Eyewear Selector

O. Deniz, M. Castrillon, J. Lorenzo, M. Hernandez, L. Anton, G. Buenco

Eyewear Selector

O. Deniz[2], M. Castrillon[1], J. Lorenzo[1], M. Hernandez[1], L. Anton[1], G. Bueno[2]

[1]Universidad de Las Palmas de Gran Canaria, Dep. Informatica y Sistemas
Edificio de Informatica, 35017 Las Palmas, SPAIN
[2]Universidad de Castilla-La Mancha, E.T.S.Ingenieros Industriales
Avda.Camilo Jose Cela s/n, 13071 Ciudad Real, SPAIN

Abstract. The widespread availability of portable computing power and inexpensive digital cameras is opening up new possibilities for retailers. One example is in optical shops, where a number of systems exist that facilitate eyeglasses selection. These systems are now more necessary as the market is saturated with an increasingly complex array of lenses, frames, coatings, tints, photochromic and polarizing treatments, etc. Research challenges encompass Computer Vision, Multimedia and Human-Computer Interaction. Cost factors are also of importance for widespread product acceptance. This paper describes a low-cost system that allows the user to visualize different spectacle models in live video. The user can also move the spectacles to adjust its position on the face. Experiments show the potential of the system.

1 Introduction

The widespread availability of portable computing power and inexpensive digital cameras are opening up new possibilities for retailers in some markets. One example is in optical shops, where a number of systems exist that facilitate eyeglasses selection. These systems are now more necessary as the market is saturated with an increasingly complex array of lenses, frames, coatings, tints, photochromic and polarizing treatments, etc. [1]. The number of clients can grow only if the selection process is shortened or automated. A number of deployed systems have already demonstrated that eyeglass selectors can increase sales and customer satisfaction.

From a research viewpoint, such systems represent an interesting application of Computer Vision, Multimedia and Human-Computer Interaction. Computer-based systems for eyeglasses selection can be roughly classified according to a) use of live video or still image, and b) 3D or 2D-based rendering.

With still images, two options are possible. Some systems use a photo of the user without spectacles and then superimpose models on the image. Other systems simply take photos of the users wearing the different spectacles, allowing them to select the frame they like by direct comparison of the captured images. The use of still images is particularly convenient for web-based software. A number of sites are currently available that allow the user to upload his/her photo

and see the spectacles superimposed on it. Some systems can automatically extract facial features from the picture. In most of them, however, the user has to mark the pupils in the photo.

3D systems model the user's head and have the advantage that a depiction can be rendered from different viewpoints, see for example [2, 3]. 2D-based rendering does not work well for large out-of-plane rotations. 3D systems can also be of great help to opticians, as they can take measurements needed to manufacture the frames. However, 3D systems use special hardware and computing power, which can make them too expensive for most optical shops.

Most commercial 2D systems use still images, see for example [4–6]. This paper describes a 2D live-video eyeglasses selection system. Live video has an advantage over the use of static photos. Even if the user remains practically still, the experience is more realistic: other people near the user appear on the image, spectacles can be placed on the face by the user, etc. Live video effectively creates the illusion of a mirror. The paper is organized as follows. Section 2 gives an overview of the system. Sections 3 to 5 describe the modules in detail. Experiments are shown in Section 6. Finally, the main conclusions and ideas for future work are outlined.

2 System overview

The hardware system has the following components: a Windows box and two Sony FCB cameras with motorized zoom, focus, white balance and shutter speed. The cameras are placed together on top of the screen (either a computer monitor or a projector can be used). Camera 1 has a resolution of 384x288 pixels and makes no use of the zoom, while Camera 2 (352x288 pixels) uses a (fixed) zoom such that only the user's face appears in the image, see Fig. 1. Camera 2 only captures gray scale frames. The monitor displays the full-screen live video of Camera 1 with overlaid spectacles and on-screen buttons for showing the Previous/Next spectacles.

Fig. 1. Left: image of Camera 1. Right: image of Camera 2.

