UNDERWATER VIDEO ANALYSIS FOR NORWAY LOBSTER STOCK QUANTIFICATION USING MULTIPLE VISUAL ATTENTION FEATURES

Paulo Lobato Correia[1,2], Phooi Yee Lau[1], Paulo Fonseca[3] and Aida Campos[3]
[1]*Instituto de Telecomunicações,* [2]*Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisboa, Portugal,*
[3]*INIAP/IPIMAR, Av. de Brasilia, 1449-006 Lisboa, Portugal*

ABSTRACT

Underwater video is being increasingly used to assess the impact of human activities in marine habitats, as a complementary tool for the assessment of commercial stocks. But, analysing video images manually to study and evaluate marine habitats is a lengthy and tedious task. This paper proposes an automatic method to detect the Norway lobster (Nephrops Norvegicus) an important east-Atlantic and Mediterranean wide-distributed commercial crustacean species, in order to reduce the time and effort it takes marine scientists to manually quantify them. Here, the detection procedure follows a human visual attention model. Three visual attention features are considered: intensity map (IM), edge map (EM), and motion map (MM). The work is composed of two main parts: first the three feature maps are extracted; then, all candidate regions are processed and categorized in view of lobsters detection. Experimental results show that the proposed methodology is able to reliably detect candidate regions after combining the partial results.

1. INTRODUCTION

During the last two decades, there has been an increasing concern over the effects of bottom-fishing activities on the benthic ecosystems in all regions where commercial fishing is practised, with the evidence that fishing gear, particularly bottom trawling, may injure benthic organisms, reduce habitat complexities, and reduce biodiversity [1, 2].

Instrumentation capable of measuring dynamic events and/or processes within and immediately above the seafloor has been developed that facilitates the collection of ecological information. Underwater video imaging is one of these techniques, becoming increasingly useful to assess the impact of human activities, including fishing in marine habitats. One of the most studied species has been the Norway lobster, *Nephrops Norvegicus*, an important east-Atlantic and Mediterranean wide-distributed commercial crustacean species. Methods for Norway lobster abundance assessment include the use of underwater video techniques, since lobsters live in burrows that can be identified and enumerated, thus resulting in quantified video estimates of their density [3, 4, 5]. But, processing video images manually to study and evaluate marine habitats are lengthy and tedious tasks. In manual operations, highly trained marine scientists scan the video for meaningful information such as burrows and lobsters. Also landings information, available from commercial fishing, can be used for verification purposes. Nevertheless, because wide varieties of marine animals have low contrast and the structure of those marine animals may be very complex, the required processing is not an easy task, and the results may depend on the skill and concentration of the human operator. The application of automatic digital image processing techniques provides a valuable help on these procedures, e.g., to minimize the uncertainty of the results.

Nowadays, an increasing amount of research teams are using underwater video imaging devices, also contributing to the emergence of research in the area of automatic underwater video image analysis, namely for detection and classification of the relevant organisms [6, 7]. Even though many techniques have been developed and improved regularly as computer-based procedures, the biggest difficulties are related to the illumination at high depths, low contrast of the underwater images and the 'marine snow' that often clutters the visual scene, making identification difficult even for professionals. Moreover, a lobster may move during the detection and tracking process, and different viewing angles give a different perspective with respect to its size and length. Among the various image analysis methods proposed in the literature, color-based approaches are among the more robust methods, providing an efficient means of detecting Norway lobster due to its colorful shell. However, when only monochrome video images are available, as it is the case, other visual information such as luminance, edge, motion or depth should be explored.

The remainder of the paper is organized as follows: Section 2 outlines the underwater video acquisition procedure. Section 3 describes the proposed visual attention model and explains how each feature map is obtained. Section 4 discusses the lobster detection process. Section 5 concludes the paper and discusses future work.

2. UNDERWATER VIDEO ACQUISITION

This work explores underwater video images acquired by the Instituto de Investigação Agrária e das Pescas (INIAP/IPIMAR), along the Portuguese coast. The videos were taken during a research survey carried out on board the R/V "Noruega" within the scope of the EU Project "NECESSITY", in which selective trawl gear were tested to minimize the capture of undersized Norway lobster. The camera used was a Konsgsberg Maritime OE1324 mono-

chrome low-light SIT camera, with a light sensitivity (limiting) of 2e-4 lux, associated to a recording and powering system able to work up to 1500 metres depth. With this equipment, the first sequences of these benthic habitats were taken at depths around five hundred meters off the Portuguese coast. The camera was positioned either at the immediate front, or further at the back, hanging on the trawl headline while trawling at three knots speed (see Figure 1), during selectivity trials. The selected position affects the type of images obtained. Using camera location 1, lobster get better illuminated as they approach the camera. At camera position 2 outgoing lobsters are less illuminated as they move away from the camera. Figure 2 shows a sample image captured with the available monochrome camera, when placed in position 1.

