational workload at a bearable level for a PC equipped with relatively unexpensive off-the-shelf image acquisition, A-D/D-A conversion and data communication hardware.

2.1 Processing the dominant images

Motion detection, localization of moving regions, target segmentation and centroid estimation result from the processing of images acquired by the dominant camera during tracking. Motion detection is done by thresholding the subtraction of a pair of consecutive images. Typically, image subtraction indicates the relative differences in both consecutive images. Differences arising from the previous image are discarded by means of a logical "AND" between the thresholded subtraction image and the gradient of the present image. The underlying assumption is that a moving target possesses a much richer texture than its surrounding background. In this case, the motion detector outputs sets of pixels, called here "*moving pixels",* that belong to moving image regions. Prewitt masks and a convenient threshold are used for the gradient operation (Fairhurst, 1988).

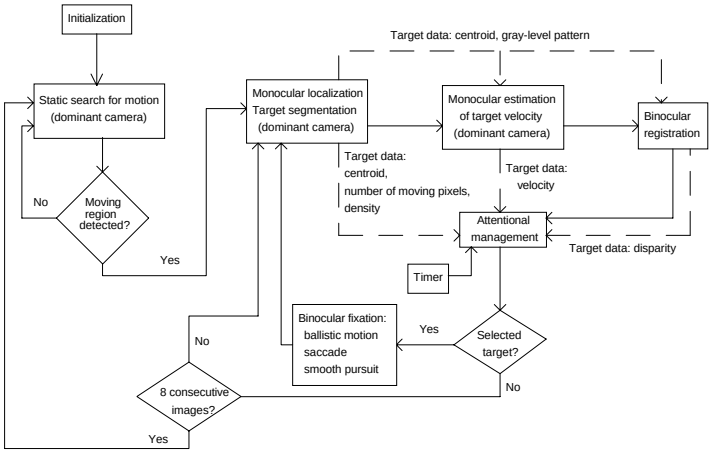


Figure 1 - System implementation with information flow (dashed lines).

As the camera tracks a selected target, adequate motion detection calls for image stabilization as seen in figure 2 (Bispo and Waldmann, 1998; Murray and Basu, 1994). It is accomplished by reading measurements from encoders that are fixed to the tilt and vergence axes. Measurement errors, delays between image acquisition and encoder reading, uncertainties in the optical parameters and undesired camera translation due to the distance from the rotating axes to the optical center cause inaccurate stabilization and thus noisy motion detection. It is almost completely remedied by the morphological gray-level opening operator defined in the following sequence of operations (Murray and Basu, 1994):

$$\forall (i,j) \in \text{image array and } (k,l) \in \text{structure element R centered on } (i,j):$$

$$I_{eroded}(i,j) = \min_{k,l}\left(I_k(k,l) - I_{k-1,stabilized}(k,l)\right)$$

$$I_{opened}(i,j) = \max_{k,l} I_{eroded}(k,l) \qquad (1)$$

The result of equation (1) is thresholded and a binary image with the moving pixels result. Many tiny regions falsely detected as moving because of incorrect image stabilization are thus removed. The remaining regions are assumed to be parts of the actual targets. The clustering of moving pixels into classes and subsequently into superclasses employs a distance criterion which is defined on the image plane. Each superclass centroid represents a target location on the dominant image. It is regarded as a candidate for selection and tracking according to its interest value relative to those of the other targets.

2.2 Target segmentation

Often independently moving targets exist in the scene. In general, moving regions on the image plane might result from either the same target or from different ones. The clustering of such regions into classes employs a distance-based criterion. The clustering algorithm operates on the binary image that results from thresholding the output of equation (1). There is

no sort of motion analysis. It reflects the hypothesis that nearby moving regions belong to the same target and it fails if neighboring regions move very differently because they actually belong to distinct nearby targets.

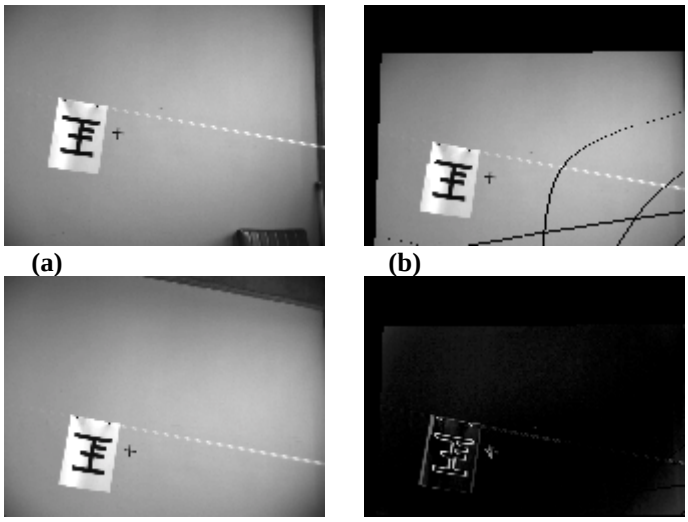


Figure 2 - Image subtraction with compensation of camera motion. a) Previous image. b) Image after motion compensation. c) Present image. d) Subtraction result.

Clustering is a computationally heavy task. It is divided in two phases. Initially, classes are derived from sets of nearby pixels which belong to moving regions. The same target may spread over a number of classes if its projection on the image plane covers a large area. Afterwards, the derived classes are grouped into superclasses which are analogous to classes but at a macroscopic level. Therefore, as nearby pixels are clustered into classes, likewise are nearby classes into superclasses. The latter is the adopted representation for the multiple targets on the image.

Since the purpose of visual tracking in primates seems to be the stabilization of the projection of the selected target on the high-resolution fovea, it seems reasonable to expect that for machine vision systems a corresponding area on the image should have a higher resolution and receive a larger share of attention. A region of fixation is then defined on each stereo image. Targets that raise attention should be maintained within such a region with the highest resolution available for processing. The resulting foveated images are employed to alleviate the computational burden, as in Bandera *et allii* (1996) and Scott and Bandera (1990). The region of fixation employs the original image resolution and subsampling generates a periphery with a coarser resolution.

2.2.1 Foveated images by subsampling

Some topologies for multiple resolution images are shown in Andersen (1996). In general, it consists of a central fovea with its maximum resolution surrounded by rings of decreasing resolution. To alleviate the computational workload of clustering the moving regions, the partition of the image in a fovea and a single-ring periphery is considered satisfactory. The central 80x60 fovea is selected to coincide with the region of fixation. Subsampling is done by averaging within 4x4 windows and then thresholding. Only the binary output of equation (1) which corresponds to moving regions on the image periphery is subsampled, thus further reducing the computional workload. An instance of the fovea-aided motion detector operating on images acquired with full resolution is shown in figure 3.

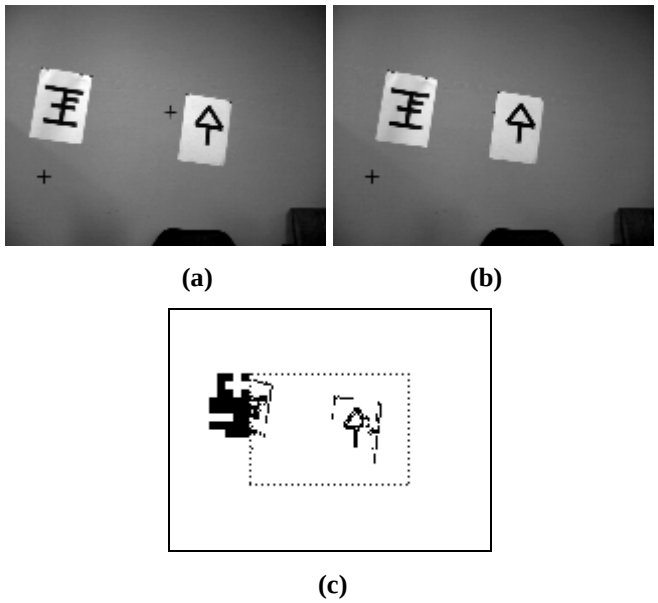


Figure 3 - Fovea-aided motion detector. a) Previous image. b) Present image. c) Binary result.

2.2.2 Clustering moving pixels into classes

The clustering algorithm employed here is a modified version of that described by Fairhurst (1988). Basically, it scans the output of the motion detector for pixels labeled as moving and attempts to assign them to classes according to a distance metric. Class attributes are the number of moving pixels in a class, #pix(class), the size of the smallest rectangle enclosing such pixels, its corresponding area and the density of moving pixels, dens(class), in this rectangle. Three cases may occur:

1 - None among the n already existing classes $c_1,...,c_n$ is found sufficiently close to the moving pixel $\mathbf{c} = (x, y)$ under consideration. In this case, it becomes the seed of a new class c_{n+1} with its centroid $\bar{\mathbf{c}}_{n+1} = (\bar{x}_{n+1}, \bar{y}_{n+1})$. Its coordinates are described in the high-resolution coordinate frame regardless of whether it is on the fovea or on the periphery. The quantity of pixels in this new class is 1 if it is on the fovea or 16 on the periphery, the latter corresponding to a 4x4 window of high-resolution pixels. Likewise, its area is defined as 1 if on the fovea or 16 on the periphery. The density is one corresponding to the ratio between the quantity of pixels in the class and its area. The smallest enclosing rectangle is a square with dimensions 1x1 on the fovea or 4x4 on the periphery.

2 - n classes have been already initiated, each one with its respective moving pixels given by $(x_p, y_p), p = 1,2,...,\#pix(c_k); k = 1,...,n$ and at least one of them is sufficiently close to the moving pixel $\mathbf{c} = (x, y)$ under consideration. Proximity to the k-th class centroid $\bar{\mathbf{c}}_k$ is defined according to Euclidean distance on the maximum resolution image and independent of whether the fovea or the periphery are involved:

$$d = \min_k \left\| \mathbf{c} - \bar{\mathbf{c}}_\mathbf{k} \right\|_2 \le L_d, k = 1,2,...,n \qquad (2)$$

If the above criterion is true, then the k-th class that minimizes the above equation and does not violate the

inequality accepts $\mathbf{c} = (x, y)$ and the class attributes are updated, i.e., its number of moving pixels #pix(c_k) is incremented, the size of the smallest enclosing rectangle, its area and density, each of them as required and according to whether the location of $\mathbf{c} = (x, y)$ is on the fovea or on the periphery. Class centroid is not updated in this case.

3 - Classes have been initiated but none satisfies $d \leq L_d$. In this case, all class centroids are recomputed to reflect the most recent state of clustering and a new attempt at minimizing equation (2) is made. If successful, the procedure is analogous to case 2. If not, the procedure continues according to case 1. The purpose of case 3 is to reduce the amount of required computations.

