Design and implementation of an affect-responsive interactive photo frame

Dibeklioğlu, H.; Ortega Hortas, M.; Kosunen, I.; Zuzánek, P.; Salah, A.A.; Gevers, T.

DOI
10.1007/s12193-011-0057-5
Publication date
2011
Document Version
Final published version
Published in
Journal on Multimodal User Interfaces

J Multimodal User Interfaces (2011) 4:81–95
DOI 10.1007/s12193-011-0057-5

ORIGINAL PAPER

Design and implementation of an affect-responsive interactive photo frame

Hamdi Dibeklioğlu · Marcos Ortega Hortas · Ilkka Kosunen · Petr Zuzánek · Albert Ali Salah · Theo Gevers

Received: 10 February 2011 / Accepted: 16 March 2011 / Published online: 16 April 2011

Abstract This paper describes an affect-responsive interactive photo-frame application that offers its user a different experience with every use. It relies on visual analysis of activity levels and facial expressions of its users to select responses from a database of short video segments. This ever-growing database is automatically prepared by an offline analysis of user-uploaded videos. The resulting system matches its user's affect along dimensions of valence and arousal, and gradually adapts its response to each specific user. In an extended mode, two such systems are coupled and feed each other with visual content. The strengths and weaknesses of the system are assessed through a usability study, where a Wizard-of-Oz response logic is contrasted with the fully automatic system that uses affective and activity-based features, either alone, or in tandem.

Keywords Affective computing · Facial expression · Optical flow · Interactive photograph · Human behavior understanding · Automatic video segmentation

1 Introduction

Affective household artifacts belong to a near future, where people are surrounded by technology and demand more intelligent interaction from simple objects. Some of these objects, by virtue of their traditional role in the everyday environments, are more open to such enhancement than others. In this paper we describe one such object, a dynamic affect-responsive photo frame. This system replaces a traditional static photograph with a video-based frame, where short segments of the recorded person are shown continuously, depending on the multimodal input received from the sensors attached to the interactive frame. The photo frame is an object of emotional focus. Our purpose is to bring interaction into it to raise its affective value and 'presence', as well as to re-define it as an object of communication.

The prototypical scenario we consider is the interactive photograph of a baby, set up in a different location, for instance in the living room of the grandparents. While there is no one around, the baby is asleep in the photo frame. Once the grandmother arrives, the baby 'wakes up', and responds to the grandmother. These responses are matched to the displayed affect of the grandmother; a smile from her may elicit a smile from the baby, or an abrupt gesture may prompt the

This paper is based on [9].

H. Dibeklioğlu · A.A. Salah (✉) · T. Gevers
Intelligent Systems Lab Amsterdam, Informatics Institute, University of Amsterdam, Science Park 904, 1098XH, Amsterdam, The Netherlands
e-mail: a.a.salah@uva.nl

H. Dibeklioğlu
e-mail: h.dibeklioglu@uva.nl

T. Gevers
e-mail: th.gevers@uva.nl

M. Ortega Hortas
University of A Coruna, Campus de Elviña, 15071, Corunna, Spain
e-mail: mortega@udc.es

I. Kosunen
Helsinki Institute for Information Technology HIIT, University of Helsinki, PO Box 68, 00014, Helsinki, Finland
e-mail: ilkka.kosunen@hiit.fi

P. Zuzánek
Faculty of Information Technology, Czech Technical University in Prague, Kolejní 550/2, 160 00, Prague 6, Czech Republic
e-mail: zuzanpet@fel.cvut.cz

baby to be more active as well. The responses of the baby come from previously recorded videos. The parents of the baby record new videos from time to time, and these are added to the system automatically. Thus, the system may have novel content, perhaps indicated by a visual cue similar to received mail signals.

To realize such a system, we propose methods to automatically analyse and segment video sequences to create response dictionaries, combined with methods for real-time affect- and activity-based analysis to select appropriate responses to the user. We then propose a number of system extensions and conduct a usability study to understand the limits of the proposed system.

The contributions of this work are the following:

- It presents the first fully operational affect-responsive photo frame system, which relies on visual cues of arousal and facial affect.
- It proposes a way of bringing constant novelty to the photo frame by specifying a method of enriching the response library via automatic offline segmentation of uploaded content.
- It describes a very challenging setup, where the system's response is not engineered, but automatically matched to the users.
- It assesses a number of features for real-time valence-arousal measurement from a monocular camera in the context of a simple interaction logic.
- Based on a usability study, it offers qualitative conclusions to guide the design of similar systems.

This paper is structured as follows. Section 2 describes the related work, with particular emphasis on actual systems rather than on the algorithmic tools. In Sect. 3 the proposed system, its separate modules, and its use-cases are detailed. Section 4 explains the experimental methodology and the assessment of the proposed algorithms within the application context. Section 5 details the usability study, and discusses its implications. Finally, Sect. 6 presents our conclusions.

2 Related work

2.1 Interactive artifacts

There are a number of interactive artifacts that can be considered as precursors to the proposed approach. A project which brought some interaction to photographs is the Spotlight project of Orit Zuckerman and Sajid Sadi (2005), developed at MIT MediaLab.[1] In this project, 16 portraits are placed in a 4×4 layout. Each portrait has nine directional

Fig. 1 The Spotlight installation of Zuckerman and Sadi (2005)

temporal gestures (i.e. one of nine images of the same person can be displayed in the portrait at any given time), which give the appearance of looking at one of the other portraits, or to the interacting user. The user of the system can select a portrait, at which point the remaining portraits will 'look' at it (see Fig. 1). This project demonstrates the concept of an interactive photograph with static content. While the combination of portraits create novel patterns all the time, the language of interaction is simple and crisply defined.

In 2006, an interactive photo frame is created in "Portrait of Cati" by Stefan Agamanolis, where the portrait in question can sense the proximity of the spectator, and act accordingly [1]. When no one is close to the portrait, Cati displays a neutral face. When someone approaches, it selects a random emotion, and displays it in proportion to the proximity of the spectator. If the selected expression is a smile, for instance, the closer the spectator comes, the wider Cati will smile. A similar project is the Morface installation, where an image of Mona Lisa was animated based on the proximity of the interacting person [20]. In this project camera-based tracking is used to determine proximity and head orientation of the user.

A responsive interactive system in which virtual characters react to real users was proposed in [26], called an Audiovisual Sensitive Artificial Listener. Facial images and voice information in the input videos are used to extract features, which are then submitted to analysers and interpreters that understand the user's state and determine the response of the virtual character. The sequential recognition and synthesis problems are handled with Hidden Markov Models (HMMs). In [6], a dialogue model is proposed that is able recognize the user's emotional state, as well as decide on related acts. The observed user's emotional states and actions are modeled with partially observable Markov decision processes (POMDPs).

The system described in this work is different in several aspects from the systems discussed in the literature.

[1] http://ambient.media.mit.edu/people/sajid/past/spot-light.html.

In our model the responses of the system are not fixed or engineered, but are automatically created, and their volume grows in time as the user uploads new videos. In this manner, the system maintains novelty. The two interactive systems we just described are proposed as art installations, but we target a home application, for which novelty plays an important role. We use real videos in the systems output, with no manual annotation of the contents. This is much more challenging than producing appropriate responses through a carefully engineered synthesis framework, where the system has control over the output. A fully automatic segmentation procedure is proposed to create self-contained response patterns based on affect and activity cues sensed from the user, for which the precise semantics is not known at the onset. In the absence of such precision, we require our system to show consistency, where a given type of user behavior produces system responses in a consistent manner, and where the user is the primary driver of the interaction semantics. Yet another challenge is the real-time interaction, which makes it necessary to work with lightweight features.

