Real-Time Motion Attention and Expressive Gesture Interfaces

Matei Mancas (1), Donald Glowinski (2), Gualtiero Volpe (2), Antonio Camurri (2), Pierre Bretéché (3), Jonathan Demeyer (1), Thierry Ravet (1), Paolo Coletta (2)

(1) IT Reseach Center, FPMS, University of Mons, Belgium
(2) Casa Paganini/InfoMus Lab, University of Genova, Italy
(3) Laseldi Lab, University of Franche Comté, Montbéliard, France

Abstract
This paper aims at investigating the relationship between gestures' expressivity and the amount of attention they attract. We present a technique for quantifying behavior saliency, here understood as the capacity to capture one's attention, by the rarity of selected motion and gestural expressive features. This rarity index is based on the real-time computation of the occurrence probability of expressive motion features numerical values. Hence, the time instants that correspond to rare unusual dynamic patterns of an expressive feature are singled out. In a multi-user scenario, the rarity index highlights the person in a group which shows the most different behavior with respect to the others. In a mono-user scenario, the rarity index highlights when the expressive content of a gesture changes. Those methods can be considered as preliminary steps toward context-aware expressive gesture analysis. This work has been partly carried out in the framework of the eNTERFACE 2008 workshop (Paris, France, August 2008) and is partially supported by the EU ICT SAME Project (www.sameproject.eu) and by the NUMEDIART Project (www.numediart.org).

Index Terms—computational attention, saliency, rarity, expressive gesture

Introduction

Emotions lead our attention toward those objects and situations that are ecologically relevant, enhancing the adaptation to the environment. Neuroimaging studies converge with behavioral studies to suggest that emotional signals may affect the allocation of attentional resources in order either to facilitate performance in a current task or to interrupt ongoing activity and redirect attention towards a more relevant event [26]. Spatial attention in particular can be preferentially drawn to the location of emotional stimuli [9] [24].
In the context of social communication, body gestures appear to be a relevant channel in the human judgment of affective behavior. Discrete emotions like anger or attitudinal state like boredom can be communicated through full-body or body-parts movements such as the hands and head's ones. These types of gesture that convey an emotional message are called expressive gestures [5].
A better understanding of bodily communication processes can actually lead to the development of intelligent/affective computing that could anticipate users' intention without request of explicit instructions [23]. Affective gestural analysis however often applies to a single user which is statically selected (e.g., at the start-up of the system or when the user enters the area the system is operating on). In addition, the dynamics of the expressive gesture features is rarely considered. The possibility of dynamically selecting the person to carry analysis on or to adapt and personalize analysis to the context and to the current behavior of a user could represent a decisive improvement for the development of multimodal and interactive systems. We hypothesize that a system which aims to recognize emotions on the basis of expressive gesture could be enhanced and applied in multi-user scenarios if it reproduces some of the attentional mechanisms present in humans.
Applications of computational models of human attention in the field of emotion and expressivity could also be extended to video conferencing and remote communication in order to find the person which has the most outstanding behavior. This might be a sign that this person want to take part or to react to the conversation. Also virtual agents could show emotions (as laugh for example) when the people they try to communicate with perform strange gestures. Virtual agents which pay attention should show higher natural and believable behavior. Finally, expressivity and emotion can be associated to one person but also to several persons or a crowd. Strange behavior in parts of a crowd should attract the attention of a system which can locate different emotions in those places. The location of these interesting areas can detect groups of people which are more

expressive (in a festive or violent way depending on the context of the crowd). This information could also be exploited by surveillance systems pointing out potentially dangerous groups within a crowd.
The goal of this paper is to investigate the relationship between part of the human attention which is here computationally modeled and the way to automatically extract expressive cues from human gestures.
We will first present a state of the art of computational attention models and we will introduce the notion of expressive gesture. A second section will describe our motion attention model based on the detection of the rarity of an event. We will show how an automatic rarity index is modeled and implemented, which highlights which movements should be the most salient for a human observer. Attention is computed both in a spatial (collective) and in a temporal (individual) context on expressive motion features, low-level ones such as speed of head trajectories and higher-level expressive features such as the motion index and the contraction index. The rarity index has been tested in several scenarios and validated with perceptual evaluation by subjects. Finally, we conclude by a discussion on the use of the rarity index and with possible future work. This work has been partly carried out in the framework of the eNTERFACE 2008 workshop (Paris, France, August 2008). Part of the developed source code and some video demos can be found on the eNTERFACE 2008 workshop website [10].

State of the art

Computational attention (automatic modeling of human attention)

