Sensing and Reacting to Users' Interest: an Adaptive Public Display

Gianluca Schiavo
CIMeC, University of Trento & FBK
Trento, Italy
gianluca.schiavo@unitn.it

Eleonora Mencarini
FBK
Trento, Italy
mencarini@fbk.eu

Kevin B. A. Vovard
FBK
Trento, Italy
vovard@fbk.eu

Massimo Zancanaro
FBK
Trento, Italy
zancana@fbk.eu

CHI 2013 Extended Abstracts, April 27–May 2, 2013, Paris, France.
ACM 978-1-4503-1952-2/13/04.

Abstract
In this paper we describe a public display system that detects the users' interest and adapts the on-screen content accordingly. An interest estimation algorithm based on the analysis of the users' non-verbal behaviour, including the users' position, their orientation and the social context, is proposed. A preliminary field study suggests that an adaptive public display may be more appealing than a control condition, where the same content is offered without any adaptation. We argue that behavioural-based measures are valuable data to inform and adapt a public display in a social-aware way, improving the users' engagement.

Author Keywords
Public Displays; Adaptive Interfaces; Interest estimation

ACM Classification Keywords
H.5.1 Multimedia Information Systems: Evaluation/Methodology.

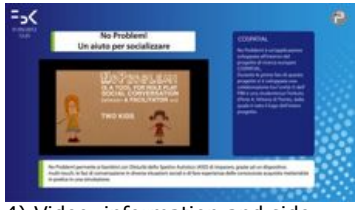
Introduction
Public display systems are an encouraging technology for public and semi-public spaces because of their: (i) *ubiquitous potential*, they can provide ubiquitous access to information; (ii) *social-aware potential*, they are media that support both individual and group

Figure 1. The stages of the media player: each stage is displayed according to the audience level of interest

1) The attractor

2) Video

3) Video and information



4) Video, information and side panel

interactions in public and social contexts; (iii) *context-aware potential*, they are situated artefacts deeply embedded in their specific physical and social environment. In this work, we mostly focused on the two latter points, proposing a social-aware public display that provides different level of information accordingly to the perceived interest of the user(s) and the social context. One of the main challenges in this research line is to expand context-aware systems' capabilities for sensing and model social signals [5]. Our work contributes to this topic by proposing a public display system capable of tracking the surrounding visual scene, by means of a 3D depth sensor, and collecting information from the users' non-verbal behaviour. Behavioural information, including users' spatial position as well as orientation and social context, are then used to estimate the level of attention and interest and finally to automatically adapt the interface to provide a more rewarding experience.

Related Work

This work focuses on the research of ubicomp technology; in particular, it explores the notion of proxemic interaction [1]. Proxemic dimensions, such as distance and orientation, have been used in ubiquitous systems and ambient displays to support user's interaction [6]. In our system, detailed information is presented to the users depending on the estimated interest level and not just on their physical proximity to the screen. Recently, several studies have investigated interaction with public display systems using 3D depth sensors. Müller and colleagues [4] used a Kinect sensor to study how users noticed the interactivity of a public display using visual feedback provided by the sensor itself, while Gollan and colleagues [2] used a 3D depth and RGB camera on a public display to track people movements and orientation in order to estimate passers-by's level of attention.

In our approach, spatial depth information is used in a two-fold way: (*i*) to provide information to the display in order to adapt the content to the users' level of interest and (*ii*) to collect ecological data about the behavior of individuals and groups in front of the screen. In the first case, the system estimates the attention and the interest considering different cues (e.g. spatial position, orientation, number of users, and time passed watching the display). The second point was pursued by quantitative and qualitative analysis of the depth videos collected during a field study.

System Description

The system is composed of (*i*) a 32" LCD screen deployed as an informative public display showing videos and text information; (*ii*) a Microsoft Kinect sensor, fixed on the ceiling at a height of 3.5 m; (*iii*) an algorithm (described in the next section), used to estimate a model of the audience's attention and interest from the data captured by the sensor; (*iv*) a media player application.