2.2 Measuring visual affect

The two modalities we use for real-time analysis are facial expression and the activity of the user, respectively. Our facial expression analysis is based on the eMotion system, which recognizes six basic facial emotional expressions in realtime [28, 32]. This method uses a Bézier volume-based tracker for the face and a naïve Bayes classifier for expression classification.

For the second channel, we use rapid optical flow tracking and derive a host of simple features. The activity levels of the user can be described efficiently through these features. Optical flow descriptors were previously used in the literature (e.g. [11]) to model the arousal levels of users. Arousal is a feature that is prominent in several models of affect [22, 24]. We do not explicitly follow any of these models, but if we position the proposed system with respect to Russell's circumplex model of affect [24], facial expressions we detect will (mostly) indicate valence, and optical flow features will indicate arousal levels of the user. Taken together, these two sets of features guide the interaction of the system in real-time.

In the last few years, facial expression recognition and recognition of bodily gestures have seen great improvements, and there are several systems proposed for determining affect using both modalities. Here we will mention a few key works, and refer the reader to [34] and references therein for a more general overview of affect sensing methods.

Kaliouby et al. previously proposed a MindReader API which models head and facial movements over time with Dynamic Bayesian Networks, to infer a person's affective-cognitive states in real time [16]. Shan et al. used spatio-temporal features extracted from bodily gestures to recognize affect, and combined these cues with facial expressions using canonical correlation analysis [29]. In their approach a set of annotated affective gestures are used for guiding supervised learning. A large number of visual features are extracted, and support vector machine (SVM) classifiers are used to determine the closest template, which subsequently indicates the affect class of a novel sample. In [7] Gabor features are used with SVMs to map facial expressions to pleasure, arousal, and dominance dimensions of Mehrabian's affect model [22].

In [2] a real-time emotion analysis system was proposed that used an efficient facial feature detection library in conjunction with a number of physiological signals. Gunes and Piccardi presented a system in which feature level and decision level fusion is considered together for facial and bodily expression of affect [12]. They point out to the fact that affective face and body displays are simultaneous, but not strictly synchronous.

The approach presented in this work is different from the work discussed in this section in that it uses an unsupervised approach for grouping spontaneous activity into affect-based clusters. During the operation of the system, matching is performed in a straightforward way, but the assignments do not bear labels, and the clustering is fully-automatic. The next section describes the algorithmic aspects of the proposed system.

3 Design of the affect-responsive photo frame

In this section we describe the logic and the design choices for the affect-responsive photo frame, and introduce the separate modules.

3.1 Overview

The system is composed of five modules, which we briefly summarize here before a more detailed exposition. The first of these is the offline segmentation module. The purpose of this module is to create a response segment library, composed of short video fragments. The input is any number of uploaded videos or recordings. In the prototypical use-case, these are the videos of the baby. These videos are analyzed in an offline fashion, and the segments are stored in a segment library. During interaction, the system will choose one of these segments at any given time.

The second module is the affect and activity analysis module. Here the visual input from the user is analyzed in real-time, and a feature vector is generated. This is the module that processes the behavior of the grandmother in the use-case.

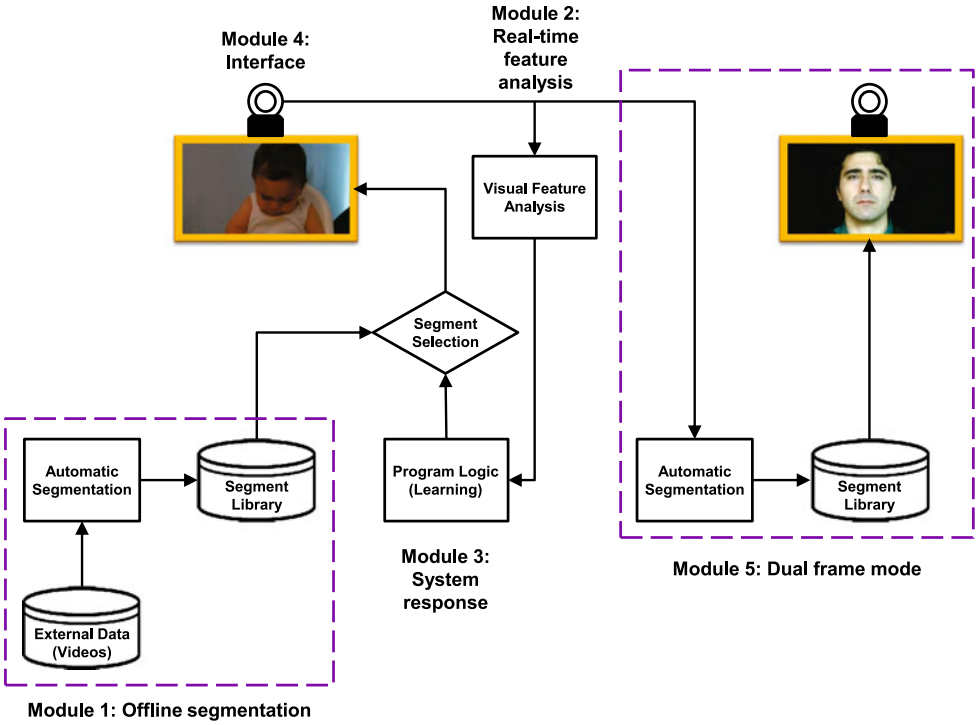

Fig. 2 The overview of the operation of the system. In the dual-frame mode, each frame is used to record new videos that are automatically segmented and added to the segment library of the other frame, establishing an asynchronous communication channel

The feature vector is used by the third module, which is the system response logic. The features computed in the second module are used to select an appropriate video segment from the segment library. This module also incorporates learning, to fine-tune its response over time. The system uses its offline period to execute an unsupervised learning routine for this purpose.

The fourth module is the user interface. The segments are displayed to the user in the photo frame, depending on the user input. In the design phase we have experimented with different types of additional feedback to the user, to indicate what the system perceives of the user during its operation, or to show how close the user is to eliciting a certain response. These external cues (see Fig. 7(b) for one example of these cues) diminish the perception of the system as an affective artifact, and were removed eventually.

Finally, the fifth module is the implementation of the dual frame mode. In this extension there are two coupled frames, in different locations. Each frame records new segments when it is interacting with a user, analyses those segments, and sends them over the Internet to the other frame system. These segments are added to the segment library of the other frame. They also come with some ground truth; it is already known what kind of input elicited these responses in the first place, so their activation can be associated with similar input patterns. This design takes care of content management, and provides constant novelty to the system. Alternatively, the dual frame can be replaced by a simple interface to upload new videos. Figure 2 shows the overall design of the system, complete with the dual-frame mode. We now describe each module separately.

3.2 Offline segmentation module

The task of the offline segmentation module is to automatically generate meaningful and short segments from collected videos. These are stored with indicators of affective content and activity levels. The system is robust to segmentation mistakes, as the synthesis module eventually uses all footage material.

The segmentation module uses optical flow (detailed in Sect. 3.3.2) to find calm and active moments in the video. The magnitudes of the resulting optical flow is then summed up to produce a total amount of activity for each frame. Because we are interested in events that last for several seconds, this accumulated feature is then smoothed using a moving average window to get rid of noise and large fluctuations.

Extended periods of activity and calm are determined by the minima and maxima of the smoothed total optical flow function. If certain amount of time expires while the calculated optical flow stays below a threshold (τ_c) then the period is labeled as calm. Similarly, if the optical flow is above a threshold (τ_a) for some time, then the period is labeled as active. In the working prototype we used a 75-frame window for calm segments, and a 25-frame window for active segments. The thresholds are calculated as percentages of the average optical flow per frame, computed over the whole video that is being segmented. This way a single algorithm

works on videos with relatively different amounts of optical flow. The working prototype uses $\tau_c = 0.5$ and $\tau_a = 0.6$. Our experimental results (Sect. 4) have shown that this segmentation method creates segments similar to manual segmentation.