The aim of computational attention is to automatically predict human attention on multimodal data such as sounds, images, video sequences, smell or taste, etc… The term *attention* refers to the whole attentional process that allows one to focus on some stimuli at the expense of others. Human attention is mainly divided into two main influences: a bottom-up and a top-down one. Bottom-up attention uses low-level signal characteristics to find the most salient or outstanding objects. Top-down attention uses a priori knowledge about the scene or task-oriented knowledge in order to modify (inhibit or enhance) the bottom-up saliency. The relationship and the relative importance between bottom-up and top-down attention is complex and it can vary depending on the situations [18]. While numerous models were provided for attention on still images, time-evolving two-dimensional signals as videos have been less investigated. Nevertheless, some of the authors providing static attention approaches generalized their models to the time dimension: Dhale and Itti [8], Yee and Pattanaik [29], Parkhurst and Niebur [22], Itti and Baldi [13], Le Meur [15] and Liu [16]. Motion has a predominant place and the temporal contrast of its features is mainly used to highlight important movements. Zhang and Stentiford [30] provided a 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 [2] provided a model which is able to compare the current movements with others from the video history or a database. The major problem of this approach is in its high computational cost. Bruce [3] also recently proposed an approach mixing information theory and an image database which seems to provide results close to eye-tracking tests. Nevertheless this method remains complex and needs natural images databases. Most of those methods provide bottom-up attention approaches. To our knowledge, a majority of these computational models focuses on low-level motion features (e.g., displacement of people in a scene). We suggest in this paper that computational models would gain considering higher-level motion features related to full-body movements to better capture expressive gesture that characterize the communication of an emotion.

Gesture expressivity

According to Kurtenbach and Hulteen gesture can be defined as "a movement of the body that contains information" [14]. Thus, gestures can be named expressive gestures since the information they carry is an expressive content, i.e., an "implicit message" [7]. That is, they are responsible of the communication of a kind of information that is different and independent, even if often superimposed, to a possible denotative meaning, and that concerns aspects related to emotions. A multilayered framework for automatic expressive gesture analysis was proposed by Camurri et al. [5]. In this framework, expressive gestures are described with a set of motion features that specify how the expressive content is encoded.

Different attempts can be found in the literature to map a set of expressive gesture features with one of the emotional dimensions as valence and activation that are considered to describe the entire space of conscious emotional experience [28]. The valence dimension measures how negative or positive a human feels. The activation dimension measures whether humans are more or less likely to take an action under the emotional state, from active to passive. Activation dimension is sometimes mixed with the arousal dimension which indicates the intensity of an emotion, from low to high. This latter dimension has been mapped to expressive feature such as the amount of energy of a person [4].
However, the main shortcoming of expressive gesture analysis is the scarce consideration of the context in which expressive gestures take place. The context we are speaking here has to be considered at two levels: the temporal dynamic of the motion features (i.e., how they evolve over time) and the spatial context where the behavior analysis of more than one user has to be performed. For example, if somebody is walking slowly and smoothly for a while and suddenly, he performs a rapid movement of the arm, this will likely be noticed by an external observer. In a multi-user scenario where the majority of angry people move in the same energetic way, a sad user which remains still would be likely to be noticed again.
These examples point out that the saliency of an event can be related to its novelty. [1] established a relationship between the arousal level of an emotion and the uncertainty of a visual stimulus. Mehrabian and Russell formulated the information rate-arousal hypothesis and confirmed a linear correlation between information rates of a real environment and emotion arousal [21]. Next section exposes our model which attempts to integrate these findings related to research on emotions, attention and information theory.

The computational model of motion attention

As we already stated in [17] and [20], 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 bottom-up attention is the rarity and the contrast of a feature in a given context.
The features used here are mainly the speed, motion and contraction indexes and they will be defined within the experiments where they are used in the next sections.
A low-computational cost quantification of rarity was achieved referring to the notion of self-information. Let us note m_i a message containing an amount of information. This message is part of a message set M . The bottom-up attention attracted by m_i is quantified by its self-information $I(m_i)$ which will be called rarity index in this paper:

$$I(m_i) = -\log(p(m_i)) \quad \textbf{(1)}$$

where $p(m_i)$ is the occurrence likelihood of the message m_i within the message set M. We estimate $p(m_i)$ as a combination of the global rarity of m_i within M and its global contrast compared to the other messages from M. Mathematically, $p(mi)$ is the result of a two-terms combination:

$$p(m_i) = A(m_i) \times B(m_i) \quad \textbf{(2)}$$

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:

$$A(m_i) = \frac{H(m_i)}{Card(M)} \quad \textbf{(3)}$$

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:

$$B(m_i) = 1 - \frac{\sum_{j=1}^{Card(M)} \left| m_i - m_j \right|}{(Card(M) - 1) \times Max(M)} \quad \textbf{(4)}$$

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.

The model: the three level approach for collective context

The rarity index (or motion attention index) operates at three levels corresponding to three different time scales: up to 1s (instantaneous motion attention), from 1s to 3s (short-term motion attention), more than 3s (long-term motion attention). These three levels are briefly described in this section. Our objective is to measure the rarity index in this scene, i.e. to simulate where and when the attention of an observer is attracted.

Instantaneous level
Let us consider a collective context, e.g., a group which contains interacting persons. Motion features (e.g., speed, direction) characterizing each moving person are compared at each instant. A rare motion behavior (e.g., one person speed very different from the others) immediately pops-out and attracts attention. This refers to pre-attentive human processes, usually faster than 200 milliseconds. In our approach we compute an approximation of the instantaneous component of the rarity index in intervals of 200ms to 1s.

Short-term level
Each selected component of instantaneous rarity index is analyzed over short-term time intervals from 2 to 3 seconds. This allows confirming the rarity on a larger time interval. This level refers to the human short-term memory (STM), in the range of 2 to 3 seconds. This goal of this stage is to ensure that the selected object remains outstanding compared to its past behavior or not. The capacity of STM, in terms of tracked objects, is limited to about 3 simultaneous occurrences of instantaneous rarity [6].

