Stefan Kopp
Ipke Wachsmuth (Eds.)

Gesture in Embodied Communication and Human-Computer Interaction

8th International Gesture Workshop, GW 2009
Bielefeld, Germany, February 25-27, 2009
Revised Selected Papers

Springer

Gesture Saliency: a Context-aware Analysis

Matei Mancas[1], Donald Glowinski[2], Gualtiero Volpe[2],
Paolo Coletta[2], Antonio Camurri[2]

[1] University of Mons, F.P.Ms/IT Research Center/TCTS Lab
31, Bd. Dolez, 7000 Mons, Belgium
Matei.Mancas@umons.ac.be
[2] University of Genoa, INFOMUS Lab, Italy
{Donald.Glowinski, Gualtiero.Volpe, Antonio.Camurri, Paolo.Coletta}@unige.it

Abstract. This paper presents a motion attention model that aims at analyzing gesture saliency using context-related information at three different levels. At the first level, motion features are compared in the spatial context of the current video frame; at the intermediate level, salient behavior is analyzed on a short temporal context; at the third level, computation of saliency is extended to longer time windows. An attention/saliency index is computed at the three levels based on an information theory approach. This model can be considered as a preliminary step towards context-aware expressive gesture analysis.

Keywords: Visual attention, expressive gesture, context-aware analysis.

1 Introduction

Objects and situations can lead our attention because of their emotional values. Neuroimaging and behavioral studies suggest that emotional signals may affect the allocation of attentional resources either to facilitate performance in a current task or to interrupt an ongoing activity and redirect attention towards a more relevant event [17]. 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 states 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 [4].
A better understanding of bodily communication processes can actually lead to the development of intelligent/affective computing that could anticipate people intention without request of explicit instructions by considering the spatial or temporal context of their behavior [16]. Affective gestural analysis, however, often applies to a single user which is manually 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 naturalness of the human-computer interaction could highly benefit from the possibility to dynamically select the person to carry analysis on or from the possibility to adapt and personalize analysis to the context and to the current behavior of a user. 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. 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.

After a state of the art of computational attention models, we recall the notion of expressive gesture. A second section will describe how an automatic saliency index is modeled and implemented to highlight which movements may be the most salient for a human observer. In a third section, the application of motion attention is achieved on several scenarios that exemplify the three contexts of the analysis (spatial, short and long-term). Finally, we conclude by the findings on the relationship between gestures expressivity analysis and computational attention algorithms.

2 State of the Art

2.1 Computational Attention or 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 mainly consists of two processes: a bottom-up and a top-down one. Bottom-up attention uses low-level signal features 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. While numerous models were provided for attention on still images, time-evolving two-dimensional signals such as videos have been less investigated. Nevertheless, some of the authors providing static attention approaches generalized their models to videos, but very few to audio or time-evolving feature signals (for a detailed review, see [11]). Most of these 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). We suggest in this paper that computational models would gain considering higher-level motion features related to full-body movements to better capture the expressive gestures that characterize the communication of an emotion. Our approach is able to easily adapt to different spatial, short and long temporal contexts.

2.2 Gesture Expressivity and Attention

According to Kurtenbach and Hulteen gesture can be defined as “a movement of the body that contains information” [8]. Thus, gestures can be named expressive since the information they carry is an expressive content, i.e., content related to the emotional sphere. A multilayered framework for automatic expressive gesture analysis was proposed by Camurri et al. [4]. 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 literature to map a set of expressive gesture

features with one of the emotional dimensions that are considered to describe the entire space of conscious emotional experience [18] which are valence and activation. For example activation dimension has been mapped to expressive features such as the amount of energy of a person [3]. However, the main shortcoming of expressive gesture analysis is the scarce consideration of the context in which expressive gestures take place. The context we focus on has to be considered both related to the temporal dynamics of a motion feature and to the spatial context of this feature if the behavior analysis of more than one user is performed.