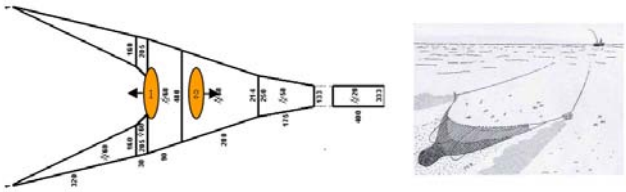


Figure 1: Trawl showing both camera positions

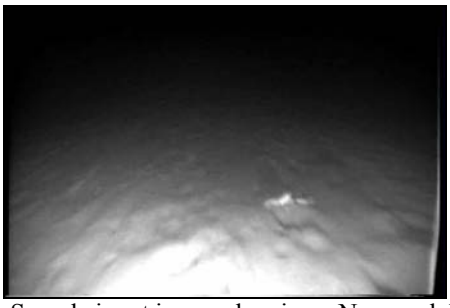


Figure 2: Sample input image showing a Norway lobster out of its burrow

3. VISUAL ATTENTION FEATURE MAPS EXTRACTION

Visual search for particular objects involves an active scan of the visual environment. Object recognition is frequently based upon features such as contrast, color, intensity, and orientation. Here, the concept of a saliency map, representing the visual attention, or saliency, at every location of the visual input is explored. This saliency map provides a reliable indication of the focus of attention by scanning the feature map in order of decreasing saliency. In an initial human analysis of the available underwater monochrome videos, lobster can be detected by their contrast to the background and higher intensity values.

In 2003, Edgington proposed a computer vision system combining a saliency-based attention module and a recognition module, both designed to mimic the human visual system for automatic recognition of mid-water organisms [8]. The proposal included several visual filters sensitive to colour, intensity and orientation, at different spatial scales. These features are combined into a single saliency map, the attention being directed to the area with highest activity. On the other hand, Walther proposed an object recognition system based on the biological plausible hierarchical feed-forward model of object recognition in the cortex for detecting faces [9]. This work shows that using cortical feedback connections and top-down processes, simple features such as hue and shape can be utilized to bias visual attention to locations with higher probability of containing the target object. Both techniques perform visual search to detect and track the salient objects in the images.

In this paper, a conceptually simple visual attention model is proposed, exploring visual features such as intensity, shape and motion. The work is composed of two main parts: first the extraction of three visual attention feature maps is described in the remainder of this section; then, in section 4, the resulting candidate regions are processed and categorized in view of lobster detection, also taking into account the characteristics of the operating environment.

3.1 Intensity Map

The first feature exploited in this proposal is image intensity, leading to the creation of a visual attention intensity map (IM), according to the system architecture shown in Figure 3.

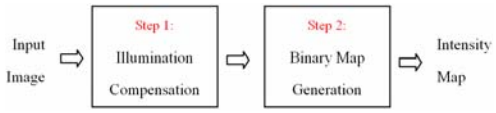


Figure 3: Computation of the proposed Intensity Map

The assumption made here is that lobsters correspond to higher intensity regions in a uniformly illuminated image area. In fact, visual attention tends to focus on the higher intensity pixels. Using a dynamic threshold, regions that closely model the lobsters can be selected, resulting in the desired intensity map.