Third level: Long-term attention modulation
Long-term memory (LTM) component of the model deals with the computation of the rarity index in a time interval from several seconds to minutes. The output is a modification of the instantaneous attention indexes in some spatial locations according to their recurrence in the considered spatial locations. A "motion model" is progressively built along time in different locations of the observed scene. This leads to the definition of areas in the scene which concentrate attention more than others: e.g., a street accumulates more attention than a grassy area. The scene can thus be segmented into 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 is 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 high attention. This attention level is therefore able to modulate the instantaneous bottom-up attention.

Real Time Motion Attention Implementation: experiments and validation

Several experiments and real time implementations for instantaneous and short-time attention were achieved to get qualitative and quantitative results. Those experiments focused on two kinds of features: trajectory-related information about head movements and higher-level features of human expressivity based on the analysis of full-body movements. The analysis of human motion was carried on following an appearance-based approach (i.e., based on 2-D information such as color/grayscale images or body silhouettes and edges). Feature extraction, analysis and tracking procedures were selected and implemented in the EyesWeb XMI open software platform (www.eyesweb.org). Preliminary results were obtained in a pilot study carried out at the occasion of the one-month eNTERFACE 2008 workshop held in Paris (enterface08.limsi.fr) and additional results were obtained after the workshop at Casa Paganini, Genoa, Italy.
As for the trajectory-related features mainly qualitative results were obtained, the experiment providing more artistic applications which concerns the instantaneous attention level is described in the discussion section. The main focus of the following sections is a full-body features short-term attention experiment where quantitative tests were carried out.

System overview

Tests were achieved in a constrained environment with relatively small changes in terms of illumination, background, and occlusions. A single JVC gy-hd251 video camera was used in a fixed position in front of the actor (1280x720 pixel resolution, 60 fps in progressive scan). The signal coming from the video stream was processed 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 (the blob corresponding to the full-body silhouette).

Motion features

We used two categories of full-body motion features: the Motion Index and the Contraction Index. Both these full-body features have been used in the field of psychology to characterize emotional expression [27]. The Motion Index (MI) is extracted from body silhouettes. This index is a measure of the overall amount of motion detected by a video camera and is obtained by integrating in time the variations of the body silhouette (called Silhouette Motion Images - SMI). The Contraction Index (CI), measures the amount of contraction of the body with respect to its baricenter (i.e., contraction is high when the posture is such that limbs are kept near to the baricenter, e.g., arms along the body).

Experiment description

This experiment considers the application of the rarity index to full-body descriptors of expressivity such as the contraction index (CI, figure 1) and the motion index (MI, figure 2). An actor performed two sequences of movements. Each one of these sequences emphasizes a particular gestural characteristics: (i) movement activity (MI-performance, duration: 1min12) and (ii) arms' extension with respect to the body (CI-performance, duration:1min25).

Frame 686 *Frame 814* *Frame 902*

***Figure 1**. Snapshots of the CI-performance corresponding to three motion patterns*

Frame 501 *Frame 641* *Frame 755*

***Figure 2.** Snapshot of the MI performance corresponding to three motion patterns*

The two videos were presented to 16 subjects (six males and 10 females, with a mean age of 26, ranging from 20 to 44). They all had normal or corrected-to normal vision. Subjects were asked to indicate the novelty of a motion pattern they perceived. The term novelty was used because its meaning is commonly well understood in comparison with other terms as saliency. Written instructions were provided. Subject's task was to point out moments of novelty in the sequence of movements by pressing the space bar of a computer keyboard. Both performances were shown twice to each subject in a randomized order. Stimuli were displayed and participants' responses were recorded using the EyesWeb-Mobile platform on a 15-inch monitor screen (60 Hz frame rate, 1440×900 pixel resolution) [11]. Within the EyesWeb-Mobile graphical interface, stimuli were presented at 30 frames per second, on a uniform black background display with a frame size of 9 cm in height and 16 cm in width viewed from a distance of about 50 cm, which

subtended approximately 10.2 degrees of visual angle. Subject's performances were collected and processed and then compared with the results obtained with the rarity index algorithm. Human subjects' judgments were recorded at the frame at which they indicated the detection of the rare/key events (on a 30 frames per second video).

First approach: direct computation of the rarity index

In a first approach the rarity index is directly applied to the temporal feature mono-dimensional signal. The outputs resulted from setting empirically-tested categories (bins) on the raw MI and CI data to individuate three categories/bins of motion pattern: low, mid, high (e.g., for the CI: low: arms along the body, mid: both arms extended, high: arms up) as it can be seen in Figures 1 and 2. In parallel of these perceptive tests, the ground truth based the mean result of the human participants was also set-up. Figures 3 and 4 summarize the results we obtained compared to the ground truth. Eq. 2 was used here only with the occurrence likelihood factor *A*.
In both figures, the top images show that there are many variations on the raw CI or MI data (solid line). The exact number of maxima is of 634 for CI and 438 for the MI.
The middle images show that only 21 events for the CI and 11 events for the MI are really perceived as important by most of the subjects (solid line) and correspond with the ground truth (dotted line). Human attention is able to filter the huge quantity of acquired information. The middle images of the figures also show a very good correspondence between the ground truth (dotted lines) and the testers' events. All the ground truth events are well detected with very few (3) false positive detections.
The bottom images show a good correspondence between the automatic attention algorithm (solid line) and the ground truth events (dotted line). Detections are validated after a thresholding of the very small values dues to noise in the automatic attention graphs. 19 of the 21 key events were detected while few (3) false positive values ("false alarms") occurred for the CI which leads a precision score of 86% and a recall score of 90%. Concerning the MI measures, all the 11 key/rare events were detected. Several false positive values were observed but they have low rarity index values (in particular around the frame 500). The precision score was lower than in the CI performance (41%), however, the rarity index reaches a higher recall score (100%) because there are no false negative values.