The threshold L_d is chosen to bias the clustering results towards generating many small classes from one large target in opposition to a number of targets ending up within one large class. The clustering algorithm is applied to the output of the fovea-aided motion detector - the moving pixels - and yields classes - moving regions formed by concatenating such pixels. Figure 4 shows an example.

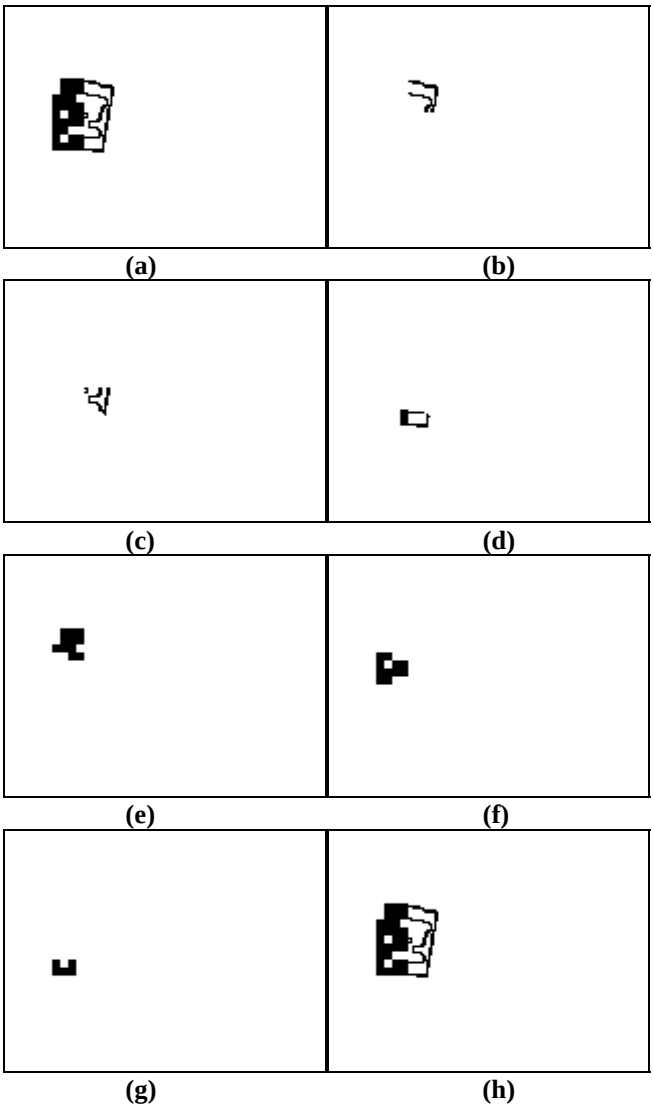


Figure 4 - Clustering moving pixels into classes. a) Fovea-aided detected moving regions. b-g) Resulting classes. h) Superclass resulting from class concatenation.

2.2.3 Clustering of classes into superclasses and target formation

The previous section showed how moving pixels are clustered into classes according to distance. However, such classes cannot be regarded as distinct targets yet because one target can be spatially spread over a number of classes. It is assumed that nearby classes are likely to belong to the same target. Once more a distance metric is the basis of a criterion for accepting or not the merger of classes into a superclass. Additionally, the resulting density is also considered before accepting a merger. Each i-th superclass Sc_i i=1,...,n is a subset of the n existing classes. Each class is to be contained by only one superclass. Each superclass is initialized as an empty set and each non-empty one that results by the end of all mergers becomes a target. Let $d(\bar{\mathbf{c}}_i, \bar{\mathbf{c}}_j)$ represent the Euclidean distance on the image plane between the centroids of classes c_i and c_j and dens(c_i) the i-th class density. The logical-valued criterion $C(c_i,c_j)$ for accepting a merger of c_i with c_j into the k-th superclass is as follows (see Appendix A):

$$C(c_i, c_j) = (d(\bar{\mathbf{c}}_i, \bar{\mathbf{c}}_j) = \left\| \bar{\mathbf{c}}_i - \bar{\mathbf{c}}_j \right\|_2 \leq 2{,}5\, L_d) \quad \text{AND}$$

$$(\text{dens}(c_i \cup c_j) = \frac{\text{dens}(c_i).\#\text{pix}(c_i) + \text{dens}(c_j).\#\text{pix}(c_j)}{\#\text{pix}(c_i \cup c_j)} \geq (2/3).\text{dens}(c_i)) \quad (3)$$

The density attribute enters the accceptance criterion to reduce the detrimental effect of image noise. In general, stabilization errors cause image artifacts that, when not eliminated by the gray-level opening operation, contain a few sparsely distributed pixels incorrectly detected as moving. A merger with such a region erroneously accepted as a class would not significantly alter the resulting number of pixels. The same reasoning may however be incorrect for the density attribute, since there could be a significant increase in target area. If such a merger were accepted into a superclass, the resulting target would be large and composed of sparse pixels within some of its parts. Targets with low density should not raise the attention of the vision system.

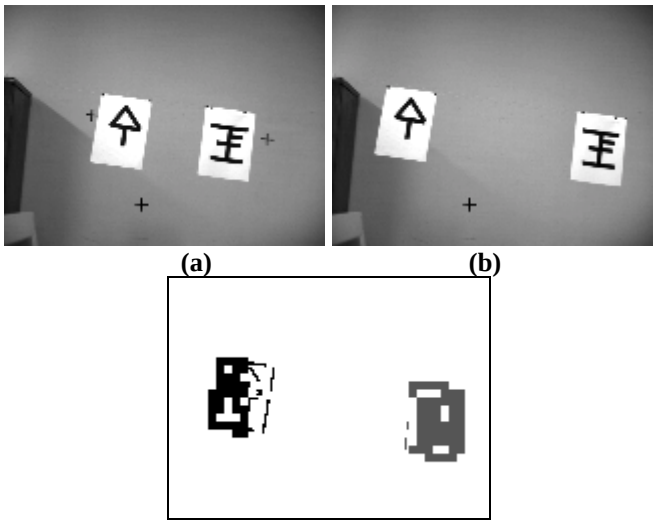
No sort of motion analysis is used to decide whether a merger should be accepted or not. Figure 5 shows that adequate segmentation is achieved for a pair of targets. As they approach one another the previous result blends into one target though. Target attributes and data available for further processing are the number of targets, the density and number of moving pixels in each target, the estimated centroid and the corresponding gray-level patterns around the centroid of each target.

2.3 Monocular velocity estimation

Target velocity on the image plane is a feature that enters the computation of the target interest value. Rapidly moving objects are more likely to leave the visual field before the system has an opportunity to track. In terms of a machine vision system, fast objects may bring about the danger of collision and in terms of biological vision systems, they may represent an agile prey or predator. Hence, one should expect that fast objects should raise more interest than slow ones.

Image velocity estimation is based on target displacement between consecutive images. As a number of targets may exist within the field of view at a certain time, the establishment of target correspondence in consecutive images is required. The correspondence criterion employs a variation of the sum-of-

squared-differences (SSD) method: the sum-of-squared-pixels (SSP) is used for normalization. Let I_t and I_{t-1} represent consecutive images acquired by the dominant camera, R_2 an MxN window in I_t containing a gray-level pattern around a target centroid and R_1 likewise in I_{t-1} .



(a) **(b)**

(c)

Figure 5 - Clustering classes into superclasses and generation of targets. a-b) Consecutive images. c) Resulting targets.

Consider further a search area of size PxQ around a target centroid in I_{t-1} , P>M and Q>N. The need for a search area originates in the displacement of the estimated centroid within the target borders from image to image due to changes in illumination, object pose and fluctuations in the motion detector output. Then, for each combination of R_2 and R_1:

$$SSD'(R_1,R_2)=\min_{a,b}[\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}(R_2(i,j)-R_1(i+a,j+b))^2 / \sum_{i=0}^{M-1}\sum_{j=0}^{N-1}(R_2(i,j))^2]$$

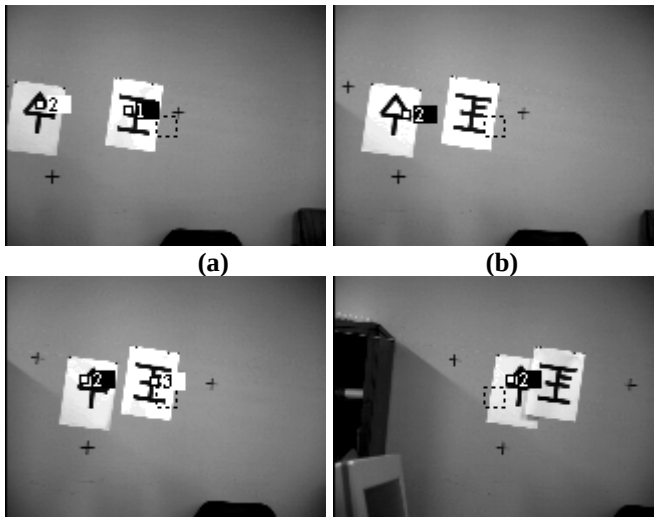
$$a\in\{0,1,...,P-M\},\ b\in\{0,1,...,Q-N\} \qquad (4)$$

The criterion for establishing a corresponding pair of gray-level patterns is:

$$\text{corresp}(R_1^*,R_2^*)=\min_{R_1\in S_1,R_2\in S_2}[SSD'(R_1,R_2)]\le L_{ac} \qquad (5)$$

where $S_1(S_2)$ is the set of remaining $R_1(R_2)$ windows left for correspondence. Whenever a correspondence is established the corresponding $R_1^*(R_2^*)$ is removed from $S_1(S_2)$. The inequality refers to a threshold set to 0.15 for accepting a correspondence as valid. Targets in the most recent image that could not find an acceptable correspondence are labeled as new and their velocities are set to null. A list of corresponding targets is built and the associated velocity estimates are stored for later use by the attentional management module. Empirically tuned window dimensions are M=N=5 and P=Q=21. It is a compromise between the preservation of the information content and the desired robustness to unexpected image variations and to changing background texture around the target. The former precludes too small windows whereas the latter too large ones.

Figure 6 displays an instance of target correspondence between images. Target 2 is properly corresponded along the image sequence until its occlusion. Segmentation produces only one target as both approach one another. Minute motion causes target 1 to be missing in figure 6b. Minute motion, though detected by image subtraction, is further deleted by the gray-level opening operation described in Section 2.1. Target 1 reappears as new target 3 in figure 6c because the correspondence process does not store a gray level pattern for more than one image.



(a) **(b)**

(c) **(d)**

Figure 6 - Monocular SSD-based correspondence along an image sequence.

