Tracking-dependent and interactive video projection

Matei Mancas (1), Donald Glowinski (2), Pierre Bretéché (3), Jonathan Demeyer (1), Thierry Ravet (1), Gualtiero Volpe (2), Antonio Camurri (2), Paolo Coletta (2)
(1) FPMS, Mons, Belgium (2) Casa Paganini/InfoMus Lab, Genova, Italy
(3) Laseldi Lab, Montbéliard, France

Abstract

Gestures' expressivity, as perceived by humans, may be related to the amount of attention they attract. In this paper, we present three experiments that quantify behavior saliency by the rarity of selected motion and gestural features in a given context. The first two ones deal with the current quantity of motion of a person's silhouette compared to a brief history of his quantity of motion values and with the current speed compared to a brief history of the person's speed. The third one focuses on the motion speed of a person compared to the motion speed of other persons around him. Considering both features (speed and quantity of motion) and contexts (space and time), we compute an attention index providing cues on the behavior novelty. This can be considered as a preliminary step to an expressive gesture analysis based on behavioral. In order to achieve accurate tracking, a fusion between color and IR camera streams is achieved. This fusion let us have a robust tracking system with respect to illumination and partial occlusions issues.

***Index Terms*—computational attention, saliency, rarity, data fusion, tracking, gestures.**

I. PRESENTATION OF THE PROJECT

A. Introduction: towards a context-based gestural analysis

A lot of research effort has been devoted to robustly track humans in a scene and to analyze their gestures in order to individuate and characterize their behavior. Gestural analysis, often applies in situations where either the human on which the analysis is carried on is previously selected or the same kind of analysis is performed to all the subjects that can be distinguished in the scene. A recent field of research aims at investigating collective behaviors [1]. Still, the object of the analysis is already defined and the work mainly focuses on characterizing collective displacements. The possibility of dynamically selecting the person to carry analysis on or to adapt analysis to the current behavior of a person in a context-dependent way would open new directions for gesture research. Human beings naturally show the capacity to dedicate limited perceptual resources to what is of particular interest. However, computers capacity to exhibit a behavior worthy of attention remains very limited.

B. Computational attention interest in Human-Computer Interfaces (HCI)

In many real situations, where people interact freely together, it can be difficult to select the participants that exhibit a behavior worthy of attention.
The design of (expressive) gestures interface can gain from a better understanding of individual- and context-dependent human behavior. It can ensure their usability in a more naturalistic environment. Automatic attention cues are also able to simplify information access in those complex-situations which leads, in the HCI domain at foster interaction, anticipating focus of attention as automatic zoom on the Region Of Interest (ROI).

C. Work overview

The project consisted in two main steps:

- *A robust tracking system*

This system uses both color video cameras and infra-red cameras to be robust enough to illumination changes and partial occlusions. Infra-red (IR) camera is much less sensitive to light changes if those lights are not directly in the camera field of view (FOV). The color camera FOV is larger and it is able to keep on the tracking process if infra-red markers are occluded or out of the infra-red camera FOV.

- *Motion attention: human-like reactions*

Once participant tracking is robust enough to handle naturalistic scenarios, an automatic attention index can be computed which highlights which movements should be the most "interesting" for a human observer. Attention is computed both in a spatial and in a temporal context on several features: speed and quantity of motion.

In section II, after a hardware and software overview, we will describe the video data acquisition and processing (blob segmentation) for each one of the two video modalities. In section III, the fusion mechanism which leaded to a robust tracking system will be explained. Section IV deals with attention computation both in spatial and temporal context. Finally, we will conclude by a discussion in section V. The source code of this project and some video demos can be found on the eNTERFACE 2008 workshop website [2].

II. SIGNAL ACQUISITION

A. *Material and system overview*

We developed a setup that analyzes human behavior in a flexible environment with regard to illumination changes. Three cameras were used to capture video: two "Eneo VKC-1354" color analogical cameras with a 752*582 pixel resolution at 25 frames per second (fps) and one "Imaging Source DMK 31BF03" monochrome digital camera delivering a 1024*768 pixel resolution at 30 fps. They were equally placed on the side of a 3 x 3 meters area, at a height of 2.5 meters, downward looking above the participants and recording with constant shutter, manual gain and focus. The participants used each one a red hat (for color segmentation) and a halogen light which emits visible but also infra-red light (for IR segmentation) in all space directions. Figure 1 shows the setup configuration.

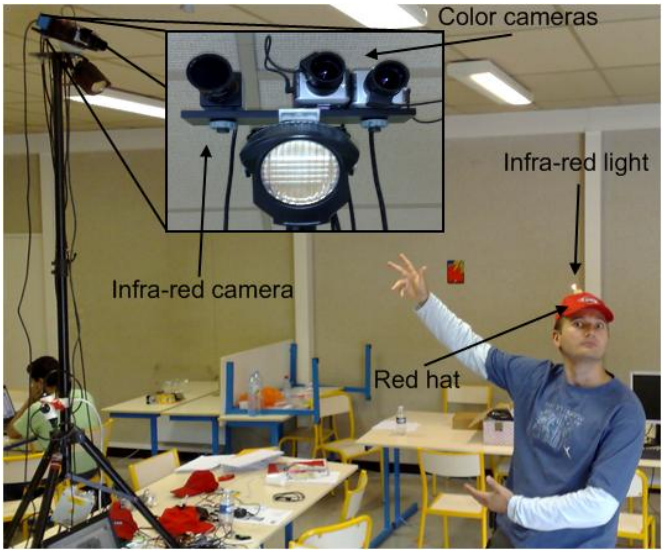


Fig. 1: Experimental setup

As described on Figure 2, two computers were used to perform the IR video and color video stream acquisition and processing (IR-and Color-based blob detection and tracking). The third PC was used to achieve data fusion between the IR and color stream data and to further higher-level processing and rendering.