3.3 Real-time feature analysis module

The real-time analysis module is motivated by the need to analyze and characterize user behavior in order to provide an appropriate response at any given time. Keeping this in mind, this module can be considered as the data source of the system, as it receives signals from the user and processes them to determine affect- and activity-based content.

The feature analysis module maps the input to a pre-fixed feature space, where each point characterizes the user's affective state. The feature extraction considers arousal and facial expression based cues for this purpose. The feature space representation also allows the system to improve its responses over time by learning interaction patterns from each user.

We have focused on the following aspects of the user behavior to be able to model a significant and useful set of actions:

- Face: The location of the face allows the system to detect the presence of a user to initiate a session, and at the same time it offers information during the session such as movement with respect to camera's frame of reference, and proximity of the user.
- Motion: The activity level of the user is a lightweight feature that can be usefully employed to characterize actions. Taking the face area to be the center of a 3×3 grid, motion features are extracted from each of the nine grid cells. This way, all the motion is encoded with respect to the location of the face.
- Expression: Facial expression analysis is computationally costly. We detect the six prototypical facial expressions (joy, sadness, anger, fear, surprise, disgust) using a Bayesian approach based on motion of facial features. The expression analysis part of the system gives soft membership values for each category (including neutral expression) at 15 frames per second.

Taken together, these features form a reasonable and rich set of cues for implementing a responsive system. True to the nature of typical photo frames, the audio channel is removed from the videos. Consequently, is not considered as an input modality. While audio remains to be a potentially useful modality to add to the system, subjects interacting with the system remarked that the lack of audio appeared natural, and they did not seek interaction in this modality.

3.3.1 *Face analysis*

We rely on face information for both offline and online processing, and for normalization of features. Because of its proven reliability, the well-known Viola & Jones algorithm is used for face detection [33]. For better accuracy we have used the improved version as proposed by Lienthart and Maydt [18], which uses 45° rotated Haar-like features in addition to the original set of Haar features. The rotated Haar-like features increase the discrimination power of the framework, and a post optimization of the boosted classifiers provides for reduced false alarms. Only frontal face cascades were used to recognize nearly frontal faces.

The presence of a face in the field of view of the camera is the main cue used to arouse the system from its sleep mode. The features extracted from the face allows for quantification of changes in different aspects. For instance the change in the scale of the facial area is indicative of movement towards the frame or away from it.

The system proposed in [32] is adapted and implemented for facial expression analysis. In this approach, the face is tracked by a piecewise Bézier volume deformation (PBVD) tracker, based on the system developed by Tao and Huang [31]. A three dimensional facial wireframe model is used for tracking. The generic face model consists of 16 surface patches, and it is warped to fit the estimated facial feature points, which are simply estimated by their expected locations with respect to the detected face region boundary. These expected locations are learned on a separate training set of faces, and the locations of the wireframe nodes incorporate this knowledge.

The surface patches are embedded in Bézier volumes to generate a smooth and continuous model. A Bézier curve for $n+1$ control points can be written as:

$$x(u) = \sum_{i=0}^{n} b_i B_i^n(u) = \sum_{i=0}^{n} b_i \binom{n}{i} u^i (1-u)^{n-i}, \quad (1)$$

where the control points b_i and $u \in [0, 1]$ control model shape according to Bernstein polynomials, denoted with $B_i^n(u)$. The Bézier volume is an extension of the Bézier curve, and the displacement of the mesh nodes can be computed as $V = BD$, where B is again the mapping in terms of Bernstein polynomials, and D is a matrix whose columns are the control point displacement vectors of the Bézier volume.

After initialization of the facial model, head motion and facial surface deformations can be tracked. 2D image motions are estimated using template matching between frames at different resolutions. Previous frames are also used for

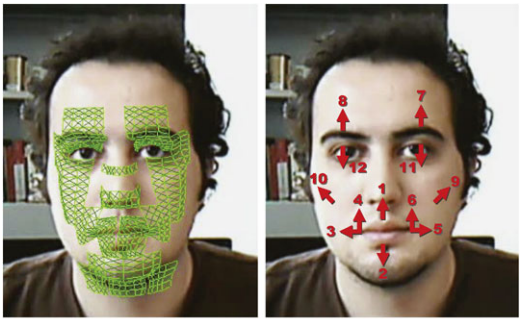

Fig. 3 (**a**) The Bézier volume model. (**b**) The motion units

Fig. 4 Example of the optical flow vectors obtained in a frame using the pyramidal implementation of the Lucas and Kanade algorithm. Optical flow vectors are represented as *red arrows* in the picture

better tracking. Estimated image motions are modeled as projections of true 3D motions. Therefore, 3D motion can be estimated using the 2D motions of several points on the mesh.

Expression classification is performed on a set of motion units, which indicate the movement of several mesh nodes on the Bézier volume with respect to the initial, neutral/frontal frame. 12 different motion units are defined as shown in Fig. 3. Unlike Ekman's Action Units [10], motion units represent not only the activation of a facial region, but also the direction and intensity of the motion. A naïve Bayes classifier is used to compute the posterior probabilities of seven basic expression categories (neutral, happiness, sadness, anger, fear, disgust, surprise).

3.3.2 Motion energy and activity levels

The real-time estimation of activity levels in a particular frame is computed by means of optical flow. The Shi-Tomasi corner detection algorithm is used to select the tracked points [30]. We use a pyramidal implementation of the Lucas-Kanade algorithm [19], developed by Jean-Yves Bouguet [4]. This method assumes that the flow is essentially constant in a local neighborhood of pixels under consideration, and solves the basic optical flow equations for all the pixels in that neighborhood under a least squares criterion. By combining information from several nearby pixels, the Lucas-Kanade method can often resolve the inherent ambiguity of the optical flow equation. It is also less sensitive to image noise compared to point-wise methods. Figure 4 shows a graphical example of the optical flow algorithm output for a particular frame.

The optical flow algorithm can be controlled in various ways depending on the type of segmentation that is desired. First there is the question of whether the optical flow should be calculated between two consequent frames, or a longer period, which might be necessary if the video footage is very static. Secondly, the number of tracked features can be adjusted: in videos with lots of small, uninteresting motion, the algorithm could be set to track only the most important features. Furthermore, the distance between two unique features can be scaled, and the maximum effect of a given feature can thereby be made greater or smaller. This provides robustness against outliers, so that a single large deviation in a given feature, which may be the result of an outlier or noise, does not overly affect the result. With all these options, the module can be used to segment a wide variety of video content.

3.4 System response module

The system response determines the quality of interaction. Following the automatic segmentation, the system has a number of short segments that can be played in any sequence. This forms a realistic baseline for the operation of the system, as current commercial systems for digital photo frames play the existing content sequentially (or randomly) for fixed amounts of time. The purpose of the system response module is to improve on this baseline by evaluating the user input in real-time, and by producing consistent and meaningful responses. We have selected finite state machines for the abstract representation of the system's operation, where input and output relations can be probabilistically indicated.

3.4.1 Simple prototype: peek-a-boo

As a simple first prototype, we have used a system that can be represented with a 2-state machine, where each state represents one possible output segment of the system. The input part implements the *peek-a-boo* game: a smile is elicited when the user shows his or her face, and a sad response is shown otherwise (see Fig. 5). We have used two expressive face action sequences ('Sad' and 'Smile', respectively) from the Cohn-Kanade database [17]. The advantages of using these sequences are that they are well-illuminated, normalized with respect to face location and size, and that the expressions start from a neutral face and evolve into the full manifestation of the expression.