Targets with similar textures may yield false correspondences and the SSD criterion is known to be susceptible to rotations, geometric deformations and illumination changes (Hager and Belhumeur,1996). Moreover, ambiguities arise due to the aperture problem. The function to minimize in equation (4) generates a surface whose principal curvatures are useful to determine the directions of the largest and the smallest uncertainties when establishing a correspondence between targets. A small curvature indicates large uncertainty and vice versa. Once a correspondence is produced, the target displacement has more uncertainty in its component along the direction of smallest curvature. This effect is often due to a rather homogeneous target texture along this direction. The uncertainty analysis regarding target correspondence will be further elaborated in Section 4 which discusses the results.

2.4 Binocular Registration

Image disparity originates in the image acquisition from distinct points of view of a stereoscopic assembly as shown in figure 7. It depict the baseline b, the 3D point P and its perspective projection Pl (Pr) on the left (right) image plane according to the projection center Ol(Or) and the change on the epipolar constraint caused by verging cameras. Disparity is related to object depth. Depth estimation requires knowledge of parameters such as focal length, baseline and vergence angle. For the purpose of managing the system attention, disparity estimation suffices and normalized correlation-based registration operating on a pyramidal data structure is used. The vergence angle of the non-dominant camera is controlled aiming at nulling the estimated disparity. Ideally, the cameras are oriented in such a way that the projection of the tracked object should fall upon their respective regions of fixation. This control employs correlation which requires a rich texture around the target centroid. Normalization reduces the sensitivity of the correlation method to illumination changes, but the method suffers performance limitations due to rotation and scale factor changes. The use of a pyramidal data structure allows a coarse-to-fine solution to the registration problem

which further helps to reduce the computational workload. A pyramidal data structure is built for each of the stereo images. It consists of three levels of resolution: the original 160x120 (level 0), 40x30 (level 1) and 20x15 (level 2). Levels 1 and 2 are generated by partitioning level 0 into 4x4 and 8x8 windows, respectively, and computing the average gray level at each window.

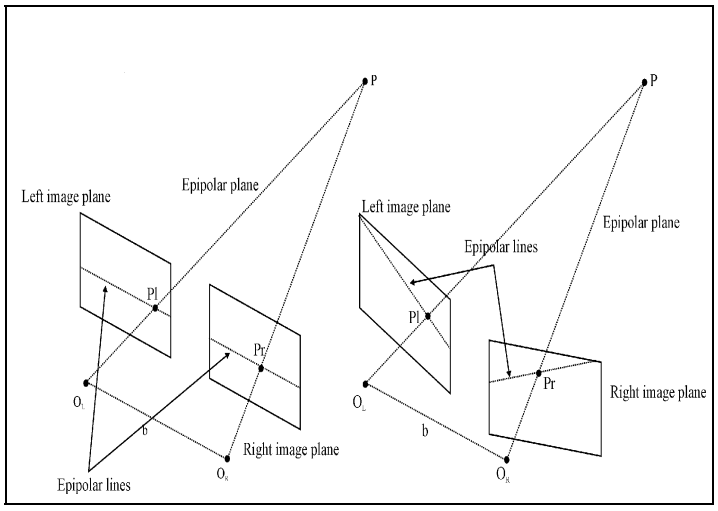


Figure 7 - Effect of camera vergence on the location of the epipolar constraint.

The registration of a target previously detected on the dominant image is made by correlating the gray-level pattern around its computed centroid with patterns within a search area. This area is positioned around the epipolar constraint on the non-dominant image, as seen in figure 8. Due to the vergence angle control for binocular fixation, the epipolar constraint moves and does not coincide with a row of pixels. This particular case only occurs when the cameras are aligned with parallel optical axes. To cope with the varying position of the epipolar constraint on the non-dominant image, a 3-pixel-wide search area in level 2 is defined. The corresponding search area on the non-dominant image in level 0 is 13 pixels wide about the row of pixels which, on the dominant image, contains the target centroid. Thus, the required computational effort is reduced and the rate of correctly estimated disparities is improved.

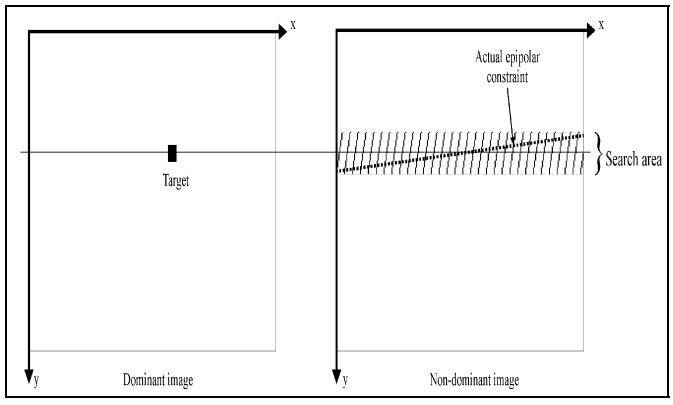


Figure 8 - Search area encompassing the epipolar constraint.

Normalized correlation is as follows:

$$c_{ij,d,e}(f,g) = \frac{\mathrm{cov}(f(i,j), g(i+d, j+e))}{\sigma(f(i,j))\sigma(g(i+d, j+e))} \qquad (6)$$

where f(.,.) and g(.,.) are the dominant and non-dominant images, respectively, cov(.,.) indicates covariance and *s(.)* the standard deviation. The position of a target centroid on the dominant image array is given by indices i,j and the candidate disparity components on the non-dominant image by d,e. The following statistics estimators circumvent redundant computations and yield the correct result when used in equation (6) (Sun, 1997):

$$\mathrm{cov}(f(i,j), g(i+d, j+e)) = \left(\sum_{m=i-K}^{i+K} \sum_{n=j-L}^{j+L} f(m,n) g(m+d, n+e) \right) - (2K+1)(2L+1)\bar{f}\bar{g} \qquad (7)$$

$$\sigma^2(f(i,j)) = \left(\sum_{m=i-K}^{i+K} \sum_{n=j-L}^{j+L} f^2(m,n) \right) - (2K+1)(2L+1)\bar{f}^2 \qquad (8)$$

$$\sigma^2(g(i+d, j+e)) = \left(\sum_{m=i-K}^{i+K} \sum_{n=j-L}^{j+L} g^2(m+d, n+e) \right) - (2K+1)(2L+1)\bar{g}^2 \qquad (9)$$

with $\bar{f}$ and $\bar{g}$ as the average gray level of each pattern. Pattern sizes at each resolution level of the pyramid are 7x7 (K=L=3) at level 0, 5x5 (K=L=2) at level 1 and 3x3 (K=L=1) at level 2. Candidate disparities in level 2 are ordered from the largest to the smallest correlation value computed from equation (6). A coarse-to-fine procedure follows as the best candidate disparity in level 2 is refined at level 1 and subsequently at level 0 by using as initial guess the best disparity found at the previous level with a coarser resolution. The best candidate at level 0 is accepted as the best registration (d^*, e^*) according to the following criterion, which removes the average gray level in order to reduce the sensitivity to illumination changes:

$$e_{ij,d^*,e^*} = \sum_{m=i-2}^{i+2} \sum_{n=j-2}^{j+2} \left(\tilde{e}_f(m,n) - \tilde{e}_g(m+d^*, n+e^*) \right) \le L_{disp} \qquad (10)$$

$$\tilde{e}_f(m,n) = f(m,n) - \frac{1}{25} \sum_{a=i-2}^{i+2} \sum_{b=j-2}^{j+2} f(a,b)$$

$$\tilde{e}_g(m+d^*, n+e^*) = g(m+d^*, n+e^*) - \frac{1}{25} \sum_{a=i+d^*-2}^{i+d^*+2} \sum_{b=j+e^*-2}^{j+e^*+2} g(a,b)$$

If the inequality in equation (10) is violated then the second best candidate disparity of the ordered list at level 2 undergoes the coarse-to-fine procedure and is tested against the above criterion. The process continues until a candidate disparity is accepted. If none passes the inequality then the disparity is labeled void. This label is also employed when a small standard deviation in equation (6) indicates a rather homogeneous texture which does not suffice for the purpose of binocular registration.

2.5 Attentional Management

The presence of multiple targets within the field of view raises the need for an adequate allocation of attentional resources. Some criteria for such allocation are found in the literature. Balkenius and Kopp (1996) employed edge motion intensity to evaluate an attentional field over the image. Culhane and Tsotsos (1992) proposed an attentional model which Andersen (1996) combined with a scale-space representation similar to the image pyramid employed here. A combination of 42 normalized feature maps - contrast, color and edge orientation among them - yielded a saliency map whose peaks were selected via a winner-take-all neural net in the work by Itti *et allii* (1998). Ahuja and Abbott (1993) presented some

experimental results on the psychophysics of human vision which support proximity to the fovea, close fixation points and close range to objects as attentional attractors.

The attentional features available for the i-th target, i=1,...,n, are its velocity magnitude |veloc(i)| on the image plane, number of moving pixels #pix(i), density dens(i) and binocular disparity disp(i) if the latter is not void. They are linearly combined in the attention function I(i) along with a time measurement of how long a target occurs in consecutive images:

$$I(i) = \alpha_1 \#pix(i) + \alpha_2 |veloc(i)| + \alpha_3 dens(i) + \\ + (1 - \alpha_4 timer(i)) + \alpha_5 disp(i) \quad (11)$$

where the weights were tuned during the experiments to the following values:

$$\alpha_1 = 8/(160.120), \quad \alpha_3 = 0.3, \quad \alpha_4 = 0.05;$$

$$\alpha_2 = \begin{cases} 0 \text{ if max\_veloc[pixels/s]} = 0 \\ 1/(3.\text{max\_veloc}) \text{ otherwise} \end{cases}; \quad \alpha_5 = \begin{cases} 0 \text{ if disp(n)} = \text{null} \\ -1/160 \text{ otherwise} \end{cases}$$

$$\text{max\_veloc} = \max_{t=1,2,...,n} |veloc(t)|$$

Tuning was carried out according to the following considerations:

a) The maximum possible value for #pix(.) is the total number of pixels on the image. The factor 8 in the numerator of α_1 provided an adequate balance between the influence of the #pix term and that of the other terms in the attention function. The normalization by max_veloc and the factors 1/3 in α_2 and 0.3 in α_3 follow a comparable justification;

b) The timer term has the purpose of gradually decrementing the interest on old targets. New targets should raise the system attention because they might bring new important information about the evolving scene contents within the field of view;

c) The normalization of α_5 uses the maximum value which can be attained by the horizontal component of a disparity. A target depth smaller (larger) than that of the fixation point in the 3D space yields a negative (positive) disparity. The fixation point is here the 3D point with null disparity because it is, ideally, located at the intersection of the optical axes. The negative sign of α_5 results in a higher attentional value for a near object than that of an object beyond the fixation point. Therefore, the control of the vergence angle for binocular fixation, in spite of altering the disparity magnitudes, still preserves their relative effect on the attention function.