The media player use the algorithm's output to adapt the content displayed across 4 different stages (Figure 1): (1) when no user is detected, an attractor consisting of a rotation of videos' screenshots is displayed; (2) when at least one person enters the sensor's field of vision, the system immediately plays a video; (3) more textual information about the topic of the video is provided if the users show more interest to the display; (4) at the further increasing of interest, another side panel with information related to the video is presented.

A strong limitation of the Kinect system could be the assumption of a standard location for the sensor which

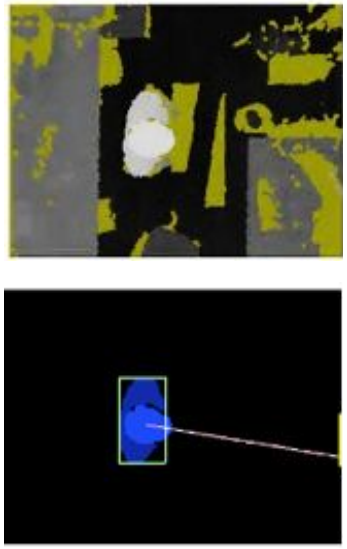


Figure 2. The depth scene analysis: the depth view scene (above) and the same scene analyzed (below): the image shows the user's blob (in blue), the screen position (in yellow, on the right) and the direction of the user toward the screen (the line)

has to be located in front of the screen. Yet, with that configuration, the presence of several users may be problematic since a user can occlude the visual scene. To avoid this problem, we placed the Kinect sensor on the ceiling in order to have a bird's-eye view. We therefore implemented dedicated algorithms to analyze the scene and detecting the users' presence as well as their distance and orientation towards the display (Figure 2 & 4). In the next section, the algorithm to estimate the users' interest is described.

Estimating users' attention and interest

From the related literature [1, 6], we defined the user's attention considering the following rules:

- Attention is expected to decrease when the angle of the head with the screen increases (*orientation function*)
- Attention is expected to decrease when the distance with the screen increases (*distance function*)
- Distance is expected to have a smaller impact compared to the orientation
- Attention is expected to decrease with the presence of other users (*social function*)

We then analytically defined the following formula to estimate the attention of a single user as the combination of the *orientation*, the *distance* and the *social function*:

$$attention = \underbrace{\left(1 - \left(\frac{\propto}{180}\right)^{k_\propto}\right)}_{orientation\ function} * \underbrace{\frac{k_d}{d}}_{distance\ function} * \underbrace{\min(bo)}_{social\ function}$$

The formula includes 3 variables (α, *d*, *bo*) and the parameters $k_\propto$ and k_d that were estimated by trial and error.

The *orientation function* was calculated with $k_\propto = 0.5$ and α as the angle (in degrees) between the direction of the user's orientation and the center of the display. When the user is watching the screen, α is equal to zero; while when the user is looking away, α is different from zero and the formula results in a decrease of the attention value.

The *distance function* was designed under the assumption that the attention is inversely proportional to the distance. The parameter k_d was set to 140 cm (i.e. the maximum distance reached by the sensor).

The *social function* results in a decrease of the attention value when someone is between a viewer and the display (e.g. two people are talking). We considered the obstruction variable (*bo)*:

$$bo = \tanh\left(\frac{(\beta - 45) * k_{bo} * \frac{\pi}{180}}{2}\right) + 0.5\,, with\ k_{bo} = 6$$

The angle β is the angle between the direction from the user to the screen center and the direction from the user toward any other user. The function *bo* works as a high-pass filter behaviour and returns a value near 0 when β is under 45 degrees and a value near 1 when β is above 45 degrees.

In estimating the interest of a given user, the value of *bo* is computed for each other user in the field of view and the smallest value of *bo* corresponding to the most obstructing people is kept.