First studies related to the context-aware analysis of expressivity which established a relationship between the arousal level of an emotion and the uncertainty of a visual stimulus can be found in [2]. Mehrabian and Russell formulated the information rate-arousal hypothesis and confirmed a linear correlation between information rates of a real environment and emotion arousal [15]. These studies put in evidence that the saliency of an event can be related to the novelty of an expressive content.

3 The Model of Motion Attention

3.1 A Rarity-based Approach

As we already stated in [9] 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 context. The main cue which involves bottom-up attention is the rarity and contrast of a feature in a given context. The features considered in this paper are speed, and the motion and contraction indexes. They are described in the presentation of the experiments in the next section.

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 here *saliency index*:

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

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 a 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 close to each others to be seen as the same.

$B(m_i)$ quantifies the global contrast of a message m_i on the context M:

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

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 saliency index (or motion attention index, $I(m_i)$) 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).

3.2 Instantaneous Level

Let us consider a collective context, e.g., a group with interacting people. Motion features (e.g., speed, direction) characterizing each moving person are compared at each instant. Salient 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, motion saliency detection at instantaneous level is computed over time intervals of 200ms – 1s.

3.3 Short-term Level

Each participant selected in the previous instantaneous level may have his motion features analyzed over short-term time intervals from 2 to 3 seconds. This level refers to the human sensory memory (SM), in the range of 2 - 3 seconds [1]. This stage goal is to ensure that the selected object remains outstanding compared to its past behavior or not. Information from SM goes then to the short-term memory (STM). The capacity of STM, in terms of tracked objects, is limited to about 4 simultaneous occurrences of instantaneous rarity [5].

3.4 Long-term Attention Modulation

The long-term memory (LTM) [1] component of the model processes the saliency index in a time interval from several seconds to much longer periods (related to the application time scale). The output is a modification of the instantaneous attention indexes in such interval according to their considered recurrence. The attention amplitude map in the different locations of the observed scene along time is progressively built. This leads to the definition of areas, which capture 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 (usual motion). 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 with high attention (novel motion).

4 Application to Analysis of Expressive Gesture

4.1 Instantaneous Motion Attention

We tested the motion attention model and the saliency index it provides during a dance master-class to consider the emergence of a salient behavior in the components of a group. The feature taken into account here was Motion Index. 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). In this dance application, the value of the saliency index was computed for each dancer and compared in the spatial context of the current video frame. This salient index value controlled the transparency of the silhouette of the dancer, which was extracted from the live video from an infra-red video-camera using a multi-blob tracking technique. The higher was the dancer's saliency index, the more opaque was its silhouette. Figure 1 shows some results. On the left image, the dancer located in the middle stays still whereas the two others are running: his behavior is salient 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 most salient behavior. A following discussion with 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. A saliency 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 [7].

Fig. 1 Two snapshots of 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.

If a minority of participants exhibit a different, salient behavior, this is worthy of attention because it shows at least a higher perseverance in delivering their expressive message.

4.2 Short-term Motion Attention

The saliency index was tested over short temporal periods on expressive features such as the motion index (MI) and the contraction index (CI) related to individual full-body movements. The 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). An actor performing two sequences of movements was recorded. Each one of these sequences emphasizes a particular gestural characteristics: (i) movement activity (MI-performance: figure 2, right box) and (ii) arms' extension with respect to the body (CI-performance: figure 2, left box).

