Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Computer vision based eyewear selector

Oscar DÉNIZ†[1,2], Modesto CASTRILLÓN[1], Javier LORENZO[1], Luis ANTÓN[1], Mario HERNANDEZ[1], Gloria BUENO[2]

([1]*Instituto Universitario de Sistemas Inteligentes y Aplicaciones Numéricas en Ingeniería, Universidad de Las Palmas de Gran Canaria, Edificio Central del Parque Científico-Tecnológico 35017 Las Palmas, Spain*)

([2]*E.T.S. Ingenieros Industriales, Universidad de Castilla-La Mancha Campus Universitario, Avda. Camilo José Cela, s/n 13071 Ciudad Real, Spain*)

†E-mail: Oscar.Deniz@uclm.es

Received June 26, 2009; Revision accepted Nov. 28, 2009; Crosschecked Dec. 8, 2009

Abstract: 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. 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 glasses models in live video. The user can also move the glasses to adjust its position on the face. The system, which runs at 9.5 frames/s on general-purpose hardware, has a homeostatic module that keeps image parameters controlled. This is achieved by using a camera 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 includes a face and eye detection module and a glasses management module.

Key words: Face detection, Eye detection, Perceptual user interfaces, Human-computer interaction
doi:10.1631/jzus.C0910377 **Document code:** A **CLC number:** TP391.4

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 increasingly necessary as the market is saturated with an increasingly complex array of lenses, frames, coatings, tints, photochromic and polarizing treatments, etc. (Roberts and Threlfall, 2006). The number of clients can increase only if the selection process is shortened or automated. A number of deployed systems have already demonstrated that eyeglasses selectors can increase sales and customer satisfaction (Kuttler, 2003; Morgan, 2004).

From a research viewpoint, such systems represent an interesting application of Computer Vision, Multimedia and Human-Computer Interaction. The Augmented Reality (see a survey in Azuma (1997)) of watching ourselves and trying different 'virtual' spectacles can be achieved by combining computer vision and graphics. The Magic Lens and Magic Mirror systems, for example, use the ARTag toolkit (Fiala, 2004), which mixes live video and computer-generated graphics. People can wear cardboard patterns that ARTag can detect. The graphics are placed in the visible positions of the patterns. The system can work in real time on general-purpose hardware, although people have to wear the cardboard. Another approach was taken by Lepetit *et al*. (2003), where virtual glasses and moustaches are added to live video. Although the system works with an impressive frame rate of 25 Hz, the user must start the tracker by reaching a position close to a generic triangle-based face model shown on the screen. The ARMirror is a kiosk-based entertainment setup that shows live video overlaying virtual hats,

skulls, etc. (Lyu *et al*., 2005). Only the face as a whole, however, is tracked.

Commercial systems for eyeglasses selection can be roughly classified according to (1) the use of live video or snapshots, and (2) 3D- or 2D-based rendering. With snapshots, two options are possible. Some systems use a photo of the user without glasses and then superimpose models on the image. Other systems simply take photos of the users wearing the different glasses, allowing them to select the frame they like by a direct comparison of the captured images.

The use of snapshots 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 glasses superimposed on it (see a brief list in GlassEyes (2009)). 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. In some cases the pupillary distance in millimeters has to be entered by the user.

Three-dimensional systems model the user head and have the advantage that a depiction can be rendered from different viewpoints (Activisu, 2007; Rodenstock, 2007; Visionix, 2007). 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. The system described in Visionix (2007), for example, uses six digital cameras; the system in Activisu (2007) requires the user to wear a special plastic frame with markers.

Other possible features include: visual effect of tinted lenses on the whole displayed image, simulation of colored contact lenses, touch screens, compactness of the system, active sensing of the eyes (i.e., infrared illumination for eye localization), etc.

Most commercial 2D systems use static pictures (ABS, 2007; Carl Zeiss Vision, 2007; CBC Co., 2007; CyberImaging, 2007; OfficeMate Software Systems, 2007; Practice, 2007). This paper describes a 2D live-video eyeglasses selection system. Live video has an advantage over static photos. Even if the user remains practically still, the experience is more realistic: other people near the user appear on the image; glasses can be placed on the face by the user; etc. Live video effectively creates the illusion of a mirror.

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 384×288 pixels and makes no use of the zoom, while camera 2 (352×288 pixels) uses a (fixed) zoom such that only the user's face appears in the image (Fig. 1). Camera 2 captures only gray scale frames. The monitor displays the full-screen live video of camera 1 with overlaid glasses and on-screen buttons.

(a) (b)

Fig. 1 (a) Image of camera 1; (b) Image of camera 2

The software system has the following modules (Fig. 2): homeostatic image stabilization, face and eye detection, and glasses management. The first module tries to keep image characteristics stable using the motorized parameters of the cameras. The face and eye detection module localizes the position of the user's eyes. The glasses management module is in charge of overlaying glasses and controlling glasses fitting and on-screen buttons. The following sections describe the modules in detail.