The interest is defined as the sustained attention over time. The value of interest is proportional to the value of attention, as defined above, and it takes into account the interest's variation over time. At each time stamp (i.e. 166ms):

$$Interest = k_{int} * attention + (1 - k_{int}) * lastInterest$$

The parameter k_{int} is empirically set at 0.009.

In order to estimate the group's interest, we chose a simple approach, defining it as the maximum value of the interest among the users. In other words, the

system adapted the content to the most interested user (which can change during the course of the interaction). In this way, the interface changes smoothly and it always offers the best content for the more interested user. Although the model is simple, it offers an automatic adaptation of the level of content in a dynamic public context and it may represent an interesting baseline to investigate acceptance of this kind of technologies in an ecological setting.

The Field Study

A first preliminary field study was carried out during a public cultural event that took place in the city of Trento (Italy). The system was deployed for 3 days, from 9am to 6pm in a pavilion dedicated to the event in the city's main square. The display was located at the entrance of an exhibition area (Figure 3) and it had the function of presenting videos related to the exhibitors' projects. The objective of the study was to explore, in an ecological setting, how users interact with an adaptive display (Figure 4).

We compared the adaptive system with a control condition, consisting of a non-adaptive system that randomly chose the videos and consistently presented all the available textual information. Both conditions used the same content database. In the control condition, the information about the visual scene was collected as described above, but not used by the system. In order to minimize the influence of time, people affluence and light conditions on the results, the two conditions were counterbalanced during the 3 days, switching automatically every 60 minutes.

Figure 3. The setting of the first field study.

	Passers-by Users		Engaged Users	
	Adaptive	*Non-Adaptive*	*Adaptive*	*Non-Adaptive*
Day 1	69	24	36	11
Day 2	40	10	25	24
Day 3	16	23	25	24
Total	**125**	**57**	**86**	**59**

Table 1. Distribution of the 327 clips over the three days and between the two different modes.

Data Collection and Analysis

Quantitative data about the users' behaviours were recorded by the depth sensor's log file (containing the number of users, distance, duration, attention and interest values for each time stamp).

Similar to Müller and colleagues [4], an analysis algorithm was implemented to automatically search the log files for scenes in which at least one user was detected for more than 1 second, hereafter named *clips*. The data was segmented in clips each one containing an interaction: from the arrival of one user in an empty setting until the last user left the setting. A total of 327 clips were retained for the analysis, showing the interactions of about 400 people. Overall, more users approached the display in the adaptive mode (223 vs 118 users; χ^2= 18.19, *df*= 2, p<.05). As shown in Table 1, we categorized the users in two different types depending on whether they passed in front of the screen for less than 5 seconds (passer-by users) or stayed longer (engaged users). Both passers-by and engaged users were more frequent in the adaptive condition (respectively χ^2= 18.16, *df*= 2, p<.01 and χ^2= 6.54, *df*= 2, p<.05).

Subsequent analyses were focused on engaged users' data and thus 145 interactions of about 200 users were considered, resulting in roughly 2 hours and 9'' of depth view videos (for screenshots see Figure 5).

Results

Firstly, we present the results of a human validation aimed to measure the accuracy of the interest algorithm. Then, the findings from the field study are summarized according to the different metrics collected.

ACCURACY OF THE INTEREST ALGORITHM

Two independent observers (1M, 1F) were involved in a validation to understand the level of agreement between the algorithm and the human ability to rate people's interest by observing the depth view videos. The observers watched each of the 145 clips and rated