Fig. 2 Snapshots of the CI (left box) and MI (right box) performance

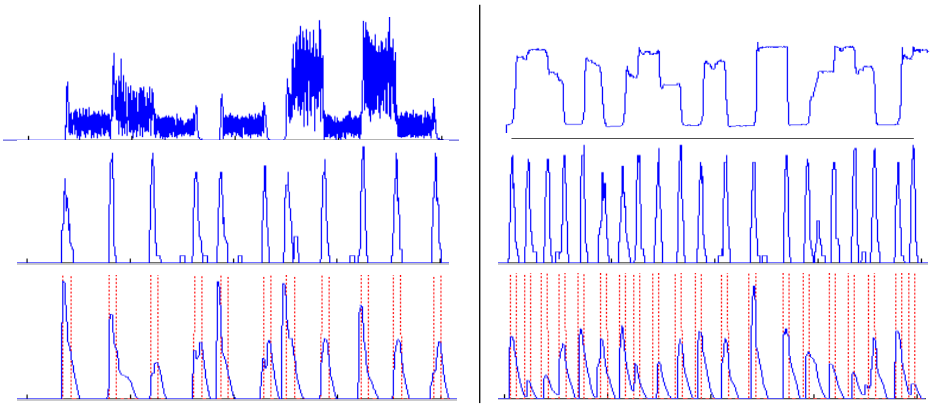
The two videos were presented to 16 participants (six males and 10 females, with a mean age of 26) who pointed out moments of novelty in the sequence of movements by pressing the space bar of a computer keyboard. Stimuli were displayed and participants' responses were recorded and time-logged to the video using the EyesWeb-Mobile platform [6]. Participant's motion sequence segmentations were collected and then compared with the automatic segmentation obtained with the saliency index algorithm. More details about the experiment can be found in [13].

As proposed in [12] and [13], the mono-dimensional signal that characterizes saliency over time was computed on the spectrogram of each expressive feature (CI and MI). The following procedure was applied:

- computation of the signal Fourier transform on a 50 ms sliding temporal window
- division of the resulting spectrogram into 128 frequency bands
- quantification of each frequency band into 16 bins
- selection of a time window on which applying the saliency index (Eq. 1): 128 saliency indexes corresponding to each band are obtained
- integration on the lower frequency bands in order to neglect the effect of noise and to obtain a mono-dimensional signal characterizing feature signal saliency

A preliminary analysis presented in [13] showed that for both expressive gestural features (MI and CI) the automatic saliency index with a temporal window of 2 s and a bin number of 16 provided a segmentation close to the human's one.

A comparison between the human observers' segmentation variability (figure 3, bottom row, dotted red line) and the one provided by the saliency index (figure 3, bottom row, blue line) showed a very high correlation with 100% of precision (a measure of fidelity) and recall (a measure of completeness) for the MI feature and 100% and 95% of precision and recall respectively for the CI feature.



Saliency index applied on MI Saliency index applied on CI

Fig. 3. From top to bottom: initial feature measures (top images), participants' mean segmentation (middle images) automatic saliency index segmentation (in blue) and participants' mean variability in dotted red line (bottom images).

In this paper, we completed the evaluation of the algorithm segmentation with an in-depth statistical analysis. We wanted in particular to observe how the performance can be related to the bin number used for quantifying the frequency bands in the algorithm. This bin number specifies the sensibility of the algorithm to distinguish between different events. Receiver operating characteristics (ROC) curve were employed to assess the saliency index algorithm performance with respect to human segmentation (*ground truth*).

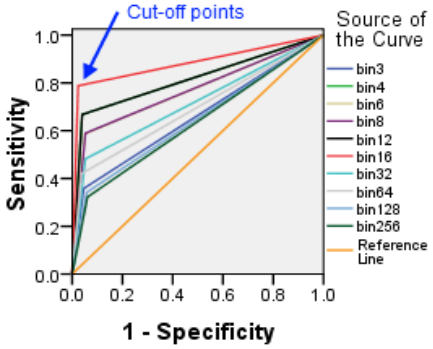


Fig. 4. ROC curve of the saliency index implemented with ten different bin values (MI feature)