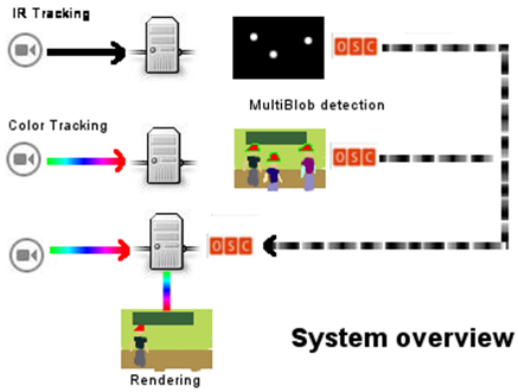


Fig. 2: System overview

A non-invasive video-based approach was adopted based on the EyesWeb XMI free software platform. We were interested in automatically extracting the displacement of people moving in front of the camera and computing their motion features. Head detection and tracking solutions were privileged to fully exploit the reduced available space, obtain depth information and avoid occlusions due to the interactions between participants.

The EyesWeb XMI (www.eyesweb.org) is a free software platform [3]. It consists of two main components: a kernel and a graphical user interface (GUI). The GUI manages interaction with the user and provides all features needed to design patches. It allows hand fast development of custom interfaces for use in artistic performances and interactive multimedia installations. The kernel manages real-time data processing, synchronization of multimodal signals. It supports the integration of user-developed plugins; an SDK (Software Development Kit) is provided to simplify the creation of such plugins by means of Microsoft Visual C++. The user-developed plugins, together with the ones provided with EyesWeb are the building blocks that the end user can interconnect to create an application (*patch*).

B. *Blob Detection*

The present system's analysis of human activity starts from foreground segmentation based on the analysis of the color and infra-red video streams. This analysis provides a binary mask of the spatial extension of the region of interest through time (blob detection).

- *Color video stream*

A skin color detection algorithm was used on the signal coming from the color camera. We developed a modified version of the Continuously Adaptive Mean Shift Algorithm (CamShift) which is itself an adaptation of the Mean Shift algorithm for object tracking. Our method consisted in manually selecting the color of interest (COI) which is converted into a HSV colorimetric system. The set of colored pixel is quantized in a one-dimensional histogram to create the COI's model. Furthermore, a bandwidth of acceptable Hue and Saturation values are defined to allow the tracker to compute the probability that any given pixel value corresponds to the selected color. In order to enhance the system robustness to illumination changes, the color model was updated in several areas of the scene which have different illumination. In that way the color model was resistant to moderate illumination changes as those present on our scenario visible on Figure 3 between the left-side and the right-side of the scene.

Fig. 3: Color-based red hat blob detection

- *Infra-red video stream*

The signal coming from the IR camera is not affected by illumination variations but might be artefacted by light reflections or some static infrared sources. We processed the video stream with a background subtraction to eliminate static elements. Then, we binarized the signal with an empirically-tested threshold value to extract the moving regions of interest (blobs). Figure 4 shows that the IR lights located on the top of the red hats are very clearly detected.

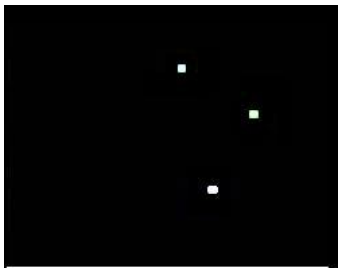
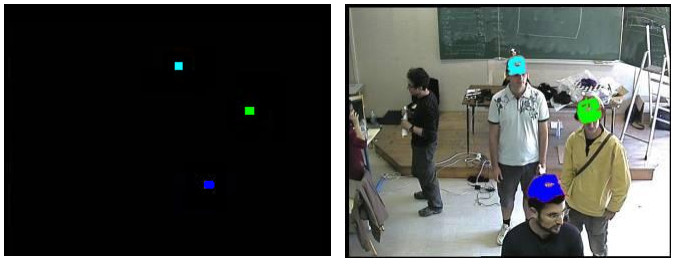


Fig. 4: Infra-red lights located on the top of the red hats detection

C. Single and Multi-Blob Tracking

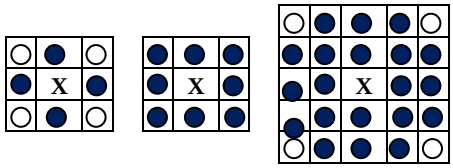
Resulting from the pre-processing step (color or IR based), we obtained a binary image where white represents the foreground objects. The next step is to assign a label that identifies the different white blobs, and to track them. The tracking result can be seen on Figure 5.



(1) **(2)**

Fig. 5: Multi-blob tracking : from left to right the labeling (identification) of (1) the three blobs of the IR video stream (2) the three blobs of the color video stream

To achieve the image tracking, we defined an adjacency measure based on the n-connectivity of two pixels (Figure 6).



(1) **(2)** **(3)**

Fig. 6. From left to right, illustration of (1) a 4-connectivity: the four filled circles (pixels) are connected to the one of interest (the cross); we defined this case as adjacency 1. (2) 8-connectivity, defined as adjacency = 2; (3) 20-connectivity defined as adjacency = 3. This adjacency measure can be generalized to n-connectivity

Basing on this definition, we have used the following algorithm for image segmentation and tracking:

- *Definitions*

Image segmentation *image_sgm* (video frame *f*) returns the set of distinct connected components *cc* (*f*) such that each pixel of an item in *cc* (*f*) is at distance x ≥ threshold from any other pixel belonging to other items in *cc* (*f*)
Valid region *valid_reg* is user defined as the maximum Euclidean distance between two blobs' baricentres in two consecutive frames.
Minimum difference *min_diff* returns the following:

$$\text{min\_diff} = \alpha \times \text{dist}(b_1, b_2) + \beta \times |\, \text{area}(b_1) - \text{area}(b_2) \,| \qquad \textbf{(1)}$$

where *dist* (b1, b2) is the Euclidean distance between the baricentres of blobs b1 and b2 and α and β are the weights of area and position, comprised between:
A = [0,1] and β = [0,1]
These values are manually decided according to the cam's FOV and the foreground objects to track (e.g. humans)

- *Procedure*

Initialization: $t = 0$
cc(*t*=0) = *image_sgm(frame(0));*
store *cc*(0) and assign a new label to each item in *cc*(0)
For $t = 1, 2, \ldots$
cc(*t*) = *image_ sgm(frame(t));*; i.e. the set of distinct blobs in current frame (frame(t));
store *cc*(t)
for each item *x* in *cc*(*t*)
find *x* in *cc*(*t*-1) which minimizes *min_diff*
if *x* is in *valid_reg*, then
{
assign *y* the same label as *x*
if *x* is recognized in N consecutive frames,
then the item is considered to be trackable (i.e., we can measure its velocity, direction, etc.)
else
the item is unstable (recognized but not tracked)
}
else
y is assigned a new distinct label

III. DATA FUSION

A. A common reference

Prior to any fusion algorithm, we need to provide a common reference to the signals to be fused. As we worked with different cameras, their fields of view (FOV) were different. A robust transformation was necessary to match the position of a point computed on a frame from one video camera, to the position of the same point computed on a frame from another video camera. The projective transformation meets our need because the perspective is subject to change with respect to the camera focal distance. A correction of the radial distortion was not necessary as the distortion due to wide-angle lenses was not very important.