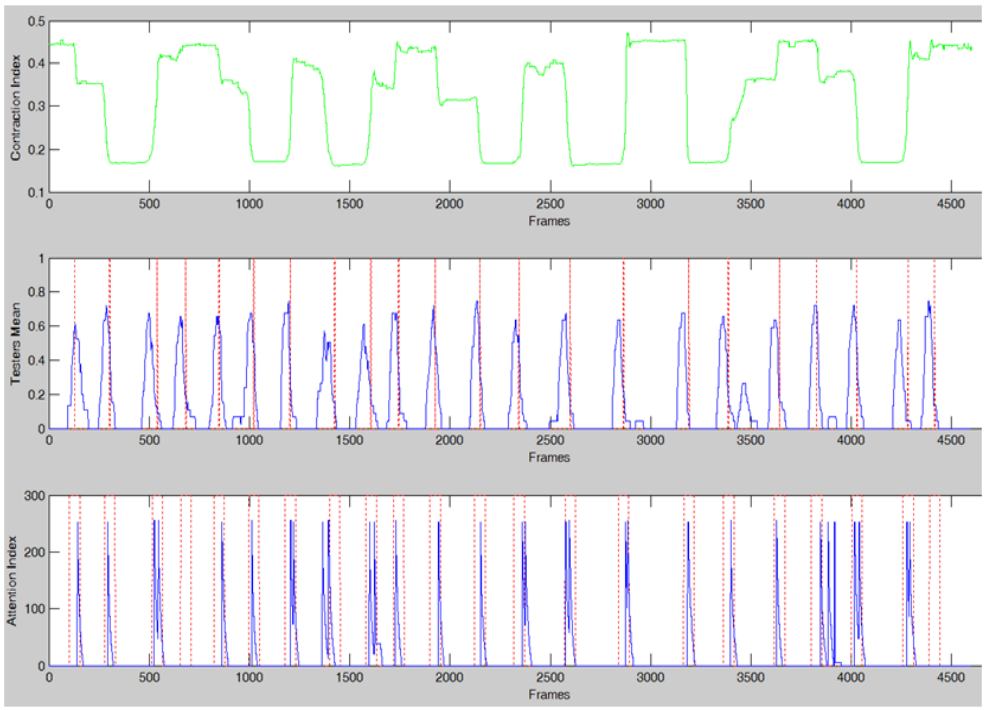

Figure 3. *Contraction Index results. Top image: CI values, Middle image: mean tester results (solid line) and ground truth (dotted line), Bottom image: rarity index (solid line) and ground truth with the mean human variability (dotted line)*

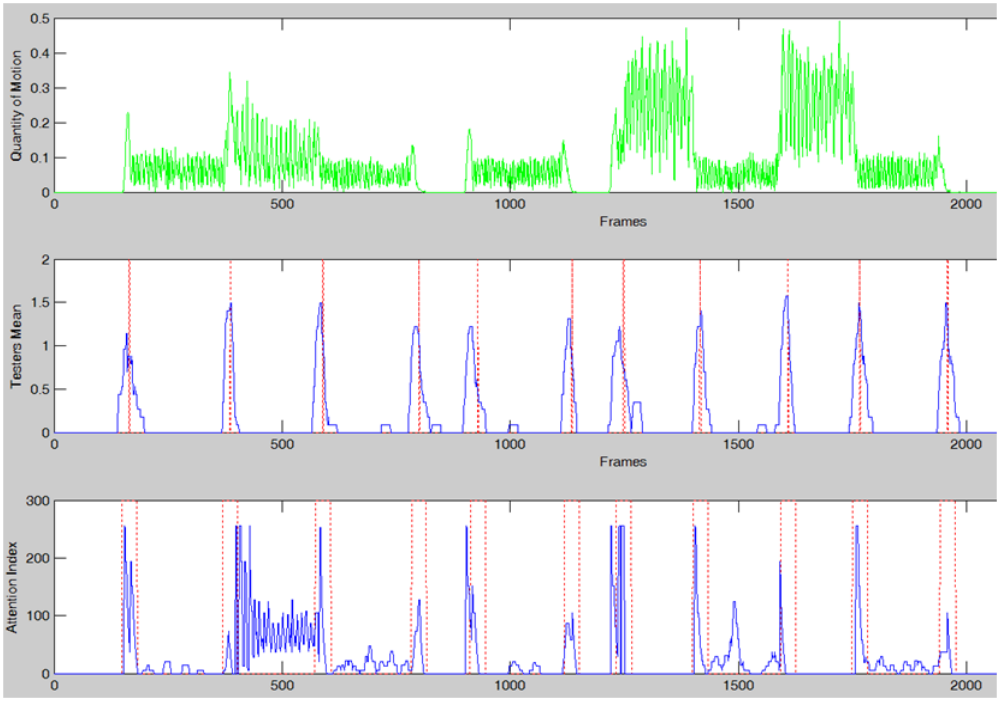

Figure 4. *Motion Index results. Top image: MI values, Middle image: mean tester results (solid line) and ground truth (dotted line), Bottom image: rarity index (solid line) and ground truth with the mean human*

