2011 Ninth International Conference on Creating, Connecting and Collaborating through Computing

Supporting Multi-user Participation with Affective Multimodal Fusion

Céline Coutrix, Giulio Jacucci, Ivan Advouevski
HIIT
Helsinki, Finland
Celine.Coutrix@hiit.fi

Valentin Vervondel, Marc Cavazza, Stephen W. Gilroy
Teesside University
Middlesbrough, UK
Marc.O.Cavazza@tees.ac.uk

Lorenza Parisi
Sapienza University of Rome
Rome, Italy
Lorenza.Parisi@gmail.com

Abstract—In this paper, we present an application of affective computing as an art installation designed for group interaction. *The Common Touch* utilises a large multi-touch display, presenting interactive visualisations of emotive slogans. The artistic brief is to engage participants in the exploration, touching and manipulation of slogans. Participants reveal the missing words of slogans by touching them. *The Common Touch* utilises several input modalities to build an affective representation of the group interactions: emotional speech recognition, video feature extraction, multi-keyword spotting and touch events. The output of affective fusion is used to refine the selection of slogans presented. We include results of a series of experiments using *The Common Touch* with 24 subjects in groups of 3 using video analysis, logs and questionnaires for data collection. We describe, through interaction analysis, how users utilised the different modalities, suggesting implications for implementing multimodal aesthetic applications to support multi-user participation: tracking multi-user engagement, coverage of possible affective cues in each modality, multimodality in temporal and event analysis, dramaturgy and performative interaction.

Keywords-Multimodality, Affective Interaction, Adaptivity, Aesthetic.

I. INTRODUCTION

Interactive art installations have recently begun to include sophisticated multimodal interfaces in order to track bodily interactions of visitors, together with their utterances. These are fruitful settings for investigation into how multimodal interfaces feature in public settings, the experiences they support and their use in social situations. The latest studies of group interaction with interactive art installations have described the emergence of engaging experiences, suggesting opportunities for adaptive and affective computing—in particular, addressing the influence of social and participative aspects [15]. Multimodality in artistic installations differs from task-based multimodal interfaces in that it serves both as a means of interaction and as a measure of users' engagement. Recent computational models have made the tracking of aesthetic aspects of experience available in real-time by identifying trajectories in dimensional models of affect. These can be used to create affective art installations [10]. We aim to contribute to this stream of work by presenting an application of affective computing as an art installation supporting group interaction on a large multi-touch surface. *The Common Touch* utilises a large multi-touch display, presenting interactive visualisations of emotive slogans.

This raises several questions: How to conceptualise engagement in aesthetic interaction? Which modalities to use to track affective cues and how to fuse modalities? How to make the content adaptive? More importantly, how do multiple users engage in interaction on different modalities?

In this paper, we first consider the conceptual underpinnings of engagement as a basis for discussion about an aesthetic approach to multimodality. We then review several works that present analyses on group interaction and engagement in installations, examining the use of modalities. We then present *The Common Touch* which utilises several input modalities build an affective representation of the group interactions: emotional speech recognition, video feature extraction, multi-keyword spotting and touch events. The output from affective fusion of the different modalities is used to refine the selection of slogans to be presented. This fusion and selection use a dimensional model of affect: Pleasure-Arousal-Dominance (PAD) [21]. Finally, we discuss results of a series of experiments analysing questionnaires, logs and interaction using video based interaction analysis.

II. AESTHETIC APPROACH TO MULTIMODALITY

Different approaches to multimodality have been used for adaptivity. Input synergic multimodality [19] enables the user to complete a task in which the different modalities are combined to allow the system to determine the desired command and parameters [3]. Other approaches have considered multimodality as a way to increase the robustness of potential interaction [25], proposing a fusion of data at a relatively low level of abstraction. However, in

978-0-7695-4363-5/11 $26.00 © 2011 IEEE
DOI 10.1109/C5.2011.14

an artistic installation, aesthetic aspects, such as expression and experience, are more relevant than task orientation. Explicit goals are less of a determinant of the interaction within such systems than they are in, for instance, theory of action [24]. In an installation, action and goals are often constructed dynamically and pragmatically, given the perceived affordance and interaction possibilities offered by the system. Indeed, it has been shown experimentally in [16] that artistic installations require testing phases for the users, which can be explained conceptually by the performative approach on interaction [14]. In these less strongly task- and goal-oriented systems, the design should be based on observations during experimental trials that are as realistic as possible. This also mirrors recent discussion on the need to collect a corpus for multimodal processing of affect [20], justifying the importance of studies of multimodal interaction and participation presented in this paper. From a software architecture viewpoint, we wish to highlight the fact that the absence of pre-defined tasks implies that such systems are different to other commonly described multimodal systems. Indeed, in this context, input modalities and appropriate fusion mechanisms are directly connected to the output modalities, without the need for a functional core in the system like that in [23].