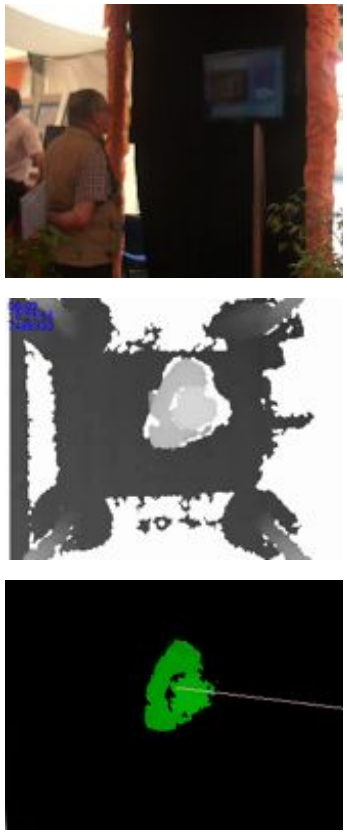


Figure 4. User in front of the display (1), the related depth image (2) and the final elaborated image (3).

the level of interest of the user(s) using a 3-point scale. The means values calculated by the algorithm were recorded in 3 equal-width intervals to allow the comparison with the human rates. The observers' scores were highly correlated between them (r_s= .596, p<.01) and significantly correlated with the estimations provided by the algorithm (r_s= .470, p<.01 and r_s= .541, p<.01). Although less accurate than human annotation, we can conclude that the proposed algorithm provides a good estimation of the users' interest in an ecological setting.

Interest

Interest values were averaged for each clip. Means values were tested with an ANOVA considering two between-subject factors: *mode* (adaptive vs. non-adaptive condition) and *number of users* (single users vs. pairs). Pairs were selected because only few cases with more than 2 users were observed (specifically, 6 clips of 3 users and 1 of 4 users were not considered, reducing N= 138). The ANOVA showed a significant interaction between *mode* and *number of users* ($F_{1,134}$= 10.06, p<.01), and a significant effect of the *mode* ($F_{1,132}$= 4.88, p<.01). Higher values of interest were computed for the adaptive (M= 0.20, SD= 0.018) compared to the non-adaptive system (M= 0.14, SD= 0.021). Pairwise comparisons showed that, with the adaptive system, pairs exhibited a higher level of interest compared to single users (p<.01); this difference did not emerge in the control condition.

Duration and Distance from the screen

Considering the duration of each clip, people spent on average 53.2s in front of the display, but with a high variability SD= 68.7s. The ANOVA showed a significant effect for the *number of users* ($F_{1,132}$= 4.92, p<.01): pairs stood in front of the display longer than single users (M=72.5s, SD= 11.2s vs. M= 43.2s, SD= 6.9s). No effect of the factor *mode* emerged from the analysis. Users' space proximity from the device is a relevant behavioural cue that has been used in the context of interactive public display [1, 6]. According to the data, users on average kept a distance of 198.7cm from the screen (SD= 45.9); no significant differences between *modes* or *number of users* were found. Considering duration and distance metrics alone, significant differences between the two modes were not observed. However, combining their information and including the social context, as done in the interest value, a difference between the two modes emerged.

User experience (Questionnaire)

To evaluate the user experience, we administered the AttrakDiff questionnaire (the pragmatic and the hedonic scales) [3] to 81 users, randomly selected after they had interacted with the display (39 with the adaptive and 42 with the non-adaptive system). The 7-point Likert scales ranged from 1 (positive) to -1 (negative) and they had good reliability (Cronbach's α= 0.67 and 0.86). The adaptive system performed better in both scales compared to the control condition (pragmatic: M= 0.27, SD= 0.39 vs. M= 0.18, SD= 0.38; hedonic: M= 0.25, SD= 0.41 vs. M= 0.20, SD= 0.41), suggesting that the users considered the experience slightly better (or at least equal) in the adaptive condition. Moreover, the pragmatic scores were lower for groups compared to single users (respectively M= 0.16, SD= 0.43 and M= 0.28 SD= 0.34), while the hedonic scores were similar (M= 0.23, SD= 43 and M= 0.24, SD= 0.37). Since the pragmatic scale included dimensions related to usability, this result suggested that groups still have difficulties in interacting with this public display.