The software system has the following modules: homeostatic image stabilization, face and eye detection and spectacle management. The first module tries to keep image characteristics stable by using the motorized parameters of the

cameras. The face and eye detection module localize the position of the user's eyes. The spectacle management module is in charge of overlaying spectacles and controlling spectacle fitting and on-screen buttons. The following sections describe the modules in detail.

3 Homeostatic Image Stabilization

Homeostasis is defined in the Merriam Webster dictionary as *"a relatively stable state of equilibrium or a tendency toward such a state between the different but interdependent elements or groups of elements of an organism, population, or group"*. The state of equilibrium is normally related to the survival of the organism in an environment. Organisms are endowed with regulation mechanisms, generally referred to as homeostatic adaptation, in order to maintain this state of equilibrium. This idea has been used by some authors for building systems that carry out their activity in a complex environment [7, 8]. Arkin and Balch in their AuRA architecture [9] propose a homeostatic adaptation system which modifies the performance of the overall motor response according to the level of internal parameters such as battery or temperature. Another work which includes a homeostatic adaptation mechanism is the proposal of Hsiang [10] who introduces it to regulate the dynamic behavior of the robot during task execution.

In most computer vision systems, performance heavily depends on the quality of the images supplied by the acquisition subsystem. Face detection systems that make use of the skin color depend on the white balance, tracking systems based on edge detection depend on image contrast and so on. On the other hand, image quality is affected by environmental conditions, namely lighting conditions or distance from the object of interest to the camera. The typical scenario for an eyewear selection system is an optical shop, where environmental conditions change significantly throughout the day. Homeostatic adaptation will try to compensate for these effects on the image by making use of the adjustable parameters of the cameras.

Since the proposed system does not have a body, we simulate the physiological changes that influence the homeostasis mechanism. Cañamero [11] proposes synthetic hormones to imitate physiological changes. This approach was adopted in our system by implementing synthetic hormones that reflect the internal state of the vision system (Fig. 2).

The internal state of the vision system is represented by four hormones associated to luminance (*h_luminance*), contrast (*h_contrast*), color constancy (*h_whitebalance*) and size of the object (*h_size*), see Figure 2. The homeostatic mechanism will be in charge of keeping this internal state into a regime which will allow the system to operate with acceptable performance. An important element in a homeostatic mechanism is its adaptive aspect. When the internal state of the body is too far away from the desired regime, the homeostatic mechanism must recover it as soon as possible. The adaptive response of the homeostatic mechanism is governed by the hormone levels which are computed from the controlled variables by means of a sigmoid mapping. In this way, we can implement

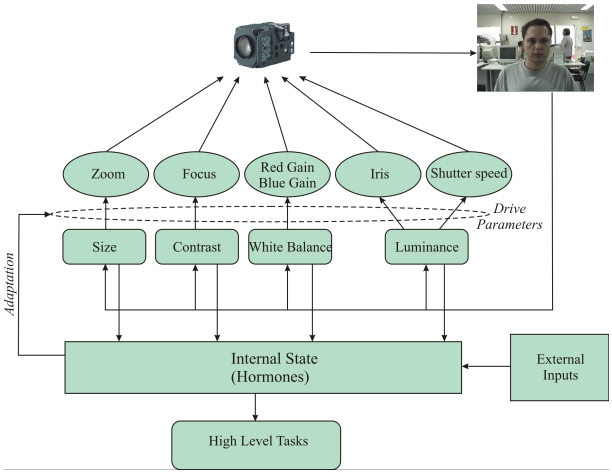

Fig. 2. Elements of the homeostatic adaptation mechanism.

adaptive strategies more easily in the drives since the hormone levels that define the normal and urgent recovery zones are always the same independently of the values of the controlled variables. Some methods that can be used to compute these variables are described in [12].