The projective transformation conserves the proportions between two sets of two points: a quadrilateral is projected onto another one. Since it is not a linear transformation, we used homogeneous coordinates to compute the transformation with a matrix product.

The transformation matrix H_{ab} performing the projection of the points P_a in image "a" to the points P_b in image "b" can be written in homogeneous coordinates as follows (points are located on the same subjective plane):

$$P_a = \begin{pmatrix} x_a \\ y_a \\ 1 \end{pmatrix}, P'_b = \begin{pmatrix} x'_b \\ y'_b \\ w_b \end{pmatrix}, \mathbf{H}_{ab} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}$$

Then:

$$P'_b = \mathbf{H}_{ab} \cdot P_a \quad \text{and} \quad P_b = {}^{P'_b}/_{w'} = \begin{pmatrix} x_b \\ y_b \\ 1 \end{pmatrix} \qquad \textbf{(2)}$$

In order to visually test the accuracy of our transformation, we developed EyesWeb blocks which perform projective image transformation according to an input matrix H_{ab}. We also developed a block which composes this projective matrix from the correspondence of four points in the original and the referent image.

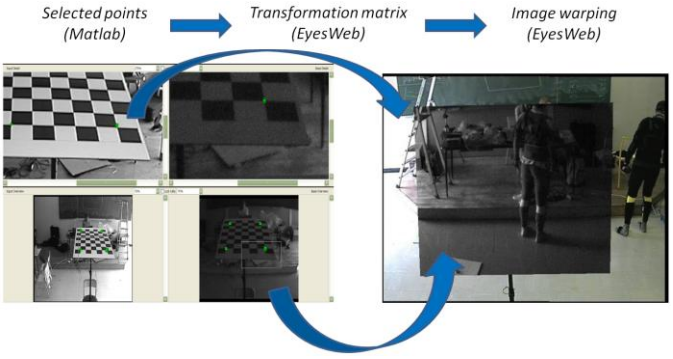


Fig. 7. IR and color video stream registration process. From left to right: selection of the four corresponding points in a color and IR snapshot in Matlab, computation of the transformation matrix in EyesWeb, image warping in EyesWeb

Figure 7 displays the entire transformation process. We first selected four points in a snapshot of the IR and color video streams using Matlab. Then, the coordinates of those points are used by EyesWeb to compute the transformation matrix. Finally, this matrix is used to warp the initial image: the superposition of the color and IR image show a good registration in the plane where the points in the two images were chosen. Figure 8 shows how the initial video stream is transformed into a new video stream according to a projection matrix H_{ab}.

Fig. 8. Left: initial video frames, right: real-time projective warping using OpenCv-based EyesWeb block

Once the registration was visually validated, the matrix H_{ab} was used to compute the projective transformation for the blob baricenters and not for the whole image. The use of a newly implemented EyesWeb block which performs matrix multiplication greatly reduced the computational cost: this solution avoided the computation of the projective transformation for all pixels in each frame of the video stream. Figure 9 shows the rendering of the final video stream with the surimposition of IR (green markers) and color video tracking (red markers) which converges onto the participants hats.

Fig. 9. IR tracking of the lights on top of the red hats (green), color tracking of the top of the red hats (red

The projective transform blocks were achieved by using some functions already implemented in OpenCV (Open Computer Vision) [4]. OpenCV is a library developed by Intel with a BSD license. EyesWeb can easily wrap OpenCV functions to handle them as blocks in the software platform.

B. Confidence Level

The second step performed before the fusion was to compute a confidence level on each modality (IR and color).
The weight (confidence level) changes according to the participants' visibility with respect to the camera's field of view (FOV) and obstruction occurrences. Data range from 0 when the participants are not visible by the camera to 1 when they are visible. For the color modality, the confidence level also gradually changes between 0 and 1 depending on blob area variations: abrupt variations due for example to sudden illumination changes decrease the confidence level whereas stable blob areas over time increase it.

- *Color video stream confidence level*

Blob tracking with skin color detection can have a lack of accuracy due to illumination variation or even a loss of coordinates due to visibility issues (obstruction or disappearance from the camera's FOV).
This confidence level is built on two hypotheses. The first one is to consider location information obsolete after a short delay (mobile objects tracking). The second hypothesis is that blob's area abrupt variation can be related either to some undetected surface or to unwanted detections. According to these assumptions, the confidence level CL_{VID} is computed as the conjunctions of two terms:

$$CL_{VID} = (\textit{blob in FOV}) \wedge (\textit{stable blob's area}) \quad \textbf{(3)}$$

where $CL_{VID} \in [0, 1]$.
For the "*blob in FOV*" indicator computation, we used a clock generator to check that the elapsed time since the last coordinates' acquisition keeps below an acceptable delay. The "*stable blob's area*" indicator is computed on a temporal sliding window. We compared the current blob's area value with its mobile average and returned an inverse variation rate (minimum value between the ratio and its inverse).

- *Infra-red video stream confidence level*

The IR light tracking might be interrupted in two situations: the blobs might be obstructed by the occlusion of one participant with respect to the others or they can come out from the camera FOV. The blob detection could also be corrupted when the tracked participants move closely. A binary confidence level was developed to handle these tracking issues. We measured the elapsed time since the last valuable detection. If this delay exceeded an empirically tested threshold, the confidence degree was set to 0 until a new detection occurred. The confidence level CL_{IR} is computed as:

$$CL_{IR} = (\textit{blob in FOV}) \quad \textbf{(4)}$$

where $CL_{IR} \in \{0, 1\}$.
We measured the elapsed time since the last valuable detection. If this delay is over a threshold that we fixed, "*blob in FOV*" is fixed at 0 until a new detection occurs.