Second approach: rarity index computation on the motion feature spectrogram

In addition to the direct application of the rarity index on the raw features data, it is also possible to apply it on the spectrogram of the data. A spectrogram is the computation of the Fourier transform amplitude of the signal on a sliding temporal window and it is extensively used in mono-dimensional signal processing (e.g., for audio signals). This idea was already described in [19] and figure 5 shows the 4-step process used in the particular case of expressive gesture features: the raw feature data (top-left) is used to compute its spectrogram on a 50 ms sliding temporal window (top-right). The resulting spectrogram contains 128 frequency bands (lines). After quantification on 16 bins, on each line of the spectrogram (i.e., for each frequency), the rarity index of the frequency with respect to a temporal window is computed. The rarity index is computed here using the full Eq. 2 with both the occurrence likelihood factor *A* and the global contrast factor *B*. The result is shown in the bottom-right image of figure 5.
In order to neglect the effect of noise and to obtain a mono-dimensional signal characterizing feature signal saliency, an integration on the lower frequency bands is achieved in the bottom-left image of figure 5. On this image we can see some peaks which correspond to the rarity index peaks.

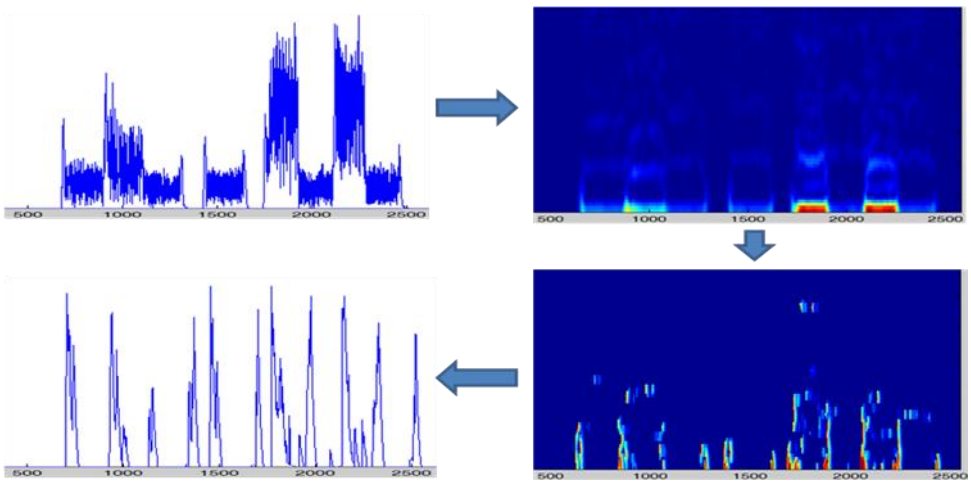

Figure 5*. The spectrogram of a motion feature is used to compute the attention score for each frequency line. The final result is the attention integration over the low frequencies.*

We first tested the variation of the two main parameters of this approach: the length of the temporal window on which the rarity index is computed and the number of bins (data quantification). Concerning the length of the temporal window, figure 6 shows for the CI feature the results on a 0.1 second (left image), 2 seconds (middle image) and 10 seconds (right image) time window. If the time window is too short (0.1 second), the method misses some onsets which have a weaker variation. If the time window is too long (10 seconds) the method may miss gestures changes if those changes have shorter duration than the window length (which is possible as 10 seconds is quite long for human gestures). Finally, we found that 2 seconds is a good time scale which correctly detects the gestures changes. This time scale is hopefully the one of the STM which shows that STM seems to be well tuned for the understanding of human gestures.

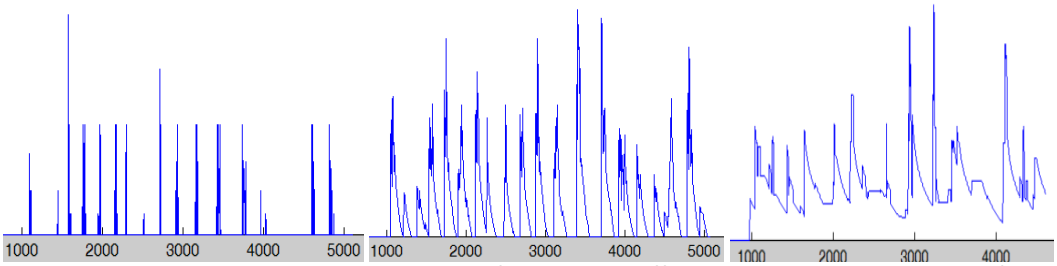

***Figure 6**. Rarity index computation on the CI feature with different time windows length. Left: 0.1 seconds, Middle: 2 seconds, Right: 10 seconds*

The second important parameter here is the number of bins used to the spectrogram of the feature quantification. Several bin values are displayed in table 1 and the precision and recall factors are computed for both MI and CI features. The best results are obtained for 16 bins which are usually enough to describe well a signal with very few loss of data and with a small amount of noise. For bins under 14 the loss of data induces more errors and above 20, there is more and more noise and the precision of complex feature signals (as for MI) drastically decreases.