4 Face and Eye Detection

Several approaches have recently appeared tackling reliable face detection in real time [13–15], and making face detection less environment dependent. Cue combination usually provides greater robustness and higher processing speeds, particularly for live video stream processing. The face detection system used in the selector (see [16]) integrates, among other cues, different classifiers based on the general object detection framework by Viola and Jones [15], skin color, multilevel tracking, etc. The detection system provides not only face detection but also eye location in many situations. This additional feature reduces the number of false alarms, due to the fact that it is less probable that both detectors, i.e., face and eyes, are activated simultaneously with a false alarm.

In order to further minimize the influence of false alarms, we extended the facial feature detector capabilities, locating not only eyes but also the nose and the mouth. For that reason, several Viola-Jones' framework based detectors have been computed for the chosen inner facial elements. Positive samples were obtained by annotating manually the eye, nose and the mouth location in 7000 facial images taken randomly from the Internet. The images were later normalized by means of eyes information to 59×65 pixels, see Figure 3 (left). Five different detectors were computed: 1-2) Left and right eye (18×12 pixels), 3) eye pair (22×5), 4) nose (22×15), and 5) mouth (22×15).

Fig. 3. Normalized face sample and likely locations for nose and mouth positions after normalization.

The facial elements detection procedure is only applied in those areas which bear evidence of containing a face. This is true for regions in the current frame, where a face has been detected, or in areas with detected faces in the previous frame. For video stream processing, given the estimated area for each feature, candidates are searched in those areas not only by means of Viola-Jones' based facial features detectors, but also by SSD(sum of squared differences)-tracking previous facial elements. Once all the candidates have been obtained, the combination with the highest probability is selected and a likelihood based on the normalized positions for nose and mouth is computed for this combination, see Fig. 3.

The face and eye localization system described above works with images provided by Camera 1. The zoom camera (Camera 2) is used to capture the user's face with larger resolution than Camera 1. This can potentially provide a more precise and stable localization. Both eyes are searched for in the images taken by the zoom camera. A Viola-Jones detector is used along with tracking of eye patterns. Complex eye localization methods were discarded in order to keep an acceptable frame rate of the whole system. As the spectacles will have to be superimposed in the images taken from Camera 1, the localizations found in each Camera-2 frame have to be mapped onto the Camera-1 frame. Whenever an eye pair localization is obtained, the eye patterns in those localizations are scaled down. The scale factor is the ratio of intereye distances found in frames of the two cameras. The scaled eye patterns are then searched for in the images captured by Camera 1. This search is carried out in the vicinity of the last eye pair localization obtained for Camera 1. Fig. 4 shows the process.

5 Spectacle Management

Spectacles are superimposed on Camera 1 images through alpha blending. This process is basically a mixing of two images, with the mixing weights given by a third image. The models are made up of two images: the spectacles and the alpha channel. The alpha channel defines the zones of the spectacles that are translucent (i.e. the mixing weights). The spectacle models were obtained by taking frontal photographs of real spectacles of a local optical shop. The pictures were cropped and the alpha channels extracted using image editing software.

Spectacle models are scaled according to the intereye distance, rotated, and finally placed on screen according to the eye midpoint. Blending is performed

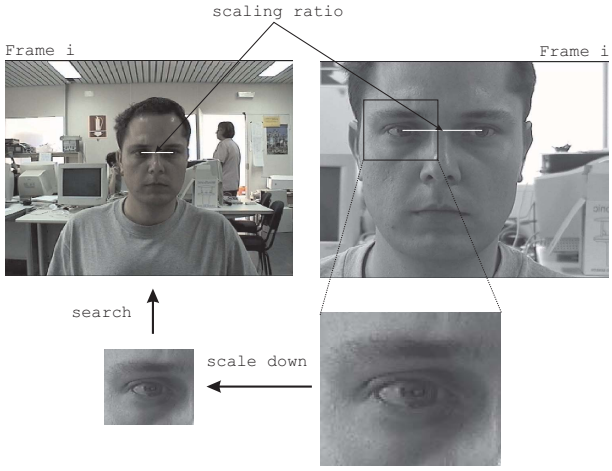

Fig. 4. Eye localization with the zoom camera.

only in the affected image region. Note that eye localization has an inherent error, which is also present in the midpoint. The eye midpoint has to be obtained robustly. The spetacles should move with the face, otherwise the rendition will appear unrealistic. Thus, a Lucas-Kanade pyramidal tracker [17] tracks strong corners within the face region. The average displacement vector of the tracking points is used in each frame to correct the displacement of the eye midpoint.

