Available online at www.sciencedirect.com
ScienceDirect
Interacting with Computers 20 (2007) 17–28
Interacting with Computers
www.elsevier.com/locate/intcom

EmoPlayer: A media player for video clips with affective annotations

Ling Chen [a,b,*], Gen-Cai Chen [a], Cheng-Zhe Xu [a], Jack March [b], Steve Benford [b]

[a] College of Computer Science, Zhejiang University, Hangzhou 310027, PR China
[b] School of Computer Science, The University of Nottingham, Nottingham NG8 1BB, UK

Received 27 November 2006; received in revised form 7 June 2007; accepted 12 June 2007
Available online 8 November 2007

Abstract

The development of multimedia annotation technique provides the possibility to redesign the interfaces of widely used media players, and EmoPlayer is such a media player that can be used to play video clips with affective annotations. A user can select a character in a video clip and view the distribution of his/her emotions along the video timeline through a colour bar based interface. Two experiments were conducted to evaluate the efficiency of affective annotation. The results of these experiments indicate that affective annotation is effective in both improving the speed of locating a specific scene within a video clip and helping comprehend a video clip in a limited viewing time period. Based on the analysis of recorded operations of participants, the strategies employed by participants and the factors that might influence the utilization of affective annotation are also highlighted.

Keywords: Affective computing; Multimedia annotation; Media player; User interface

1. Introduction

With the rapid development of computing and network technologies, multimedia and networked multimedia technologies have proliferated quickly in recent years. Nowadays, more and more computer users utilize various media players to watch movies and TV programmes stored on their personal computers or in real-time over the Internet. In addition to watching a video clip from the beginning to the end, users sometimes just view the scenes they are interested in or browse the video clip to get a summary. Although this is not what the director wants, users often employ these methods when they do not have enough time, or just want to determine whether spending further time on the video clip is worthwhile. On a media player, the slide bar, designed for users to seek a specific scene within the playing clip, is one of the most important interface components. Currently most users utilize the slide bar to perform just this function. Since the slide bar in a traditional media player only provides limited information, such as the total duration of the playing clip and the position of current scene, it is difficult for users to rapidly locate a specific scene within a video clip or comprehend a video clip in a limited viewing time period. When watching streaming media, it is even more difficult, because a drag operation will cause a short term pause for the media to be buffered.

In order to help users locate a specific scene rapidly, most DVDs provide index menus. A menu item in an index menu should be related to a specific scene, and the representative picture, short video clip, or brief description of the scene might be included in the menu item. If users select an item, they are navigated to the corresponding scene. In this area, Norman (2001) studied the usability problems of index menu design and Költringer et al. (2005) conducted some usability tests and identified several usability issues. Additionally, storyboards (Christel and Warmack, 2001), which are common to most digital video libraries, can provide scene navigation by presenting an ordered set of representative thumbnail images simultaneously on a computer screen. From a technical perspective, both index

* Corresponding author. Address: College of Computer Science, Zhejiang University, Hangzhou 310027, PR China. Tel.: +86 571 85975656.
E-mail addresses: lingchen@cs.zju.edu.cn (L. Chen), chengc@cs.zju.edu.cn (G.-C. Chen), jason_azhe@163.com (C.-Z. Xu), jzm@cs.nott.ac.uk (J. March), sdb@cs.nott.ac.uk (S. Benford).

0953-5438/$ - see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.intcom.2007.06.003

menus and storyboards are based on multimedia annotation and hyperlink, with multimedia annotation assuming a more important role.

Since multimedia content is becoming a prevalent information source and the volume of such content is continually growing, annotating multimedia content with semantic information, such as scene/segment structures and metadata about visual/auditory objects etc., is necessary for advanced multimedia content services, such as multimedia summarization and multimedia retrieval etc. In recent years, multimedia annotation has been widely studied. Related work of traditional multimedia annotating methods and tools includes: Costa et al. (2002) implemented VAnnotator which allowed users to annotate audio-visual content through a timeline model; Abowd et al. (2003) provided a method to archive and annotate large collections of informal family movies; and Bulterman (2003) discussed applying facilities in SMIL 2.0 to solve the problem of annotating multimedia presentations, and proposed a method to represent annotations within a common document architecture. Related work of innovative multimedia annotating methods and systems includes: Qian and Feijs (2004) integrated photo annotating tasks into instant messaging applications; Fogli et al. (2004) proposed an electronic annotation method for two way exchange of ideas among humans pursuing a common goal; Bottoni et al. (2004) presented a multimedia digital annotation system named MADCOW which allows Internet users to create, read, modify, save, search, and filter public or private annotations related to the contents of web pages; and Carter et al. (2004) designed methods for users to post and acquire digital information to and from public digital displays, and to modify and annotate previously posted content to create publicly observable threads. Additionally, some support techniques for multimedia annotation have also been studied, for example, Ramos and Balakrishnan (2003) explored a variety of interaction and visualization techniques for fluid navigation, segmentation, linking, and annotation of digital videos.