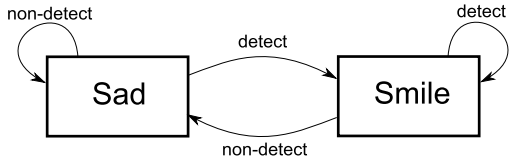

Fig. 5 Scheme of the two-state *peek-a-boo* machine that changes the response of the system according to the results of the Viola-Jones face detector

This prototype helped us to inspect the behavior of the system under very simple operating principles, and led to the following observations:

– **State transitions:** When we have a transition between segments that naturally follow each other, the state transition is very smooth, as expected. However, switching to a distant segment of the same video session, and even more prominently, switching to a segment of another video session can be sharp and unnatural. These transition artifacts should be eliminated using a smoothing or blur function during the transitions. In [23] a subspace method is proposed to control real-time motion of an object or a person in a video sequence. The low-dimensional manifold where the images are projected can be used to define a trajectory, which is then back-projected to the original image space for a smooth transition. While this method is promising for controlling transitions between segments, the subspace projection will not be very successful with the dynamic and changing backgrounds we deal with. We have experimented with different approaches to switch from one video segment to the next, including simple fade-in/fade-out, blending and morphing, for ensuring perceptually smooth transitions. The lack of manual intervention (for instance in specifying correct anchor points for morphing) makes elaborate schemes unwieldy. Subsequently, we use a much simpler scheme. If we have a transition from frame $\mathcal{F}_1$ to frame $\mathcal{F}_2$, we use an exponential forgetting function to synthesize transition frames, given by equation:

$$\mathcal{F}_3 = \alpha \cdot \mathcal{F}_1 + (1 - \alpha) \cdot \mathcal{F}_2 \tag{2}$$

where $\alpha \in (0; 1)$.

– **System responsivity:** Long and uninterrupted video segments reduce the perception of responsivity of the system. Experiments with the prototype showed that short segments of 3–5 seconds produce the best results, and they should be interrupted as soon as the user generates a new visual event.

– **Consistency:** To ensure consistency in the interface, we define several segments manually. These are fixed for the life of the system, and include the standby segment (e.g. baby sleeping), a wake-up segment, a neutral segment, and a session termination ("bye-bye") segment. The neutral segment is the first element in a pool of low-arousal segments selected from the segment library.

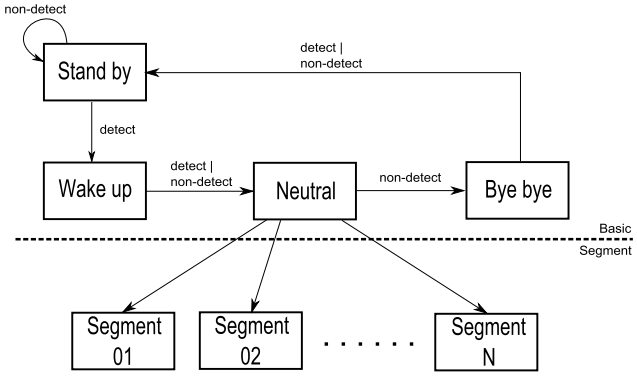

Fig. 6 Scheme of the finite state machine for system response. The two kinds of states are distinguished by the *line separator*

3.4.2 *Design of the finite state machine*

According to the observations we made following the prototype experiment, we have developed a more extensive finite state machine for the affect-responsive photo frame, depicted in Fig. 6, where we distinguish between *Basic states* and *Segment states*, respectively.

– **Basic states** are used to provide a general and consistent outlook to the system. When the system is not in use, a default state of low key activity is played in loop. In the prototypical use-case, this state would depict the baby asleep in the frame. When a person is present, the baby wakes up, and normal operation is resumed. When the interacting user is absent for a long period, the system returns to the sleep mode. The basic states make sure that this skeleton response is properly displayed. They are assigned manually, although their segmentation need not be manual. The transition from one basic state to another basic state depends solely on the input from the face detector.

– **Segment states** constitute the dynamic part of the finite state machine. Each segment state S_i is associated with one video segment V_i from the segment library, as well as an expected feature vector F_i that will guide the activation of the segment. The segment V_i is activated when the feature vector describing the user's activity is close to the expected feature vector F_i. The 'closeness' here is described statistically, by specifying a Gaussian distribution around each expected feature vector, and admitting activation if the feature vector computed from the user's activity is close to the mean by one standard deviation.

3.4.3 *Adaptation to the user*

The stored segments are presumably recorded under different conditions than the operating conditions of the system.

This raises an issue in matching the observed user affect to the stored segments. If the system is not trained for a specific person, there is a possibility that only a few segments will be activated during the lifetime of the system, and other response possibilities are left unexplored. Also, the response dictionary of the system is not static, and grows each time a new video is added to the system.

To solve this issue, action-response pairs are stored during interaction. The system then periodically updates its response function by analysing the existing action-response pairs. This serves a two-fold purpose. (1) The response of the system becomes consistent over a period of usage, in that the user becomes able to trigger a certain response by a certain action, and these triggering actions are suitably idiosyncratic. (2) The system induces certain actions, yet if the user is not able to produce the expected valence or arousal levels, the learning process will shift the required activity to an appropriate level suitable for the user's activity range. In other words, the user and the system simultaneously adapt to each other, and for each user, the final response pattern of the system will be different [25].

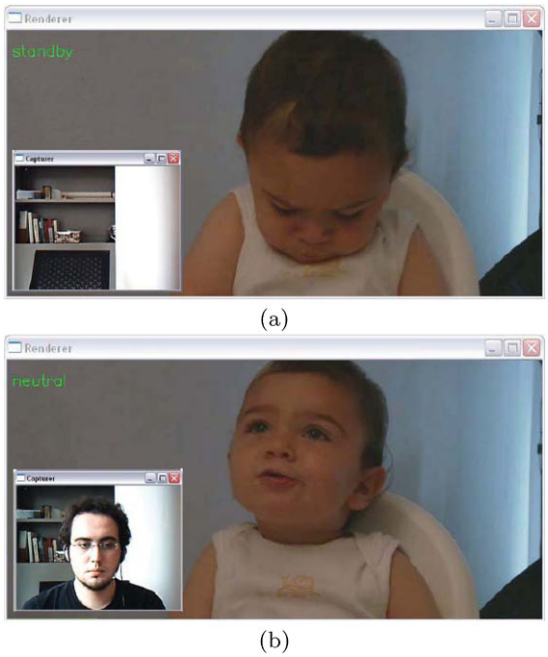

Fig. 7 The system during (**a**) stand-by mode and (**b**) interaction mode. The *lower left corner* shows the current camera input to the photo frame for diagnostic purposes

Let F^t denote the feature responses collected during a session of interaction with a user. At a specific moment T of the session, if there are k active segments, and one additional segment that the user seeks to activate at the moment of analysis, there will be $k+1$ feature distributions, represented as $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, with $i = 1 \ldots k+1$. Here, each segment is activated by a feature response that is close to its distribution, as measured by the Mahalanobis distance between $\boldsymbol{\mu}_i$ and F^t.

We can take into account the idiosyncratic variations that are conditioned to users by letting the system adapt its response to the user. The terms that determine the system response are F^t, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$. Since F^t is computed from the camera input recording user's behavior, the adaptation of the system is not concerned with it, but rather involves changing $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$. The idea is to update these variables for an improved modeling of user behavior. Figure 8 illustrates this idea on a toy example.