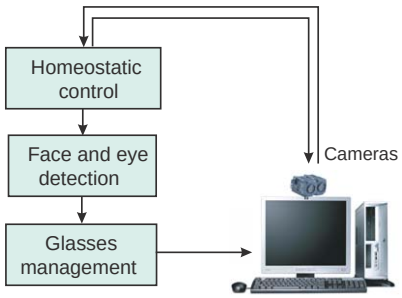


Fig. 2 Schematic of our proposed 2D live-video eyeglasses selection system

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 (Velasquez, 1997; 1998; Breazeal, 1998; Gadanho and Hallam, 2001). Arkin and Balch (1997) in their AuRA architecture proposed a homeostatic adaptation system that modifies the performance of the overall motor response according to the level of internal parameters such as battery or temperature. Another work that includes a homeostatic adaptation mechanism is the proposal by Low *et al*. (2002) who introduced 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; Edge detection based tracking systems depend on image contrast; etc. On the other hand, image quality is affected by environmental conditions, namely lighting conditions or the distance from the object of interest to the camera. A 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 using the adjustable parameters of the cameras.

In the affective computing framework (Picard, 1997), systems must be 'embodied' because human emotions involve both the body and the mind. As the proposed system does not have a body, like an anthropomorphic robot, we simulate the physiological changes that influence the homeostasis mechanism. Cañamero (1997) proposed synthetic hormones to imitate physiological changes in the body of a robot which evolves in a 2D world and whose motivations respond to the levels of the synthetic hormones. We adopted this approach in our system by implementing the synthetic hormones that reflect the internal state of the vision system (Fig. 3).

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). 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. The internal state of the system also modifies high level behaviors (Fig. 3). If the image is too dark, for example, it makes no sense to carry out any visual process on it.

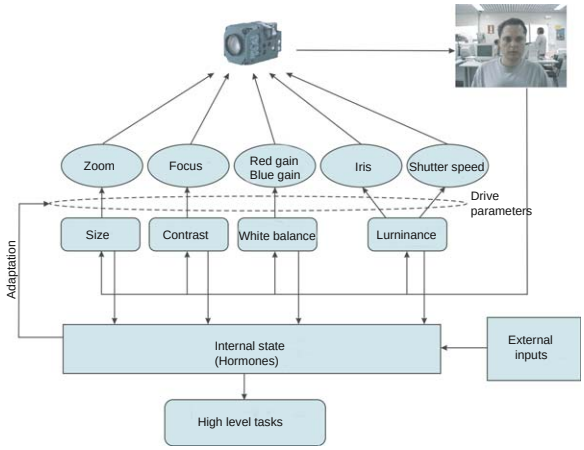


Fig. 3 Elements of the homeostatic adaptation mechanism

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 that are computed from the controlled variables by means of sigmoid mapping (Fig. 4). In this way, we can implement adaptive strategies more easily in the drives since the hormone levels that define the normal and urgent recovery zones are always the same, independent of the values of controlled variables.

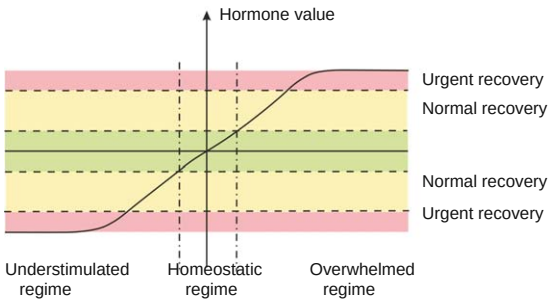


Fig. 4 Hormone value mapping from the variable of interest

The luminance of the image is computed by dividing the image into five regions, similar to the method proposed by Lee *et al*. (2001): an upper strip (R0), a lower strip (R4), and the central strip is divided into three regions (R1, R2 and R3 from left to right) (Fig. 5). These five regions allow us to define different auto exposure

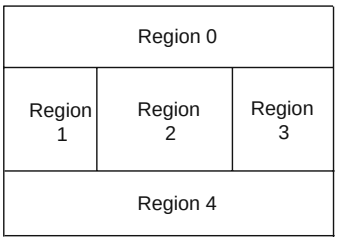


Fig. 5 Regions used to compute the luminance of the image

(AE) strategies according to the nature of the object of interest giving different weights to the average luminance in each region.

We have tested three different strategies for auto exposure that we have called uniform, centered and selective. The luminance for each of these strategies is computed as follows:

$$L_{\text{uniform}} = (L_0 + L_1 + L_2 + L_3 + L_4)/5, \quad (1)$$

$$L_{\text{centered}} = 0.8L_2 + 0.2(L_0 + L_1 + L_3 + L_4)/4, \quad (2)$$

$$L_{\text{selective}} = 0.8(L_2 + L_4)/2 + 0.2(L_0 + L_1 + L_3)/3, \quad (3)$$

where L is the total luminance of the image and L_i denotes the average luminance of region i. The L_{centered} strategy is suitable for tracking tasks where the object of interest is in the center of the image, whereas $L_{\text{selective}}$ is suitable for human-computer interaction because it considers the part of the image where normally a person appears when he/she is sitting in front of a computer. As for white balance, we assumed a 'grey world' scenario (Nanda and Cutler, 2001), which tries to make the average amount of green, blue and red in the image constant by adjusting the red and blue gains.