Human emotion, in its most general definition, is a neural impulse that moves an organism to action, prompting automatic reactive behavior that has been adapted through evolution as a survival mechanism to meet a survival need. Davidoff (1980) defines emotion as a feeling that is expressed through physiological functions such as facial expressions, faster heartbeat, and behaviors such as aggression, crying, or covering the face with hands. Since emotion is more explicit than human feeling, it can be recognized through technical methods, for example facial expression recognition from the computer vision community (Wang and Ahuja, 2003), speech emotion recognition from the speech processing community (Kwon et al., 2003), and textual affect sensing from the natural language processing community (Liu et al., 2003a). Based on the advances in multimedia content analysis and the fact that emotion is a basic aspect of human functioning, researchers have attempted to realize more semantic annotation, resulting in the specific research focus of affective annotation (Xia et al., 2005) coming into being. The results of affective annotation could be used to perform information indexing and retrieval, for example, the affective annotations of films were utilized to query specific scenes in films (Chan and Jones, 2005).

Emotions are of increasing interest to the HCI community (Cockton, 2002). Within the last decade, emotion research in HCI grew from an eccentric hobby of some visionary scientists to a widely accepted field of research. Picard and Klein (2002) presented prototypes of interactive computer systems that can detect and label aspects of human emotional expression, and that respond users experiencing frustration and other negative emotion with emotionally supportive interactions, demonstrating components of human skills such as active listening, empathy, and sympathy. Additionally, user needs of such systems were defined in a broader perspective. Partala and Surakka (2004) investigated the psychophysiological effects of positive and negative affective interventions in human–computer interaction during and after the interventions, the results suggested that both types of affective intervention had beneficial effects over ignoring the user. Peter and Herbon (2006) proposed a suitable approach to structure and represent emotions for use in digital systems, which is an applicable model of emotions suitable for designing emotion-aware systems or performing HCI-related emotion studies. Besides designing interactions upon the emotional states of users, HCI community also works on utilizing the results of affective annotation. For example, Liu et al. (2003b) presented an interface to visualize the emotional states extracted form sentences in a text document, with different colours to represent different emotions and a colour bar to present the emotional distribution of the document. Evaluation results indicated that with affective annotation and the colour bar based interface, within-document information foraging performance was improved.

In most movies or TV programmes, different events or scenes might be associated with the emotions displayed by the actors and actresses playing each of the characters. For example, when a person is given a promotion or wins a prize, this is most possibly accompanied with happy emotion. The emergence of affective annotation gives a possibility to further improve the efficiency of media players while performing tasks like locating a specific scene. Inspired by this, we developed a media player named EmoPlayer which can play video clips with affective annotations. By selecting a character in a video clip, users can view that character's emotional distribution along timeline through the colour bar based interface of EmoPlayer. In addition, two experiments were conducted to evaluate the efficiency of affective annotation. The first experiment studied the effects of affective annotation on locating a specific scene within a video clip. The second experiment studied the effects of affective annotation on the comprehension of a video clip in a limited viewing time period and the user satisfaction with respect to affective annotation. To our knowledge, EmoPlayer is the first media player designed for video clips with affective annotations.

Affective annotation, especially automatically annotating a video clip through facial expression or speech recognition, itself is quite difficult and needs further research. The primary aim of this paper is to discuss design considerations for creating a media player exploiting the results of affective annotation and study the efficiency of affective annotation while performing given tasks. The remainder of the paper is organized as follows: Section 2 describes the design and implementation of EmoPlayer; Section 3 discusses the designs, procedures, and results of the experiments conducted with EmoPlayer; Section 4 provides some conclusions and suggests areas of future work given this research.

2. EmoPlayer

2.1. Emotion categorization

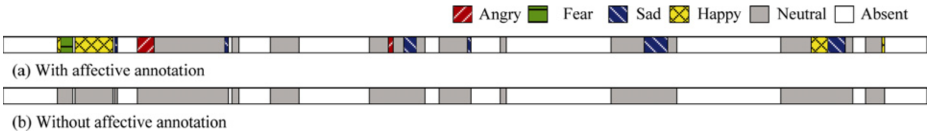
Humans have a large vocabulary with which they can describe emotion, such as "happy", "angry", and "sad" etc. While metrics concerning human emotion can be instrumentally measured from facial expression, speech, and heart rate etc., people still describe emotion and interpret it in impressionistic terms which can be highly subjective (Cowie et al., 2001). Since there does not exist a single universally accepted human emotion categorization (Russell, 1991), researchers have proposed different categorizations for different purposes, for example Ekman (1982), Craggs and Wood (2004), Murray and Arnott (1993) proposed categorizations for facial expression, the linguistic content of dialogue, and voice, respectively. In order to improve the accuracy of emotion classification, our study employs a simple existing categorization, in which human emotion is categorized into five types: happy, sad, fearful, angry, and neutral. Thompson et al. (2004), Williams et al. (2005) employed an analogous categorization in their studies.