Spectacle models are stored as images. The center of the spectacle image is placed on the eye midpoint. This may lead to undesired results if the spectacle image is not well centered. Horizontal centering is not difficult to achieve, though the vertical center is subjective. Besides, each user's facial characteristics may require different placements over his/her nose. In order to tackle this, spectacle placement gesture detection was added to the system.

Considering the unrestricted context of this application, in terms of hands gestures the use of multiple or complex detectors would produce an approach not suitable for real time processing. In consequence, we have chosen the skin color approach for faster processing. However, instead of using a predefined color space definition, the information obtained from the face blob (see the previous section) is used to estimate the an histogram-based skin color model for that individual. The skin color model is used to locate other skin-like blobs in the image, and in consequence to find the hands for a given face.

The spectacle placement algorithm is based on detecting the placement gesture on the skin-color image. Along the vertical sides of the face rectangle a series of small lateral rectangles are considered, see Fig. 5. Their size is proportional to the size of the detected face, considering anthropomorphic relations. The skin-color image is softened using a Gaussian filter with an aperture that equals the detected face width. Thus, isolated pixels and small blobs are removed, while the face and hand blobs create consistent high-valued regions. The

hand vertical position is given by the position of the rectangle R containing the highest sum of skin-color pixels (the integral image is used to speed up sums of skin pixels). However, the hand must be in contact with the head. In order to check the "touching-the-head" condition, pixel values are analysed in R. Skin-color continuity is checked from the face side through half the width of R. Every column should contain at least one pixel with a high enough skin-color value (32 on normalized conditions). Otherwise, the hand may be still approaching the face or leaving it. Once the hand is detected as "touching the head", its relative displacement is used to move the spectacles upward and downward. When the hand no longer touches the head the final relative displacement is stored with the current spectacle model. Fig. 6 shows a sequence in which the user is fitting the spectacles.

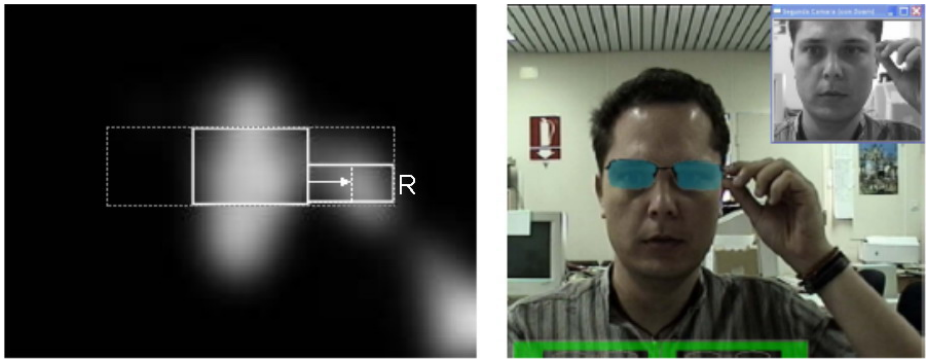

Fig. 5. Spectacle placement gesture detection. Left: the biggest white rectangle represents the face area.

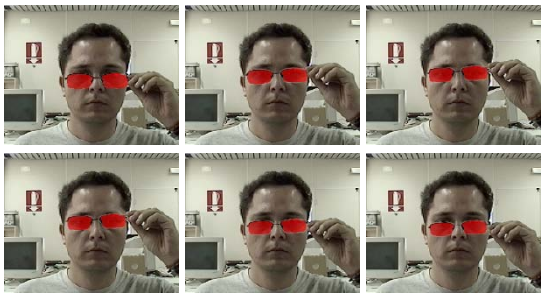

Fig. 6. Spectacle placement sequence.