4 Face and eye detection

Several approaches have recently been described to tackle reliable face detection in real time (Schneiderman and Kanade, 2000; Li *et al*., 2002; Viola and Jones, 2004), making face detection less environment dependent. Cue combination usually provides greater robustness and higher processing speeds, particularly for live video stream processing. This is evidenced by the fact that a face detector based on cue combination (Castrillón *et al*., 2007) outperforms single cue based detectors (Viola and Jones, 2004), providing a more reliable tool for real-time interaction.

The face detection system developed for the selector (called ENCARA2, see Castrillón *et al*. (2007) for details) integrates, among other cues, different classifiers based on the general object detection framework by Viola and Jones (2004): skin color, multi-level 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, because 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, the 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 (Fig. 6a). Five different detectors were computed: (1)–(2) Left and right eyes (18×12), (3) eye pair (22×5), (4) nose (22×15), and (5) mouth (22×15). These detectors are publicly available (Reimondo, 2007).

Fig. 6 Normalized face sample (a) and likely locations for nose (b) and mouth (c) positions after normalization

The facial element detection procedure is applied only in those areas that 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 for in those areas not only by means of Viola-Jones based facial features detectors, but also by SSD-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 the nose and the mouth is computed for this combination. To this end, two probability maps were computed for the location of the nose and the mouth (given a normalized location for both eyes). These two estimated probability maps (Figs. 6b and 6c) were computed based on the information achieved after manually annotating around 6000 frontal face images.

Thus, during detection, different eye, nose and mouth candidates are obtained making use of trackers and detectors in specific areas. For each eye, nose and mouth candidate their transformation to the normalized location and scale is computed. The positions with the highest total likelihood (using the estimated probability maps) are selected. Fig. 7 (see p.84) shows the possibilities of the described approach with a sequence extracted from DaFEx (Battocchi and Pianesi, 2004).

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 a larger resolution compared with 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. ENCARA2 and complex eye localization methods are discarded to keep an acceptable frame rate of the whole system. As the glasses 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 inter-eye distances found in frames of the two cameras. The scaled eye patterns are then searched for in the images captured by camera 1 (the cameras are not synchronized, though the speed of head movement is not high enough to cause significant differences in the images). This search is carried out in the vicinity of the last eye pair localization obtained for camera 1. In order to avoid strange positions and sizes of the glasses for significant out-of-plane head rotations, the glasses are not shown whenever the inter-eye distance is smaller than a threshold. Fig. 8 shows the process.

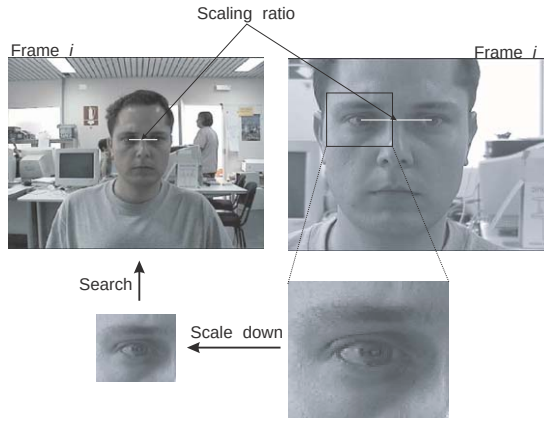


Fig. 8 Eye localization with the zoom camera

5 Glasses management

Glasses are superimposed on camera-1 images through alpha blending. This process is a mixing of two images, with the mixing weights given by a third image. The models are made up of two images: the glasses and the alpha channel. The alpha channel defines the zones of the glasses that are translucent (i.e., the mixing weights) (Fig. 9). The glasses models are obtained by taking frontal photographs of real glasses of a local optical shop. The pictures are cropped and the alpha channels extracted using image editing software.

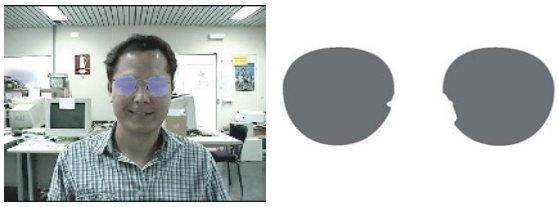


Fig. 9 (a) Glasses model; (b) Associated alpha channel

Glasses models are scaled according to the inter-eye distance, rotated, and finally placed on the screen according to the eye midpoint. Blending is performed 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 glasses should move with the face; otherwise, the rendition will appear unrealistic. Thus, a Lucas-Kanade pyramidal tracker (Bouguet, 1999) tracks strong corners within the face region (Fig. 10). The average displacement vector $\boldsymbol{p}_{\mathrm{m}}$ of the n tracking points is used in each frame to correct the displacement of the eye midpoint:

$$\boldsymbol{p}_{\mathrm{m}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{p}_i(t) - \boldsymbol{p}_i(t-1)), \tag{4}$$

$$\boldsymbol{e}^* = \boldsymbol{e} + \boldsymbol{p}_{\mathrm{m}}. \tag{5}$$

The current glasses model can be changed with on-screen previous-next buttons (Fig. 11). Button pressing detection is achieved by detecting skin color blobs within the button zones. Each button press is followed by a feedback sound. With on-screen buttons there is no need to use keyboard or mouse to control the system (note that the system is activated as soon as a face is detected by the system; no other user interaction is necessary). In our case the subjects are separated from the screen. If the subject can be closer, a touch screen can be used instead. Additional buttons may be added to change tints, coatings, frame thickness, etc.