In order to compute the precision and recall a small median filtering (with a window smaller than the rarity index time window length) was applied to the signal to better highlight the main peaks without eliminating them. Results are summarized in Table 1. True positives were detected if a rarity index peak was within the mean (ground truth) +/- the variability range of the human observers. False positives were all the peaks out of this range and the false negative those peaks from the ground truth which were not detected by the algorithm. As a difference with the direct computation of the rarity index on the raw feature data, no thresholding was achieved here to eliminate low attention peaks: all the attention peaks were taken into account.

BINS	**3**	**4**	**5**	**6**	**8**	**10**	**12**	**14**	**16**	**20**	**32**	**64**	**128**	**256**
Precision in % (MI)	100	85	63	91	92	92	73	92	100	73	65	50	50	50
Recall in % (MI)	64	100	90	91	100	100	100	100	100	100	100	100	100	100
Precision in % (CI)	92	93	92	100	100	83	95	100	95	100	100	100	95	100
Recall in % (CI)	55	68	63	90	90	75	95	100	100	90	90	90	100	90

***Table 1**. Precision and recall for the MI and CI features at several bins*

Figure 7 shows the results compared to the manual segmentation result for a 2 seconds window length and 16 bins for the rarity index computation. An improvement compared to the direct application of the rarity index (figures 3 and 4) on the raw features is definitely observed mainly for the noisy MI feature signal. The precision and recall are both here of 100% for the MI feature and the precision is of 95% while the recall is of 100% for the CI feature which is comparable with human subjects' results.

Another advantage is in the reproducibility of this approach. While the number and the value of the bins used in the direct approach depend on the application, here a fixed regular 16 bins quantization of the data was used.
This experiment shows (for direct and spectrogram-based approaches) that for both gestural features (MI and CI) the automatic rarity index provides a segmentation of the features behavior over time. A change in the feature behavior exhibits a change in the perceived movement, thus a change of the expressive message or emotion which is transmitted. The rarity index seems to provide expressive or emotional phrase segmentation by attracting human attention in case of gesture changes.

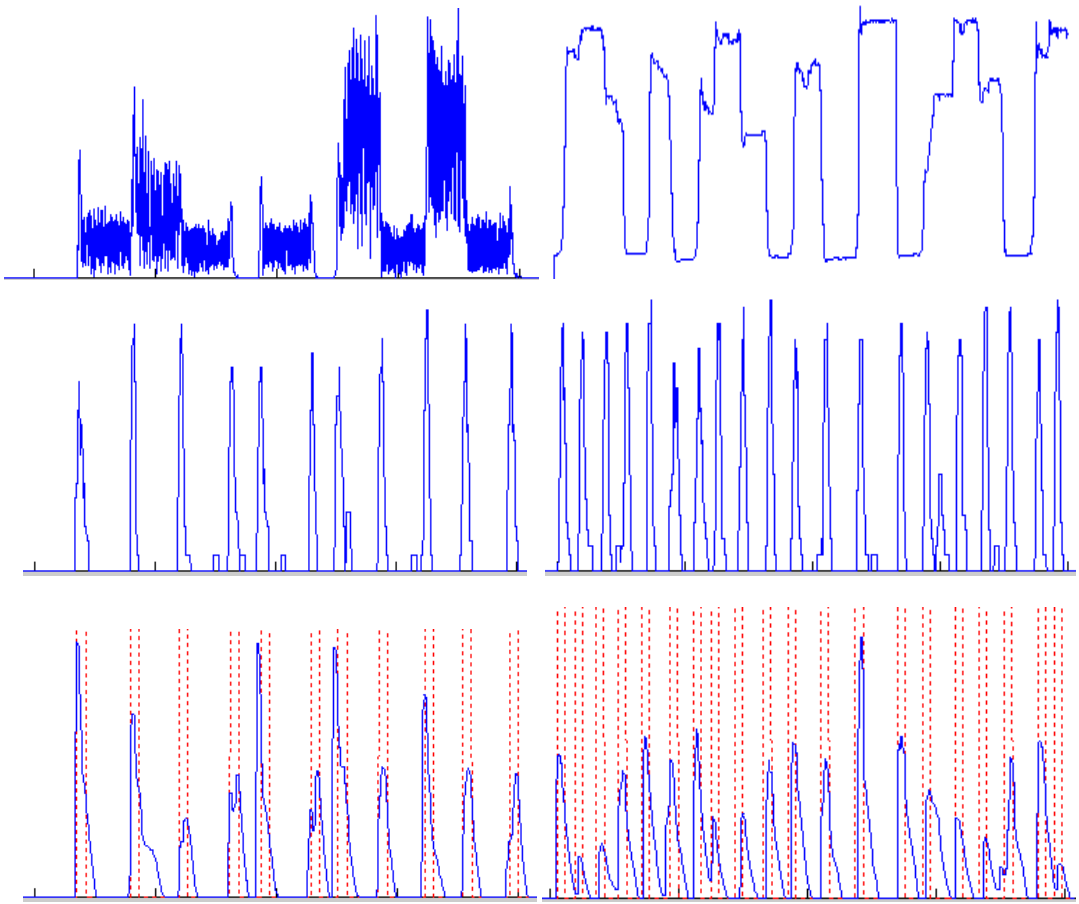

***Figure 7**. Initial feature measures (top images), mean observer gesture change detection (middle images) and automatically detected rarity index in blue and mean observer variability in dotted red line (bottom images). Left column: Motion index feature, Right column: Contraction index feature.*