Moreover, affective multimodal interfaces aim to achieve a specific type of adaptation to the user based on their affective expressions. As explained in [2], affective parameters are rarely incorporated in the functional view on interactive systems. The authors propose moving from a model where emotion is an internal state that can be detected and transmitted to the system in the same manner as any other data, to a model where emotion is dynamic, cultural and socially co-constructed. Affective expressions are complex and ambiguous, and multimodality can support sensing of these expressions in order to dynamically adapt the output of the system, and help users interpret the affective expression within their cultural and social context.

A. Recent Psychological Determinants of Engagement

Recent studies of engagement, for example in media and games, make use of different conceptual frameworks and evaluation instruments—among these are presence [31], flow [5][13] and intrinsic motivation [8]. These address, in different ways, engagement within an interactive installation. In such installations, artists attempt to immerse participants in both solo and joint activities, enabling social experiences in mixed reality environments where participants can interact 'as if it was real' (as seen in presence research). Similar phenomena occur in accounts of flow [5]—optimal experiences in which 'attention can be freely invested to achieve a person's goals'. This results in the merging of action and awareness, as well as a consequent lack of self-awareness and distorted sense of the passage of time. The original concept of flow [6] was focussed on achieving optimal (i.e., most enjoyable) experiences in work and leisure situations, where enjoyment is derived from activities that are challenging and require an element of skill. Recent authors have also posited the existence of group flow, such as when participants are engaged both with the product at hand, and with others in the collaboration [28].

B. Multi-user Participation and Multimodality in Installations

The latest studies of group interaction with interactive art and entertainment systems have described the emergence of engagement, detailing phases, patterns, and trajectories. The coverage of modalities in these systems is broad, encompassing a range of interface technologies, such as: augmented reality [22], multi-touch and gestural interaction [15], and real time analysis of voice and speech coupled with the position of users [16]. Processing of multimodality is rare, but can be found in [10] where several input modalities are considered. In [16] the installation projects visualisations of galaxies, which are generated by and move according to the motion of visitors, whilst changing colour depending on their voices. The study employed emotion questionnaires, and results indicated positive feelings were most dominant. Subjective verbalizations refined these positive feelings as showing aspects of interest, ludic pleasure and transport. However, video analysis shows the contribution of multiuser participation in engagement, evident in phases of circumspection, testing and play, including experience sharing and imitation, also found in subject verbalizations.

These studies illustrate the decisive contribution of multiuser participation in engagement, suggesting that a visitor's experience and ludic pleasure are rooted in the embodied, performative interaction with the installation, negotiated with the other visitors in social interaction.

C. A Performative Approach to Adaptivity

The current frontier in engagement, from a computational point of view, is in developing adaptivity within installations. Understanding of adaptivity informs methods for changing user models and aspects of the user interface in response to recorded traces of interaction [17]. This has also been addressed from an affective computing standpoint, based on persuasive feedback theories. The persuasive feedback approach also utilises user interaction to analyse how far the user's state of behaviour deviates from normative expectations and issues appropriate feedback to correct it. Other approaches, such as affective loop experiences [12], have users first express their emotions, which the system responds to by generating affective expressions, to which uses themselves respond; step-by-step, users are led to feel more involved with the system.

However, these approaches lack consideration of the imaginative work of users in identifying themselves in the fictional space, their connection to the physical circumstances and more importantly the contribution of social setting and interaction [15]. In these respects, performative approaches provide a useful tool in designing for adaptivity.

Performative approaches [7] involve observing how the user is simultaneously operator, performer, and spectator; considering imaginative aspects with theatrical concepts

such as fictional space [14] and phases; and taking into account the character of experiences with dramatic structures.

The opportunity for multimodality is to consider how each multimodal utterance or act is not just processed or directed to the system, but is also available as a resource in social interaction (hence the notion of non-interacting spectators in an installation).

III. THE COMMON TOUCH: AN ARTISTIC AFFECTIVE INSTALLATION

The Common Touch is an artistic installation drawing on the analogy between political revolt and advertising in the use of affective engagement of the audience. It uses multiple modalities to provide affective interaction with a group. After introducing the artistic brief, we present the modalities and explain the affective model in which modality data is interpreted, and then present the affective fusion of the modalities and how this is used in the system.

A. Artistic Brief

The Common Touch—that is, figuratively, the ability to appeal to ordinary people—is a large interactive wall (Figure 5 to Figure 9). Common words of an incomplete sentence are displayed in a conspicuous way that appeals to passers-by, who lightly touch the wall and notice that the touch reveals hidden words and causes another sentence to appear. This attracts passers-by, who begin to reveal new sentences and the installation is soon overwhelmed by slogans. The group and the installation itself compete for control of the display. Users form a group of engaged people with their hands raised, revealing advertising slogans conveying an idea of political contestation (e.g., "Be the revolution of you"—Nike and Foot Locker). The slogans that are displayed are determined by the affective expressions of the group, derived from input from multiple modalities and exploited by the installation as in an advertisement or charismatic rhetoric. The group becomes an integral part of an adaptive, controlled demonstration that is presented to passers-by, fulfilling the meaning of *The Common Touch*.