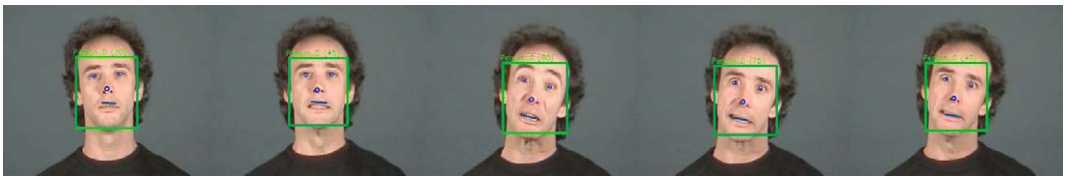


Fig. 7 Facial element detection samples for a sequence extracted from DaFEx

Fig. 10 Facial tracking

Fig. 11 Selecting the next glasses model

Glasses models are stored as images. The center of the glasses image is placed on the eye midpoint. This may lead to undesired results if the glasses 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. To tackle this, glasses placement gesture detection is added to the system.

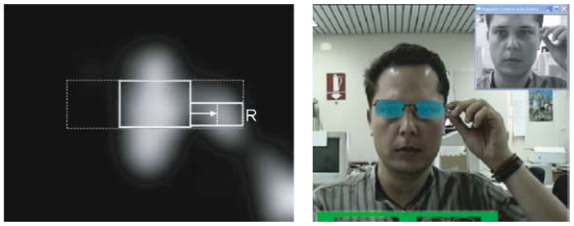
Robust and efficient hand pose detection in video is a challenging problem, mainly due to the inherent variability and flexibility of the articulated hand structure, the large domain of gestures, the restriction of real-time performance, varying illumination conditions, and complex background clutter. Therefore, different restrictions are commonly considered or even manual initialization is performed for this task.

The literature is rich in hand detection approaches. These approaches have traditionally been based on skin color segmentation (Storring *et al*., 2004), basically due to their reduced processing cost. Recent approaches (Kölsch and Turk, 2004; Stenger *et al*., 2004; Just *et al*., 2006; Wagner *et al*., 2006), however, have utilized Viola-Jones' object detection framework (Viola and Jones, 2004). Although frontal faces share common features (eyes, eyebrows, nose, mouth, hair), hands are not that easy to describe. They are highly deformable objects, so training a single cascade classifier for detecting hands is a complex and arduous task. For that reason, a different classifier for each recognizable gesture has been trained (Stenger *et al*., 2004), but also a single classifier for a limited set of hands has been proposed (Kölsch and Turk, 2004).

Considering the unrestricted context of this application, in terms of hand gestures, the use of multiple detectors would produce an approach not suitable for real-time processing. In consequence, we have chosen the skin color approach for faster processing. Instead of using a predefined color space definition, the information obtained from the face blob (see the previous section) is used to estimate the histogram-based (Swain and Ballard, 1991) skin color model for that individual (see a similar technique in Sanchez-Nielsen *et al*. (2005)). The skin color model is employed to locate other skin-like blobs in the image, and in consequence to find the hands for a given face using coherent skin blobs and considering anthropomorphic dimensions. This is done only for a robustly detected face, for which at least three facial features—the face itself, its eyes, and the nose or the mouth—have been detected as a trusted result. This consideration is used to reduce the possibility of false detections, i.e., false positives. A color model is then learned or updated (if already created for that individual) only from these trusted faces, reducing the probability of using erroneous face detections.

The glasses 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 (Fig. 12). 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.



(a) (b)

Fig. 12 Glasses placement gesture detection. (a) The biggest white rectangle represents the face area. Along its sides (dotted area), the rectangle *R* with the highest sum of skin pixels is found. Inside *R*, columns are analyzed from the face side through half this rectangle to check skin-color continuity. The same operation is performed on the other side of the face. An integral image is used to speed up the summing of skin pixels. (b) Gesture detection in action

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. However, the hand must be in contact with the head. In order to check the 'touching the head' condition, pixel values are analyzed 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 glasses upward and downward. When the hand no longer touches the head, the final relative displacement is stored with the current glasses model. Fig. 13 shows a sequence in which the user is fitting the glasses.

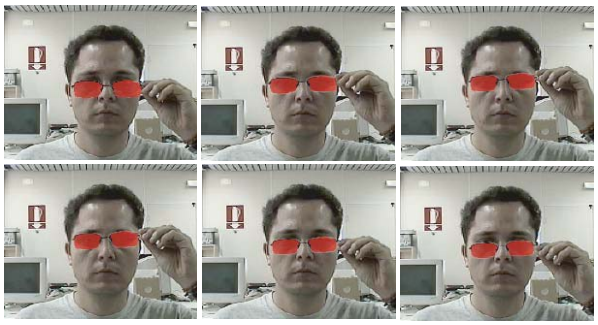


Fig. 13 Glasses placement sequence

6 Experiments

For completeness, we first show results for the face detection task. Seventy-four sequences corresponding to different individuals, cameras and environments with a resolution of 320×240 were recorded. They represented a single individual sitting and speaking in front of the camera or moderating a TV news program. The face pose was mainly frontal, but it was not controlled; i.e., lateral views and occlusions due to arm movements were possible. The eyes were not always visible. The total set contained 26 338 images.