A dynamic list is built with entries being created or deleted as targets on the images appear and disappear. Only the attentional features of the most recent pair of consecutive images are stored in the list to limit its complexity. If a target on the most recent image is not in registration with any target on the previous image then it is labeled a new target. If no motion at all is detected then the attentional features are kept in the list for a maximum of eight consecutive images during which the vision head continues with its smooth pursuit in an attempt to find the target again. Such caution derives from the possibility of target minute motion being filtered out by the grey-level opening operation or by target occlusion. If no motion is detected after that then the system halts its motion and static search resumes.

The target selected for tracking is the one that maximizes equation (11). However, incorrectly detected tiny targets that arise due to errors in background motion compensation and unsuccessful filtering by the opening operation often possess a high density. Such is the case of a 1-pixel false target. Moreover, when such a case occurs, the respective timer term in equation (11) attains its maximum value. To cope with such an artifact it is required that the selected target occurs three times, not necessarily on consecutive images, before the vision head performs a ballistic motion to reorient and resume tracking. If no target is in agreement with this constraint it is reduced to two occurrences and, if necessary, to only one occurrence. This means that the restriction is temporarily either softened or disabled. In either situation, the target maximizing equation (11) is selected but not tracked. Selection in spite of a no-track status is required to update the state of a centroid position filter which smooths the head motions during monocular tracking and binocular fixation.

2.6 Binocular Fixation

Binocular fixation aims at correctly orienting the cameras in tilt and vergence in such a way that the selected target stays within a 10x10 region of fixation about the center of the stereo images. Control of either the camera angular position or its velocity depends on the type of motion which is engaged.

Ballistic motion occurs when the camera attitude should rapidly change as the attentional management selects a new target for tracking. It consists of a very fast angular displacement with a strong acceleration at the motion onset and offset. It is activated only after a newly selected target occurs three times. As mentioned in the previous Section, it helps to reduce the impact of image noise and of background compensation errors. Overshoot and oscillations at motion onset and offset result in image blur and erroneous encoder reading which prevents image processing during ballistic motion. Image processing is resumed at motion completion. Figure 9 shows the tilt and vergence angles θ and γ, respectively, the error components e_x and e_y, the focal distance f_u in pixels and the projection center O, all of which enter the computation of the angular increment to be attained by a ballistic motion.

Required angular increments in tilt and vergence in encoder units are given by:

$$inc\_pos\_verg = 180\gamma/(166.67\pi);$$
$$inc\_pos\_tilt = 180\theta/(44.4\pi);$$
$$\gamma = tg^{-1}(e_x/f_u); \quad \theta = tg^{-1}(e_y/f_u) \quad (12)$$

Saccades are corrections to the camera angular velocity so that the selected target is ideally kept within the region of fixation. A strong acceleration occurs at the motion onset and offset but image processing is not interrupted. Required increments to angular velocities in tilt and vergence are computed as:

$$inc\_vel\_verg = inc\_pos\_verg/(q.\Delta t);$$
$$inc\_vel\_tilt = inc\_pos\_tilt/(q.\Delta t) \quad (13)$$

where q is the specified number of frames ahead and Dt the average time interval between frames. The computation of the increments to angular velocity requires the specification of the

number of frames ahead in time to fulfill the required angular increments.

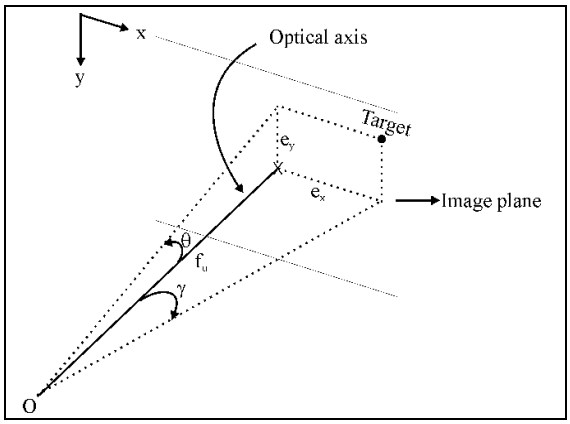


Figure 9 - Tracking error on the image plane and required angular increments.

Smooth pursuit maintains the camera's angular velocity as long as the selected target stays within the region of fixation. Ideally, this target stabilization should produce a sharper image of the tracked target. The size of the region of fixation cannot be too large because large image displacements would result during tracking. The justification for that is that a large target motion would be allowed before the engagement of a correcting saccade. The non-dominant camera does not perform smooth pursuit. A saccade correcting its angular velocity is performed after each image acquisition. Whenever the disparity of the tracked target is void, binocular fixation assumes the last valid disparity in spite of changing vergence angles.

Fluctuations in the output of the target segmentation process due to illumination changes and background texture cause the computed centroid of a target to suffer displacements which are not related to its motion. However, attentional management copes with it as a target slipping away of the region of fixation. To reduced this undesired effect, the centroid coordinates of each target on the dominant image is filtered by:

$$cg\_f(k) = 0.6cg(k) + 0.3cg(k-1) + 0.1cg(k-2) \qquad (14)$$

where k indexes an image in a sequence. If the last three occurrences of the selected target are not consecutive or less than three occurrences are observed then the system resorts to the unfiltered computed centroid for tracking purposes.

3 SYSTEM IMPLEMENTATION

The binocular tracking system was evaluated with the Helpmate Robotics Bisight vision head, nicknamed Otelo, at the ITA/INPE Active Computer Vision and Perception Lab - LVCAP-ITA/INPE - (http:\\www.ele.ita.cta.br/~labvisao/) described in Viana *et allii* (1999). Image processing and the computation and transmission of command signals to Otelo were carried out by a 166Mhz Pentium MMX PC equipped with 64 MB RAM, 2 GB SCSI hard-disk, a DataTranslation DT3152 monochrome video acquisition board and two AD-DA conversion boards. The latter exchanges data with the drives that control the Fujinon KPM1 servoactuated lenses. Off-line calibration produced the estimate f_u=221.8 pixels (Viana *et allii*, 1999). Targets were made with paper and black tape to provide texture. The targets underwent translation by means of a setup of wires and pulleys. The physical dimensions of the lab precluded experiments with significant motion along the depth direction and thus a more effective evaluation of the disparity term in the attention function. Eventually, moving people were included in the experiments.

The system was developed on the Windows for Workgroups 3.11 platform and coded in three distinct programming languages: the graphic user interface in Visual Basic 3.0, the tracking routines in the Visual C++ 1.51 16 bit environment and the most computationally intensive image processing routines in 80486 Assembly and compiled with Turbo Assembler 5.0. For comparison purposes, image subtraction for motion detection took about 11 ms when coded in C and 1 ms in Assembly. The application of the gradient operator on a 160x120 image consumed about 38 ms when coded in C and 11 ms in Assembly. Routines coded in Assembly are: image subtraction, gray-level opening, thresholding, logical ANDing of binary images, the generation of foveated images and normalized correlation. Image capture consumed 130 ms as the only acquisition board first selects a camera and then acquires the image for each camera of the stereo assembly.

Table 1 depicts typical time intervals elapsed when the main processing tasks are performed. Binocular fixation shows significant variations that depend on the performed motion. Ballistic motion requires a full stop by the vision head before resuming image processing. It is followed by a saccade which issues an angular velocity update command to the head drive. The host computer awaits a confirmation that the command was successful to resume the image processing. Finally, smooth pursuit by the dominant camera includes a saccade by the non-dominant camera.

Static search (ms)	553
Monocular motion localization and target segmentation (ms)	721
Monocular estimation of target velocity (ms)	195
Binocular registration (ms)	215
Attentional management	< 1
Binocular fixation (ms)	ballistic motion: 709 saccade: 177 smooth pursuit: 65

Table 1 - Typical computational workload of the main processing tasks.

Static search for motion and monocular target segmentation, in spite of mostly implemented in Assembly, are in a listed entry which includes the acquisition of stereo images and encoder reading. The latter required issuing a command to the head drive and await until the measurement became available. The frame rate attained during operation was about 0.8 stereo frame pair/second due to limitations of both off-the-shelf hardware and the software platform. This rate severely limited interframe target motion during the experiments but did not invalidate the assessment of the overall system potential.

4 RESULTS AND ANALYSIS

Performance analysis is presented both quantitatively and qualitatively. A qualitatively successful performance is characterized by keeping the selected target within the field of view of both cameras. A proposal for a quantitative performance assessment requires the definition of a metric which should be independent of a particular system implementation. The proposed metric is the lock-on rate P within a 30x30 test window about the center of the stereo images which is given by:

$$P = \frac{N_c}{N_t} \quad (15)$$

where N_c represents the number of frames in which the tracked target centroid is within the test window and N_t the number of frames in which this target is selected for tracking. The latter does not include the three occurrences before which a ballistic motion is not engaged for attentional reorientation.

Four experiments were devised. They consisted of stereo sequences lasting 120s. Experiment A consisted of a rigid target under mainly horizontal translation. Experiment B consisted of two translating rigid targets aiming at an evaluation of the attentional management process. Experiment C consisted of a moving person which led to various targets undergoing non-rigid motion. Finally, experiment D added a rigid target to the latter experiment. Excerpts of the stereo images acquired in the experiments are found in Appendix B. Each detected target received an identification tag for as long as it was tracked. The numerical tag on the selected target is white over a black blackground whereas other targets have their tags with inverted gray levels. Image pairs corresponding to dominant and non-dominant cameras are displayed right and left, respectively. The 10x10 region of fixation and the 30x30 test window are delimited in the display by a black trace and a white one, respectively, about the dominant camera image center. Since the non-dominant camera does not undergo smooth pursuit, no region of fixation is displayed on its image.

Frame rate is affected mainly by the elapsed time required to complete a ballistic motion, saccade or smooth pursuit and by the number of moving pixels which form the targets. Generally, good results were obtained in sequence A. Sequence B presented less accurate tracking mainly due to occasional failures of the SSD-based target correspondence along the sequence.The occurrence of numerous targets - caused by rotation and non-rigid motion - in sequences C and D caused disputes for the system attention which resulted in a rather erratic behavior. Such performance degradation was caused by significant interframe displacements as moving people went in front of the cameras with their gait. Dealing with this condition pushed the system to the limits of its implementation. The subjects were maintained within the field of view during most of the sequences though. The workload of sequentially processing the images and the sluggish vision head control cycle are the main reasons for the present bottleneck. The control cycle initiates with the issuing of commands from the host computer to the vision head drives and ends with the emission by the drive to the host computer of a message that the requested command was accomplished. Table 2 presents the number of stereo pairs in each experiment and the equivalent stereo pair rate. Uhlin (1996) and Andersen (1996) reported image processing and head control with special-purpose hardware at a rate of 25 and 10 frames per second, respectively.