The artistic brief requires the system to (1) mirror the affective expressions of the audience and (2) engage people, encouraging participation in the interaction for a few minutes or more. We show in this paper how multiple modalities can contribute to these goals.

B. Interaction Modalities and Affective Model

Based on our experience during pilot experiments and previous works, we chose to use four different modalities to detect affective expressions of the audience as a group: tactile input, number of people facing the screen (extracted from video), affective keyword spotting and emotional speech recognition (Figure 1). These are mapped to appropriate PAD dimensions so that the affective model supports the brief.

The tactile modality[1] records the variation in the number of touches, which then contribute to an *interest* variable (with a value of 50×Δtouches). The video feature extraction modality records variations in the number of people facing the screen and is also mapped to interest. (If a face enters/leaves the camera field, interest is incremented/decremented by 200/50.) These decisions are based on the artistic brief, which says that the progressive touching and crowding of audience indicates their curiosity, and experimental fine-tuning. As explained in [9], interest then constitutes the arousal and dominance dimensions of a PAD vector, adjusted with an offset (-150), a scale (1.3) and limited between 0 and 600:

$$Arousal = -150 + 1.3 \times Interest$$
$$Dominance = -150 + 1.3 \times Interest$$

The video modality also maps the closeness of the faces to pleasure, based on postural interpretations described in [11].

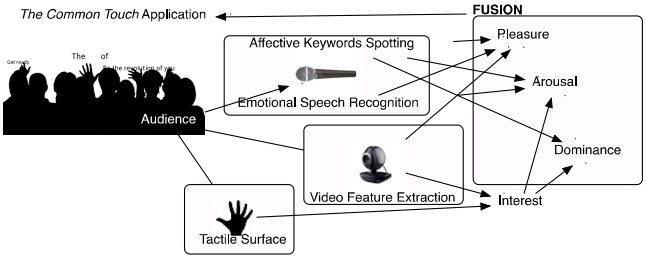
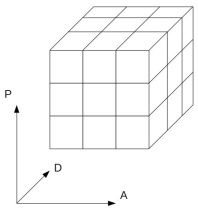


Figure 1: System overview, with its modalities and their contribution to PAD affective representation.



Average rating	PAD class
1 to 3	Negative
4 to 6	Neutral
7 to 9	Positive

Figure 2: PAD space of slogans and mapping between SAM rating and PAD space of slogans.

Affective vocabulary used during conversation by the audience (e.g., "awesome" or "boring") is taken interpreted by the multi-keyword spotting modality. Vocabulary is categorized into categories that are mapped to combinations of PAD values drawn from affective descriptions [27]. For instance, "awesome" is assigned to the enjoyment category that maps to a pleasure of +760, arousal of +480 and dominance +350.

Emotional speech recognition [30] is trained to distinguish Negative-Passive, Neutral and Positive-Active classes from acoustic features, which are then mapped to Pleasure and Arousal.

C. Fusion

Most of the work in affective multimodal fusion has focussed on the recognition of a discrete emotional state

[1] www.multitouch.fi

(such as the combined analysis of facial expression and speech to detect joy or anger). By contrast, we are interested here in fusion as a means to aggregate continuous affective information available across different modalities. Whilst it is unclear how much of the conceptual framework of traditional multimodality [19], [23], [25] can be transposed to the context of affective multimodality (e.g., in terms of complementarity and redundancy of modalities), we can still identify both semantic and temporal aspects. The former depends on the semantics of mapping affective input to PAD space, and the latter to the sampling rate of modalities and the updating rate of fused data. Since the PAD space offers a unified affective representation for the various affective modalities, each of which can be represented as a vector in PAD space, we posit that fusion can be achieved through a linear combination of individual modality vectors (v), the resulting vector characterising user experience at each point in time:

$$PAD_{rep}(t) = \sum_{i=0}^{n} v_i(t) \cdot w_i$$

The individual PAD weighting for each modality (w) normalises their individual contribution based on the distribution of modalities across dimensions. These weightings have been determined using input from the literature, subsequently refined through calibration experiments.

In addition, the temporal aspects of affective fusion should take into account natural decay of affective values (in line with previous research in affective interfaces) and ensure a smooth transition between successive updates. This is achieved by calculating a difference vector to update the current PAD(t) vector, this difference vector being scaled by a time-dependent smoothing function s(t):

$$PAD(t) = PAD(t_0) + (PAD_{diff}(t) \cdot s(t))$$

As a result, the fusion mechanism continuously outputs a PAD vector reflecting the instantaneous affective state of the group of users interacting with *The Common Touch*. The PAD space thus constitutes an elegant representation for aggregating information across modalities and over time. The trajectory of the resulting PAD(t) vector within the PAD space can also be used to characterise users' experience.