Since, for this type of environment, illumination is much stronger near the camera than further away, the first step is to compensate for this effect. Notice that pixels near the light source show very high intensities due to excessive illumination, which should not be part of the IM as candidate objects for detection. In view of the above scenario, illumination compensation, by subtracting the local average luminance from the input images is proposed. The vertical luminance gradient of input images, illustrated in Figure 4, is calculated using equation (1), where W and H are the width and height of the input image, respectively. The result of illumination compensation is illustrated in Figure 5 (a)

$$G_{lum}(h) = \frac{1}{W}\sum_{w=0}^{W-1} I(w,h) \qquad w \in \{0,1,2,...,W-1\}, h \in \{0,1,2,...,H-1\} \tag{1}$$

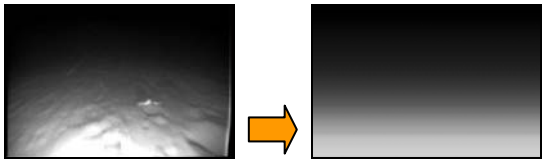


Figure 4: Luminance gradient of the input image

The next step is to generate a binary map, separating candidate object pixels from the background. Due to the dataset huge variety of lighting conditions, a dynamic thresholding method is applied. Bernsen's [10] dynamic thresholding of

gray scale images proposal uses region values and a contrast limit as parameters, according to equations (2) and (3).

$$T(x,y) = \frac{1}{2}\left(Z\_low + Z\_high\right) \quad (2)$$

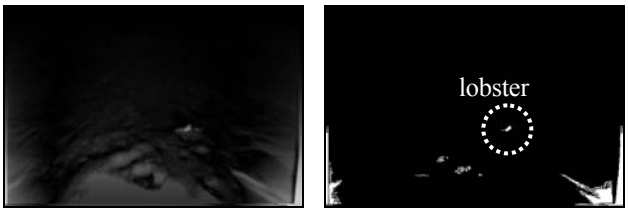
$$C(x,y) = \left(Z\_high - Z\_low\right) < contrast\_limit \quad (3)$$

A pixel *(x,y)* is marked as object if its value is higher than *T*(*x,y*), with *Z_low* and *Z_high* being the lowest and highest intensity values in a window around that pixel. However, if the contrast measure *C*(*x,y*) is lower than *contrast_limit*, then all pixels in the analysis window are assumed to belong to a single class, all pixels being labelled as object or background. A slightly different thresholding method is proposed here to simultaneously filter image noise and reduce segmentation errors. Instead of using the local maximum and minimum intensities, the second-maximum (*2_max_N*) and second-minimum (*2_min_N*) values in the pixel's 8-neighborhood are used – see equations (4) and (5).

$$T_{modified}(x,y) = \frac{1}{2}\left(2\_max\_N + 2\_min\_N\right) \quad (4)$$

$$C_{modified}(x,y) = \left(2\_max\_N - 2\_min\_N\right) < contrast\_limit \quad (5)$$

In this case, when a pixel belonging to the background has a contrast lower than *contrast_limit*, and additionally its luminance value is higher than a second local threshold, T_{low}, then that neighbourhood is said to consist only of object pixels. The resulting binary image is the intensity map (IM). An example is shown in Figure 5 (b).



(a) (b)

Figure 5: (a) Illumination compensated image, and (b) resulting intensity map (IM)

3.2 Edge Map

For the second feature map, the contour of the observed objects, in an illumination compensated image, is considered. The output of an edge detector is thresholded to obtain salient points creating the edge map (EM), according to the architecture shown in Figure 6.

Figure 6: Computation of the proposed Edge Map

The shape of lobsters can vary a lot due to its elongate structure and the viewing angle of the camera. Also, lobsters may appear in any area of the image. Edge information can be used as salient features, as lobsters often present some contrast with the seabed. Here, the first processing step is illumination compensation, using equation (1). Then, edges are identified using the Sobel detector, according to equation (6). I_1 and I_2 correspond to the vertical and horizontal edges, obtained using the filters h_x and h_y, as shown in equation (7).

$$I_E(x,y) = \sqrt{I_1^2(x,y) + I_2^2(x,y)} \quad (6)$$

$$h_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad h_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \quad (7)$$

Finally, the edge image is thresholded to get a binary map, named the edge map, as illustrated in Figure 7.

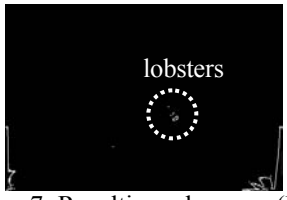


Figure 7: Resulting edge map (EM)

3.3 Motion Map

The third feature map corresponds to the analysis of the motion observed in the images. For this purpose, a change detection mask is computed, which after appropriate thresholding results in a motion map (MM), according to the architecture shown in Figure 8.