2.2. Colour bar based interface

How to show a character's emotional distribution along timeline through the interface was a key problem in the design of EmoPlayer. The slide bar for seeking a specific scene within the playing video clip intuitively shows the position of that scene as well as potentially showing time related information. EmoPlayer utilizes this characteristic of the slide bar and adds an additional bar, which is positioned above and aligned to the slide bar, to show a character's emotional distribution along the timeline.

How to show different types of emotion in the added bar was another key problem in the design of EmoPlayer. Inspired by the research of Liu et al. (2003b), EmoPlayer employs different colours to show different types of emotion, and analogous design was used to visualize the embodying information of a playing video clip (Barbieri et al., 2001). Although there are many studies about the psychology of colour, for example the effects of colour on emotions (Valdez and Mehrabian, 1994), the results of these studies vary largely over different situations. EmoPlayer employs a mapping between different emotions and colours, in which yellow, red, blue, green, and grey indicate happy, angry, sad, fearful, and neutral emotional states, respectively. This mapping is partially consistent with generally accepted findings of colour psychology, for example yellow, red, and blue are used to indicate happy, angry, and sad emotional states (Tokitomo et al., 2004), and attempts to maximize the distances between different colours, for example primitive colour green is used to indicate fearful.

Fig. 1(a) illustrates the added colour bar which shows a character's emotional distribution along the timeline with the colour white indicating the absence of that particular character. This version of the added colour bar is referred to as *with affective annotation*. Compared with the interfaces of traditional media players, this interface can provide information on when a specific character is present in the playing clip and the emotional state of that character whenever he/she is present. The first step in employing this interface is for the user to select a character to be displayed. The user can then utilize the slide bar to jump to a scene in which that character is present and given his/her specific emotional state. Since the main aim of this study is to investigate the effects of showing emotion related information on human performance, not the effects of showing presence related information, EmoPlayer also employs another version of the added colour bar, which only provides an indication of when a character is present. Fig. 1(b) illustrates this version of the added colour bar, in which the colours grey and white are employed to indicate the presence and absence of a character, respectively. Since emotional state information is not conveyed, this version of the added colour bar is referred to as *without affective annotation*. Note that for the purpose of clarity and



(a) With affective annotation

(b) Without affective annotation

Fig. 1. Different kinds of the added colour bar. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

printing in black and white, all illustrated colour bars are drawn as vector graphs, and the implementation of the interface can be seen in Fig. 2.

2.3. Implementation

EmoPlayer is implemented with VC++ and DirectShow is employed to provide multimedia related functions. Fig. 2 shows a screen snapshot of EmoPlayer which can provide most functions of a traditional media player, for example play, stop, and pause etc. What make this interface different are the following additional features: (i) an added colour bar over the slide bar to show the emotional distribution of a character (or just the presence of a character when the bar is without affective annotation); (ii) a combo box which can be used to switch between different characters; and (iii) a legend which indicates the relationship between different emotion types (or presence states when the bar is without affective annotation) and colours. The components named *event cue* and *process* are added for the purpose of displaying corresponding information in the experiments.