D. Displaying the Slogans

The installation should induce, as well as respond to, emotional reactions in the audience. To this end, the slogans mirror the combination of sensed affective expressions provided by fusion, and are stored in a three-dimensional data structure (Figure 2). Each dimension has three discrete levels: negative, neutral, and positive. Slogans haven been previously rated along PAD dimensions by seven different persons using the 9-point scale of the SAM method [4]. The average of the rating is transformed to one of the states for each dimension (Figure 2).

When a new slogan is to be displayed, the PAD output of fusion is normalized in order to produce PAD values in {-1, 0, 1}. The system displays one of the slogans in this state that was not displayed before. However, it is possible that the emotional state remains the same for a longer period and after a while all the slogans corresponding to that state would have been displayed. To overcome this problem, slogans from nearest cells will be displayed until the number of shown slogans reaches a threshold experimentally defined. The threshold value is a compromise between showing users new slogans and showing a slogan that correspond to the current emotional state. For threshold we have chosen 35 slogans, which is around one third of the overall number of available slogans.

IV. Evaluation

A. Approach and Data Collection

Initial evaluation consisted eight sessions of interaction with groups of three people. We recruited from our lab 24 subjects, 15 male and 9 female, from 22 to 54 years old. They were first introduced to the concept of the installation, informed that it is an artistic installation where the aim is to explore the interactions proposed by the installation.

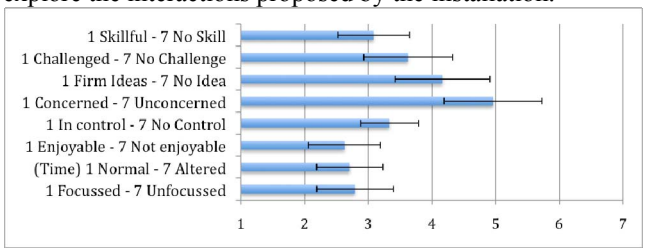


Figure 3: Subjects evaluation of their flow during the experience.

We collected logs from the software (four modality components, fusion and application output) in order to evaluate the multimodal affective interaction, and captured video clips from the front of the installation in order to analyse facial expressions of the subjects and also from the side of the installation, in order to have a wide-angle view on the scene. We collected answers from a questionnaire evaluating subjective affective experience and engagement in terms of Flow [6], Interest and PAD. Subjects rated their pleasure, arousal and dominance according to the 1-9 SAM scale [4]. They found themselves happy overall (P=3.83±0.64, α = 0.05), stimulated (A=4.42±0.74, α = 0.05) but neither in control nor controlled (D=5.37±0.62, α = 0.05), which is consistent with the artistic brief. Figure 3 presents the subjects' detailed rating of the flow during the experience. Surprisingly, there was no significant alteration in the perception of time. We can also see in Figure 3 that control, firm idea of goals and challenge are the most neutral, confirming that subjects overall felt neither in control nor controlled, that they had neither firm ideas nor completely no idea about how they wanted the system to behave—expected in an art installation—and felt to be challenged by the installation, but not too much or too little. In all Figures, confidence intervals are calculated with α = 0.05.

B. *Exploration of Slogans*

The slogans are clearly a critical resource for engagement, as a user explains at the end of the nine-minute-long session 2: "*There is too much of the same thing. They are just repeating. […] I am bored enough.*" Therefore, it is important for the system to display a variety of slogans. We analysed exploration of the slogans and found that on average 45% (±7%, $\alpha = 0.05$) of the 108 slogans were displayed. A single slogan was repeated 8.35 times on average (±1.98 times, $\alpha = 0.05$) during a session. On average over all sessions, the system displayed 0.91 new slogans per second (±0.10 slogan, $\alpha = 0.05$). This shows opportunity for enlarging the list of slogans even more.

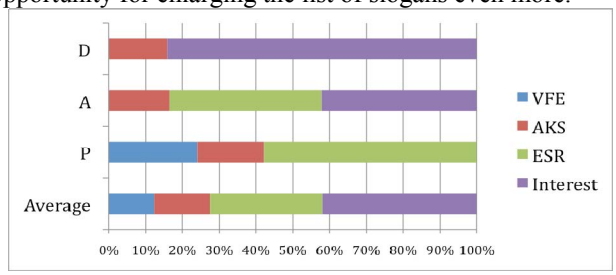


Figure 4: Contribution of modalities to the PAD expression: Video Feature Extraction (VFE), Affective Keywords Spotting (AKS), Emotional Speech Recognition (ESR) and Interest from tactile and video modalities.

C. *Multimodal Group Interaction*

We note from software logs of the three last sessions that all modalities contributed to the PAD expression according to the affective model defined for the installation (Figure 4).