The procedure we use for improving the adaptation of the system is simple. At periodical intervals, the parameters of the system are updated as follows:

$$h_i(F^t) = \frac{p(F^t|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{j=1}^{k+1} p(F^t|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{3}$$

$$\boldsymbol{\mu}_i' = \alpha\boldsymbol{\mu}_i + (1-\alpha)\frac{\sum_{t=1}^{T} h_i(F^t)F^t}{\sum_{t=1}^{T} h_i(F^t)} \tag{4}$$

$$\boldsymbol{\Sigma}_i' = \alpha\boldsymbol{\Sigma}_i + (1-\alpha)\frac{\sum_{t=1}^{T} h_i(F^t)(F^t-\boldsymbol{\mu}_i)(F^t-\boldsymbol{\mu}_i)^T}{\sum_{t=1}^{T} h_i(F^t)}. \tag{5}$$

Here, $h_i(F^t)$ denotes the normalized membership probabilities of a particular set of features F^t for behavior segment i, $p(F^t|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is computed from the Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, and α is a control parameter. Small values of α will result in small adjustments in the systems behavior,

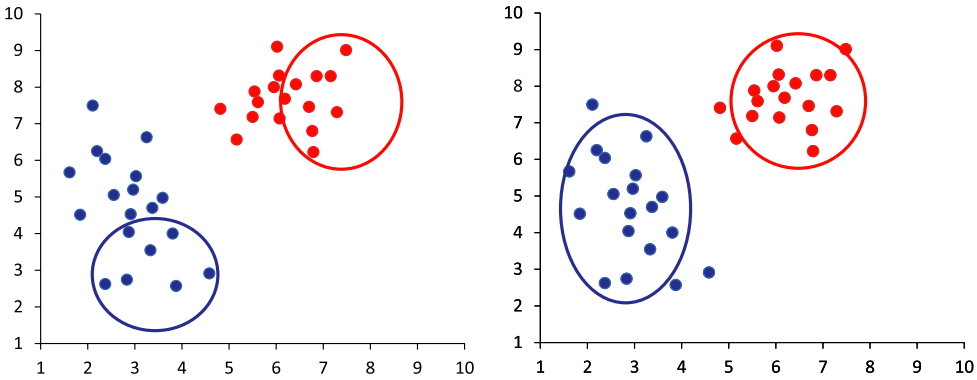

Fig. 8 (Color online) The user responses (each point is one frame) projected to two dimensions. The response thresholds of the system are shown as ellipses for two segments (*red* and *blue*), (**a**) before adaptation (**b**) after adaptation

Fig. 9 The manual segmentation of videos and the corresponding automatically determined segmentation

making it more responsive to the type of activities displayed by the user, as opposed to activities expected by the system. Large values of α may cause inconsistent behavior in the system, and abrupt changes in response.

3.5 Dual frame mode

The principle behind the dual-frame mode resembles that of the PhotoMirror appliance [21]. In PhotoMirror, a camera is hidden behind a mirror in a home setting, which can record segments of the inhabitants life, and play them back on the surface of the mirror (or another mirror). Similarly, the dual-frame mode of our system implies an asynchronous communication between two persons.

Consider our example scenario with the baby and the grandmother, and add to it a time-differential, where the baby lives in another continent. While the grandmother uses the interactive photo frame in her house, the system will record short segments of her activity (where the face detector is active) and create a segmented behavior library for the grandmother. These segments will be played on a second frame, placed in the baby's room. Through this symmetrical setup, we will also have a kind of action-response ground truth; the segments recorded from the grandmother's frame will be associated to particular segments of the baby. Then, these associations can be used to weakly guide the response patterns. Furthermore, each usage of the frame will send a sequence of new segments to the other frame, taking care of automatic content update for improved novelty.

4 System assessment

We have constructed a working prototype of the system, and we assessed it quantitatively, as well as through usability studies. In this section we describe the assessment of several modules separately.

4.1 Offline segmentation

To assess the offline segmentation, human operators have manually segmented a number of video sequences. The system segmentation is then contrasted with manual segmentation, provided by five different persons for each video sequence. During manual segmentation, segments were also assigned labels. We have not constrained these labels in any way; the only constraint was conciseness. The freely available ANVIL multimedia annotation tool[2] was used.

Figure 9 shows a video sequence being processed in the ANVIL tool. Five different segmentations are displayed as rows at the bottom of the video image. The temporal dimension is represented in each row in a left-to-right fashion. Labeled segments are represented as boxes, with the custom

[2] http://www.anvil-software.de/.

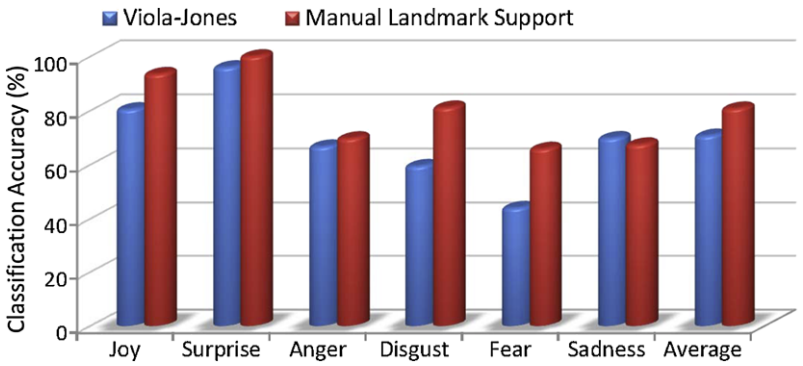

Fig. 10 Classification accuracy of the facial expression analysis module for different emotional expressions with and without manual landmark correction

label written inside. The smoothed optical flow graph that is appended to the figure (aligned in the temporal axis) is not a part of the annotation tool. It displays the result of automatic offline segmentation (as vertical bars) and the optical flow illustrates the 'reasoning' of the system in choosing these segments. The bars are elongated to intersect all five manual segmentations, so as to allow visual comparison. As it is evident from the figure, the most important segment boundaries (as evidenced by consensus among the taggers) is found by the automatic algorithm.

4.2 Facial expression analysis

We have assessed the accuracy of the facial expression module on the Cohn-Kanade AU-Coded Facial Expression Database [17]. In this database, there are approximately 500 image sequences from 100 subjects. These short videos each start with a neutral and frontal face display, and with little overall movement of the face display an emotional expression. Cohn-Kanade dataset has single action unit displays, action unit combinations, as well as six universal expressions, all annotated by experts.

Without any manual facial landmark correction, the facial expression recognition module provides 70.68 per cent average classification accuracy for six basic emotional expressions on this database. We have used 249 of the emotional expression sequences (46 joy, 49 surprise, 33 anger, 37 disgust, 41 fear, 43 sadness sequences) with three-fold cross validation to obtain the accuracy. To give a better notion of what the system can achieve with better alignment, we also report results with additional (manual) landmarking, which is a more typical setting in facial expression research. Warping the generic face model of the module into a more accurate face representation anchored by seven manually annotated facial feature points (outer eye corners, inner eye corners, nose tip, and mouth corners) by a Thin-Plate Spline algorithm [3] increases the average classification accuracy to 80.72 per cent. Figure 10 shows the classification accuracy for different emotional expressions, with and without manual landmark correction.

4.3 Real-time feature analysis

The computation burden of real-time feature computation is high, and this is a common problem we have noted in similar systems. The SEMAINE API [27], which is developed for building emotion-oriented systems, and which provides a rich set of tools for this purpose, was considered for usage in an early stage of development. Our initial experiments have shown that enabling the facial feature analysis module in this system required a lot of computational resources. The information provided by the API in this modality is quite detailed, which led us to pursue a computationally cheaper system that would nonetheless be useful in guiding the interaction.