Discussion

Rarity index applied on motion speed or silhouette motion and contraction measurements proved successful in relieving on behavior saliency. We believe that, at all the time scales described here, an algorithm which aims in predicting part of the human attention is useful in expressivity analysis.
Instantaneous attention level should help in selecting the person which has the higher motivation to express a message out of a collective context. The ability to select the person which should deliver expressive messages with the higher probability definitely improves the feasibility of a system working on ecological scenarios with collective interactions.
Concerning the short term attention level, the ability to detect at a given time the most salient expressive feature may help in providing a dynamic relative importance between the expressive features which are simultaneously analyzed in order to characterize gestures expressivity. We could determine features importance based on their rarity score and assign greater weight to

relevant features (high rarity index values) as compared to less relevant features (low rarity index values) at a given time.

Additional features: getting closer to human attention

Rarity index applied on motion speed or silhouette, motion and contraction measurements proved successful in relieving on behavior saliency. Additional expressive features could be further tested. Camurri and all [5] revealed that bounding box variations or ellipse inclination, that approximate 2D translation of body, can account for expressive communication. Glowinski et al. [11] showed that the head and hands dynamics, symmetry or trajectories smoothness for example can reveal expressive information related to emotions. Motion direction could also be further considered. Literature on human attentional processes suggest that the sudden direction change of one participant in a group moving in the same direction is perceptually salient. These additional features could be in turn integrated, and processed by the rarity index for a more pertinent context-based analysis of expressivity.

Towards artistic applications

We tested the rarity index in a collective situation (instantaneous attention level) during a dance master-class directed by choreographer Giovanni Di Cicco in the framework of Digifestival (Casa Paganini, Genoa, Italy, November 2008, www.casapaganini.org). The feature taken into account here was the Motion Index (MI). The experiment was performed on the stage of the 250-seats auditorium at Casa Paganini, an international center of excellence for research on sound, music, and new media, where InfoMus Lab has its main site. The installation covered a surface of about 9 m × 3.5 m. A single infra-red video-camera observed the whole surface from the top, about 4 m high, and at a distance of about 10 m from the stage. A white screen covered the back of the stage for the whole 9 m width and was used to project the video feedback.
In this dance application, the value of the rarity index controlled the transparency of the silhouette of each of three dancers which was extracted from the live video from an infra-red video-camera through a multi-blob tracking. The higher was the dancer's rarity index, the more opaque was its silhouette. Figure 8 shows some results. On the left image, the dancer, located in the middle, stays still whereas the two others are running: his MI is rare relatively to the others. On the right image, the dancer, located in the right, is moving at a higher speed than the two others, thus having the rarest MI.

Figure 8. *Two snapshots corresponding to two situations observed during the dance master-class. In both situations the silhouette which appears on the video in the background is the one of the dancer which has the rarest behavior with respect to the two others.*

Following discussion with dancers to get their feedback, promising potential applications of the rarity index in the field of performing arts emerged. The dancers put in evidence that this algorithm provide telltale signs of the onset or progression of their movements and forced them to be aware of the other's motion pattern as they were performing themselves their own sequence of movement. A rarity index based feedback may foster a higher interaction in social and collective behaviors. Moreover, from a psychological point of view, in a collective context all the participants naturally tend to reach the dominant emotion through emotional contagion processes [12]. If a minority of participants have a different message through a different expressive gesture this is worthy of attention because it shows at least a higher perseverance in delivering their expressive message.

Conclusion

A real time rarity index was developed to highlight motion saliency at several time scales and made experiments for both instantaneous and short time periods. Starting from a motion tracking system, we analyzed spatio-temporal profiles of people activities at different levels of detail. 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 [25]. A more-detailed level of person's activity was analyzed in terms of full-body motion features (e.g., Motion Index, Contraction Index) and found more relevant to characterize motion sequences and related gestures. The rarity index selected the participant which exhibited the highest saliency though instantaneous comparison of participants' motion features. Then, the selected participant was tracked and the salient changes of his behavior were detected over short time intervals.

The obtained outcomes showed that such a rarity index better matches human motion perception rather than what would achieve simple motion detection: depending on the context, the lack of motion for example appeared to be more salient than motion itself.

Our study also showed that context is a key aspect in the analysis of the expressive content conveyed by gestures. This context-related information is naturally captured by human through attentional mechanisms and help to focus limited visual resources on the most salient aspects of the visual scene. The rarity index draws upon these human bottom-up attentional processes. It relieves on the saliency of user's behavior by computing the probability of occurrence and contrast of the expressive features values during instantaneous and short-term time periods. Our algorithm has been successfully tested in applications dealing with one or three participants simultaneously. The rarity index algorithm can be considered as a first step to give real-time multimodal interfaces context-aware abilities and to adapt efficiently to multi-user scenarios.