C. Fusion algorithm

- *Single blob fusion*

In order to fuse the 2D coordinates coming from the IR and the color video stream, a weighted mean rule was applied. The weights are the respective confidence levels previously computed for both modalities.

- *Extension to multi-blob fusion*

In the case of multi-blob tracking, each blob coordinates set was extracted and computed in each modality (color and IR) together with their confidence level.
Nevertheless, the fact that the blobs from both modalities will remain linked to the same target during the experiment is not obvious. To prevent this issue, our method needed to test continuously the relation between the coordinates of the same blobs located in the two modalities. We created a new fusion EyesWeb block where we considered only the non null confidence level elements and we matched each one of them in a modality with the nearest neighbor point (following a Euclidian distance) in the other modality. We fused these couples with the same weighted mean rule described in the previous section. Figure 10 shows the fusion results.

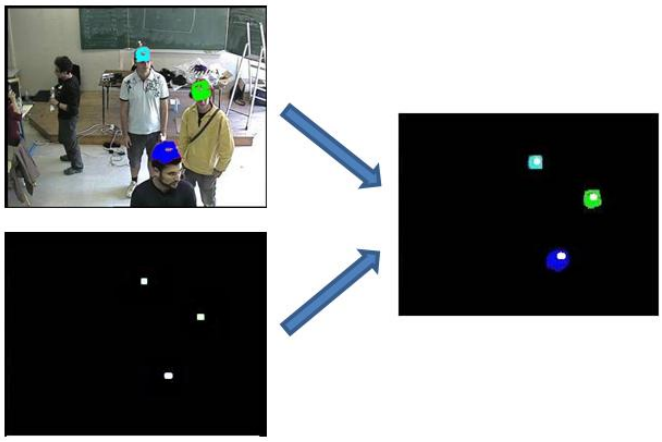


Fig. 10: Top-left: color image tracking, bottom-left: IR image tracking, right: modality fusion

IV. Salient Gestures: an attention filter

A. Computational Attention

The aim of computational attention is to automatically predict human attention on different kinds of data such as sounds, images, video sequences, smell or taste, etc… This domain is of a crucial importance in artificial intelligence and its applications are numberless from signal coding to object recognition. Intelligence is not due only to attention, but there is no intelligence without attention.
Attention is also very closely related to memory through a continuous competition between a bottom-up or unsupervised approach which uses the features of the acquired signal and a top-down or supervised approach which uses observer's a priori knowledge about the observed signal. We focused here only on bottom-up attention due to motion.

While numerous models were provided for attention on still images, time-evolving two-dimensional signals as videos have been much less investigated.
Nevertheless, some of the authors providing static attention approaches generalized their models to the time dimension: Dhale and Itti [5], Yee and Pattanaik [6], Parkhurst and Niebur [7], Itti and Baldi [8], Le Meur [9] and Liu [10]. Motion has a predominant place and the temporal contrast of its features is mainly used to highlight important movements. Zhang and Stentiford [11] provided a motion analysis model based on comparing image neighborhoods in time. The limited spatial comparison led to a "block-matching"-like approach providing information on motion alone more than on motion attention. Boiman and Irani [12] provided an outstanding model which is able to compare the current movements with others from the video history or video database. Attention is related to motion similarity. The major problem of this approach is in its high computational cost.
As we already stated in [13] and [14], a feature does not attract attention by itself: bright and dark, locally contrasted areas or not, red or blue can equally attract human attention depending on their context. In the same way, motion can be as interesting as the lack of motion depending on the scene configuration. The main cue which involves attention is the rarity or the contrast of a feature in a given context. A pre-attentive analysis is achieved by humans in less than 200 milliseconds. How to model rarity in a simple and fast manner?
The most basic operation is to count similar areas in the context. Within information theory, this simple approach based on the histogram is close to the so-called self-information. Let us note m_i a message containing an amount of information. This message is part of a message set M. A message self-information $I(m_i)$ is defined as:

$$I(m_i) = -\log(p(m_i)) \quad \textbf{(5)}$$

where $p(m_i)$ is the probability that a message m_i is chosen from all possible choices in the message set M or the occurrence likelihood. We obtain an attention map by replacing each message m_i by its corresponding self-information $I(m_i)$. The self-information is also known to describe the amount of surprise of a message inside its message set: rare messages are surprising, hence they attract our attention.
We estimate $p(m_i)$ as a two-terms combination:

$$p(m_i) = A(m_i) \times B(m_i) \quad \textbf{(6)}$$

The $A(m_i)$ term is the direct use of the histogram to compute the occurrence probability of the message m_i in the context M:

$$A(m_i) = \frac{H(m_i)}{Card(M)} \quad \textbf{(7)}$$

where $H(m_i)$ is the value of the histogram H for message m_i and $Card(M)$ the cardinality of M. The M set quantification provides the sensibility of $A(m_i)$: a smaller quantification value will let messages which are not the same but quite close to be seen as the same.
$B(m_i)$ quantifies the global contrast of a message m_i on the context M:

$$B(m_i) = 1 - \frac{\sum_{j=1}^{Card(M)} |m_i - m_j|}{(Card(M) - 1) \times Max(M)} \quad \textbf{(8)}$$

If a message is very different from all the others, $B(m_i)$ will be low so the occurrence likelihood $p(m_i)$ will be lower and the message attention will be higher. $B(m_i)$ was introduced to avoid the cases where two messages have the same occurrence value, hence the same attention value using $A(m_i)$ but in fact one of the two is very different from the others while the other one is just a little different.
In order to get a fast model of motion attention we have here a three-level rarity-based approach. While the first two approaches are bottom-up and use motion features context to attract computer's attention, the last one is mostly top-down and it learns a model of the scene which will be able to modify bottom-up attention by inhibiting some movements for example. The three levels of motion attention we propose here are:

- *Low-level instantaneous motion attention:*

Motion features are compared in the spatial context of the current perceived frame. Rare motion behaviors should immediately pop-out and attract attention. This low-level approach is pre-attentive (reflex) and it uses no memory capacities.

- *Middle-level short-term motion attention:*

Once a moving object was selected using low-level attention, its behavior into a short temporal context is than observed. Short-term memory (STM) is here used to save an object motion during 2 or 3 seconds (for longer time periods, motion details of an object are forgotten). Rare behaviors of an object through time will be quoted as interesting while repeating motion will be less important.