The complete set of features derived in the system allow content matching based on valence and arousal. Figure 11 shows the projection of individual frames from the segment library along three features that represent this space. The horizontal axis shows valence through two facial expression features, 'happiness' on the positive direction and 'sadness' on the negative direction, respectively. The vertical axis shows arousal through average optical flow (static images do not adequately represent the movement in these frames).

4.4 Learning and adaptation

The benefit of learning is illustrated in Fig. 12(a). This figure shows the features extracted from the input during a single session, as well as features from the response library segment frames, projected onto a two dimensional space obtained with principal component analysis (PCA). Each point corresponds to a single frame, and marked with different symbols to indicate the closest segment from the segment library. The first principal direction here roughly corresponds to arousal, and loads most heavily onto mean optical flow magnitude around the face area.

The input from the user forms five clusters in Fig. 12(a). The neutral and static frames are clustered tightly on the left (shown in black), and the rest of the clusters can be easily discerned. In this example, there are six segments in

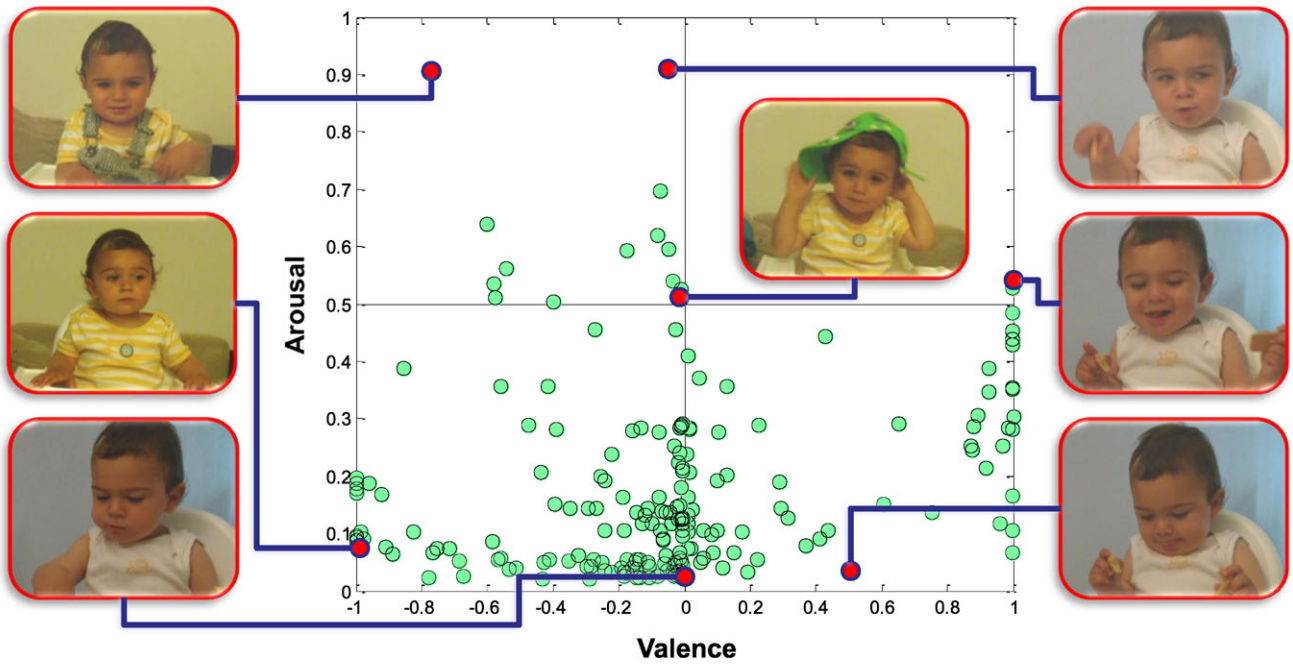

Fig. 11 Projection of individual frames on valence and arousal dimensions. See text for details

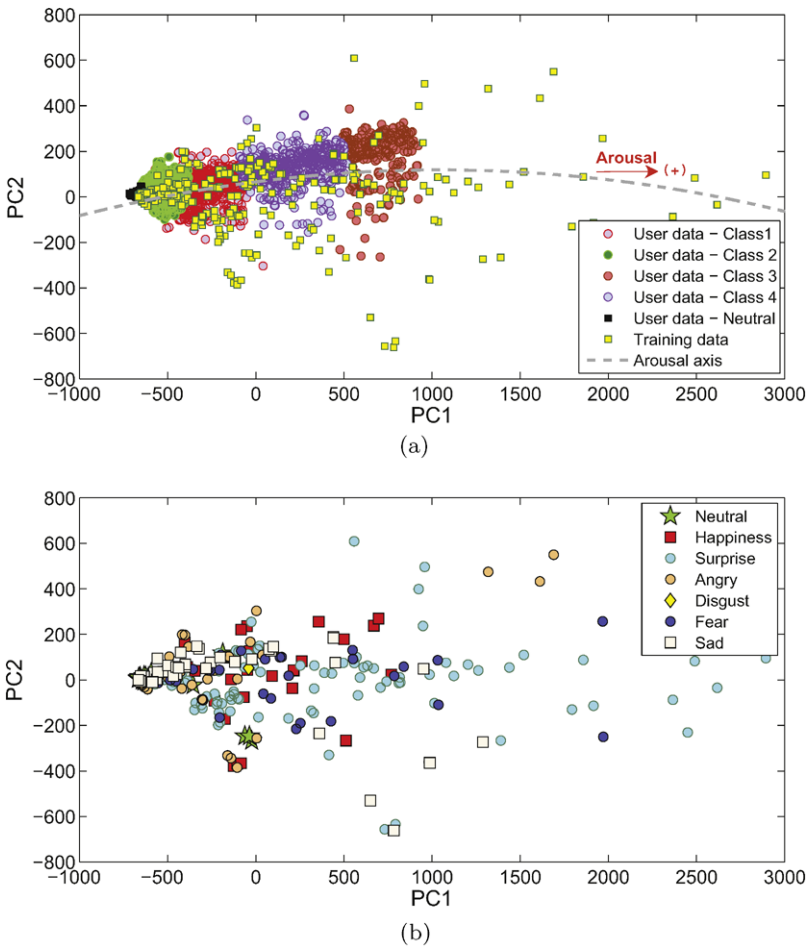

Fig. 12 (**a**) The projection of extracted input features from individual frames onto a two-dimensional PCA-space. The symbols denote matching segments from the response library for each frame. The first principal axis mostly reflects motion in the face area. (**b**) The dominant emotional expressions of individual training frames for the same projection. The surprise expression correlates most with the arousal, but there are no discernible clusters for facial expressions

Table 1 Experimental conditions

Condition	Response type
1	Random
2	Facial affect based
3	Facial affect and activity based
4	Manually selected (Wizard of Oz)

the response library, including the neutral segment. However, the user activity does not span the same space spanned by the principal components obtained from the response library. The sixth segment is marked by very high activity (in which the baby gesticulates wildly), and the user never shows arousal to this degree. However, the continuous adaptation eventually pulls the expected feature vector corresponding to this segment inside the region defined by the user's activity, and the segment is activated.

Figure 12 also shows the relative importance of motion in terms of explaining the variance. The second principal component predominantly loads onto motion features complementary to the first principal component. Features that relate to facial motion units do not automatically receive loadings that are consistent with a particular expression. Figure 12(b) shows the emotional expression labels for the points that belong to the training set. The surprise expression correlates most with the arousal, as expressed by the first principal component. Looking at the distribution of points with happiness and sadness expressions, it can be seen that the second principal component (PC2) weakly correlates with valence.