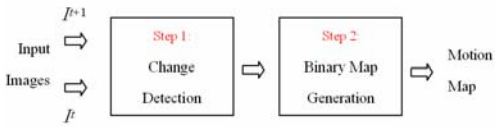


Figure 8: Computation of the proposed Motion Map

The motion of the camera during image acquisition results in the variation of lobster position – as a first approach it is assumed that lobsters are not moving and the camera moves at a constant speed. The parameters describing the motion between consecutive frames can be analysed. Szeliski and Coughlan [11] assumed that when image intensities are conserved, $I_t(x,y)$ can be formed locally by displacing the reference image *I(x,y)*, according to equation (8):

$$I_t(x+u_t, y+v_t) = I(x,y), \quad (8)$$

where u_t and v_t are the x- and y-axis components of the object's 2D velocity field after projection onto the image plane.

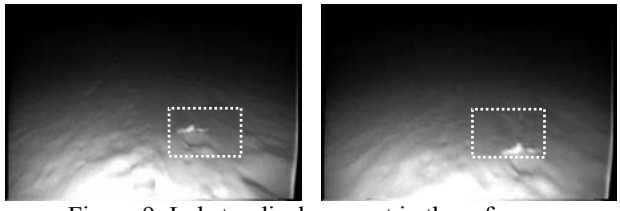


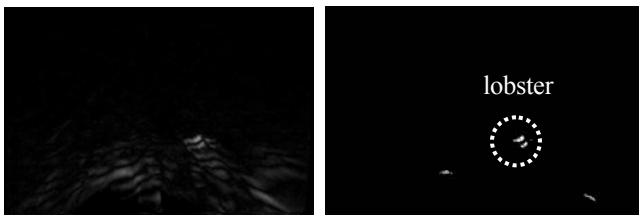
Figure 9: Lobster displacement in three frames

For the considered scenario, the aim is to detect lobsters by detecting their motion. Figure 9 shows a lobster and its displacement towards the camera after a few frames. Because of the lobsters' nature, their motion can be distinguished from that of other image structures such as trawl marks or burrows. In the present case, taking into account that consecutive images $I_{t1}(x,y)$ and $I_{t2}(x,y)$ are from the same source, a

change detection analysis can be performed to gather the desired motion information, applying equation (9).

$$CD(I_{t1}, I_{t2}) = \left| I_{t1}(x,y) - I_{t2}(x,y) \right| \quad (9)$$

To obtain the binary motion map, from the change detection result illustrated in Figure 10 (a), the dynamic thresholding specified by equations (4) and (5) is applied. Figure 10 (b) shows the final motion map.



(a) (b)

Figure 10: (a) Change detection and (b) motion map images

4. LOBSTER DETECTION

Once the individual feature maps are available, the next step is to combine these maps to produce a single salient map. Using a training set selected from the available test videos, captured as described in section 2, a number of experiments were conducted to determine the default values for the parameters used to compute the edge map, namely *contrast_limit* and T_{low}, which were set to 20 and 90, respectively.

A combined salient map was obtained by merging the three individual maps, thus classifying a pixel as foreground if that was its original classification in any of the individual maps, shown in Figure 11 (a).

The lobster detection proceeds by further analysed of the combined map, taking into account the expected lobster regions' features, such as their area, centroid and bounding box positions, to decide on their suitability as candidate lobsters.

As expected, the map may include non-lobster candidate regions, resulting from image acquisition noise or due to other limitations of the operating conditions.

At first, a morphologic close operation, with a disk-shaped structuring element of radius 9, is used to remove part of the existing noise. The resulting 8-connected regions are identified and several features are extracted for each candidate region, including its area, its centroid, and its bounding box. An example of the regions identified on the combined map, together with the corresponding bounding boxes, is shown in Figures 11 (b) and (c), respectively.