To test the detector performance, the sequences were manually annotated; therefore, the face containers were available for the whole set of images. However, eye locations were available only for a subset of 4059 images. The eyes location allows us to compute the actual distance between them, which will be referred to 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:

(1) Correct face criterion: A face is considered correctly detected if the detected face overlaps at least 80% of the annotated area, and the sizes of the detected and annotated areas do not differ by more than a factor of 2.

(2) Correct eye criterion: 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. The threshold considered was EyeDist/4, similar to Jesorsky *et al*. (2001).

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 was notably slower than the others, but it provided eye detection for 78% of the detected faces, feature which is not considered by Viola-Jones' detector. As for ENCARA2, it is observed that it performed at least twice as fast as Viola-Jones' detector, and almost ten times faster than Rowley's. Speed is the main goal in our application, and the face detector is critical for the live-video selector. More details of ENCARA2 can be found in Castrillón *et al*. (2007).

Another important part of the system is homeostatic regulation. Fig. 14 shows the effect of the uniform, centered and selective strategies for luminance control (see Section 3). With the uniform strategy the face appeared darker because the bright zone on the left made the average luminance level larger. With both the centered and selective strategies the face was more clearly visible.

To test the effect of homeostatic regulation in the

Table 1 Results for face and eye detection processing using a Pentium IV 2.2 GHz CPU*

Detector	TD (%)			FD (%)			Processing time
	Faces	Left Eye	Right Eye	Faces	Left Eye	Right Eye	(ms)
Rowley	89.27	77.51	78.18	2.16	0.80	1.00	422.4
Viola-Jones	97.69	0.00	0.00	8.25			117.5
ENCARA2	99.92	91.83	92.48	8.07	4.04	3.33	45.6

* Taken from Castrillón *et al*. (2007). TD: correct detection ratio; FD: false detection ratio

face detection task, a performance measure was defined as the ratio between the number of detected faces in a second and the number of images per second. As ENCARA2 depends heavily on skin color and pattern matching to detect faces, we studied the influence of luminance and white balance on performance.

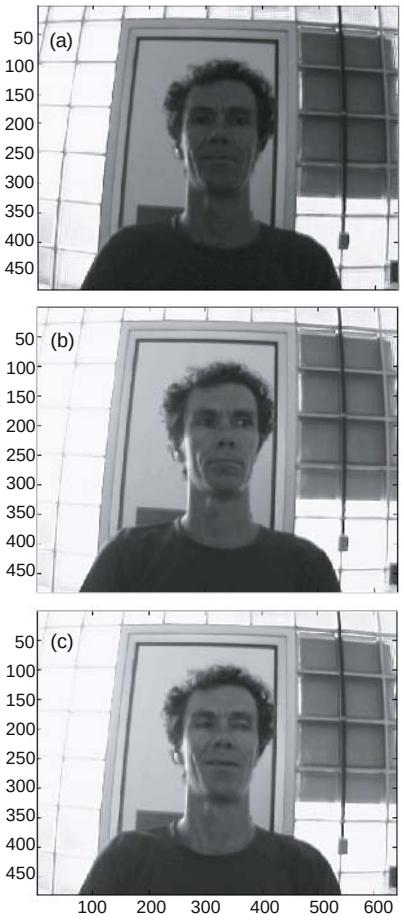


Fig. 14 Luminance regulation using (a) the uniform strategy, (b) the centered strategy, and (c) the selective strategy

Fig. 15 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). When the system started, the detection rate was high, and it decreased slightly when more lights were switched on (30–57 s). When the lights were switched on, both the h_luminance and h_whitebalance hormones went out of their desired states but the homeostatic mechanism recovered them after a delay, larger for the h_whitebalance hormone than for the h_luminance one.

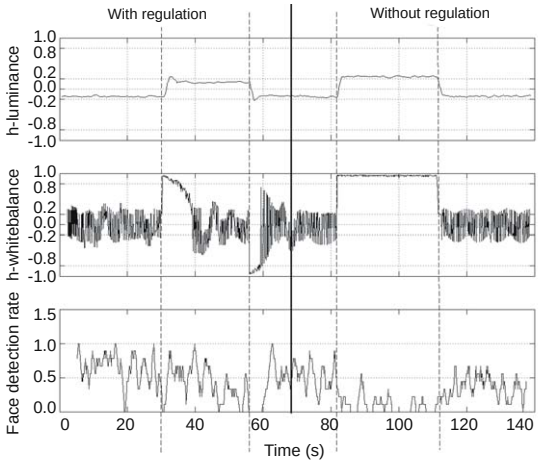


Fig. 15 h-luminance, h-whitebalance, and the face detection rate with and without homeostatic mechanism

The homeostatic mechanism was deactivated after 70 s, so when the conditions changed again, the state of the hormones was not recovered and the performance of the system decreased with a low rate of detections.