In order to go beyond this observation and contribute to the understanding of the group participation to the installation through multimodality, we conducted video analysis of the interaction during the experiments. The approach adopted is based on previous work, drawing from interaction analysis ([18], [29], and see for example [15]).

Videos were examined qualitatively, based on human coding of events, to find the phenomena that arise in the use of our multimodal affective system. We were looking for patterns of interaction for each modality independently, as well as the combined use of those modalities, as evidences of participation.

1) *Body attitude and proxemics.*

Whilst gestures and touch appear as the dominant interaction mode, body attitudes convey precious information about the user experience. Subjects tended to remain standing at the same location *along the screen*: we observed that users mostly tend to respect each other's personal working space. However, on at least 3 occasions in 8 of the sessions, subjects changed their spatial arrangement. As an example in Figure 5 shows, a user changed his position in space to better help the team to complete a sentence. If a subject left the installation, the others tended to occupy this empty space, slowly reconfiguring their position (Figure 6).

Subjects stood at different *distances from the display*. They mostly stood close to the display in order to touch it. However, they sometimes stepped back (Figure 7a), to observe others and would either approach closer again to help/collaborate or would leave.

We noticed twice that subjects *crouched down*, shown in Figure 5. Also they changed the *orientation of their body* in the space. As shown in Figure 7b, they sometimes turned around and talked to the people in the background.

In the current version of the system, the video modality counts the number of people facing the screen as well as their distance from it. We think this is still relevant since, as previous studies show [15], it is an indicator of the audience's participation and interest. Nevertheless, these experiments in the lab showed that there is opportunity for a more fine-grained design of bodily interaction in the installation space.

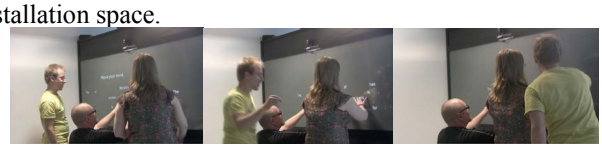
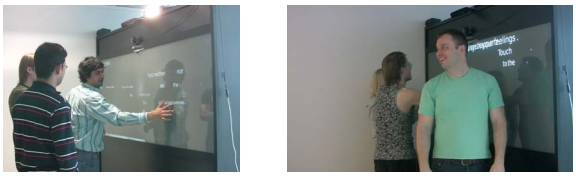


Figure 5: A user changes location while another crouches down (session 6).

Figure 6: Reconfiguring the space after someone left.



(a) **(b)**

Figure 7: (a) A subject invades a fellow spectators' space. (b) A user turning around to comment about the installation with people watching in the background.

2) *Affective keywords in speech*

We observed that the amount and subject of conversation depended strongly on the individual subjects and their relationships with each other. As an extreme example, there was a session where subjects talked throughout the whole interaction (session 2), whereas in session 4 no one spoke for almost the entire session. In a similar manner to movements in space, this shows the importance of including several modalities as inputs to affective fusion since some individual sessions were characterized by relatively very small contributions from one of the modalities.

In most cases, subjects spoke aloud as though they were speaking to others next to them. In general, there were few replies to these comments. We observed use of the vocabulary that was detected by our keyword spotting component (e.g., *"Nice", "That's funny", "Come on! Join the fun!", "I actually like these slogans"*).

Many subjects read aloud sentences once they completed them, for example, saying: *"Don't let anything*

stop you!" as though thinking aloud. Some also read the incomplete sentences, trying to figure out the missing word(s) (e.g*., "we win when you....*"). In session 2, one subject even explicitly proposes this as a game to the others: "*Let's try to guess!*". This vocabulary is not currently taken into account. However, it shows additional interest from the audience, and could contribute to the interest variable of the affective fusion too.

Subjects also talked to the system itself. We observed in session 6 one user saying: *"Don't go away!",* referring to a sentence that was leaving the top of the screen. Nearly half the subjects exchanged comments about how the system works. For instance, in session 6 (from 1'00" to 1'25" in a 6'30" long session), we find the following dialog:

- *So, how do you complete the sentences?*
- *You have to have all the words there, then they stay*.

This phased approach in engagement has been observed previously [16] and the dialogues could be used to help track the users through these phases.

3) Emotional speech

We noticed that the speech was mostly neutral. However, we found some episodes where expressive speech was used extensively. Also, particular users had very expressive speech (in the example of session 2, Positive-Active) and enticed others next to them to speak in the same manner.

Users talked directly to others in an expressive manner: for example, [negative-passive, disappointed] *"you are not helping me at all"* (session 2), [positive-active]"*So I KEEP my fingers there*" (session 5), [positive-active]"*Come on! Join the fun!*" (session 2).