6 Experiments

The face detection system was tested with video sequences. Seventy-four sequences corresponding to different individuals, cameras and environments with a resolution of 320x240 were recorded. They represent a single individual sat and speaking in front of the camera or moderating a TV news program. The face pose is mainly frontal, but it is not controlled, i.e. lateral views and occlusions due to arm movements are possible. The eyes are not always visible. The total set contains 26338 images.

In order to test the detectors performance, the sequences were manually annotated, therefore the face containers are available for the whole set of images. However, eye locations are available only for a subset of 4059 images. The eyes location allows us to compute the actual distance between them, which will be referred below as *EyeDist*. This value will be used to estimate the goodness of eye detection.

Two different criteria have been defined to establish whether a detection is correct: a) Correct face criterium: A face is considered correctly detected if the detected face overlaps at least 80% of the annotated area and the area difference is not doubled, and b) Correct eye criterium: The eyes of a face detected are considered correctly detected if for both eyes the distance to manually marked eyes is lower than a threshold that depends on the actual distance between the eyes, *EyeDist*.

Table 1 shows the results obtained after processing the whole set of sequences with different detectors. The correct detection ratios (TD) are given considering the whole sequence, and the false detection ratios (FD) are related to the total number of detections. Rowley's detector is notably slower than the others, but it provides eye detection for the 78% of detected faces, feature which is not considered by Viola-Jones' detector. As for the face detector, it is observed that it performs more than twice faster than Viola-Jones' detector, and almost ten times faster than Rowley's. Speed was the main goal in our application, the face detector is critical for the live-video selector.

	Rowley [18]		Viola-Jones [15]		face detector used here [16]	
	TD	FD	TD	FD	TD	FD
Faces	89.27	2.16	97.69	8.25	99.92	8.07
Left Eye	77.51	0.8	0.0	-	91.83	4.04
Right Eye	78.18	1	0.0	-	92.48	3.33
Proc. time	422.4 msecs.		117.5 msecs.		45.6 msecs.	

Table 1. Results (%) for face and eye detection processing using a PIV 2.2Ghz.

Face detection depends heavily on skin color and pattern matching to detect faces. Therefore, experiments were carried our to study the influence of luminance and white balance on face detection performance. Fig. 7 shows the values

of the *h_luminance* and *h_whitebalance* hormones along with the face detection rate for an individual moving in front of the camera. The dashed lines represent the changes in the environmental condition (lighting).

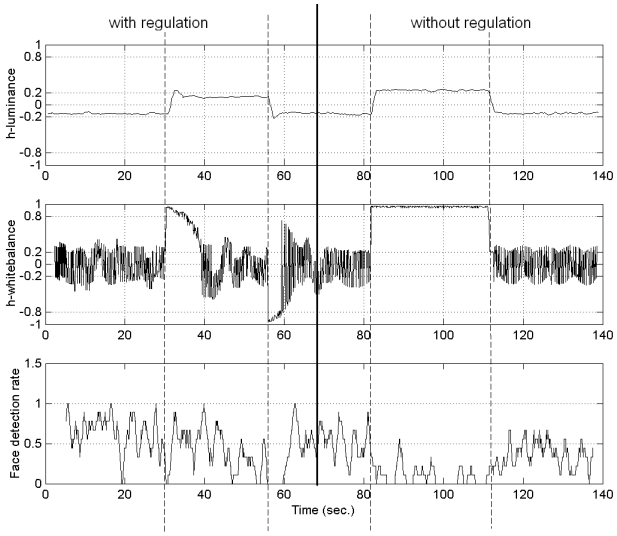

Fig. 7. Face detection rate with and without homeostatic mechanism.

When the system starts the detection rate is high and it decreases slightly when more lights are switched on (30-57 secs.). When the lights are switched on, both the *h_luminance* and *h_whitebalance* hormones go out of their desired states but the homeostatic mechanism recovers them after a delay, larger for the *h_whitebalance* hormone than for the *h_luminance* one. In the experiment the homeostatic mechanism is deactivated after 70 seconds, so when the conditions change again the state of the hormones is not recovered and the performance of the system decreases.