To speed up the recovery, we implemented an adaptive strategy using the two recovery levels of the hormones. When a hormone went into the urgent recovery zone we applied a more aggressive recovery strategy (increasing the aperture of the iris or the red and blue gains) than when hormones were in the normal recovery zone (Fig. 4). We repeated the experiment, but we were interested in only the delay until the hormones returned to the homeostatic regime. Fig. 16 shows the results with the adaptive strategy and it can be noted that the recovery time has been reduced.

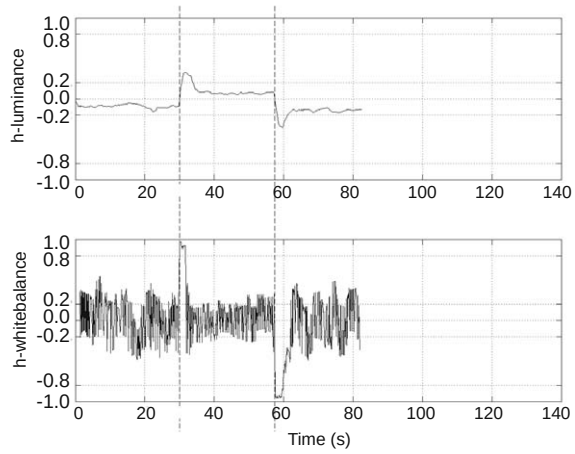


Fig. 16 Effect of the introduction of an adaptive strategy on the recovery time

To measure the localization capabilities of the system, seven video sequences were recorded in which a subject moved his head, from almost no motion to extreme movements. The eyes of the subject were manually located to have 'ground truth' data. Table 2 shows the number of frames, average inter-eye distance, and amount of motion of each video sequence. In sequences 6 and 7 the head movements were exaggerated for testing purposes and they did not represent a typical situation (most of the time the individual was not even looking at the screen). Fig. 17 shows the ground truth eye positions for two of the videos.

Table 2 Video sequences used in the experiments, ordered by increasing head motion*

Video sequence	Number of frames	Average inter-eye distance (std. dev.) (pixels)	Variance of eye position (pixels)
1	126	42.7 (1.6)	8.2
2	175	42.9 (2.2)	11.1
3	176	44.1 (1.8)	11.3
4	148	40.0 (2.8)	27.1
5	119	42.9 (2.8)	37.7
6	129	42.9 (4.4)	120.8
7	208	41.6 (3.1)	164.4

* Taken from images captured by camera 1

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) were larger than those using camera 1. Despite the higher resolution available in the zoom camera, this finding is explained by two reasons:

(1) The eye localizer of the second camera is much simpler (though 80% faster) than ENCARA2.

(2) Head motion often makes the eyes go out of the zoomed image.

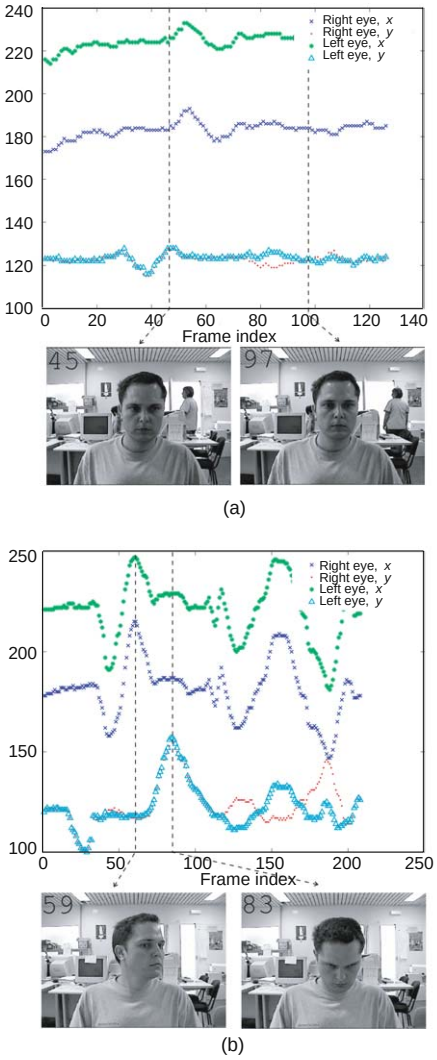


Fig. 17 Ground truth eye motion. (a) Sequence 1; (b) Sequence 7

As camera-1 data were better for localization, they had priority in the combination. That is, whenever localization data were available from ENCARA2, they were used to place the glasses on the screen. Data obtained with camera 2 were used only when ENCARA2 could not provide an eye pair localization.

The combined use of information of the two cameras did not improve either of them alone, except in the cases where numbers are marked with bold in Table 4. The use of the second camera, however, was advantageous in terms of the number of frames with available eye localization (Table 4). This allows the glasses to remain longer on the screen (currently glasses are not

Table 3 Average eye localization errors in pixels

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

REE=right eye error, LEE=left eye error. In parentheses is the standard deviation. Note that all values are referenced to camera 1 (for camera-2 localizations the mapping depicted in Fig. 8 is used). The bold number represents an improvement from the use of either camera-1 or camera-2 data alone

Table 4 Number of frames with available eye pair localization

Video	Number of frames		
sequence	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

The bold number represents an improvement from the use of either camera-1 or camera-2 data alone

drawn if no eye localization can be obtained, although it is also possible to draw them using the last known position), even if the individual is moving (Fig. 18).