- *High-level long-term motion attention:*

This third top-down attention approach uses long-term memory capacities and it is a first step through motion and scene comprehension. The attention level of each pixel through time is accumulated which leads in areas of the scene which concentrate attention more than the others : a street accumulates more attention through time than a grassy area close to it, a tree which moves because of the wind or a flickering light will also accumulate attention through time. The scene can thus be segmented in several areas of attention accumulation and the motion in these areas can be summarized by only one motion vector per area. If a moving object passes through one of these areas and it has a motion vector similar to the one summarizing this area, its attention will be inhibited. If this object is outside those segmented attention areas or its motion vector is different from the one summarizing the area where it passes through, the moving object will be assigned a very high attention score. This third attention step builds an attention model learnt from instantaneous and short-term attention steps which is able to inhibit bottom-up attention if it corresponds to the model or to enhance it if the motion does not match with this model.

Within this project we developed an implementation of the two bottom-up motion attention comparing current motion features with a spatial and a short-term temporal context. The third top-down attention model is further discussed in section V of this article.

B. *Instantaneous motion attention*

An implementation of the spatial motion rarity was achieved as an EyesWeb XMI patch by using Equation (5) with $p(m_i) = B(m_i)$. In the scenario three people were tracked and their instantaneous speed was used. As only 3 motion vectors were available, the computation of the rarity $A(m_i)$ had not much sense from a statistical point of view.
Fig. 11 shows part of the tested scenario. Three people moving or not are present behind the cameras. Their instantaneous velocity vectors V_1, V_2, and V_3 are computed.

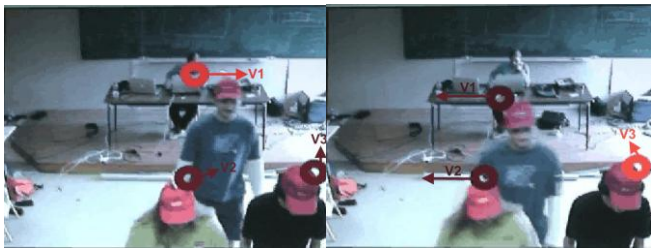


Fig. 11: Left: fastest moving object (V_1) is the most important (hot red), Right: slower moving object (V_3) is the most important.

On the left-side V_1 is very different from V_2 and V_3: V1 has high speed amplitude while V_2 and V_3 are very slow or stopped. In this case the faster object has a higher attention score (hot red) compared with the lower attention score (darker red) of V_2 and V_3. This situation can be compared with the one which often occurs if a classical motion detection module is used and where faster motion is well highlighted. That is not the case on the right image where the most different speed is V_3 (stopped) while V_1 and V_2 are quite the same (very fast). In this case, the attention score is higher (hot red) on V_3 which does not move while fast moving objects do not attract a large amount of attention. The result of this approach shows a different behavior compared to a simple motion detection algorithm. This first step enables the computer to choose the most "interesting" moving object very efficiently and then to apply to it short-term motion attention.

C. *Short-term motion attention*

- *Trajectory-related features*

The moving object speed was computed on a history of 3 seconds and the speed mean was computed on a 10 frames sliding window in order to avoid a too high variability of the speed due to segmentation and tracking noise.
The speed range was divided into 3 bins: static or very low speed, normal speed and high speed. The Equation (5) was applied with $p(m_i) = A(m_i)$ because there was enough data within the 3 seconds history to get a statistically reliable occurrence likelihood $p(m_i)$. Moreover, the contrast between the 3 speed bins is always very high so the $B(m_i)$ term is here less important.

Fig. 12: The brutal speed change is more important (red) then the speed value: the amplitudes of V(T1) and V(T4) are the same but they have a different attention score. Similar behavior can be seen with V(T2) and V(T3).

Fig. 12 shows the tested scenario. One participant moves his head from left to right very fast, then normally and finally he stops. When a change in the speed is detected (stop to normal speed, normal speed to high speed, high speed to stop, etc...), the attention score is very high, but it decreases exponentially when a stable speed occurs.

- *Silhouette-related features*

The same algorithm as in the previous point was applied but the feature used here was the Quantity of Motion (QoM). This measure is obtained by integrating in time the variations of the body silhouette (called Silhouette Motion Images - SMI).

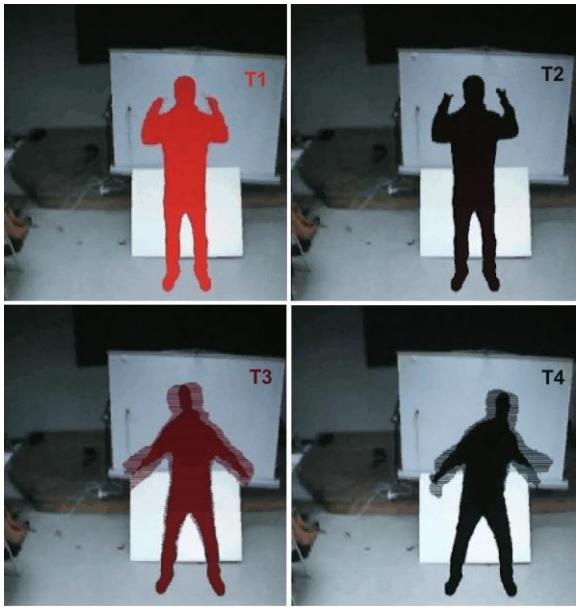


Fig. 13: The brutal QoM change is more important (red) than its value: the QoM at T1 and T2 are the same but they have a different attention score. The same behavior can be seen between T3 and T4.

Fig. 13 shows the test case. One participant is stopped, moves normally and moves a lot. If he moves fast while within his temporal context he is majorly stopped, this will be interesting. But if he moves fast while within his temporal context he also majorly moves fast the computer will classify this movement as uninteresting because it is repetitive and no new information is brought.
As in section IV.B, the short-term motion attention algorithm was implemented as an EyesWeb XMI patch.