We also observed expressive reading of slogans. This followed different patterns: neutral speech (like thinking aloud) followed by expressive speech (directed towards others) or vice versa. For example in session 2, one user says [neutral] *"touch... power..."* [exaggeratedly happy] *"Touch Power!"*

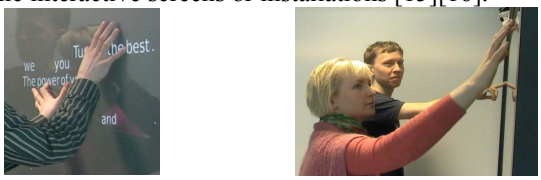
A recurring event in the sessions was the use of expressive sounds: positive ones like "*Aaah!*" or "*Wooo!*" to express enjoyment, surprised or encouragement and negative ones to express frustration or effort.

Based on these experiments, we see that the neutral and positive-active categories for emotional speech recognition are relevant.

4) Tactile modality

Our study confirms the results of previous studies such as [15], showing that users tend to start touching with one hand and one finger first. They then used two hands, one finger each. After a while, they made more complicated gestures for revealing the slogans and moving them once revealed. We observed them using the elbow when no finger is left and no opportunity for collaboration, testing with the whole palm of the hand or the external side (from the wrist to the little finger, in contact with the glass) and stretching between thumb and little finger (Figure 8a). One subject even commented, "*you'll need at least 8 hands!*" In terms of dynamics, we observed mostly discrete touches, but also a sliding of the fingers along the sentence, tapping with fingers tips or palm. A large majority of the subjects eventually end up throwing the slogans towards the top.

Some users learned how to use the system not by trying it out themselves, but by observing the first user to do so. They then started interacting right away without much exploration of techniques. Several times one user was seen to teach another how to interact, by showing or by collaborating on one of their workspaces. This type of co-learning and imitation is a recurrent finding of usage studies of public interactive screens or installations [15][16].



(a) (b)

Figure 8: (a) Stretching fingers to reveal the most slogans simultaneously. (b) A user touching while paying attention to another's activity and commenting on it: "You can move these?!"

(a) (b)

Figure 9: Collaborating to complete a slogan (a). Crossing arms between subjects (a) and within a single subject (b).

Subjects often collaborated to reveal slogans: one touches two words and the other the rest (Figure 9a). Some even asked explicitly: "*Can you do* 'The Wind of change' *with me?*", "*Can you press this?*" or *"One more finger there!*". On one occasion, three users collaborated together to reveal a long slogan. Subjects also sometimes moved the slogan amongst themselves once they had completed it. For instance, in session 7, three users collaborated by moving a slogan from the right to the left of the screen.

Users sometimes interacted in somebody else's space. Possible reasons for this are (1) to teach someone how to interact, (2) to help or collaborate, (3) to intervene if the other subject stepped back and (4) when no more slogans are displayed in one's own 'area'. Interacting in anothers' space sometimes made them step back (Figure 7a).

Several times, we observed two users' arms crossing whilst collaborating (Figure 9a) or invading another's workspace, or crossing of arms by a single user (Figure 9b). Also, users had a characteristic pace evident in their rate of touching the screen: some touched a lot, and quickly, others took time to read and were slower.

A progressive increase in the number of touches at the beginning, detected in the system for the interest variable, seems experimentally to be a good predictor of this interest. We also observed, in almost all the interactions, "hyper"

behaviour in individuals and collective engagement during the second part of the interaction.

Finally, as with the speech modalities, evidence of collaborative participation in the experience was found. Not all instances of participation in multiuser interaction at a public display are collaborative as observed by [15]. However, whether collaborative or conflicting, the degree of participation we saw varied, ranging from loosely coupled to tightly coupled participation [1].

5) Combined use of modalities

We observed a *synergistic* [19] *use of semantic and expressive speech* together, as described in previous sections considering voice modalities, for example, "*Come on! Join the fun!*" said in a very positive-active way. We also observed *simultaneous use of emotional speech and touch*, e.g. "*Pfuuu*" (in session 7 at 8'17"), together with the throwing of a slogan with pleasure. Subjects touched and laughed at the same time too: in session 7, they moved a slogan all around the screen and laughed as the slogan bounced off the borders.

The bodily movement modality was also used in a synergistic way with the speech modalities. Indeed, subjects repeatedly turned back to spectators to comment (as in Figure 8b), for example saying, "*What is this all about?*"

We also find *simultaneous use of semantic speech and touch*. However, touch and talk sometimes occur simultaneously without being related to the same user's touching. For instance, in session 5 a subject reveals a slogan without paying attention to it but one of his neighbours (see Figure 8b) comments on it: "*You can move these?*"

We find *alternating use of speech (emotional or semantic) and touch*, for example, in session 7 at 2'20", one says *"So we have to..."* and then presses a sentence. This supports collaboration. In session 2, one subject read shyly the slogan: "*Compromise between future and equality*" after completion and another asked "*Compromise what?*" leading the first to pass the slogan to the second user by sliding it over.