Further analysis allows discarding spurious regions that result from image acquisition limitations (as illustrated in Figure 12), corresponding to:

1) the hardware containing the camera to resist to the high water pressure, visible on the image sides (*a*);
2) the insufficiently illuminated area at the top (*b*),
3) the excessively illuminated area at the bottom (*c*);

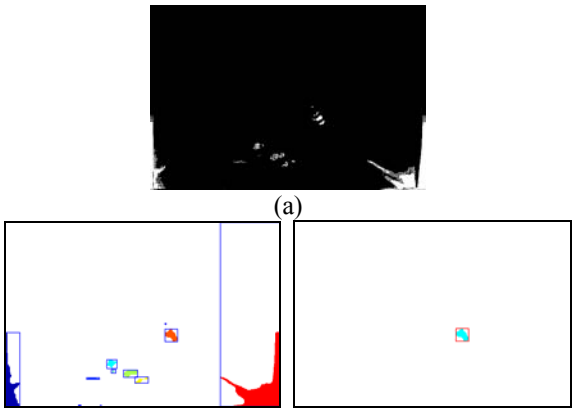
The above listed regions are considered to correspond to either contain artefacts or be of insufficient quality for analysis, thus being categorized as low confidence (LC) areas. With the videos captured as described in section 2, using the camera in position 1, the LC area (or the complementary region of interest – ROI) is defined by the following parameters:

$$\begin{cases} a = 0.07 \cdot W \\ b = 0.3 \cdot H \\ c = 0.23 \cdot H \end{cases} \quad (10)$$

where H and W are the height and width of the image, respectively.

Also the distance between the camera and the seabed should be taken into account, as it impacts on the perceived size of the observed lobsters. As such, for lobsters detection in the considered operation conditions, candidate lobsters' region area should be in the range $[A_{min}, A_{max}]$, with the default values considered being A_{min} = 0.01% and A_{max} = 0.3% of the total image area.

Applying the above set of restrictions the candidate regions after combined map analysis are as shown in Figure 11 (c).



(a)

(b) (c)

Figure 11: (a) Combined saliency map; (b) labelled regions and corresponding bounding boxes in the map (c) candidate regions after combined map analysis

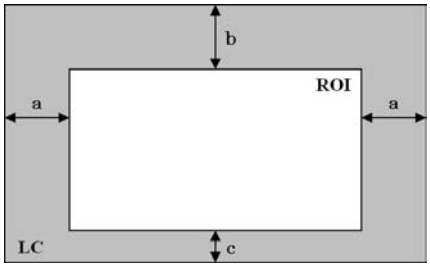


Figure 12: Confidence area for topographic features on lobster

To illustrate the behaviour of the proposed detection methodology, a monochrome video captured by INIAP/IPIMAR was considered. In particular several subsets of that video were analysed, of which some contained lobsters and some not.

It should be reminded that the main target of this paper is to develop an automatic algorithm to quantify the lobsters found in the portion of the seabed analysed. As such, it is not essential to detect the lobsters once they enter the image area, as long as they are effectively detected while in the region of interest (ROI) part of the image (each lobster should be

counted only once). An example of a final detection result is included in Figure 13.

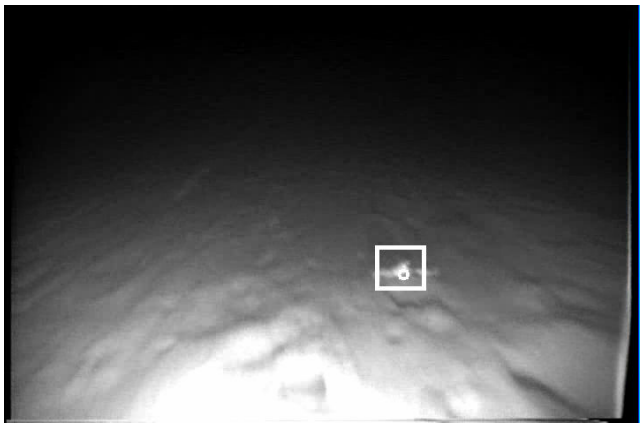


Figure 13: Final detection result

From a test set consisting of 100 video samples (11 images each), 10 of which correspond to positions in the video where lobsters appear, and the remaining 90 samples not containing lobsters, a 90% detection rate has been achieved. This means that nine out of the ten lobsters present in the video were successfully detected. Notice that less than 10% of the images in this type of videos typically contain lobsters.
The missed lobster was moving quite fast, thus presenting a somewhat undefined shape with low contrast to its surrounding, and sometimes being near the low confidence image area already showing some illumination problems, as shown in Figure 14. Also notice that the image acquisition was done at a vessel speed higher than desirable for this type of analysis, leading to somewhat blurred images.