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 are large head motions. The tracker was reinitialized every 60 frames.

Table 5 Squared errors of the eye midpoint position*

Video	Squared error (pixels)		Improvement
sequence	No tracking	With tracking	(%)
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

* The tracking of facial points was given a weight equal to the no-tracking eye midpoint localization

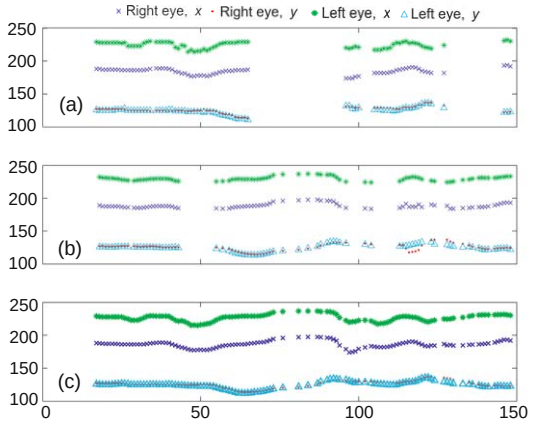


Fig. 18 Video sequence 4. (a) Camera-1 data; (b) Camera-2 data; (c) Combination of (a) and (b)

Figs. 19 and 20 show examples of situations in which no eye localization was obtained for at least two consecutive frames (only the first frame is shown). This situation happened only in sequences 6 and 8. Note that in most of the situations the subject was not frontal.

Currently, the whole system runs at 9.5 frames/s, with peaks of 12.8 frames/s, on a Core 2 Duo™ CPU at 1.86 GHz. Fig. 21 shows a subject trying six glasses models.

7 Conclusion

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

Fig. 19 Situations in sequence 6 in which no eye localization was obtained for at least two consecutive frames (only the first frame is shown)

Fig. 20 Situations in sequence 8 in which no eye localization was obtained for at least two consecutive frames (only the first frame is shown)

Fig. 21 Six glasses models

computer vision techniques. The system, running at 9.5 frames/s 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, such as in an optical shop. The system also has a face and eye detection module and a glasses management module.

Further improvements are possible. On the one hand, image resolution can be enhanced with more computing power. On the other hand, other interaction capabilities can be added to the system, such as zoom and mirror-like glasses. The system described can be easily adapted to the available hardware. Modern laptops, for example, include an integrated webcam and sufficient computing power to be used as a low-cost, portable eyewear selector.

Acknowledgements

The authors wish to thank Dr. Cayetano Guerra and J. Tomás Milán for providing the glasses models.

References

ABS, 2007. Smart Look. Available from http://www.smart-mirror.com [Accessed on July 30, 2007].

Activisu, 2007. Activisu Expert. Available from http://www.activisu.com [Accessed on July 30, 2007].

Arkin, R.C., Balch, T., 1997. AuRA: Principles and practice in review. *J. Exper. Theor. Artif. Intell.*, **9**(2-3):175-189. [doi:10.1080/095281397147068]

Azuma, R.T., 1997. A survey of augmented reality. *Presence*, **6**:355-385.

Battocchi, A., Pianesi, F., 2004. Dafex: Un Database di Espressioni Facciali Dinamiche. SLI-GSCP Workshop Comunicazione Parlata e Manifestazione delle Emozioni, p.1-11.

Bouguet, J., 1999. Pyramidal Implementation of the Lucas Kanade Feature Tracker. Technical Report, OpenCV Documents, Intel Corporation, Microprocessor Research Labs.

Breazeal, C., 1998. A Motivational System for Regulating Human-Robot Interaction. AAAI/IAAI, p.54-61.

Cañamero, D., 1997. Modeling Motivations and Emotions as a Basis for Intelligent Behavior. Proc. 1st Int. Conf. on Autonomous Agents, p.148-155. [doi:10.1145/267658.267688]

Carl Zeiss Vision, 2007. Lens Frame Assistant. Available from http://www.zeiss.com [Accessed on July 30, 2007].

Castrillón, M., Déniz, O., Hernández, M., Guerra, C., 2007. ENCARA2: real-time detection of multiple faces at different resolutions in video streams. *J. Vis. Commun. Image Represent.*, **18**(2):130-140. [doi:10.1016/j.jvcir.2006.11.004]

CBC Co., 2007. Camirror. Available from http://www.camirror.com [Accessed on July 30, 2007].

CyberImaging, 2007. CyberEyes. Available from http://www.cyber-imaging.com [Accessed on July 30, 2007].

Fiala, M., 2004. Artag, an Improved Marker System Based on Artoolkit. Technical Report, ERB-1111, NRC Canada.

Gadanho, S.C., Hallam, J., 2001. Robot learning driven by emotions. *Adapt. Behav.*, **9**(1):42-64. [doi:10.1177/105971230200900102]

GlassyEyes, 2009. Trying Eyeglasses Online. GlassyEyes Blog. Available from http://glassyeyes.blogspot.com [Accessed on Dec. 23, 2009].