D. Spatial and temporal motion attention fusion?

The previous two points (B and C) provide rarity-based attention indexes for moving objects based on spatial and on short-term temporal contexts. An interesting point is about a possible fusion of the results of those two approaches: how simultaneously take into account attention based on two contexts which do not have the same nature?
It is impossible to compare a given temporal context with a spatial one on the same basis: time and space are orthogonal which is also confirmed by the fact that time and space features are processed in two separate regions of the brain [15]. Instead of trying to fuse those two kinds of attention indexes, it seems more realistic to state that one of them (spatial context) is pre-attentive and it occurs first while the second one (short-term) is attentive and it focuses on the temporal context of an object previously selected by the spatial attention. Once this object is analyzed by the temporal attention, another one which may spatially pop-out can be tracked and its attention computed. Thus, spatial attention selects potential interesting moving targets while temporal attention verifies if those targets have also interesting behaviors through time.
In our opinion it is much more relevant to fuse the motion spatial attention map with a static image attention map based on color and gray level rarity [16]. As the context is in this case the same, it is realistic to compute rarity cues and compare static images (as background) and motion rarity (foreground) to get a final spatial attention which takes into account motion but also the other static pre-attentive features as colors or shape.

V. Results and Discussion

A. Results and rendering

We developed a solution to track human movement based on the fusion of two video modalities: a color and IR system. This solution showed robustness with respect to the light changing conditions and partial occlusions.
We analyzed spatio-temporal profiles of people activities at different levels of details. At gross level, human activity was analyzed in terms of the spatial trajectory of moving blobs corresponding to heads. However trajectory, by itself, hardly provided detailed information about the performed gestures [17]. A more-detailed level of person's activity was analyzed in terms of full-body quantity of motion (QoM) and found more relevant to characterize motion sequences and related gestures.
An attention index was developed to highlight motion saliency. Both the spatial and temporal contexts were used to compute rarity, thus attention indexes which provides additional information compared to a simple motion detection algorithm. We showed that motion perception can be radically different from simple motion detection: depending of the context, the lack of motion appeared to be more "interesting" than motion itself. Spatial context is used to select the participant which exhibits the highest saliency. Then, the selected participant can be tracked on a short period of time in order to see if he also has a salient behavior over time. This project is a first step towards systems which have more human-like reactions and perceive signals instead of only detecting them.

B. Further improvements

Several improvements could be achieved in addition to improve the already implemented system:

- *Video stream synchronization*

The blob detection data flows were not synchronized. A delay could be found between the tracking achieved on the IR video stream and the one achieved on the color video stream. This is due to the fact that the analog cameras and the digital camera had not the same time response. The output signal of the first ones must be digitalized with an extern FireWire/AD converter. We could observe a time lag between the video output and the filmed scene.
Moreover, the computation time for the color video processing is longer than the one needed for the IR video processing. Finally, the computers we used for the two video stream processing had not the same performances in terms of both CPU and graphical card.
The data flow synchronization could be improved if both cameras are digital and have the same characteristics. The synchronization could be even simpler if both cameras are acquired on the same computer using several FireWire ports. In this case, the EyesWeb XMI possibility to synchronize several processes on a single platform could be used and the two data flows would be perfectly synchronized. EyesWeb XMI highly improved the multimodal synchronization possibilities [3]: each block has two additional pins called "Sync-in" and "Sync-out" which can be used to propagate synchronization clock signals between blocks.

- *Confidence level improvement*

The present confidence level (CL) characterizes how much we should be confident in the tracked object position in each modality (IR and color video stream) in order to fuse them using a weighted mean. For the current implementation, we only used two assumptions: time since the last valuable detected position and smoothness of the detected blob variation over time. The first way to improve the modality confidence level could consist in adding other assumptions. For example, as we are tracking human heads, we could assume that abrupt speed variation is correlated with tracking inaccuracy and to decrease the positioning CL.
An additional improvement could consist in defining a confidence level for the fusion itself that we call fusion confidence level (FCL). A tracked object position could have a high CL in both modalities but the final fusion may not be

necessarily accurate if the two positions are very far one from the other for example. Other elements like speed and direction could help in defining if two positions from two different modalities can be fused. This FCL is able to point out if the data fusion should be considered or not.

- *Additional features: getting closer to human attention*

An interesting improvement could also be achieved by computing attention indexes for many other features. We can think first about motion direction which was not taken into account during this implementation: only the speed attention was computed. If the speed of a moving object is very low, the motion direction information is not very reliable because it can be due only to detection noise. Nevertheless, if the speed is high enough the information brought by motion direction is very important. From an attentional point of view, if many people have a common direction while one participant has a different direction, this is a very important cue on objects saliency. The same idea can be developed for temporal attention where brutal direction changes are very interesting.

A future work will consist in reaching a more complete description of human activities. The head tracking should be integrated by information about overall body motion. Camurri and all [18] revealed that bounding box variations or ellipse inclination, that approximate 2D translation of body, can account for expressive communication. A more detailed analysis of upper body-part should also be accomplished. Glowinski et al. [19] show that color-based tracking of head and hand can reveal expressive information related to emotion portrayals. On the basis of this refined description of human movement, we could consider new motor cues (e.g. symmetry, directivity, contraction, energy, smoothness…) accounting for the communication of an expressive content [18] that could be integrated, processed by the attention index for a more pertinent context-based analysis of expressivity.

If the fusion between spatial and temporal attention is not interesting (as discussed in section IV.D), it is crucial to fuse attention information coming from several features in the same context (space or time). This fusion can simply be done by using the maximum operator: if a feature highly attracts attention in a specific place, this area is very interesting at least from this feature point of view.

- *High-level motion attention*

In real life, when we observe a scene which does not change every time (fixed camera for example), we build knowledge models about those scenes. That model will have important influences on the final attention score and it will be able to modify attention coming from the bottom-up attention mechanisms described in this paper. A first simple implementation of this high level model is to use attention accumulation as a threshold [13]: only the objects which imply a bottom-up attention which is higher than the attention accumulation of the model are really interesting, all the other objects are inhibited.

Figure 14 displays an example of the results of this implementation: a frame of the video is extracted on the top-left image. The top-right image is an attention map of the same frame. The bottom-left image is a model of the scene: the trajectory of the moving person is visible and also is the tree which often moves because of the wind. The bottom-right image show the final attention map after the high-level attention inhibition: some noisy areas (grass which also moves because of the wind, moving person's trajectory) are inhibited. The moving tree area has also been inhibited and only few attention is focused on the tree area. The moving person remains the only area of high attention score.