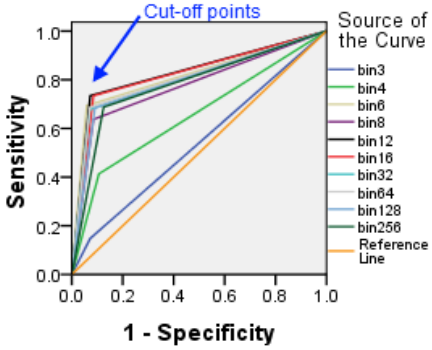


Fig. 5. ROC curve of the saliency index implemented with ten different bin values (CI feature)

The saliency index was tested with 10 different bin number values from 3 to 256. Analysis of the ROC curves (figures 4 and 5) shows that changing the bin number considerably affect the resulting classification on both MI and CI features (i.e., the distinction between rare/non-rare events). Considering in particular the area under the curve (AUC), the algorithm with a bin number of 16 (bin16) can be considered a good compromise as it shows the best performance in MI (AUC=0.88 ± 0.01 (95% C.I. (*confidence interval*) 0.86-0.91), p=0.000) and the second best performance (but very close to the best performance) in the CI (Contraction Index feature) case (AUC=0.83± 0.008 (95% C.I. (*confidence interval*) 0.80-0.84), p=0.000).

At the cut-off points for the MI feature (figure 4) the algorithm correctly identified 95.4 % of the rare event indicated by subjects (accuracy). Sensitivity (true positive rate or the rate of salient events which are detected as so by the saliency index) is of 78.8 % and the specificity (true negative rate or the rate of non salient events which are labeled as non-salient by the saliency index) is of 97.6 %. At the cut-off points for the CI feature (figure 5), the algorithm correctly identified 87.5 % of the rare event indicated by subjects (accuracy). The sensitivity is of 70.9% and the specificity of 92.4 %.

4.3 Long-term Motion Attention

For long-term motion attention, preliminary techniques were developed to compute the rarity for the direction and speed of participants observed in different regions of the space. Generalizing the work started in [10], motion history images (MHI) were used to compute the position, the direction and the velocity of participants over long-time scales (e.g., 4 minutes). The saliency index computed over these features allowed to build a model of the scene highlighting the regions where rare behaviors were observed. The model is obtained using an increment and a decrement function so that at each frame, when a participant is observed with a certain speed and direction which is already dominant for the considered region, the saliency index for the pixels in the region is decreased, otherwise, it is increased.

If current motion has the same features as the model at the same locations, the motion detection will be inhibited: it is an already seen one, it is not rare, and thus it is not worthy of attention. If motion occurs with different features as those from the corresponding model, the motion detection will not be inhibited: it is a novel movement which is rare and which should attract attention. Figure 6 shows snapshots of a recording session in a class room where most of the participants were asked to move along predetermined paths while few others moved with different velocities, directions or positions as the majority.