Jesorsky, O., Kirchberg, K.J., Frischholz, R.W., 2001. Robust face detection using the Hausdorff distance. *LNCS*, **2091**:90-95. [doi:10.1007/3-540-45344-X_14]

Just, A., Rodriguez, Y., Marcel, S., 2006. Hand Posture Classification and Recognition Using the Modified Census Transform. Proc. Int. Conf. on Automatic Face and Gesture Recognition, p.351-356. [doi:10.1109/FGR.2006.62]

Kölsch, M., Turk, M., 2004. Robust Hand Detection. Proc. Sixth IEEE Int. Conf. on Automatic Face and Gesture Recognition, p.614-619. [doi:10.1109/AFGR.2004.1301601]

Kuttler, H., 2003. Seeing Is Believing. Using Virtual Tryons to Boost Premium Lens Sales. Available from http://www.2020mag.com [Accessed on Dec. 23, 2009].

Lee, J.S., Jung, Y.Y., Kim, B.S., Ko, S.J., 2001. An advanced video camera system with robust AF, AE and AWB control. *IEEE Trans. Consum. Electron.*, **47**(3):694-699. [doi:10.1109/30.964165]

Lepetit, V., Vacchetti, L., Thalmann, D., Fua, P., 2003. Fully Automated and Stable Registration for Augmented Reality Applications. Proc. 2nd IEEE and ACM Int. Symp. on Mixed and Augmented Reality, p.93-102. [doi:10.1109/ISMAR.2003.1240692]

Li, S., Zhu, L., Zhang, Z., Blake, A., Zhang, H., Shum, H., 2002. Statistical learning of multi-view face detection. *LNCS*, **2353**:67-81. [doi:10.1007/3-540-47979-1_5]

Low, K.H., Leow, W.K., Ang, M.H.Jr., 2002. Integrated Planning and Control of Mobile Robot with Self-Organizing Neural Network. Proc. 18th IEEE Int. Conf. on Robotics and Automation, p.3870-3875. [doi:10.1109/ROBOT.2002.1014324]

Lyu, M.R., King, I., Wong, T.T., Yau, E., Chan, P.W., 2005. ARCADE: Augmented Reality Computing Arena for Digital Entertainment. Proc. IEEE Aerospace Conf., p.1-9. [doi:10.1109/AERO.2005.1559626]

Morgan, E., 2004. Dispensing's New Wave. Eyecare Business. Available from http://www.eyecarebiz.com [Accessed on Dec. 23, 2009].

Nanda, H., Cutler, R., 2001. Practical Calibrations for a Real-Time Digital Omnidirectional Camera. Proc. Computer Vision and Pattern Recognition Conf., p.3578-3596.

OfficeMate Software Systems, 2007. iPointVTO. Available from http://www.opticalinnovations.com [Accessed on July 30, 2007].

Picard, R., 1997. Affective Computing. MIT Press, Cambridge, MA.

Practice, P., 2007. FrameCam. Available from http://www.paperlesspractice.com [Accessed on July 30, 2007].

Reimondo, A., 2007. OpenCV Swiki. Available from http://www.alereimondo.no-ip.org/OpenCV/ [Accessed on Dec. 23, 2009].

Roberts, K., Threlfall, I., 2006. Modern dispensing tools. Options for customised spectacle wear. *Optometry Today*, **46**(12):26-31.

Rodenstock, 2007. ImpressionIST. Available from http://www.rodenstock.com [Accessed on July 30, 2007].

Sanchez-Nielsen, E., Anton-Canalis, L., Guerra-Artal, C., 2005. An autonomous and user-independent hand posture recognition system for vision-based interface tasks. *LNCS*, **4177**:113-122. [doi:10.1007/11881216_13]

Schneiderman, H., Kanade, T., 2000. A Statistical Method for 3D Object Detection Applied to Faces and Cars. IEEE Conf. on Computer Vision and Pattern Recognition, p.1746-1759.

Stenger, B., Thayananthan, A., Torr, P., Cipolla, R., 2004. Hand pose estimation using hierarchical detection. *LNCS*, **3058**:105-116. [doi:10.1007/b97917]

Storring, M., Moeslund, T., Liu, Y., Granum, E., 2004. Computer Vision Based Gesture Recognition for an Augmented Reality Interface. 4th IASTED Int. Conf. on Visualization, Imaging, and Image Processing, p.766-771.

Swain, M.J., Ballard, D.H., 1991. Color indexing. *Int. J. Comput. Vis.*, **7**(1):11-32. [doi:10.1007/BF00130487]

Velasquez, J., 1997. Modeling Emotions and Other Motivations in Synthetic Agents. Proc. AAAI Conf., p.10-15.

Velasquez, J., 1998. Modeling Emotion-Based Decision Making. *In*: Canamero, D. (Ed.), Emotional and Intelligent: The Tangled Knot of Cognition. AAAI Press, Springer Netherlands, p.164-169.

Viola, P., Jones, M.J., 2004. Robust real-time face detection. *Int. J. Comput. Vis.*, **57**(2):137-154. [doi:10.1023/B:VISI.0000013087.49260.fb]

Visionix, 2007. 3DiView 3D Virtual Try-on. Available from http://www.visionix.com [Accessed on July 30, 2007].

Wagner, S., Alefs, B., Picus, C., 2006. Framework for a Portable Gesture Interface. Proc. 7th Int. Conf. on Automatic Face and Gesture Recognition, p.275-280. [doi:10.1109/FGR.2006.54]