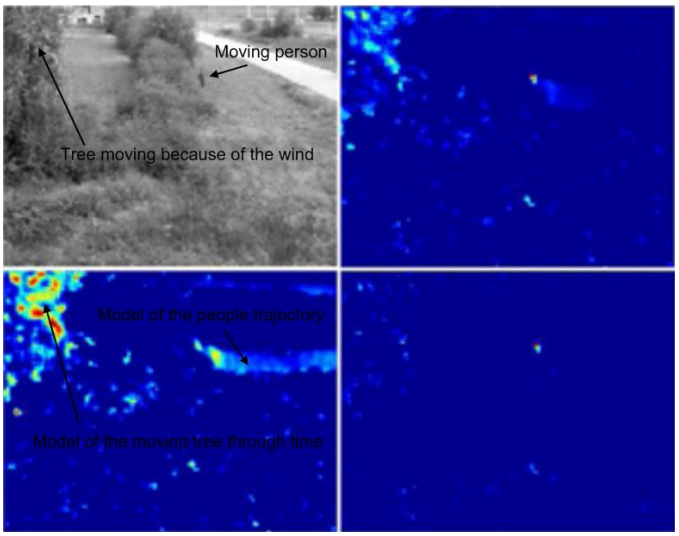


Fig. 14: Top-left: current video frame, top-right: bottom-up attention, bottom-left: scene model (top-down attention), bottom-right: final attention map after high-level attention inhibition

This simple approach already showed that it can inhibit some attention areas dues to noise or repetitive movements (trees which move because of the wind, flickering lights …).

A more complex implementation of the high-level attention model should segment the areas where a lot of attention accumulated and summarize them with a single motion vector. An object moving into these areas should be inhibited only if its motion vector is close to the motion vector which summarizes that area and its bottom-up attention score is bellow the model local amplitude. If the motion vector of this object is very different from the segmented attention area it passes through, the object will not be inhibited.

C. Discussion and Conclusion

We developed a context-based framework working in a non-controlled environment. It can deal with multiple heterogeneous situations caused by environmental disturbances and focus on relevant/rare events.

This system is more robust comparing to other applications based on video-tracking commonly developed for controlled environment. Stable environmental factors such as constant illumination and stable background greatly facilitate human activity monitoring. The present system can be a response to the growing demands, human monitoring systems in many application fields (e.g. video-surveillance, elder-people ambient assistant living, performing arts, museum spaces, etc.)

Moreover, our system open new research perspective for affecting computing and analysis of human expressivity.

Results from this study actually show that context-sensitive feature can help to better analyze gesture expressivity. The same expressive features are differently weighted depending on their spatial and temporal rarity. It put in evidence salient human motion either at the level of a single individual (rarity of behavior over time) or at the collective level (relative rarity of one member's behavior with respect to the others).

We plan to further investigate the potentialities of the rarity index in two directions: (i) by applying it to a more sophisticated set of expressive features (.e.g contraction, expansion, fluidity, impulsiveness) (ii) by analyzing how the visual feedback computed on the rarity index affect the subject behavior (e.g. whether it fosters expressive behavior).

ACKNOWLEDGMENTS

This work has been achieved in the framework of the eNTERFACE 2008 Workshop at the LIMSI Lab of the Orsay University (France). It was also included in the Numédiart excellence center (www.numediart.org) project 3.1 funded by the Walloon Region, Belgium. The authors thank to Johan DECHRISTOPHORIS who helped in IR sensor hardware components set-up. Finally, this work has been partially supported by the Walloon Region with projects BIRADAR, ECLIPSE, and DREAMS, and by EU-IST Project SAME (Sound And Music for Everyone Everyday Everywhere Every way).

REFERENCES

[1] Hongeng, S. and Nevatia, R., "Multi-agent event recognition" in ICCV, 2001, pp. II: 84-91

[2] eNTERFACE 2008 Workshop website: http://enterface08.limsi.fr/

[3] A. Camurri, A., Coletta, P., Varni, G., & Ghisio, S. "Developing multimodal interactive systems with EyesWeb XMI", Proceedings of the 2007 conference on new interfaces for musical expression (NIME07) (pp. 305-308). New York, USA, 2007

[4] OpenCV static wiki: http://opencvlibrary.sourceforge.net/wiki-static

[5] Dhavale, N., and Itti., L., "Saliency-based multi-foveated mpeg compression", IEEE Seventh International Symposium on Signal Processing and its Applications, 2003

[6] Yee, H., and Pattanaik, S., "Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments", IACM, 2001

[7] Parkhurst, D.J., Niebur, E., "Texture contrast attracts overt visual attention in natural scenes", European Journal of Neuroscience, 19:783–789, 2004

[8] Itti, L., Baldi, P., "Bayesian Surprise Attracts Human Attention", Advances in Neural Information Processing Systems, Vol. 19 (NIPS 2005), pp. 1-8, Cambridge, MA:MIT Press, 2006

[9] Le Meur, O., Le Callet, P., Barba, D., and Thoreau D. "A Coherent Computational Approach to Model Bottom-Up Visual Attention", PAMI(28), No. 5, pp. 802-817, 2006

[10] Liu, F., and Gleicher, M., "Video Retargeting: Automating Pan-and-Scan", ACM Multimedia, 2006

[11] Zhang, V., and Stentiford, F.W.M., "Motion detection using a model of visual attention", IEEE ICIP, 2007

[12] Boiman, O., Irani, M., "Detecting Irregularities in Images and in Video", International Conference on Computer Vision (ICCV), 2005

[13] Mancas, M., "Computational Attention: Towards Attentive Computers", Similar edition, CIACO University Distributors, ISBN : 978-2-87463-099-6, 2007

[14] Mancas M., Gosselin B., Macq B., "A Three-Level Computational Attention Model", Proc. of ICVS Workshop on Computational Attention & Applications, 2007, Germany.