In order to measure the localization capabilities of the system, seven video sequences were recorded in which a subject moves his head, from almost no motion to extreme movements. The eyes of the subject were manually located so as to have *ground truth* data. Table 2 shows the number of frames and amount of motion of each video sequence. In sequences 6 and 7 the head movements were exaggerated for testing purposes and they do not represent a typical situation.

The effect of the zoom camera is shown in Table 3. The first thing to note is that localization errors using information from the zoom camera (Camera 2) are larger than those of Camera 1. Despite the higher resolution available in the zoom camera, this finding is explained by two reasons: a) The eye localizer of the second camera is much simpler than the face detector (though 80% faster) and, b) Head motion often make the eyes go out of the zoomed image. As Camera

	Video sequence						
	1	2	3	4	5	6	7
N	126	175	176	148	119	129	208
V	8.2	11.1	11.3	27.1	37.7	120.8	164.4

Table 2. Video sequences used in the experiments. N=number of frames, V=variance of eye position.

1 data are better for localization, they have priority in the combination. That is, whenever localization data are available from the face detector they are used to place the spectacles on screen. Data obtained with Camera 2 are only used when the face detector can not provide an eye pair localization.

The combined use of information of the two cameras does not improve either one of them alone, except in the cases marked with bold in the table. However, the use of the second camera is advantageous in terms of number of frames with available eye localization, see Table 4. This allows the spectacles to remain longer on screen, even if the individual is moving.

Video	Camera 1		Camera 2		Combination	
sequence	REE	LEE	REE	LEE	REE	LEE
1	1.68(1.01)	1.53(0.87)	3.08(2.41)	2.85(1.78)	**1.61(1.01)**	1.53(0.87)
2	2.78(2.91)	2.81(1.21)	5.44(5.17)	4.27(3.25)	**2.71(2.92)**	**2.73(1.25)**
3	2.39(1.00)	2.03(0.80)	1.38(0.93)	2.37(1.23)	2.36(0.98)	**2.03(0.78)**
4	1.86(1.21)	2.69(1.41)	2.96(2.94)	2.22(1.43)	1.99(1.19)	2.40(1.27)
5	2.63(1.39)	2.37(1.16)	2.48(1.57)	2.69(1.78)	2.54(1.34)	**2.33(1.51)**
6	3.03(2.75)	2.64(1.76)	6.82(7.86)	9.81(10.03)	6.14(7.22)	7.79(9.27)
7	2.29(1.24)	2.22(1.55)	5.36(4.82)	7.91(11.48)	2.82(2.13)	4.81(9.93)

Table 3. Eye localization errors. REE=right eye error, LEE=left eye error. Standard deviations between parentheses.

In another experiment, facial tracking (see Section 4) as a localization aid was tested. Table 5 shows that its use is specially advantageous when there is large head motion. The tracker was is reinitialized every 60 frames in order to avoid excessive drifting.

7 Conclusions

The affordability of cameras and portable computing power is facilitating the introduction of computer vision in optical shops. Retailers are particularly interested in automating or accelerating the selection process. Most commercially available systems for eyewear selection use static pictures. This paper describes a patent pending live-video eyeglasses selection system based on computer vision

Video	Camera 1	Camera 2	Combination
1	124/126	109/126	124/126
2	173/175	57/175	173/175
3	174/176	126/176	174/176
4	76/148	81/148	**111**/148
5	100/119	89/119	**113**/119
6	51/129	71/129	**83**/129
7	165/208	157/208	**203**/208

Table 4. Number of frames with available eye pair localization.