While the affective expressions of users transmitted through all modalities, there is also opportunity to reflect on the fusion mechanism, not just the modalities themselves. Firstly, we notice that we were able to describe the use of the modalities in terms of functional multimodality [19]. However, whether they are independent or not, subjects participate equally in the overall expression of the affective state of the group, reinforcing the choice of a late, non-functional approach for affective fusion. PAD fusion allows us to deal with traditional combination of modalities but also to deal with mixed continuous and discrete data.

V. Discussion and Conclusion

Studies of public installation point to the importance of recognizing trajectories of engagement and mechanics of participation in even brief interaction sessions. We aim to contribute to a stream of work that identifies in multimodal affective fusion a promising solution for the creation of adaptive mechanisms in public installations. We have presented a system called *The Common Touch* based on an artistic brief that allows multiple users to explore artistic content via the adaptive mechanisms of the installation.

Adaptivity is implemented by real-time analysis of several modalities (number and distance of facing users from video, touch events, affective keyword spotting, emotional speech recognition) to compute a dynamic representation of the engagement of users based on affective dimensional models, such as PAD. Affective fusion is used to selectively explore the possible slogans of the installation by choosing slogans that are closest to the users' affective expressions. Our approach is based on previous interpretations of engagement that links it, in particular, to flow experiences and performative interaction. The series of experiments with *The Common Touch* serves to demonstrate the successful application of affective multimodal fusion to multiuser participation by adaptive exploration of slogans.

These experiments allowed us to infer implications for current and future implementation that track and process these modalities. The 8 sessions with 24 subjects varied considerably in usage of modalities but exhibited common properties:

Participative engagement. While traditionally affective computing has mostly addressed adapting interaction to an individual, our studies show the application of multimodal fusion in a multiuser setting. This is characterized by different social dynamics, where engagement is co-constructed [15]. Social learning, imitation, and co-play influence the engagement of the group to the installation. Users also make use of explicit interactions, such as touch, to interact with other users. Conversational utterances directed towards other users are captured by the system and used in the adaptivity loop.

Coverage of possible affective cues in each modality. Affect recognition is performed in each modality based on heuristic or training and pattern recognition. Currently, corpus collection and rule-based definition of cues in affective input processing are open challenges [20]. The analysis however shows how, for each modality, interesting cues remain to be recognised: for example, in configuration and movement of bodies in space, such as body postures and distance, or in semantic interpretation of speech.

Multimodality temporal and event analysis. Further work is required in identifying additional cues to track engagement of multiple users, including the identification of multimodal acts, in particular, the recognition of specific multimodal acts that we could call events, or the tracking of particular sequences of events that can provide additional affective cues. In this context, common types of fusion mechanisms [19] based on a functional view of interaction need to be adapted to affective interaction.

In order to improve robustness of the sensing of the affective expression, a common approach would be to use low-level fusion. Conversely, the fusion we use here supports semantic fusion while attempting to retain the robustness of interaction. Furthermore, the recognition of specific multimodal acts that we could call events, or the tracking of particular sequences of events can provide additional affective cues. In fact, previous work has

addressed how engagement in public installations follows particular trajectories, patterns and phases [10], [15], indicating the opportunity to analyse sequences of actions or situations.

Dramaturgy and performative interaction. The adaptivity and application of affective fusion followed the principle of reinforcing or reflecting on the state of the users, typical of affective or feedback loops approach [10][12]. This approach, whilst part of the artistic brief, is but one of many approaches possible. Indeed, other briefs or installation objectives might attempt to have users follow a planned dramatic structure in the interaction session [16].

Other ways of devising engaging mechanisms can also be considered. In the course of building several versions of *The Common Touch* we have seen, for example, different rhythms or size to display slogans that can be linked to users' engagement, in order to accelerate or focus their interaction.

ACKNOWLEDGMENTS

The authors would to thank members of the other partners in CALLAS who developed the modality components used in *The Common Touch*: Markus Niiranen, Tommi Keränen, Elisabeth André, Thurid Vogt, Johannes Wagner and Jérôme Urbain.

This work has been funded in part by the European Commission via the CALLAS Integrated Project. (ref. 034800, http://www.callas-newmedia.eu/)

REFERENCES

[1] Baker, K., Greenberg, S., and Gutwin, C. Empirical development of a heuristic evaluation methodology for shared workspace groupware. CSCW'02, ACM, 96-105.

[2] Boehner, K., DePaula, R., Dourish, P., and Sengers, P. 2005. Affect: from information to interaction. CC '05. ACM, 59-68.

[3] Bolt, R. A. 1980. "Put-that-there": Voice and gesture at the graphics interface. SIGGRAPH '80. ACM, 262-270.

[4] Bradley M. M., Lang P. J., 1994. Measuring emotion: The self-assessment manikin and the semantic differential, Journal of Behavior Therapy and Experimental Psychiatry, 25(1), 49-59.