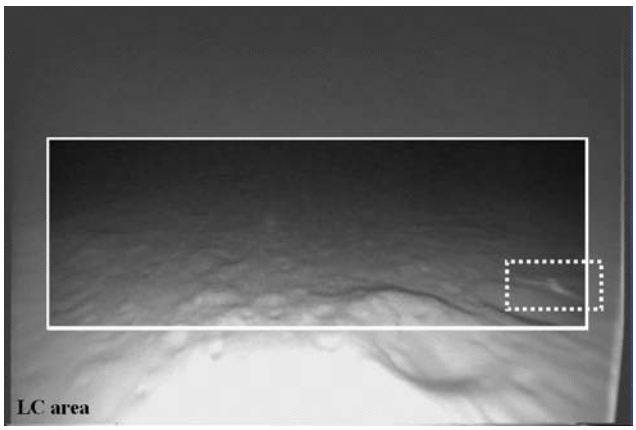


Figure 14: Missed lobster detection

Looking into the individual image detections, there were a total of 41 false negatives (meaning undetected lobsters in a single image) happening mostly in the low confidence area. There was also one false positive, as in one image a lobster was falsely detected, due to an illuminated burrow that looked like a lobster. The importance of these false negatives and false positives is very low, as 90% of the lobster were effectively counted, while the false positive positions are not temporally consistent, being easy to eliminate in a post-processing step.

5. CONCLUSION AND FUTURE WORK

In this paper, a simple visual attention model is proposed to assist marine specialists in the detection and quantification of Norway lobster stocks. The goal is to reduce the time and effort required for image analysis in stock control and management tasks. Detection of lobsters is based on the combined analysis of three different visual cues: intensity, edge and motion saliency maps. Currently the system is being developed to include a more complete temporal analysis of the video, namely to allow the tracking of detected lobsters. It is expected that this work may become a valuable complementary tool for the assessment of commercial stocks of Norway lobster. Future efforts will focus on extending analysis capabilities to include, e.g., detection of potential burrows and to allow assessing the impact of bottom trawlers in the seabed.

REFERENCES

[1] J. S. Link, "Ecological Considerations in Fisheries Management: When Does It Matter," *Fisheries*, vol. 27 (4), pp. 10-17, 2002.
[2] J. S. Link, "What Does Ecosystem-Based Fisheries Management Mean?," *Fisheries*, vol. 27 (4), pp. 18-21, 2002.
[3] N. Bailey, C.J. Chapman, J. Kinnear, D. Bova, A. Weetman, "Estimation of Nephrops Stock Biomass on the Fladen Ground by TV Survey", ICES-CM K:34, 1993.
[4] S.J. Marrs, R.J. Atkinson, C.J. Smith, J.M. Hills, "The Towed Underwater TV Technique for Use in Stock Assessment of Nephrops Norvegicus", in *Proc.* ICES CM 1998/G:8, La Coruna, Spain, pp. 88-98, 1998.
[5] I.D. Tuck, C.J. Chapman, R.J. Atkinson, "Population Biology of the Norway Lobster, Nephrops Norvegicus (L) in the Firth of Clyde, Scotland .1. Growth and Density", *ICES Journal of Marine Science*, vol. 54, pp. 125-135, 1997.
[6] D. Walther, D.R. Edgington, and C. Koch, "Detection and Tracking of Objects in Underwater Video," in *Proc. CVPR 2004*, Washington, USA, Jun 27 –Jul 2 2004, pp. 544-549.
[7] D. R. Edgington, D. Walther, C. Koch, "Automated Event Detection in Underwater Video," in *Proc. MTS/IEEE Oceans 2003*, San Diego, USA, 2003.
[8] D. R. Edgington, D. Walther, D.E. Cline, R. E. Sherlock, K. A. Salamy, A. Wilson, and C. Koch, "Detecting and tracking animals in underwater video using a neuromorphic saliency-based attention system," *ASLO/TOS Research Conference,* Honolulu, Hawaii, USA, February 15-20, 2004.
[9] D. Walther, T. Serre, T. Poggio, and C. Koch, "Modelling feature sharing between object detection and top-down attention," *Journal of Vision*, vol. 5 (8), pp. 1041, 2005.
[10] J. Bernsen, "Dynamic thresholding of grey-level images", *in Proc. 8th ICPR*, Paris, France, pp. 1251-1255, Oct. 1986.
[11] R. Szeliski and J. Coughlan, "Hierarchical Spline-Based Image Registration," in *Proc. CVPR94*, pp. 194-201, Seattle, USA, 1994