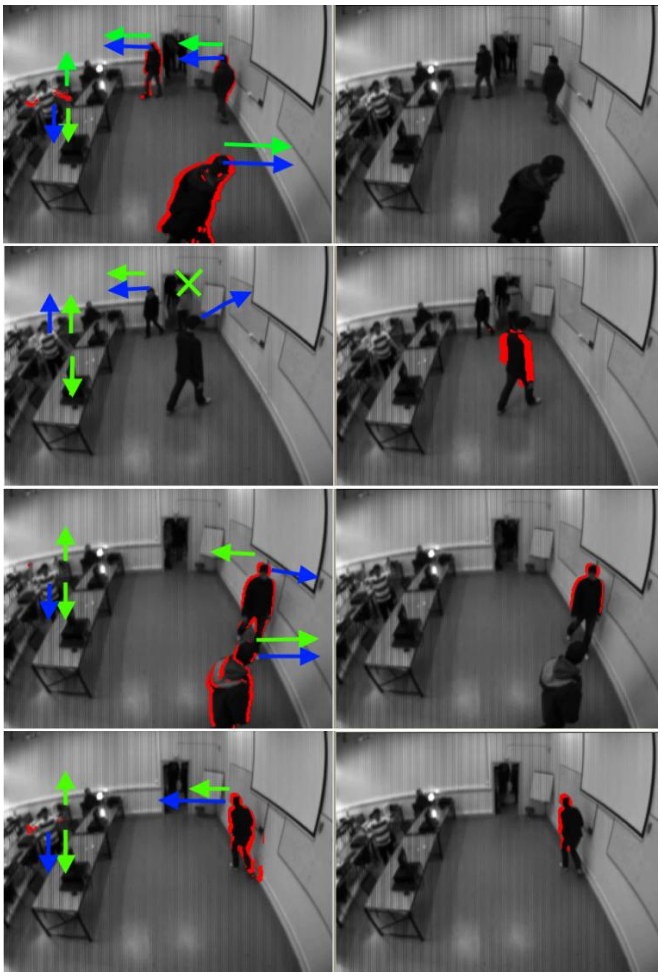


Fig. 6. Left column, the images show the detected motion of the participants (in red), the motion vector of the model (in green) and the current motion vector of the frame (in blue). Right column: salient motion of the participants (in red) detected after the model was applied (participants have different motion directions, different velocities or they are located in positions where few motion was detected).

A quantitative validation of the long-term attention model was achieved with a dancer who walked along a 6x4 meters space in various directions and at different speeds. As shown in figures 7 and 8, the dancer performed six different paths. Three models (containing information about position, motion direction, and velocity) were computed for the first three more regular paths (see figure 7). An inhibition rate (*IR*) was computed to describe how much of the initial detected motion of the dancer was inhibited by the long term attention model.

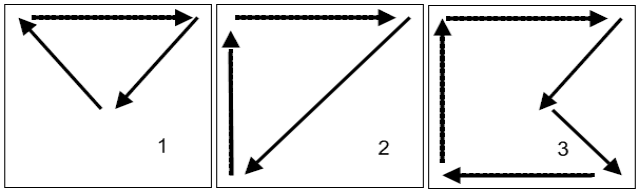


Fig. 7. The three regular paths followed by the dancer were used as models (model 1, 2 and 3).

IR values ranged from 0 (the detected motion is inhibited because it is similar to the model) to 1 (the detected motion is different from the model and it can be considered as salient). The procedure is detailed in the following pseudo-code:

```
a= 0.6, Eps = 10^-5
Nb_frames_salient_motion=0, Nb_frames_detected_motion=0
if sum(moving_pixels_after_inhibition)/sum(initial_moving_pixels)>a
        Nb_frames_salient_motion = Nb_frames_salient_motion+1
if sum(initial_moving_pixels)>0
        Nb_frames_detected_motion = Nb_frames_detected_motion+1
IR = Nb_frames_salient_motion /(Nb_frames_detected_motion+Eps)
```

Table 1 presents the inhibition rate (IR) values obtained through the application of the long-term attention models 1, 2 and 3 when the dancer moved along the paths 1, 2 and 3 (see figure 7). When the performed trajectory is close to the model used to analyze it (e.g., model 1 applied to path 1), the inhibition rate (IR) values tend to 0. On the opposite, IR values tend to 1 when the performed path differs from the model (e.g., model 1 applied to path 2).

Table 1. the comparison of the referent three models (1, 2, 3) with the three corresponding paths (1, 2, 3) shows a low inhibition rate (IR) value when matching between model and path is high and a high IR values in the opposite case.

	Path 1	Path 2	Path 3
Model 1	**0.03**	0.35	0.35
Model 2	0.15	**0.01**	0.20
Model 3	0.11	0.13	**0.07**

Table 2 shows the inhibition rate was very low for the portions of the paths 5, 6 and 7 (see figure 8), that correspond to trajectories learnt by the model (see figure 8, blue lines). The inhibition rate was on the opposite very high for novel/salient motion that was not considered by the models (see figure 8, red lines).