[15] Hubel, D.H., "Eye, brain and vision", New York: Scientific American Library, N°22, 1989

[16] Mancas, M., "Image perception: Relative influence of bottom-up and top-down attention", Proc. of the WAPCV workshop of the ICVS conference, Santorini, Greece, 2008

[17] Velastin, S., Boghossian, B., Lo, B., Sun, J., Vicencio-Silva, M.: Prismatica: toward ambient intelligence in public transport environments. IEEE Trans. Syst.Man Cybern. Part A 35(1), 164–182, 2005

[18] Camurri, A., De Poli, G., Leman, M., and Volpe, G., 2005. "Toward Communicating Expressiveness and Affect in Multimodal Interactive Systems for Performing Art and Cultural Applications", IEEE Multimedia, 12,1, 43-53, 2005

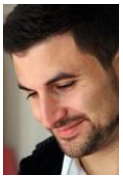
[19] Glowinski, D., Camurri, A., Coletta, P. , Bracco, F., Chiorri, C., Atkinson, A. An investigation of the minimal visual cues required to recognize emotions from human upper-body movements, (in press) Proceedings of the ACM 2008 International Conference Conference on Multimodal Interfaces (ICMI), Workshop on Affective Interaction in Natural Environments (AFFINE) Crete, 2008

Matei Mancas.

Matei Mancas was born in Bucarest in 1978. He holds an ESIGETEL (Ecole Supérieure d'Ingénieurs en informatique et TELecommunications, France) Audiovisual Systems and Networks engineering degree, and a Orsay University (France) MSc. degree in Information Processing. He also holds a PhD in applied sciences from the FPMs (Engineering Faculty of Mons, Belgium) on computational attention since 2007.

His past research interest is in signal and, in particular, image processing. After a study on nonstationary shock signals in industrial tests at MBDA (EADS group), he worked on medical image segmentation. He is now a Senior Researcher within the Information Processing research center of the Engineering Faculty of Mons, Belgium. His major research field concerns computational attention and its applications.

Donald Glowinski.



Donald Glowinski, Paris, 27-02-1977. He is doing a Phd in computing engineering at InfoMus Lab – Casa Paganini, in Genoa, Italy. (dir: Prof. Antonio Camurri). His background covers scientific and humanistic academic studies as well as high-level musical training.- EHESS (Ecole des Hautes Etudes en Sciences Sociales) MSc, in Cognitive Science, CNSMDP (Conservatoire National Supérieur de Musique et de Danse de Paris) MSc in Music and Acoustics, Sorbonne-Paris IV MSc. in Philosophy.

He was chairman of the Club NIME 2008 (New Interfaces for Musical Expression), Genoa, 2008. His research interests include multimodal and affective human-machine interactions. He works in particular on the modeling of automatic gesture-based recognition of emotions.

Pierre Bretéché.

Pierre Bretéché, Ivry sur Seine, 1981, received a MSc. degree in Computer Science in 2006 at University of Rouen (France). He is doing a PhD in Information and Communication Sciences at Laseldi Lab – University of Franche-Comté in Montbéliard (France). He previously worked in AI with a using massive multi-agent system to build semantic picturing of an environment.
He is now part of Organica research project. His research interest has moved to studying and designing new technology applications for public, cultural and artistic purpose.

Jonathan Demeyer.

Jonathan Demeyer received a MSc. in electrical engineering, Université Catholique de Louvain, Louvain-la-Neuve, Belgium in 2005 and a MSc. in applied sciences from the FPMs (Faculté Polytechnique de Mons, Belgium) in 2008.
He previously worked in the TCTS lab (Faculté Polytechnique de Mons, Belgium) developing a mobile reading assistant for visually impaired people.
He is now working on a project on automatic processing of high speed videoglottography for physicians. His main interests are medical image processing and computer vision.

Thierry Ravet.

Thierry Ravet was born in Brussels, Belgium on 31st Augustus 1976. He received an Electrical Engineering degree in 1999 in ULB (Université Libre de Bruxelles).
Afterwards, he worked as researcher in the Electronic – Microelectronic - Telecommunication department of ULB for 4 years with medical instrumentation projects.
Since February 2008, he has joined the TCTS Lab (Faculté Polytechnique de Mons). His main experiences are in non-invasive instrumentation, microprocessor system development, artefacts filtering and data fusion in the field of cardiorespiratory monitoring and polysomnographic research.

Gualtiero Volpe.

Gualtiero Volpe, Genova, 24-03-1974, PhD, computer engineer. He is assistant professor at University of Genova. His research interests include intelligent and affective human-machine interaction, modeling and real-time analysis and synthesis of expressive content in music and dance, multimodal interactive systems.

He was Chairman of V Intl Gesture Workshop and Guest Editor of a special issue of Journal of New Music Research on "Expressive Gesture in Performing Arts and New Media" in 2005. He was co-chair of NIME 2008 (New Interfaces for Musical Expression), Genova, 2008. Dr. Volpe is member of the Board of Directors of AIMI (Italian Association for Musical Informatics).

Antonio Camurri.

Antonio Camurri (born in Genova, 1959; '84 Master Degree in Electric Engineering; 1991 PhD in Computer Engineering) is Associate Professor at DIST-University of Genova (Faculty of Engineering), where he teaches "Software engineering" and "Multimedia Systems". His research interests include multimodal intelligent interfaces, non-verbal emotional and expressive communication, kansei information processing, multimedia interactive systems, musical informatics.
He was President of AIMI (Italian Association for Musical Informatics), is member of the Executive Committee (ExCom) of the IEEE CS Technical Committee on Computer Generated Music, Associate Editor of the international "Journal of New Music Research", a main contributor to the EU Roadmap on "Sound and Music Computing" (2007). He is responsible of EU IST Projects at DIST InfoMus Lab of University of Genova. He is author of more than 80 international scientific publications. He is founder and scientific director of the InfoMus Lab at DIST-University of Genova (www.infomus.org). Since 2005 he is scientific director of the Casa Paganini centre of excellence at University of Genoa on research in ICT integrating artistic research, music, performing arts, and museal productions (www.casapaganini.org).

Paolo Coletta.

Paolo Coletta born in Savona, Italy, in 1972. He received the "Laurea" degree in computer science engineering in 1997 and the Ph.D. degree in electronic engineering and computer science in 2001, both from the University of Genoa, Genoa, Italy.
From January 2002 to December 2002, he was a Research Associate at the Naval Automation Institute, CNR-IAN (now ISSIA) National Research Council, in Genoa.
In 2004 he was adjunct Professor of "Software Engineering and Programming Languages" at DIST (Dipartimento di Informatica, Sistemistica e Telematica), University of Genoa. Since 2004 he is a collaborator at DIST, University of Genoa.