[5] Csikszentmihalyi, M. Robinson, R.E. 1990. The Art of Seeing: An Interpretation of the Aesthetic Encounter. J. Paul Getty Museum Publications, Los Angeles.

[6] Csikszentmihalyi, Mihaly. 1975. Beyond boredom and anxiety. Jossey-Bass Publishers, San Francisco.

[7] Dalsgaard, P., Koefoed Hansen L., Performing Perception—Staging Aesthetics of Interaction, ACM TOCHI, 15(3).

[8] Deci, E. L., Ryan, R. M., 2000. The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. Psychological Inquiry, 11, 227–268.

[9] Gilroy, S.W., Cavazza, M., Niiranen, M., André, E., Vogt,, E. Urbain, J., Benayoun, M., Seichter, H., Billinghurst, M., 2009. PAD-based Multimodal Affective Fusion, ACII'09, IEEE.

[10] Gilroy, S.W., Cavazza, M., Benayoun, M., Using affective trajectories to describe states of flow in interactive art. ACE '09, 165-172.

[11] Hillman, C. H., Rosengren, K. S., Smith, D. P., 2004. Emotion and motivated behavior: postural adjustments to affective picture viewing. Biological Psychology, 66:51–62.

[12] Höök, K., 2008. Affective Loop Experiences -What Are They?, Persuasive Technology, Oulu Finland, 1-12.

[13] Jackson, S.A., Marsh, H.W. Development and validation of a scale to measure optimal experience: The Flow State Scale. Journal of Sport and Exercise Psychology, 18 (1996), 17–35.

[14] Jacucci, C., Jacucci, G., Wagner, I., Psik, T., 2005. A Manifesto for the Performative Development of Ubiquitous Media. CC'05, ACM, 19 –28.

[15] Jacucci, G., Morrison, A., Richard, G. T., Kleimola, J., Peltonen, P., Parisi, L., and Laitinen, T. 2010. Worlds of information: designing for engagement at a public multi-touch display. CHI '10, ACM, 2267-2276.

[16] Jacucci, G., Spagnolli, A., Chalambalakis, A., Morrison, A., Liikkanen, L., Roveda S., Bertoncini, M., Bodily Explorations in Space: Social Experience of a Multimodal Art Installation, Interact'09, Springer LNCS 5727, 62-75.

[17] Jameson, A., Adaptive interfaces and agents, The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications, Lawrence Erlbaum Associates, Inc., 2002

[18] Jordan, Brigitte and Austin Henderson. 1995. Interaction Analysis: Foundations and Practice. The Journal of the Learning Sciences, 4(1), 39-103.

[19] Lalanne, D., Nigay, L., Palanque, p., Robinson, P., Vanderdonckt, J., and Ladry, J. 2009. Fusion engines for multimodal input: a survey. ICMI-MLMI '09. ACM, 153-160.

[20] Liikkanen, L., Jacucci, G. & Helin, M. (2009) ElectroEmotion: A Tool for Producing Emotional Corpora Collaboratively. ACII 2009, IEEE.

[21] Mehrabian A., Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Current Psychology, 14:261–292, 1996

[22] Morrison, A., Oulasvirta, A., Peltonen, P., Lemmela, S., Jacucci, G., Reitmayr, G., Näsänen, J., and Juustila, A. 2009. Like bees around the hive: a comparative study of a mobile augmented reality map. CHI '09. ACM, 1889-1898.

[23] Nigay, L., Coutaz, J. 1995. A generic platform for addressing the multimodal challenge. CHI'95, ACM, 98-105.

[24] Norman, D., Cognitive Engineering. Book chapter of User Centered System Design, New Perspectives on Human-Computer Interaction, 1986, 31-61.

[25] Oviatt, S. 2000. Taming recognition errors with a multimodal interface. CACM 43(9), 45-51.

[26] It's Mine, Don't Touch!: interactions at a large multi-touch display in a city centre. CHI '08. ACM, 1285-1294.

[27] Russell, J. A., Mehrabian, A., Evidence for Three-Factor Theory of Emotions, Journal of Research in Personality 11, 273-294, 1977.

[28] Sawyer, K. Group creativity: Music, theater, collaboration. Lawrence Erlbaum Associates (LEA), NJ, USA, 2003.

[29] Suchman, L., Trigg, R. (1991). Understanding Practice: Video as a Medium for Reflection and Design. Design at Work: Cooperative Design of Computer Systems (pp. 65-90). Lawrence Erlbaum Associates

[30] Vogt T., André E. Improving automatic emotion recognition from speech via gender differentiation. LREC'06.

[31] Vorderer, P., Wirth, W., Gouveia, F. R., Biocca, F. Saari, T., Jancke, F. MEC-SPQ: Report to the European Community, Project Presence: MEC (IST-2001-37661), 2004.