Video	No tracking	With tracking	Improvement
1	1.451(0.677)	1.307(0.648)	9.91%
2	2.286(1.515)	2.189(1.467)	4.25%
3	1.505(0.626)	1.413(0.732)	6.13%
4	2.112(1.147)	1.775(0.951)	15.99%
5	2.079(1.057)	2.037(1.062)	2.00%
6	6.835(12.112)	7.026(11.346)	-2.80%
7	3.349(6.230)	3.334(5.788)	0.43%

Table 5. Squared errors of the eye midpoint position. Here the tracking of facial points was given a weight equal to the no-tracking eye midpoint localization.

techniques. The system, which runs at 9.5 frames per second on general purpose hardware, has a homeostatic module that keeps image parameters controlled. This is achieved using cameras with motorized zoom, iris, white balance, etc. This feature can be specially useful in environments with changing illumination and shadows, like in an optical shop. The system also has a face and eye detection module and a spectacle management module. Further improvements are possible like adding zoom, mirror-like spectacles, etc. Also, it may possible to eliminate the second camera, which would allow the system to be adapted to commodity hardware. Modern laptops, for example, include an integrated webcam and sufficient computing power to be used as a low-cost, portable, eyewear selector. This would also open the possibility of using the system via Internet.

Acknowledgments

This work was partially supported by Project ULPGC07-008 from Universidad de Las Palmas de Gran Canaria.

References

1. Roberts, K., Threlfall, I.: Modern dispensing tools. options for customised spectacle wear. Optometry Today **46** (2006) 26–31

2. Activisu: Activisu Expert (Last accesed: 30/7/2007) Available at http://www.activisu.com.
3. Visionix: 3DiView 3D virtual try-on (Last accesed: 30/7/2007) Available at http://www.visionix.com.
4. ABS: Smart Look (Last accesed: 30/7/2007) Available at http://www.smart-mirror.com.
5. Carl Zeiss Vision: Lens Frame Assistant (Last accesed: 30/7/2007) Available at http://www.zeiss.com.
6. CBC Co.: Camirror (Last accesed: 30/7/2007) Available at http://www.camirror.com.
7. Breazeal, C.: A motivational system for regulating human-robot interaction. In: AAAI/IAAI. (1998) 54–61
8. Gadanho, S., Hallam, J.: Robot learning driven by emotions. Adaptive Behavior **9** (2002) 42–64
9. Arkin, R.C., Balch, T.: AuRA: Principles and practice in review. Journal of Experimental and Theoretical Artificial Intelligence **7** (1997) 175–188
10. Hsiang, K., Kheng, W., Ang, M.: Integrated planning and control of mobile robot with self-organizing neural network. In: 18th IEEE Int. Conference on Robotics and Automation, Washington DC (2002) 3870–3875
11. Cañamero, D.: Modeling motivations and emotions as a basis for intelligent behavior. In Lewis, J., ed.: Procs. of the First Int. Symposium on Autonomous Agents, New York, ACM Press (1997) 148–155
12. Lorenzo, J., Castrillón, M., Hernández, M., Déniz, O.: Introduction of homeostatic regulation in face detection. In Fred, A., ed.: Proceedings of the 4th International Workshop on Pattern Recognition in Information Systems, PRIS 2004, Porto (Portugal) (2004) 5–14
13. Li, S., Zhu, L., Zhang, Z., Blake, A., Zhang, H., Shum, H.: Statistical learning of multi-view face detection. In: European Conference Computer Vision. (2002) 67–81
14. Schneiderman, H., Kanade, T.: A statistical method for 3d object detection applied to faces and cars. In: IEEE Conference on Computer Vision and Pattern Recognition. (2000) 1746–1759
15. Viola, P., Jones, M.: Robust real-time face detection. IJCV **57** (2004) 151–173
16. Castrillon, M.: On Real-Time Face Detection in Video Streams. An Opportunistic Approach. PhD thesis, Universidad de Las Palmas de Gran Canaria (2003)
17. Bouguet, J.: Pyramidal implementation of the Lucas Kanade feature tracker. Technical report, Intel Corporation, Microprocessor Research Labs, OpenCV documents (1999)
18. Rowley, H., Baluja, S., Kanade, T.: Neural network-based face detection. IEEE Trans. on Pattern Analysis and Machine Intelligence **20** (1998) 23–38