5 Usability study

5.1 Experimental methodology

We have conducted a usability study to inspect the reaction of subjects to the system. Ten subjects (age 23–33, eight males and two females) participated in the study. For each test session, the subjects received a brief description of the study, filled a consent form, and a pre-test questionnaire. This questionnaire included questions from the BIS/BAS scale [8], and the Big Five questionnaire [14, 15]. The BIS/BAS questionnaire had 20 questions answered on a 4-point scale to assess behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment. The subjects were not told that they were going to face different systems, and the interaction logic was not explained. They were just told that the system will show segments of a baby's video, and asked to see whether they liked the system.

Each test session consisted of eight short interaction rounds, where subjects faced four different systems in randomized order, twice. Table 1 lists the four different conditions used. For the last condition, a second monitor was made available to the experimenters, who manually selected responses for the system. After each round, the user completed the game experience questionnaire (GEQ) that measures different emotional responses to a game-like experience [13]. This quesionnaire includes competence, immersion, flow, tension, negative affect, positive affect and challenge related questions, answered on a 5-point scale.

The eight rounds were divided into two parts, each consisting of the four different systems in randomized order. This allowed us to measure how learning affected the perceived usability of the system. For example, the third condition (facial affect and activity based response) scored less in empathy measure than the Wizard of Oz system during the first set of four tests, but on the second round of experiments there was no statistically significant difference between the conditions. Figure 13 summarizes the GEQ results collected from the two rounds for all four conditions. During verbal debriefing the subjects reported having a positive experience with the system. Occasionally, the lack of immediate response caused frustration, but most of the users realized that the systems they were facing were different. This was also apparent in their responses, where the Wizard of Oz condition was deemed significantly more responsive during the interaction.

5.2 Results

The data were analyzed according to the Linear Mixed Models procedure in SPSS with restricted maximum likelihood estimation and a first-order autoregressive covariance structure for the residuals. User ID was specified as the subject variable, and the round id (the order in which the user received the randomized conditions) was specified as the repeated variable. Condition was selected as a factor, and a fixed-effects model that included the condition as the main effect was specified. The analysis was conducted for the first and second batches separately, to take into account the effect of learning.

Most effects failed to reach statistical significance, which may be due to the relatively restricted number of users. The Wizard of Oz system (condition 4) was perceived to be different the other three, and most of the significant differences were between condition 4 and the others. During the first four rounds, the challenge score for condition 3 was significantly lower than for the condition 4 (this and all subsequent significance results are reported at $p < 0.05$). Also, the empathy and behavioral involvement scores for condition 3 were lower than condition 4. During the last four rounds, the immersion score for condition 3 was lower than for condition 4, but condition 2 got a higher score for tension and for negative affect than condition 4. We should note that the response time of the system is shorter for condition 4, as

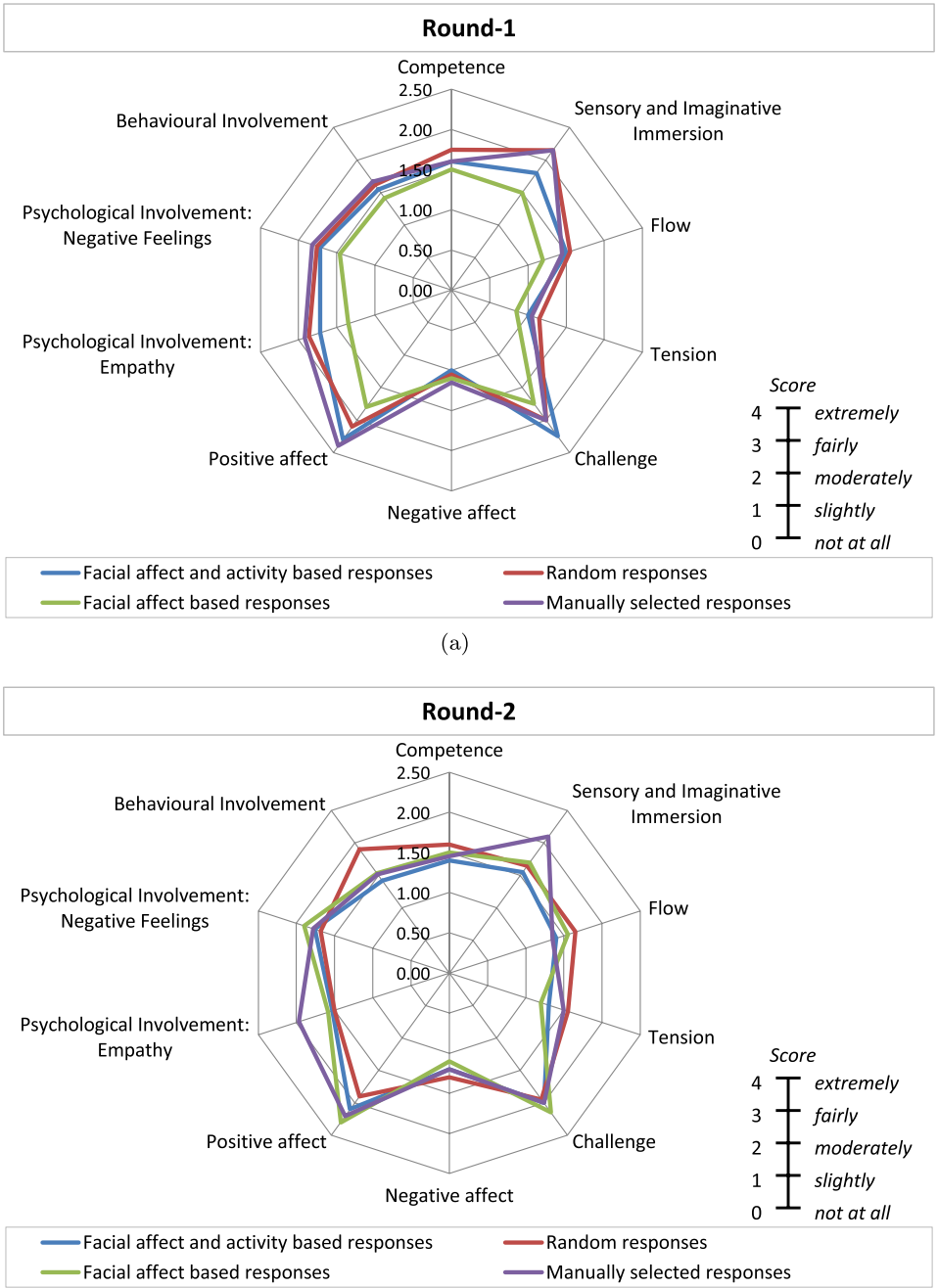

Fig. 13 The results of the GEQ study for (**a**) the first round and (**b**) the second round of the experiment. The overall responses to the system are good in all conditions, but some minor changes are observed between rounds. For instance the second round shows reduced empathy and immersion for the random condition

it involves immediate reactions of the human operator driving the system, which may have an effect on the significance of differences.

The random condition (condition 1) did not seem to generate strong response to any particular direction. The only significant finding was that the random condition was considered as the most challenging during the first four cases. On the other hand, during the last four rounds, random condition scored least on positive affect, empathy, negative feelings, behavioral involvement and flow, but these results did not reach statistical significance.