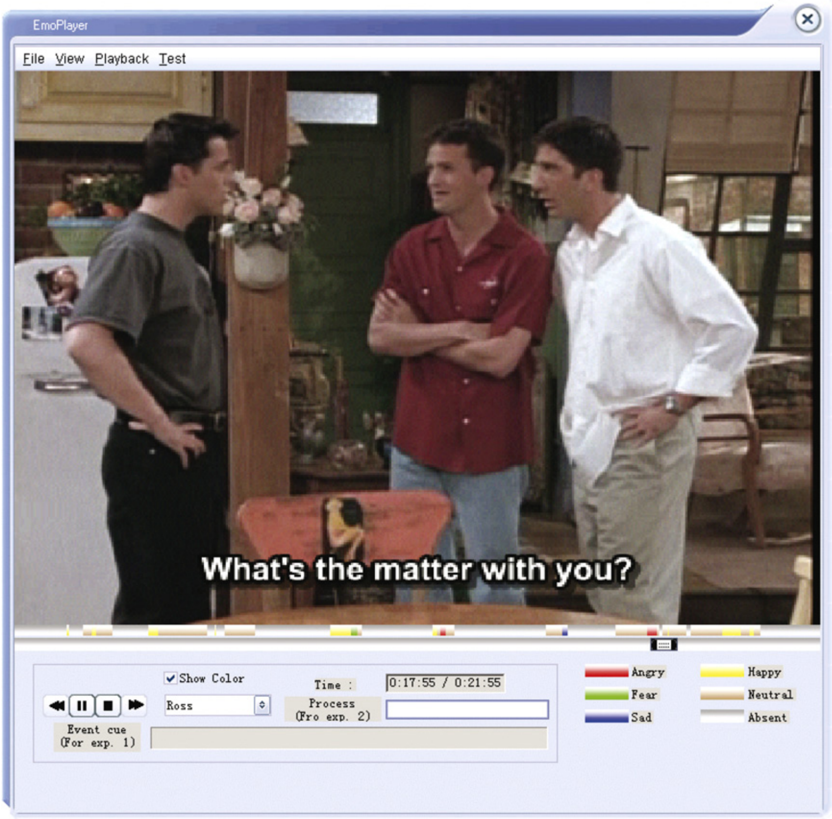
Since XML is suitable for affective annotation (Xia et al., 2005), it is employed by EmoPlayer with the affective annotation of a particular video clip being stored in an XML file. The XML file should have the same file name, and be stored in the same directory, as its corresponding video clip. When opening a video clip in EmoPlayer, its corresponding XML file is also opened and the affective annotation is read into memory. Subsequently, the colour bar of the first character is generated and displayed. Fig. 3 is a sample XML file which stores the affective annotation of a video clip containing three characters named *Ross*, *Rachel*, and *Chandler*.

3. Experiments

Two experiments were conducted to evaluate the efficiency of affective annotation. The first experiment studied the effects of affective annotation on locating a specific scene within a video clip. The second experiment studied the effects of affective annotation on the comprehension of a video clip in a limited viewing time period and the user satisfaction with respect to affective annotation.

The video clips used in the experiments met the following criteria: the characters were rich in the different emotion types; and the characters expressed these emotions in a straightforward way. According to the criteria, three video clips were selected for the experiments. All these video clips were dubbed in Chinese, and the durations were 22, 45, and 36 min, respectively. The first video clip was used for participants to practice before the formal experiments while the other two video clips were used for Experiment 1 and 2, respectively.

In order to annotate the selected video clips, a three-person group was set up and the members of the group were



☑ Show Color

Ross

Time : 0:17:55 / 0:21:55

Fig. 2. A screen snapshot of EmoPlayer.

```
<Index>
  <RoleList>
    <Block Name="Ross"/>
    <Block Name="Rachel"/>
    <Block Name="Chandler"/>
  </RoleList>
  <Ross>
    ......
    <Block Time="0:1:29-0:1:33" Duration="4" Emotion="Happy"/>
    <Block Time="0:1:34-0:2:47" Duration="51" Emotion="Neutral"/>
    ......
  </Ross>
  <Rachel>
    ......
    <Block Time="0:1:23-0:1:27" Duration="4" Emotion="Happy"/>
    <Block Time="0:1:28-0:1:46" Duration="18" Emotion="Fear"/>
    ......
  </Rachel>
  <Chandler>
    ......
    <Block Time="0:16:36-0:18:3" Duration="87" Emotion="Anger"/>
    ......
  </Chandler>
</Index>
```

Fig. 3. A sample XML file storing the affective annotation of a video clip.

familiar with emotion expression. The task of the group was to determine the emotional distributions of the main characters in the video clips. For each video clip, they first annotated the clip separately, and then they discussed where their results differed. Since the emotions in the selected video clips were expressed in a straightforward manner, there were few differences between the results of different group members and a group consensus was easily reached. The affective annotations of the selected video clips were subsequently stored as three XML files. In the video clips for Experiment 1 and 2, the numbers of happy, sad, fearful, angry, and neutral emotional states were 29, 4, 0, 15, 25 and 31, 10, 5, 7, 47, respectively. The numbers of differences between the results of group members and the final result were 4, 2, 3 and 4, 3, 4, respectively.

In the experiments, EmoPlayer was run on an IBM T42 laptop which had a 14-inch LCD screen with a resolution of 1024 × 768 pixels. Since participants used a pointing device to interact with EmoPlayer and not all of them were familiar with touch pad and pointing stick, a Microsoft IntelliMouse Optical mouse was employed as input device, and all participants used the external mouse to interact with EmoPlayer.