	Stereo pairs	Rate (stereo pairs per second)
Experiment A	97	0.8
Experiment B	106	0.9
Experiment C	78	0.6
Experiment D	78	0.6

Table 2 - Average operating rate during each experiment. (120s-long stereo sequences)

In experiment A, the SSD-based method for corresponding targets along the sequence of dominant images presented incorrect results. The only target on the image was in two occasions labeled as new and three targets were produced. Correlation-based binocular registration was more susceptible: 13 stereo pairs yielded void disparities and other three stereo pairs produced incorrect disparities due to similarities between the target texture and that of the background. As L_{disp} in equation (10) allowed for an average mismatch magnitude of 25 gray levels per pixel, it is claimed that correlation-based registration seems to be quite sensitive to the experiment conditions in spite of its normalization. It could have been caused by the combination of the changing vergence angles and a low frame rate, which could translate into illumination differences, changes in texture projection onto each camera and a varying scale factor.

Experiment B resulted in 23 detected targets. One reason for that is that the system has no memory of the targets: whenever one leaves and later reenters the visual field it is labeled as new. Another factor was erratic operation of the SSD-based method when a target had its computed centroid near its boundary with the background. In such a case, some of the background texture entered the gray-level pattern window R_1 or R_2. The target motion affected the appearance of the surrounding background texture as occluded background became uncovered. The experiment showed that the SSD-based method is effective when a target with a rich texture moves over a background with poor texture.

Experiment C resulted in 49 detected targets. The SSD criterion showed good results in spite of some targets undergoing a motion other than translation, such as target 3 in C.5-C.14. Human subjects moving in the visual field often produced a large number of targets. For instance, three targets were detected in C.8-C.10 because the subject projection on the image was large and its more distant portions moved non-rigidly. Were the target small and its motion would yield a unique target in spite of being non-rigid.

Artifacts in experiment D were caused by background compensation error. They were observed as targets 8 and 9 in D.17. Sensitivity to occlusion in subtraction-based motion detection produced targets 47 and 48 in D.75.

Vergence control of the non-dominant camera successfully maintained the selected target within the field of view in spite of inaccurate correlation-based binocular registration. The quantitative analysis in Tables 3-6 is based on the metric described in equation (15). The intermediate results are expressed for those targets in each experiment that raised the system attention the most and were thus tracked for more frames.

	Target 1	Target 2	Target 3
Target detection (frames)	63	7	22
Correct binocular registration (stereo pairs)	54	6	18
Target selected and tracked, N_t (frames)	60	4	19
Centroid within the test window (dominant / non-dominant), N_c (frames)	47/42	4/3	17/13
Lock-on rate (dominant/non-dominant), P (%)	78/70	100/75	89/68

Table 3 - Quantitative analysis. Experiment A.

	Target 6	Target 8	Target 22
Target detection (frames)	10	22	19
Correct binocular registration (stereo pairs)	7	18	17
Target selected and tracked, N_t (frames)	6	19	16
Centroid within the test window (dominant / non-dominant), N_c (frames)	5/4	19/5	9/8
Lock-on rate (dominant/non-dominant), P (%)	83/67	100/26	56/50

Table 4 - Quantitative analysis. Experiment B.

	Target 3	Target 19	Target 28
Target detection (frames)	16	12	15
Correct binocular registration (stereo pairs)	7	2	5
Target selected and tracked, N_t (frames)	8	4	10
Centroid within the test window (dominant / non-dominant), N_c (frames)	0/0	0/0	0/0
Lock-on rate (dominant/non-dominant), P (%)	0/0	0/0	0/0

Table 5 - Quantitative analysis. Experiment C.

P was often smaller for the non-dominant image because of void disparities and the acceptance of incorrect ones. Large oscillations in the attentional feature estimates produced frequent attentional shifts as can be seen in C.6-C.12. Fluctuations in the estimated location of a target centroid and the incorrect correspondence by SSD-based method - caused by a similar texture - are seen in target 20 in D.29-D.30. Such occurrences negatively impact on the tracking accuracy of targets undergoing non-rigid motion. The above support the claim that, as it is implemented, the system is not able to consistently maintain within the test window a nearby human subject with its characteristic gait which is a highly non-rigid motion. In experiment D an additional small target undergoing translation was tracked in a much smoother way than those targets that originated in the non-rigid motion of the human subject. To overcome this limitation, a higher image processing rate and a faster control cycle - encompassing command issuing, head motion control and sensor reading tasks which ultimately integrate the host computer and the vision head drive - are required. The former encourages a biologically-inspired parallellization of the vision algorithms whereas the latter depends on major changes being made to the drive hardware that presently controls the vision head. After such modifications, a target slipping away of the region of fixation would be expected to be detected before reaching a significant displacement. Furthermore, it would keep the correcting saccade amplitude small as well as improve the accuracy and robustness of both correlation and SSD-based methods.

	Target 20	Target 21
Target detection (frames)	8	14
Correct binocular registration (stereo pairs)	4	7
Target selected and tracked, N_t (frames)	5	9
Centroid within the test window (dominant/non-dominant), N_c (frames)	1/1	6/5
Lock-on rate, P (%)	20/20	67/56

Table 6 - Quantitative analysis. Experiment D.

The performance of the attentional management was impacted by the quantity and quality of the information extracted by the vision algorithms. Figure 10 shows an instance of two targets competing for the system attention. Frequent attentional reorientation as observed in C.6-C.11 are shown in figure 11 along with the contribution of each term to the attention function in equation (11). Undesired reorientation occurred because of large variations in the estimates of attentional features caused by a low frame rate. The latter can be traced back to the emphasis put on the use of off-the-shelf unexpensive hardware, on the current sequential implementation of the system and on the vision head control cycle. Attentional management is expected to become more accurate as image processing rate increases.

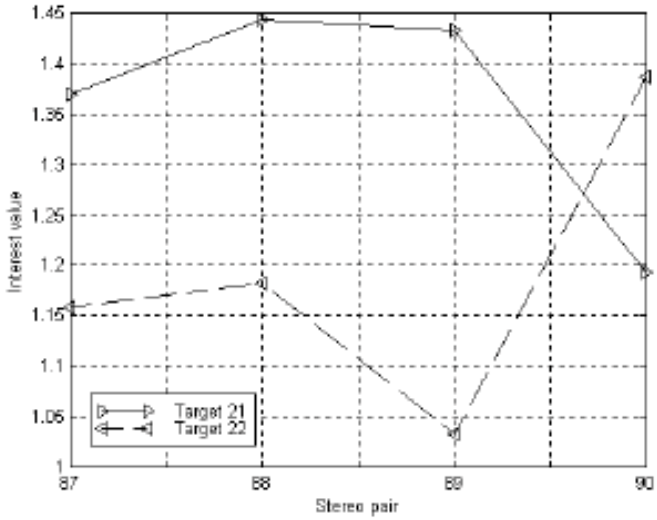


Figure 10 - Targets 21 and 22 in stereo pairs B.87-B.90 in a dispute for system attention.

Section 2 briefly mentioned the uncertainties that arise in target correspondence due to the aperture problem. The computation of either the SSD/SSP or the normalized-correlation criteria generates a surface of which the search area is its domain of minimization. The aperture problem is characterized by a lack of information which translates into a small surface curvature along a direction which is significantly aligned with the texture pattern. The analysis of the surface curvature is then useful to locate the directions along which uncertainty is either at its maximum or its minimum. Assuming that the surface generated by either criterion can be approximated by a quadratic f(x,y) in the neighborhood of the computed minimum at coordinates x_c,y_c , the principal directions along which the curvature is at a maximum or a minimum are given by the eigenvectors of the Hessian:

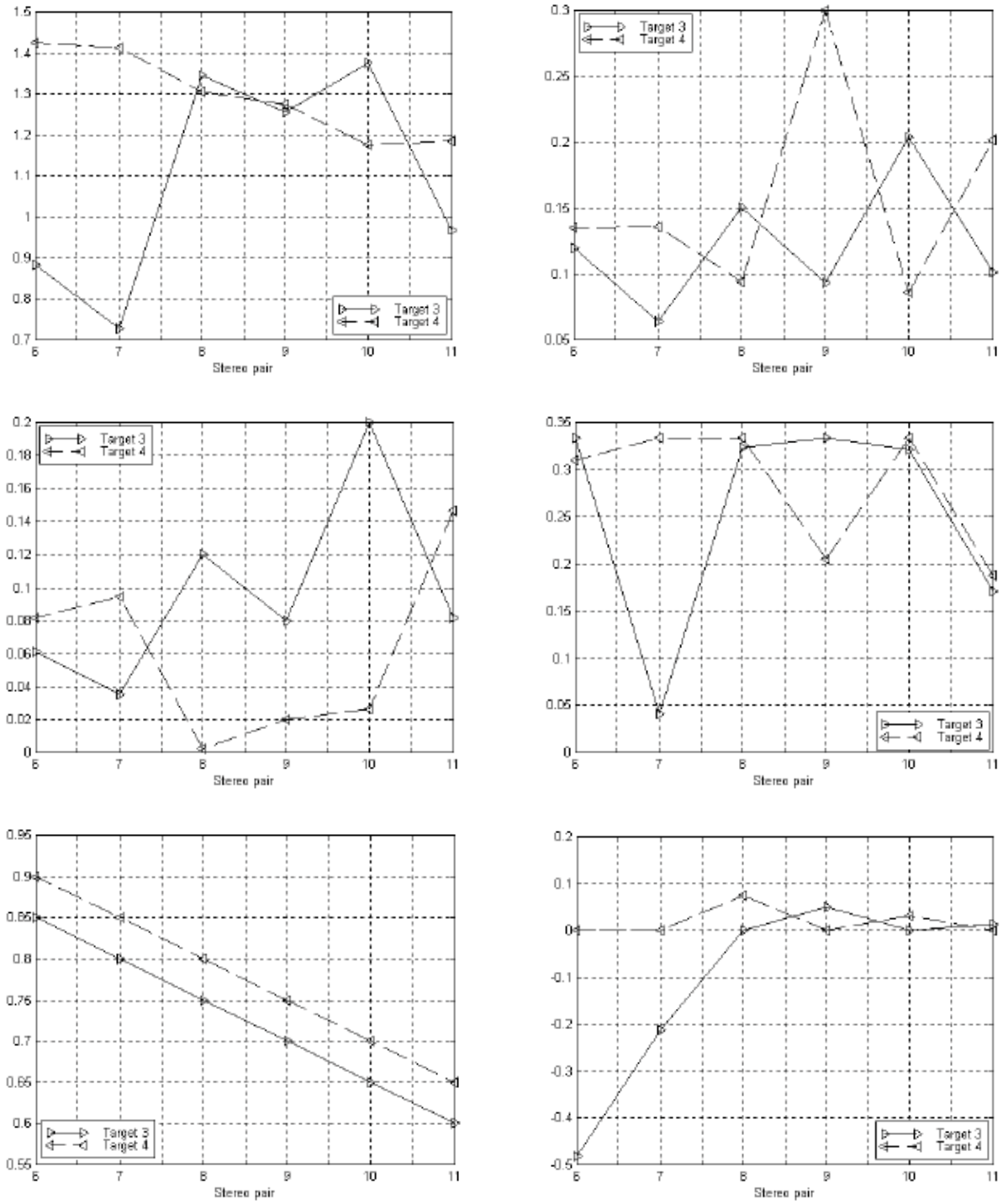


Figure 11 - Targets 3 and 4 in stereo pairs C.6-C.11 in a dispute for the system attention. From top to bottom: a) Interest value. b) Density. c) Number of moving pixels. d) Velocity magnitude. e) Timer term. f) Binocular disparity.