Table 2. IR values of paths 5, 6 and 7 when motion is already detected (see figure 8, blue line) and when motion is salient (see figure 8, red line).

	Path 5	Path 6	Path 7
Already detected motion	**0.04**	**0.03**	**0.05**
Novel/Salient motion	0.57	0.68	0.34

Top-down information let us pointing out the novel motion on one side, but it is also interesting in detecting already seen motion patterns. This second, long-term approach is related to task-driven top-down attention.

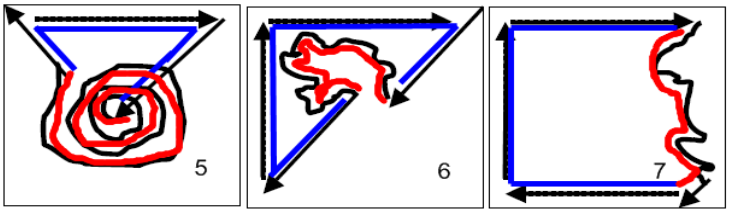


Fig. 8. In black: trajectories 5, 6 and 7. In blue: already detected motion of the models 1, 2 and 3 from figure 7. In red: novel/salient motion.

6 Conclusion

Context-related information is naturally captured by humans through attentional mechanisms and help to focus limited visual resources on the most salient aspects of the visual scene. Our saliency 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, short-term, and long-term time periods. We demonstrated that human attention related algorithms are able to set attention focus on the person with a different behavior compared to the others, to a person who exhibits changes regarding his own behavior history, but also to people whose behavior is different from the one of the majority of the people passing in the same areas. Our algorithm has been successfully tested in applications dealing with one or several participants simultaneously. The saliency index algorithm can be considered as a first step to provide real-time multimodal interfaces with context-aware abilities and to adapt them efficiently to multi-user scenarios.

We plan to further investigate the potentialities of the saliency 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 saliency index can affect user behavior (e.g., whether it fosters expressive behavior).

Acknowledgements

This work was partially supported by the Numediart project (www.numediart.org) funded by the Walloon region, Belgium. The authors thank to Loïc Reboursière and André Serre-Milan for the dancer video acquisition. Finally, this work has also been partially supported by the Walloon region with projects BIRADAR, ECLIPSE, and DREAMS, and by the EU-ICT Project SAME (Sound And Music for Everyone Everyday Everywhere Every way, www.sameproject.eu).

References

[1] RC Atkinson. Shiffrin. RM (1968). Human memory: A proposed system and its control processes. *The psychology of learning and motivation: Advances in research and theory*, 2:89–195.

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

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] G. Kurtenbach and E.A. Hulteen. Gestures in Human-Computer Communication. *The Art of Human-Computer Interface Design*, pages 309–317, 1992.

[9] M. Mancas. Computational attention: Towards attentive computers. Similar edition, 2007. CIACO University Distributors.

[10] M. Mancas. Image perception: Relative influence of bottom-up and top-down attention. 2008.

[11] 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.

[12] M. Mancas, L. Couvreur, B. Gosselin, and Macq B. Computational attention for event detection. *Proceedings of ICVS Workshop on Computational Attention & Applications (WCAA-2007)*, 2007.

[13] M. Mancas, D. Glowinski, G. Volpe, A. Camurri, J. Breteche, P. Demeyer, , T Ravet, and P. Coletta. Real-time motion attention and expressive gesture interfaces. *Journal On Multimodal User Interfaces (JMUI)*, 2009.

[14] 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.

[15] A. Mehrabian and J.A. Russell. An approach to environmental psychology. 1974.

[16] R.W. Picard. *Affective Computing*. MIT Press, 1997.

[17] 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.

[18] 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.

Stefan Kopp
Ipke Wachsmuth (Eds.)

Gesture in Embodied Communication and Human-Computer Interaction

8th International Gesture Workshop, GW 2009
Bielefeld, Germany, February 25-27, 2009
Revised Selected Papers

Springer