In order to minimize environmental disturbances, the experiments were conducted in an empty office. During the formal experiments, the operations of all participants were recorded through a screen capture tool. These records were used in the post experiment analysis to study the strategies employed by participants and the factors that might influence the utilization of affective annotation.

3.1. Experiment 1

3.1.1. Design and hsypotheses

The aim of the first experiment was to study the effects of affective annotation on locating a specific scene within a video clip. The task for participants in the first experiment was to locate a target scene as soon as possible after the description of the scene was given. Similar task was employed to investigate the searching and browsing behaviors of users employing a traditional VCR-like control set (Crockford and Agius, 2006).

In this experiment, a 2 (with affective annotation or not) × 2 (participants were familiar with the video clip or not) between group design was employed. Five target scenes were selected from the second movie clip before the experiment, and all of them had explicit relationships with different emotions, for example, “the scene in which the lovers start to dance” was the description of a target scene and this scene intuitively related to the happy emotion of evolved characters. Of the five target scenes, three had explicit relationships with happiness, and the other two had explicit relationships with angry and sad emotional states, respectively. All five target scenes were almost equally distributed in the video clip, and the order in which these scene descriptions were given was not necessarily chronological. For example, if a target scene was located the next target scene might be before or after the newly located scene. In order to ensure there were sufficient distances between consecutive target scenes, the order of these scenes was constant across participants and experimental conditions. The target scenes appeared within the timeline in the following order: 1, 3, 5, 2, and 4, with these scenes to be located in an ascending order, i.e. target scene 1, 2, 3, 4, and 5. In the experiment, the performance metric was *Task Completion Time* (TCT), which was the time duration to locate all five target scenes. Because TCT was not the time duration to locate a specific target scene, fixing the order of target scenes would not affect the results of the experiment. Since it was quite difficult for participants to locate a specific frame within the playing clip, a time range of approximately seven seconds was used for each target scene. If the slide was dragged to a position within this range, a window popped up with a message showing that the target scene had been successfully located.

Experiment 1 had the following hypotheses:

Hypothesis 1. The participants who employed the colour bar with affective annotation would locate target scenes more quickly than the participants whose colour bar was without affective annotation.

Hypothesis 2. The participants who were familiar with the video clip would locate target scenes more quickly than the ones who were not.

3.1.2. Participants

Twenty-four participants, both undergraduate and postgraduate students, were recruited from the campus of Zhejiang University. There were 6 females and 18 males and their ages ranged from 24 to 28 (Mean = 25.375, Std. Deviation = 1.313). All of the participants were familiar with traditional media players, and Chinese was their native language. None of them had watched the video clip used in Experiment 1 before. All participants were randomly divided into four equal groups, corresponding to the four experimental conditions.

3.1.3. Procedure

Before the formal experiment, the interface and operations of EmoPlayer were introduced to the participants, and they were given 30 min to familiarise themselves with EmoPlayer. The first video clip was used in this stage. The first 10 min was used for participants to learn the interface of EmoPlayer, and then they did video browsing practice with and without the help of affective annotation (10 min each). Before the start of the formal experiment, the task in the experiment was described to the participants. The names and photographs of the main characters in the video clip were also shown to each participant. In accordance with the design of Experiment 1 (some participants being familiar with the video clip), the participants in these groups were shown the entire video clip from beginning to end, before the formal experiment. During the experiment, the description of each target scene was shown to participants through the text field named event cue, as shown in Fig. 2. After a target scene was located, the description of the next target scene was shown, and all target scenes should be located in a constant increasing order, i.e. target scene 1, 2, 3, 4, and 5. The experiment ended when all five target scenes were located, and then the TCT was recorded for each participant.

3.1.4. Results

Fig. 4 shows the mean TCT over the various experimental conditions. Two-way ANOVA was used to analyze these results, which showed that affective annotation [F (1,22) = 17.313, $p < 0.05$], familiarity [F (1,22) = 66.138, $p < 0.05$], and affective annotation × familiarity [F (1,22) = 6.800, $p < 0.05$] had significant effects on TCT. For participants who were familiar with the video clip, TCT decreased 18.4% with affective annotation compared to when it was not employed. For participants who were unfamiliar with the clip, the affective annotation helped even more, decreasing TCT by 35.3%. One-way ANOVA showed that affective annotation had different effects on TCT (familiar [F (1,10) = 4.741, $p = 0.054$] and unfamiliar [F (1,10) = 13.123, $p < 0.05$]). These results support Hypothesis 1 and 2, indicating that affective annotation and the familiarity of the video clip can significantly shorten the time used to locate target scenes within the video clip.