$$\mathbf{H}(x_c, y_c) = \begin{bmatrix} \frac{\partial^2 f(x,y)}{\partial x^2} & \frac{\partial^2 f(x,y)}{\partial x \partial y} \\ \frac{\partial^2 f(x,y)}{\partial y \partial x} & \frac{\partial^2 f(x,y)}{\partial y^2} \end{bmatrix} \text{evaluated at } x_c, y_c \qquad (16)$$

The largest eigenvalue l_1 indicates the maximum curvature (smallest uncertainty) along the direction of its associated eigenvector **u** and likewise for the smallest eigenvalue l_2 , the minimum curvature (largest uncertainty) and associated eigenvector **v**. Hence:

$$\lambda_1(\mathbf{u}.\mathbf{v}) = \lambda_1 \mathbf{u}.\mathbf{v} = \mathbf{H}\mathbf{u}.\mathbf{v} = (\mathbf{u}^T\mathbf{H}^T)\mathbf{v} = \mathbf{u}^T(\mathbf{H}^T\mathbf{v}) = \mathbf{u}.(\mathbf{H}^T\mathbf{v}) \qquad (17)$$

where the dot represents the inner product operator. Because of the assumption of the quadratic approximation, f(.,.) is continuous and smooth and thus:

$$\frac{\partial^2 f(x,y)}{\partial x \partial y} = \frac{\partial^2 f(x,y)}{\partial y \partial x} \Rightarrow \mathbf{H}(x,y) = \mathbf{H}^T(x,y) \qquad (18)$$

and because **H**(.,.) is symmetric, from equation (17) the following holds:

$$\mathbf{u}.(\mathbf{H}^T\mathbf{v}) = \mathbf{u}.(\mathbf{H}\mathbf{v}) = \mathbf{u}.(\lambda_2\mathbf{v}) = \lambda_2(\mathbf{u}.\mathbf{v}) \qquad (19)$$

Assuming further, without any loss of generality in the situations usually encountered, that $\lambda_1 \neq 0$, $\lambda_2 \neq 0$ and $\lambda_1 \neq \lambda_2$, the subtraction of equations (17) and (19) yields:

$$(\lambda_1 - \lambda_2)(\mathbf{u}.\mathbf{v}) = 0 \;\Rightarrow\; \mathbf{u}.\mathbf{v} = 0$$

and thus the directions of maximum and minimum uncertainty are orthogonal under the above assumptions. The partial derivatives in equation (16) are approximated with discrete operators. Now consider, for instance, two consecutive dominant images on which two moving targets are detected as in figure 12. Target correspondence is required to estimate the velocity of each one.

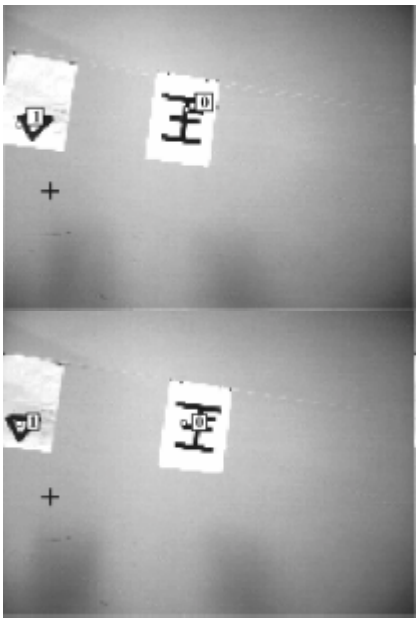


Figure 12 - Consecutive images I_{t-1} and I_t for uncertainty analysis.

Figure 13 shows the global minimum of SSD/SSP surfaces generated for each possible correspondence with a '*'-mark. Table 7 depicts the minima attained for each possibility. The search for the minimum furnishes the correct result up to fluctuations which occur in the computation of the centroid location.

The eigenvectors for the correct correspondences are in figure 14. By describing the interframe displacement according to the eigenvector base, a curvature-based confidence measure in each direction can be used to adaptively change the weights that multiply the velocity and disparity terms in the attention function (equation (11)) and thus reduce their importance when the available information is of poor quality. Moreover, the conjugation of such confidence measure with the minimization procedure seems encouraging to disambiguate non-unique correspondence solutions.

Folk and *allii* (1992) claimed that the exogenous (involuntary) allocation of attention is not the result of relatively inflexible, "hard-wired" mechanisms triggered by specific stimulus properties. Such type of allocation could be configured or set to respond selectively to a property that signaled the location of stimuli that were relevant to task performance. Such configuration was called the "attentional control setting", an endogenous (voluntary) control factor analogous to the tuning of the attention function which drives the system behavior during tracking in the present work. The setting was assumed to be determined by current behavioral goals. Once this setting was established, events that exhibited the critical properties summoned attention involuntarily, whether or not the events were relevant to task performance. Stimuli not exhibiting such properties would not summon attention. Involuntary (undesired) shifts would be thus mediated by tunable attentional control settings which would vary according to current behavior goals. In the computational model of attention

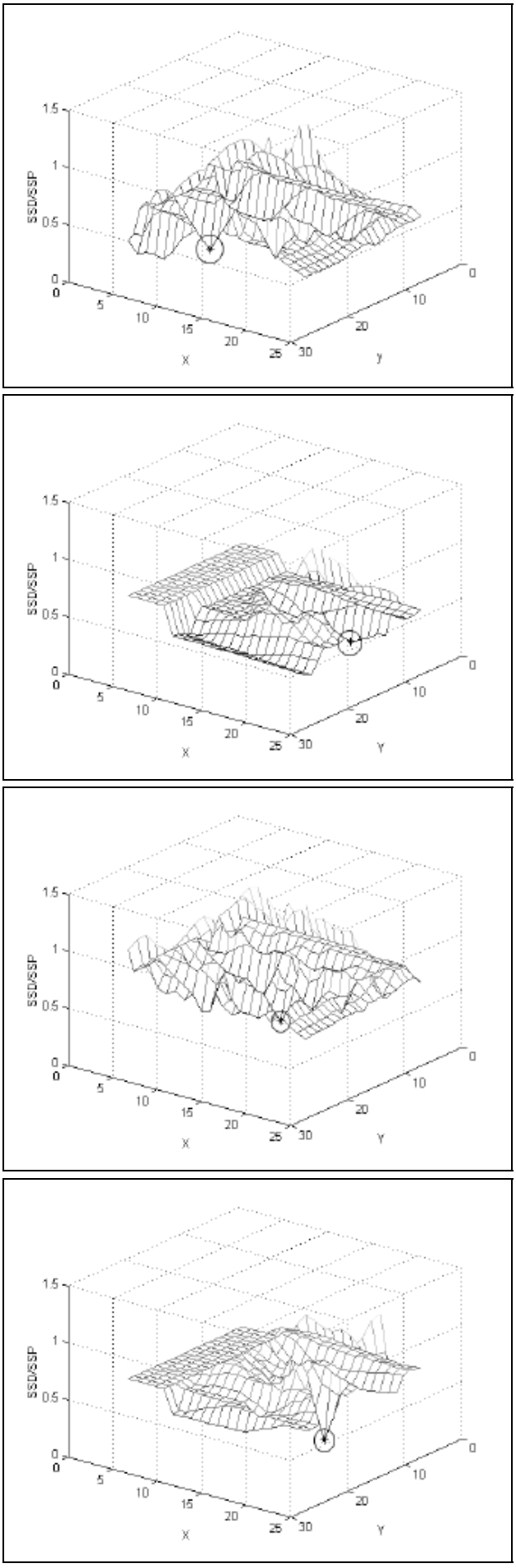


Figure 13 - Surfaces generated by equation (2.6.1) with computed minimum for each possibility of correspondence. From top to bottom: a) Target 0 in I_t with target 0 in I_{t-1}. b) 0 in I_t with 1 in I_{t-1}. c)1 in I_t with 0 in I_{t-1}. d) 1 in I_t with 1 in I_{t-1}.

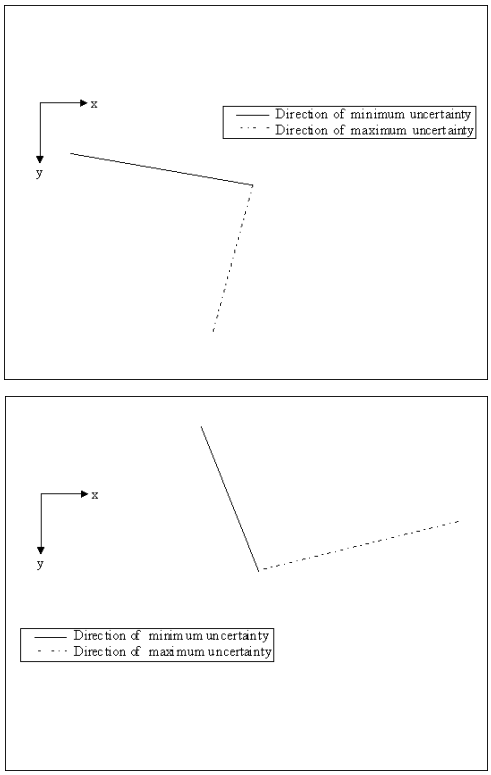


Figure 14 - Directions of maximum and minimum uncertainties. a) Target 0 in I_t with target 0 in I_{t-1}. b) 1 in I_t with 1 in I_{t-1}.

proposed by Koch & Ullman (1985), attention would be allocated to the location with the highest activation in a saliency map. Cave and Wolfe (1990) claimed that the strength of the local saliency should depend on bottom-up featural dissimilarity relative to a neighborhood and also on top-down influence dictated by current behavioral goals. The top-down influence should increase the salience signals at locations that contained task-relevant stimulus properties. The interrelation between endogenous attentional control settings (and its analogous in this work - the setting of the attention function to the conditions of the laboratory experiments with the purpose of producing an acceptable system behavior) and exogenous involuntary allocation of attention (which, similar to this work, shows as an apparently inadequate behavior, such as when attention is allocated to undesired locations which nevertheless present the expected stimuli) defines a dichotomy that possibly represents the delicate and yet efficient balance between the necessary rigidity to ensure that potentially important environmental events do not go unprocessed and the flexibility to adapt to changing circumstances and goals.