The results seem to indicate that especially the full system of facial affect and activity based responses (condition 3) took some time for users to adapt to: during the first part of the experiment it was not considered challenging, but neither was it considered as behaviorally involving and empathetic as the Wizard of Oz condition, but these effects disappeared on the second half of the experiment. It is also interesting that after learning, the system with only facial affect detection scored higher on tension and negative affect. One reason for this may be that the activity cues, when added, claim a higher contribution in the overall

variance, and mask expression-related effects to some degree.

The same analysis was repeated while controlling for the scores users received from the pre-test questionnaires, but this did not have much effect on the results. Only the tension level between condition 3 and 4 was statistically significant when controlling for neuroticism score from the Big Five questionnaire. We did not see any important effects of profiling.

6 Conclusions

We have developed a working prototype for an affect-responsive photo frame application. True to the nature of a photo frame, which can be contemplated in silence, the auditory modality was not included in the system design. Nonetheless, voice and speech modalities can be added to the system following the same principles, at the cost of higher computational complexity. We have proposed a dual-frame mode to solve the content acquisition and maintenance issues. The most important aspect of our work that separates it from similar digital constructions in the literature is that we do not assume carefully recorded and annotated response patterns, but process the input and the output of the system automatically.

Our experiments have shown that the proposed system is interesting and engaging. A formal usability study with ten users revealed strengths and limitations of the system. As expected, automatic content management results in less meaningful segments than a hand-crafted set of responses. However, the users of the system do not expect a complete and strong interaction from the system in practice, particularly in the absence of sound-based interaction, and prefer the design to a standard digital photo frame which rotates its material without interaction.

The ordinary photograph implies a spectator, who captures a moment in the act of taking the photograph, and defines it as a special moment by virtue of the selection and the subsequent production of the photograph. The latter is confirmed by the consumer of the photograph, who chooses to make it a part of his or her daily life, and gives it a space to occupy. The interactive photograph we described in this work runs the risk of removing the magic of selection from its definition by an automatic and indiscriminate production of content. On the other hand, it shifts some of the magic of the spectator to the actual consumer, by turning the immediate process of consumption into an authorship that defines the content—within certain limits—on the fly. The interactive photograph becomes thus a *focal thing*, demanding attention and engagement [5].

References

1. Agamanolis S (2006) Beyond communication: human connectedness as a research agenda. In: Networked neighbourhoods, pp 307–344
2. Bailenson J, Pontikakis E, Mauss I, Gross J, Jabon M, Hutcherson C, Nass C, John O (2008) Real-time classification of evoked emotions using facial feature tracking and physiological responses. Int J Hum-Comput Stud 66(5):303–317
3. Bookstein F (1989) Principal warps: thin-plate splines and the decomposition of deformations. IEEE Trans Pattern Anal Mach Intell 11(6):567–585
4. Bouguet J (1999) Pyramidal implementation of the Lucas Kanade feature tracker description of the algorithm. Intel Corporation, Microprocessor Research Labs, OpenCV Documents 3
5. Buchanan R, Margolin V (1995) Discovering design: explorations in design studies. University of Chicago Press, Chicago
6. Bui T, Zwiers J, Poel M, Nijholt A (2006) Toward affective dialogue modeling using partially observable Markov decision processes. In: Proc workshop emotion and computing, 29th annual German conf on artificial intelligence, pp 47–50
7. Cao J, Wang H, Hu P, Miao J (2008) PAD model based facial expression analysis. In: Advances in visual computing, pp 450–459
8. Carver C, White T (1994) Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: the BIS/BAS scales. J Pers Soc Psychol 67(2):319–333
9. Dibeklioğlu H, Kosunen I, Ortega M, Salah A, Zuzánek P (2010) An affect-responsive photo frame. In: Salah A, Gevers T (eds) Proc eNTERFACE, pp 58–68
10. Ekman P, Friesen W, Hager J (1978) Facial action coding system. Consulting Psychologists Press, Palo Alto
11. Gilroy S, Cavazza M, Chaignon R, Mäkelä S, Niranen M, André E, Vogt T, Urbain J, Seichter H, Billinghurst M et al (2008) An affective model of user experience for interactive art. In: Proc int conf on advances in computer entertainment technology. ACM, New York, pp 107–110
12. Gunes H, Piccardi M (2009) Automatic temporal segment detection and affect recognition from face and body display. IEEE Trans Syst Man Cybern, Part B, Cybern 39(1):64–84
13. Ijsselsteijn W, de Kort Y, Poels K (in preparation) The game experience questionnaire: development of a self-report measure to assess the psychological impact of digital games. Manuscript
14. John O, Donahue E, Kentle R (1991) The Big Five Inventory Versions 4a and 54. Berkeley: University of California, Berkeley, Institute of Personality and Social Research
15. John O, Naumann L, Soto C (2008) The Big Five trait taxonomy: discovery, measurement, and theoretical issues. In: Handbook of personality: theory and research, pp 114–158
16. Kaliouby R, Robinson P (2005) Real-time inference of complex mental states from facial expressions and head gestures. In: Real-time vision for human-computer interaction, pp 181–200
17. Kanade T, Cohn J, Tian Y (2000) Comprehensive database for facial expression analysis. In: Proc AFGR
18. Lienhart R, Maydt J (2002) An extended set of haarlike features for rapid object detection. In: IEEE international conference on image processing, vol 1, pp 900–903
19. Lucas BD, Kanade T (1981) An iterative image registration technique with an application to stereo vision. In: IJCAI, pp 674–679

20. Mancas M, Chessini R, Hidot S, Machy C, Ben Madhkour R, Ravet T (2009) Morface: face morphing. Q Prog Sci Rep Numediart Res Program 2(2):33–39
21. Markopoulos P, Bongers B, Alphen E, Dekker J, Dijk W, Messemaker S, Poppel J, Vlist B, Volman D, Wanrooij G (2006) The PhotoMirror appliance: affective awareness in the hallway. Pers Ubiquitous Comput 10(2):128–135
22. Mehrabian A (1996) Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. Curr Psychol 14(4):261–292
23. Okwechime D, Ong E, Bowden R (2009) Real-time motion control using pose space probability density estimation. In: Proc int workshop on human-computer interaction
24. Russell J (1980) A circumplex model of affect. J Pers Soc Psychol 39(6):1161–1178
25. Salah A, Schouten B (2009) Semiosis and the relevance of context for the AmI environment. In: Proc European conf on computing and philosophy
26. Schröder M, Bevacqua E, Eyben F, Gunes H, Heylen D, ter Maat M, Pammi S, Pantic M, Pelachaud C, Schuller B et al (2009) A demonstration of audiovisual sensitive artificial listeners. In: Proc int conf on affective computing & intelligent interaction
27. Schröder M (2010) The SEMAINE API: towards a standards-based framework for building emotion-oriented systems. In: Advances in human-computer interaction
28. Sebe N, Lew M, Sun Y, Cohen I, Gevers T, Huang T (2007) Authentic facial expression analysis. Image Vis Comput 25(12):1856–1863
29. Shan C, Gong S, McOwan P (2007) Beyond facial expressions: learning human emotion from body gestures. In: Proc of the British machine vision conference
30. Shi J, Tomasi C (1994) Good features to track. In: Proc computer vision and pattern recognition. IEEE, New York, pp 593–600
31. Tao H, Huang T (1998) Connected vibrations: a modal analysis approach for non-rigid motion tracking. In: Proc computer vision and pattern recognition, pp 735–740
32. Valenti R, Sebe N, Gevers T (2007) Facial expression recognition: a fully integrated approach. In: Proc 14th int conf of image analysis and processing-workshops. IEEE Computer Society, New York, pp 125–130
33. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Proc computer vision and pattern recognition, vol 1, pp 511–518
34. Zeng Z, Pantic M, Roisman G, Huang T (2009) A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell 31(1):39–58