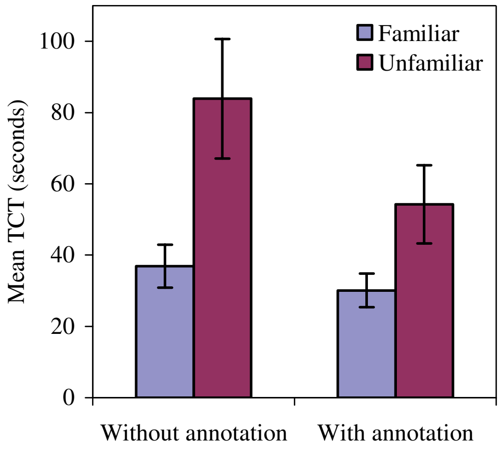


Fig. 4. Mean TCT over different experimental conditions. (Error bars show Std. Deviation).

3.1.5. Operation sequences

The operations of all participants were recorded by a screen capture tool. All records were reviewed after the experiment to analyze the strategies used by participants to locate target scenes.

In order to facilitate analyzing employed strategies, only the operation sequences between the fourth and fifth target scenes for selected participants are described here. After the fourth target scene was located, the description of the fifth target scene, "Jizheng (a character's name) made a strike while playing 10-pin bowling", was shown to participants. After seeing this description, all participants immediately switched to the colour bar of Jizheng through the combo box. Fig. 5 shows the colour bar of Jizheng with affective annotation and the locations of the fourth and fifth target scenes, which are marked as *starting point* and *target*, respectively. The fifth target scene is several seconds ahead of the thin yellow block which is at the center of the presence block, for Jizheng was very happy after the strike and celebrated it with his friends.

Sequence 1 (see Fig. 5) was the operation sequence of a participant (referred to as Participant 1), who was familiar with the video clip and affective annotation was employed. The strategy Participant 1 used might be: first located a

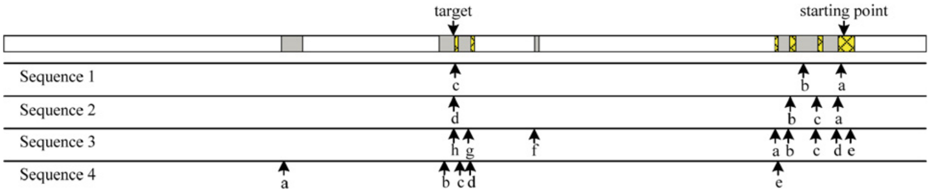


Fig. 5. The colour bar of Jizheng with affective annotation and the operation sequences of participants. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

scene that Jizheng was in a bowling alley, and then utilized the relationship between different scenes and emotions (e.g. making a strike should cause Jizheng to be happy) to find the target scene. Based on this strategy Participant 1 did not utilize the happy emotion of Jizheng first, for he dragged the slide from position a to b which did not relate to the happy indication. When Participant 1 found that the scene in position b did not occur in a bowling alley, he moved the slide to position c immediately. When Participant 1 dragged the slide to position c and found that the scene occurred in a bowling alley, he utilized the relationship between different scenes and emotions to estimate the position of the target scene. After that, Participant 1 moved the slide a little backward from the thin yellow block and located the target scene.

Sequence 2 (see Fig. 5) was the operation sequence of a participant (referred to as Participant 2), who was not familiar with the video clip and affective annotation was employed. The strategy used by Participant 2 to locate the target scene was likely to be completely dependent on the relationship between different scenes and emotions. Based on this strategy, Participant 2 directly dragged the slide to the beginning of each main yellow block (positions a, b, and c), and checked these blocks one by one. When Participant 2 dragged the slide to position d, he located the target scene immediately.

Sequence 3 (see Fig. 5) was the operation sequence of a participant (referred to as Participant 3), who was not familiar with the video clip and affective annotation was employed. The strategy used by Participant 3 was similar with the one used by Participant 2, checking yellow blocks one by one. Additionally, he checked more carefully than Participant 2. At the last presence block of Jizheng, Participant 3 checked five positions (a, b, c, d, and e), whereas Participant 2 just checked three positions. After checking position g, it is suspected that Participant 3 became impatient and lost confidence, causing a switch to the colour bar of Yinzhu (a character's name) and dragged the slide four times. After that Participant 3 switched back to the colour bar of Jizheng and dragged the slide to position h, and the target scene was located finally. Note that two positions in the colour bar of Yinzhu had already been checked before, but Participant 3 checked these positions again after switching the colour bar.