5 CONCLUSIONS

This paper describes the development and evaluation of a binocular vision tracking system augmented with an attentional management capability. It selects one among multiple targets according to the relative values achieved by an operator-tuned attention function. The main contribution is the augmentation of the monocular visual tracking concept as proposed in Murray and Basu (1994) with an attentional management capability and binocular fixation. Foveated images are used to reduce the computational workload inherent to detecting and segmenting multiple moving targets. The integration and experimental evaluation, both quantitatively and qualitatively, of various vision algorithms such as motion detection and target segmentation, SSD- and correlation-based registration along a sequence and across stereo pairs, attentional management and binocular fixation, all aiming at the extraction of information useful for controlling the head motion, are not often found in the literature. Recent related work is found in Uhlin (1996), Andersen (1996), Araújo *et allii* (1996ab), Batista *et allii* (1997), Eklund *et allii* (1995) and Molton (1998). Each employs distinct hardware configurations and experimental setups which make a comparative performance evaluation rather complicated. For instance, the dedicated hardware employed in Uhlin (1996) and Andersen (1996) was essential to attain the reported 25 and 12 frames per second, respectively, whereas 0.8 stereo pairs per second is reported here. The sequential implementation described here is a consequence of the emphasis put on the use of unexpensive off-the-shelf PC-based equipment. Eventually, such an option imposed limitations in terms of frame rate and the ensuing target motion that could be adequately dealt with by the system.

Target in I_t	Target in I_{t-1}	SSD/SSP
0	0	0.0016
0	1	0.1195
1	0	0.2756
1	1	0.0262

Table 7 - Minimum values of the SSD/SSP for each possible correspondence.

System performance is acceptable when coping with small targets undergoing translation but suffers degradation when large objects undergoing non-rigid motion enter the visual field such as the gait of human subjects moving at close range. The attentional management as proposed here displays undesired attentional reorientations due to interframe changes in the attentional features which are caused by a combination of varying illumination, scaling factor changes and rotation. Such changes cause a lock onto a different target whose features appear more similar to those of the target previously selected. The compromise between an excessive specialization by fine-tuning the attention function and an adequate flexibility to provide for the processing of new events that present a potential to raise the system interest is discussed. The analysis of the uncertainties in the estimation of attentional features and the potential of adaptively learning the attention function-either of the weights in equation (11), or in an even broader sense, of its structure - as goals change and perceptions of the environment are gathered seem to combine in a fertile field of research on mechanisms that aim at providing a robotic system with an active vision capability with true autonomy to manage its limited computational resources.

Acknowledgements:

The authors wish to acknowledge the support of Dr.Antonio Francisco Jr., the partnership with the Instituto Nacional de Pesquisas Espaciais (INPE) and the funds provided by FAPESP to the joint ITA/INPE research efforts conducted at the Active Vision and Perception Laboratory (FAPESP project no.0620/95).

6 REFERENCES

Ahuja,N. and Abbott,A.L.(1993). Active Stereo: Integrating Disparity, Vergence, Focus, Aperture and Calibration for Surface Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.15, no.10, 1007-1029.

Andersen,C.S.(1996). A Framework for Control of a Camera Head. Ph.D. dissertation, Laboratory of Image Analysis, Institute of Electronic Systems, Aalborg University

Araújo,H.;Batista,J.;Peixoto,P. and Dias,J.(1996). Gaze Control in a Binocular Active Vision System Using Optical Flow. *Proceedings of the 4th International Symposium on Intelligent Robotics Systems.*

Araújo,H.;Batista,J.;Peixoto,P. and Dias,J.(1996). Pursuit Control in a Binocular Active Vision System Using Optical Flow. *Proceedings of the 13th International Conference on Pattern Recognition*, 313-317.

Balkenius,C. and Kopp,L.(1996). Visual Tracking and Target selection for Mobile Robots. *IEEE Proceedings of EUROBOT'96*, 166-171.

Bandera,C.;Vico,F.J.;Bravo,J.M.;Harmon,M.E. and Baird III,L.C.(1996). Residual Q-Learning Applied to Visual Attention. *Proceedings of the 13th International Conference on Machine Learning*, 20-27.

Bar-Shalom,Y. and Li,X.-R.(1993). Estimation and Tracking: Principles, Techniques and Software, Artech House.

Batista,J.;Peixoto,P. and Araújo,H.(1998). Real-Time Active Visual Surveillance by Integrating Peripheral Motion Detection with Foveated Tracking. *Proceedings of the 1998 IEEE Workshop on Visual Surveillance*, 18-25.

Batista,J.;Peixoto,P. and Araújo,H.(1997). Visual Behaviors for Real-Time Control of a Binocular Active Vision System. *IFAC Algorithms and Architectures for Real Time Control AARTC'97.*

Bispo,E.M. and Waldmann,J (1998). Saccadic Motion Control for Monocular Fixation in a Robotic Vision Head: A Comparative Study, *Journal of the Brazilian Computer Society - Special Issue on Robotics*, vol.3, no.4, 61-69.

Cave,K.R. and Wolfe,J.M.(1990). Modelling the Role of Parallel Processing in Visual Search. *Cognitive Psychology*, no.22, 225-271.

Cowie,R. and Taylor,J.(1997). The Moving Eye: A Model for the 21st Century Sensor? *Proceedings of the 13th International Conference on Digital Signal Processing*, 255-260.

Crisman,J.D.;Cleary,M.E. and Rojas,J.C.(1998). The Deictically Controlled Wheelchair. *Image and Vision Computing*, vol.16, 235-249.

Culhane,S.M. and Tsotsos,J.K.(1992). A Prototype for Data-Driven Visual Attention. *Proceedings 11th International Congress on Pattern Recognition, 36-40.*

Dagless,E.L.;Ali,A.T. and Cruz,J.B.(1993). Visual Road Traffic Monitoring and Data Collection. *Proceedings of the IEEE-IEE Vehicle Navigation and Information Systems Conference*, 146-149.

Eklund,M.W.;Ravichandran,G.;Trivedi,M.M. and Marapane,S.B.(1995). Adaptive Visual Tracking Algorithm and Real-Time Implementation. *IEEE International Conference on Robotics and Automation*, 2657-2662.

Fairhurst,M.C.(1988). Computer Vision for Robotic Systems: An Introduction. Prentice Hall.

Folk,C.L.;Remington,R.W. and Johnston,J.C.(1992). Involuntary Covert Orienting Is Contingent on Attentional Control Settings, *Journal of Experimental Psychology: Human Perception and Performance*, vol.18, no.4, 1030-1044.

Hager,G.D. and Belhumeur,P.N.(1996). Real-Time Tracking of Image Regions with Changes in Geometry and Illumination. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 403-410.

Huber,E. and Kortenkamp,D.(1995). Using Stereo Vision to Pursue Moving Agents with a Mobile Robot. *IEEE International Conference on Robotics and Automation*, 2340-2346.

Itti,L.;Koch,C. and Niebur,E.(1998). A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligenc*e, vol.20, no.11,1254-1259.

Koch,C. and Ullman,S.(1985). Shifts in Selective Attention: Toward the Underlying Neural Circuitry. *Human Neurobiology*, vol.4, 219-227.

Molton,N.;Se,S.;Brady,J.M.;Lee,D. and Probert,P.(1998). A Stereo Vision-Based Aid for the Visually Impaired, *Image and Vision Computing*, vol.16, 251-263.

Murray,D. and Basu,A.(1994). Motion Tracking with an Active Camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.16, no.5, 449-459.

Rotstein,H. and Rivlin,E.(1996). Control of a Camera for Active Vision: Foveated Vision, Smooth Tracking and Saccade. *Proceedings of the 1996 IEEE International Conference on Control Applications*, 691-696.

Scott,P.D.;Bandera,C.(1990). Hierarquical Multiresolution Data Structures and Algorithms for Foveal Vision Systems. *IEEE International Conference on Systems, Man and Cybernetics*, 832-834.

Shi,J. and Tomasi,C.(1994). Good Features to Track. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 593-600.

Sun,C.(1997). A Fast Stereo Matching Method. *Digital Image Computing: Techniques and Applications*, New Zealand, 95-100.

Uhlin,T.(1996). Fixation and Seeing Systems. Ph.D. dissertation, ISRN KTH/NA/P--96/10--SE ISSN 1101-2250:TRITA-NA-P96/10, Department of Numerical Analysis and Computing Science, Stockholm University.

Viana,S.A.A.;Waldmann,J. and Caetano,F.F.(1999). Non-Linear Optimization-Based Batch Calibration with Accuracy Evaluation, *Revista Controle & Automação - Special Edition on Computational Vision*, vol.10, no.2, 89-99.

Weiman,C.F.R. and Vincze,M.(1996). A Generic Motion Platform for Active Vision. *SPIE International Symposium on Intelligent Systems and Advanced Manufacturing, Robotics and Intelligent Systems*, USA.