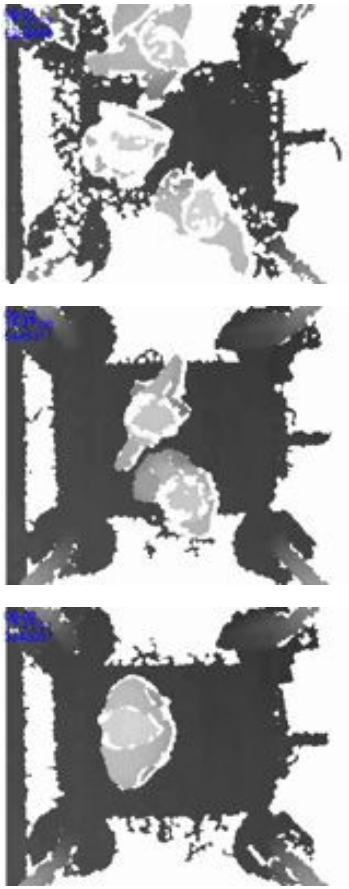


Figure 5. Screenshots from depth videos showing a group of three users, a pair and a single user.

Discussion and Future Work

In this paper, we presented a public display system that estimates the interest of individual users and groups using a 3D depth sensor to collect data and to inform the interface. We showed that information from the depth scene can be useful and insightful not just to collect information about users' behaviors but also to estimate users' interest and to adapt the information provided. This study is an initial attempt to expand the proxemic interaction framework [1,6] by including to some degrees the psychological and social dynamics of groups.

We are aware of several limitations of this work. First, the evaluation of the user experience was simplified and the findings from the questionnaire were not statistically significant. Thus, we cannot strongly conclude that our system provided a more rewarding user experience and this will be investigated in future studies. Nevertheless, a metric based on behavioural cues, as the interest level, may suggest that users were more engaged and interested in the adaptive system, especially when in groups. A second limitation lies in the fact that we used a simple model for the group interaction. In this regard, the next step of our study is to consider a more complex model, for instance a machine learning approach based on the analysis of the users' interaction across time. We also plan to include some new data input from single users (e.g. body activity and movement trajectory) along with new information from the social dynamics, including group communication, speech processes and group cohesiveness. Continuing our research, we intend to explore the application of this system in semi-public spaces, with more structured activities (e.g. work meetings, table activities). We believe that public displays and multi-users systems could express their potential as collaborative tools once they are 'aware' of the social context where they are situated. In this research line, a required step is the deep investigation of more implicit way of interactions and the development of adaptive systems able to sense and to be informed by users' behavioural cues.

Acknowledgments

We would like to thank the event organization office at FBK and Alessandro Cappelletti for their valuable help and support.

References

[1] Ballendat, T., Marquardt, N., & Greenberg, S. Proxemic Interaction: Designing for a Proximity and Orientation-Aware Environment. In *Proc. ITS 2010*, ACM (2010), 121-130.

[2] Gollan, B., Wally, B., & Ferscha, A. Automatic Human Attention Estimation in an Interactive System based on Behaviour Analysis. In *Proc. EPIA 2011*.

[3] Hassenzahl, M., Burmester, M., & Koller, F. AttrakDiff: A questionnaire to measure perceived hedonic and pragmatic quality. In J. Ziegler & G. Szwillus (Eds.), *Mensch & Computer: Interaktion in Bewegung* (2003), 187-196.

[4] Müller, J., Walter, R., Bailly, G., Nischt, M., & Alt, F. Looking glass: a field study on noticing interactivity of a shop window. In Proc. *CHI 2012*, ACM (2012), 297-306.

[5] Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., & Schröder, M. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *Affective Computing, IEEE Transactions on*, *3*-1, (2012), 69-87.

[6] Wang, M., Boring, S., & Greenberg, S. Proxemic Peddler: A Public Advertising Display that Captures and Preserves the Attention of a Passerby. In *Proc. PerDis 2012*, ACM (2012).