Sequence 4 (see Fig. 5) was the operation sequence of a participant (referred to as Participant 3), who was not familiar with the video clip and affective annotation was not employed. The search strategy employed by Participant 4 appeared to have been to check all presence blocks of Jizheng from left to right. Based on this strategy, Participant 4 first dragged the slide to position a. After finding that the scene did not occur in a bowling alley, he immediately dragged the slide to position b. Although the scenes of positions b, c and d occurred in a bowling alley, he dragged the slide to position e after brief checking. After finding that the scene of position e did not occur in a bowling alley, he dragged the slide back to position b and finally located the target scene after more careful checking.

After reviewing all of the records, a common strategy across participants and experimental conditions was identified, i.e. the participants switched to the character related to a target scene after getting the description of the scene. This strategy was quite straightforward, and all participants appeared to use it to reduce the area to be checked. Since the five target scenes in Experiment 1 had explicit relationships with different characters, all participants selected the correct characters. This observation suggested that the annotation of the presence of each character was useful for participants to shorten the time locating target scenes.

From the review of the participant records, it can be seen that most participants familiar with the video clip (11 out of 12, 5 with affective annotation, and 6 without affective annotation) appeared to use the knowledge about the clip to determine whether a block contained the target scene, and all participants with affective annotation appeared to employ the relationship between different scenes and emotions. To the participants both familiar with the video clip and with affective annotation, most of them (5 out of 6) appeared to utilize the affective annotation after they had determined the most possible presence block, while the exceptional participant only utilized the affective annotation. This observation might explain why the effects of affective annotation on TCT were different according to whether participants were familiar with the video clip and that affective annotation had a greater impact when participants were unfamiliar with the video clip.

In order to utilize affective annotation to locate a target scene, it was very important for participants to relate the target scene with a specific emotion of a character. For example, in locating the fifth target scene, participants would need to relate the scene "make a strike" with the "happy" emotion of the character. In the experiment, all target scenes could be easily related with different emotions. For example, the description of the third target scene was "Jifeng (a character's name) was talking about the moment he met his girlfriend", and it was intuitive for participants to relate the target scene with the happy emotion of the character. The colour bar of Jifeng with affective annotation is shown in Fig. 6. While the effects of affective annotation on TCT are largely depended on whether an explicit relationship between target scenes and specific emotions exists, the frequency and duration of the emotion related to a target scene can also be significant. For example, if a target scene was related with the "angry" emotion of Jifeng, it would conceivably be easier for participants to locate this target scene than a specific target scene from a video clip in which Jifeng had numerous "angry" scenes.

To summarize the above discussion, the effects of affective annotation on locating a target scene appear to be influenced by: (i) whether the target scene can be related to a character; (ii) whether the target scene can be related to a specific emotion of the character; and (iii) the frequency and duration of the related emotion.

3.2. Experiment 2

3.2.1. Design and hypotheses

The aim of the second experiment was to study the effects of affective annotation on the comprehension of a video clip in a limited viewing time period. The task for participants in the second experiment was to attain some understanding of the content of a video clip given a limited viewing time period.

In this experiment, a 2 (with affective annotation or not) × 3 (different viewing time periods) between group design was employed. Three viewing time periods (1/3, 1/2, and 2/3 of the duration of the video clip) were used during the experiment, and are referred to as *short*, *medium*, and *long viewing time*, respectively.

In this experiment a test was implemented to measure the degree of comprehension. The test consisted of ten questions, each with four choices and one correct answer, with the positions of correct answers selected randomly. Participants were instructed to answer all questions even if they were not sure about their choices. In order to minimize the effects of the understanding ability of participants on the results, all questions were designed to be straightforward, and each question corresponded to an important scene in the video clip. If a participant had viewed a scene to which a question corresponded, he/she was supposed to answer the question correctly. In other words, the test measured the understanding of the specific details of the video clip, which was an important part of comprehension. In all ten questions, seven corresponded to scenes with non-neutral emotional states of characters, for example "From whom did Qiang (a character's name) receive a gift?", while the others corresponded to other scenes, for example "What job did Shuan's father (a character's name) have?" All ten scenes corresponding to these questions were almost equally distributed in the video clip, and all participants, regardless of the viewing time period assigned, had an opportunity to view the scenes that provided answers to these questions.

The score obtained by each participant in the test was regarded as the objective metric of the experiment, although a questionnaire was also employed to gain a subjective evaluation of the exercise. This questionnaire contained three bipolar five-point Likert-type scales: (i) How helpful did you find the colour bar in understanding the video clip? (1 refers to totally helpless and 5 refers to very helpful); (ii) How accurate do you think the affective annotation was? (1 refers to totally inaccurate and 5 refers to very accurate); (iii) How sufficient do you think the given viewing time was? (1 refers to totally insufficient and 5 refers to very sufficient).