Appendix A: Algorithm for clustering classes into superclasses and target formation

For all i,j; i=1,...,n and j > i:

{ if($C(c_i,c_j)$)then

if($\exists\ k < i \mid c_i \subset Sc_k$)then

{$Sc_k=Sc_k \cup c_j$; /* merges c_j with superclass which c_i belongs to */

if($\exists\ n < i;\ n \neq k \mid c_j \cup Sc_n$)then

{$Sc_k=Sc_k \cup Sc_n; Sc_n=\{\varnothing\};$}/*merges superclasses containing classes*/

}

else if($\exists k < i \mid c_j \subset Sc_k$)then

$Sc_k=Sc_k \cup c_i$; /* merges c_i with superclass which c_j belongs to */

else if ($Sc_i=\{\varnothing\}$)then $Sc_i=c_i \cup c_j$;/* initiates superclass */

else $Sc_i = Sc_i \cup c_j$; /* merger with c_j */

else if(j=n and $Sc_i=\{\varnothing\}$)then $Sc_i=c_i$; /*superclass with isolated class*/

}

n=0; /* target count */

$\forall Sc_k \neq \{\varnothing\}$, k=1,...,n

{n=n+1; /*n-th target - moving pixels, density, centroid location*/

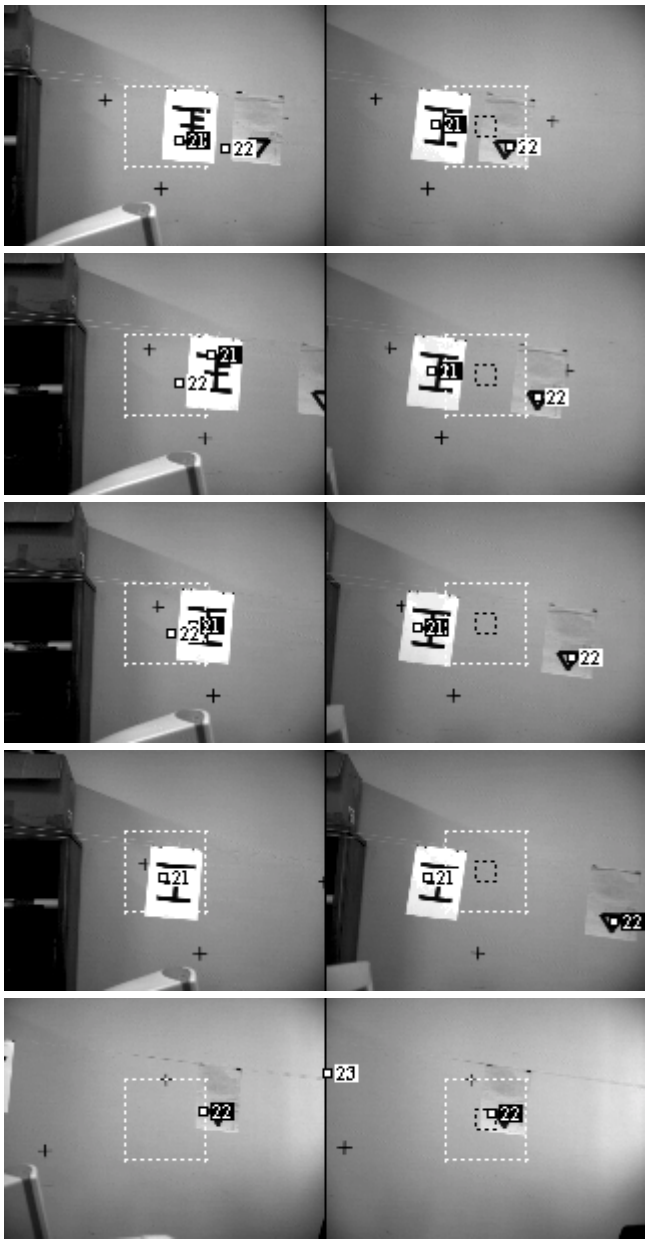
$$\#pix(n) = \sum_{c_i \subset Sc_k} \#pix(c_i)\ ;\quad dens(n) = \frac{\sum_{c_i \subset Sc_k} (dens(c_i).\#pix(c_i))}{\#pix(n)}\ ;$$

$$\bar{\mathbf{c}}^n = \frac{(\sum_{c_i \subset Sc_k} \bar{\mathbf{c}}_i.\#pix(c_i))}{\#pix(n)}\ ;$$

}

Appendix B: Excerpts from the experiments.

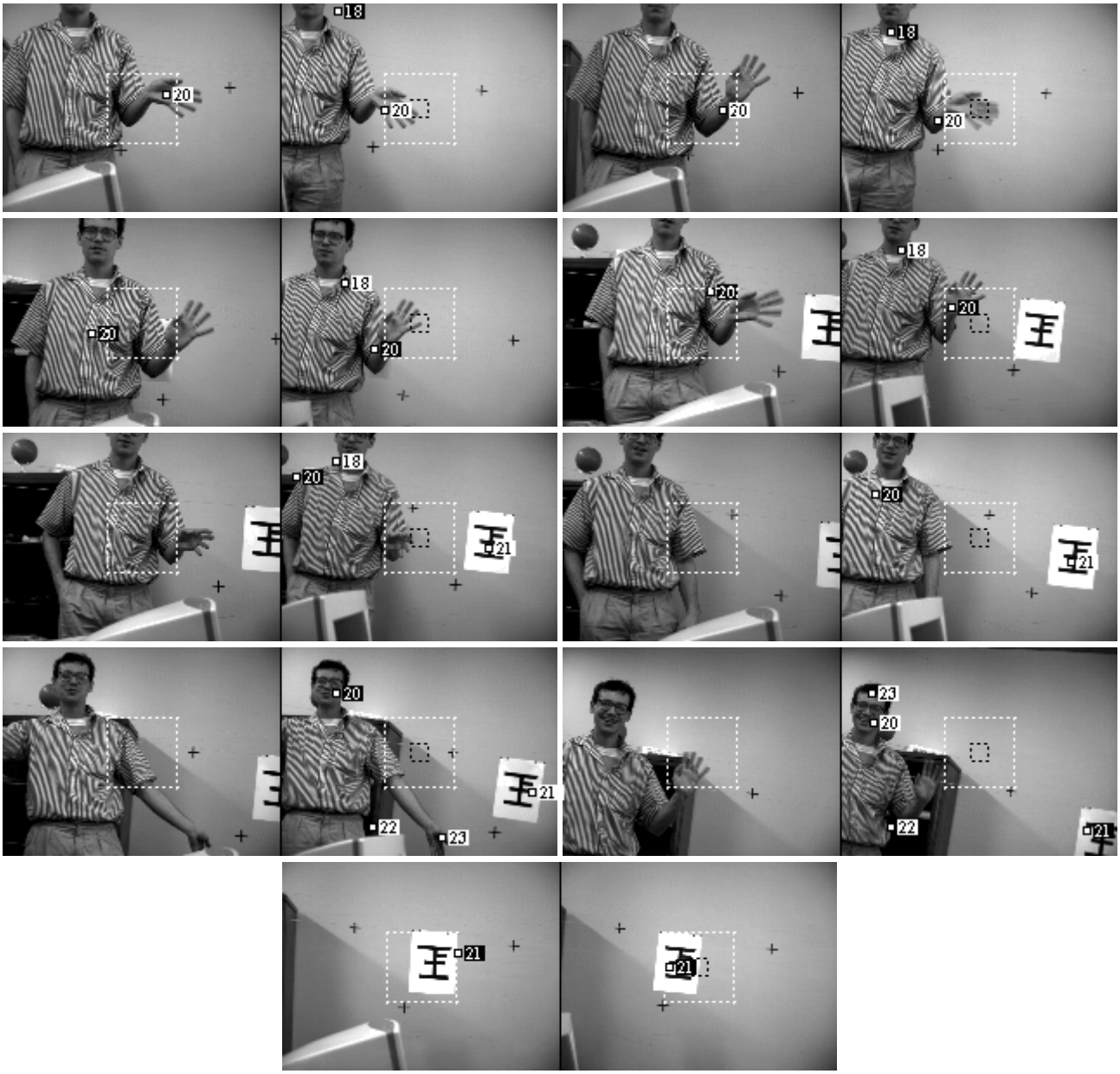
The following sequences of stereo pairs are displayed as non-dominant and dominant images (left and right, respectively) from top to bottom.



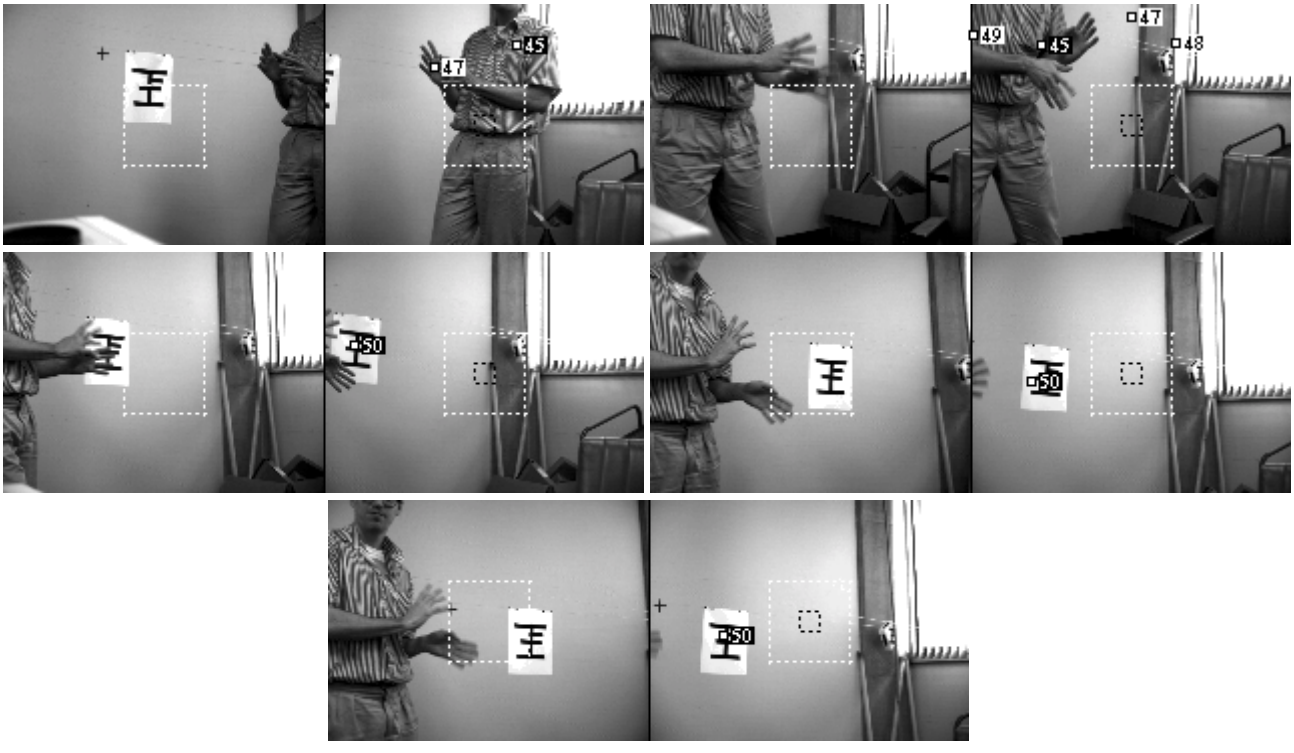
B.87 to B.91.

C.5-C.14.

D.16-D.17



D.26-D.34



D.74-D.78