We plan to further investigate the potentialities of the rarity index as a descriptor of human expressivity in three directions: (i) by applying it to a more sophisticated set of expressive features (e.g., fluidity, impulsiveness) (ii) by selecting the relevant expressive features according to their rarity score (iii) by analyzing how a visual feedback computed on the rarity index can affect user 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 NUMEDIART 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).

[1] D.E. Berlyne and DE Berlyne. Studies in the new experimental aesthetics. 1974.

[2] O. Boiman and M. Irani. Detecting Irregularities in Images and in Video. *International Journal of Computer Vision*, 74(1):17–31, 2007.

[3] NDB Bruce and JK Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3):5, 2009.

[4] A. Camurri, I. Lagerlöf, and G. Volpe. Recognizing emotion from dance movement: Comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies, Elsevier Science*, 59:213–225, july 2003.

[5] A. Camurri, G. Volpe, G. De Poli, and M. Leman. Communicating Expressiveness and Affect in Multimodal Interactive Systems. *IEEE Multimedia*, pages 43–53, 2005.

[6] N. Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(01):87–114, 2001.

[7] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and JG Taylor. Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE*, 18(1):32–80, 2001.

[8] N. Dhavale and L. Itti. Saliency-based multifoveated MPEG compression. In *Signal Processing and Its Applications, 2003. Proceedings. Seventh International Symposium on*, volume 1, 2003.

[9] JD Eastwood, D. Smilek, and PM Merikle. Differential attentional guidance by unattended faces expressing positive and negative emotion. *Perception &# 38; Psychophysics*, 63(6):1004–1013, 2001.

[10] eNTERFACE 2008. "http://enterface08.limsi.fr/".
[11] D. Glowinski, F. Bracco, C. Chiorri, A. Atkinson, P. Coletta, and A. Camurri. An investigation of the minimal visual cues required to recognize emotions from human upper-body movements. In *Proceedings of ACM International Conference on Multimodal Interfaces (ICMI), Workshop on Affective Interaction in Natural Environments (AFFINE)*. ACM, 2008.
[12] E. Hatfield, J.T. Cacioppo, and R.L. Rapson. *Emotional contagion Studies in emotion and social interaction*. Editions de la Maison des sciences de l'homme, 1994.
[13] L. Itti and P. Baldi. Bayesian Surprise Attracts Human Attention. *Advances in Neural information Processing Systems*, 18:547, 2006.
[14] G. Kurtenbach and E.A. Hulteen. Gestures in Human-Computer Communication. *The Art of Human-Computer Interface Design*, pages 309–317, 1992.
[15] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. A Coherent Computational Approach to Model Bottom-Up Visual Attention. *IEEE Transactions on pattern analysis and Machine Intelligence*, pages 802–817, 2006.
[16] F. Liu and M. Gleicher. Video retargeting: automating pan and scan. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 241–250. ACM Press New York, NY, USA, 2006.
[17] M. Mancas. Computational attention: Towards attentive computers. Similar edition, 2007. CIACO University Distributors.
[18] M. Mancas. Relative influence of bottom-up and top-down attention. *Attention in Cognitive Systems, Lecture Notes in Computer Science*, Volume 5395/2009:pp. 212–226, February 2009.
[19] M. Mancas, B. Gosselin, and B. Macq. A three-level computational attention model. In *Proc. of ICVS Workshop on Computational Attention & Applications*, Germany, 2007.
[20] M. Mancas, C. Mancas-Thillou, B. Gosselin, and B. Macq. A rarity-based visual attention map–application to texture description. In *Proceedings of IEEE International Conference on Image Processing*, pages 445–448, 2007.
[21] A. Mehrabian and J.A. Russell. An approach to environmental psychology. 1974.
[22] D.J. Parkhurst and E. Niebur. Texture contrast attracts overt visual attention in natural scenes. *European Journal of Neuroscience*, 19(3):783–789, 2004.
[23] R.W. Picard. *Affective Computing*. MIT Press, 1997.
[24] K.M. Stormark, K. Hugdahl, and M.I. Posner. Emotional modulation of attention orienting: A classical conditioning study. *Scandinavian Journal of Psychology*, 40(2):91–99, 1999.
[25] SA Velastin, BA Lo, and B.P.L.J.S. Vicencio-Silva. PRISMATICA: toward ambient intelligence in public transport environments. *Systems, Man and Cybernetics, Part A, IEEE Transactions on*, 35(1):164–182, 2005.
[26] P. Vuilleumier, J. Armony, and R. Dolan. Reciprocal links between emotion and attention. *Human brain functions (eds KJ Friston, CD Frith, RJ Dolan, C. Price, J. Ashburner, W. Penny, S. Zeki & RSJ Frackowiak)*, pages 419–444, 2003.
[27] H.G. Wallbott. Bodily expression of emotion. *Eur. J. Soc. Psychol*, 28:879–896, 1998.
[28] D. Watson, L.A. Clark, and A. Tellegen. Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(6):1063–1070, 1988.
[29] H. Yee, S. Pattanaik, and D.P. Greenberg. Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. *ACM Transactions on Graphics (TOG)*, 20(1):39–65, 2001.
[30] S. Zhang and F. Stentiford. Motion Detection using a Model of Visual Attention. In *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, volume 3, 2007.