The hypotheses for the second experiment were:

Hypothesis 1. The effects of affective annotation would change with the viewing time period assigned.

Hypothesis 2. The participants using affective annotation would achieve higher scores than those participants who did not use it.

3.2.2. Participants

Thirty-six participants, undergraduates and postgraduates, were recruited from the campus of Zhejiang University. The participants included the twenty-four participants of Experiment 1. There were 12 females and 24 males and their ages ranged from 23 to 28 (Mean = 25.083, Std. Deviation = 1.273). All of them were familiar with traditional media players, and Chinese was their native language. None of them had watched the video clip used in Experiment 2 before. All participants were randomly divided into six equal groups, corresponding to the six experimental conditions.

3.2.3. Procedure

Before the formal experiment, the interface and operations of EmoPlayer were introduced to the participants that did not attend Experiment 1, and these participants were given 30 min familiarize themselves with EmoPlayer, with the same practice procedure as Experiment 1. Before the start of the formal experiment, the task in the experiment was described and the names and photographs of

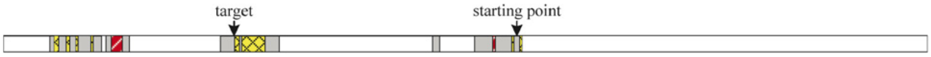


Fig. 6. The colour bar of Jifeng with affective annotation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the main characters in the video clip were shown to the participants. In the formal experiment, participants were asked to watch the video clip for a specific viewing time period, with information about elapsed and remaining time provided through a progress bar named process, as shown in Fig. 2. After watching the video clip, every participant was required to complete the test and questionnaire. Participants were instructed to guess if they were unsure of an answer. Participants who utilized affective annotation in the experiment answered all three questions in the questionnaire while the participants who did not use affective annotation only answered the third question.

3.2.4. Results

Fig. 7 shows the mean test score across the various experimental conditions. Two-way ANOVA was used to analyze the results, which showed that affective annotation [F $(1,34) = 10.028$, $p < 0.05$], viewing time [F $(2,33) = 12.012$, $p < 0.05$], and affective annotation × viewing time [F $(2,33) = 3.367$, $p < 0.05$] had significant effects on test score. Under the short viewing time condition, the mean score was less than 5 regardless whether affective annotation was used or not. Under the medium viewing time condition, the mean score was improved significantly, and with the help of affective annotation participants achieved a higher mean test score. Under the long viewing time condition, the mean test score was further improved with the help of affective annotation, whereas it even decreased slightly when affective annotation was not used. Under different viewing times, one-way ANOVA showed that affective annotation had different effects on test score (short [F $(1,10) = 0.052$, $p = 0.825$], medium [F $(1,10) = 1.875$, $p = 0.201$], and long [F $(1,10) = 16.897$, $p < 0.05$]). The above results support Hypothesis 1 suggesting that the effects of affective annotation change with the viewing time, while support for Hypothesis 2, that participants using affective annotation would achieve higher comprehension test scores, is unclear. It can, however, be suggested that the beneficial effects of affective annotation on the comprehension of the video clip increased with viewing time.

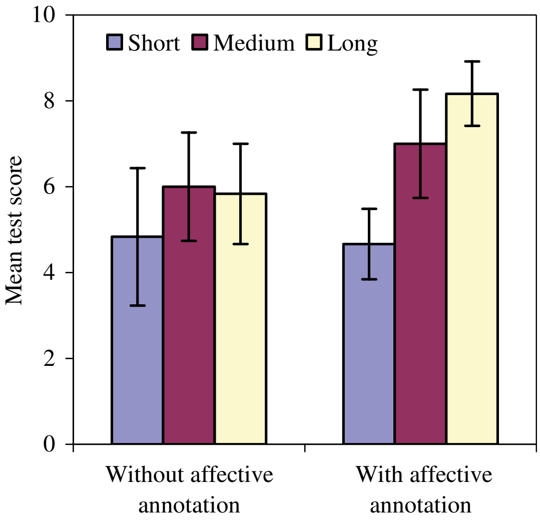


Fig. 7. Mean test score over different experimental conditions. (Error bars show Std. Deviation).

The results of the subjective evaluation were: to the first question (How helpful did you find the colour bar in understanding the video clip?), the mean evaluation was 4.111 (Std. Deviation = 0.676), and one-way ANOVA showed that viewing time did not have significant effects on the value [F $(2,15) = 1.711$, $p = 0.214$]; to the second question (How accurate do you think the affective annotation was?), the mean evaluation was 4.056 (Std. Deviation = 0.539), and one-way ANOVA showed that viewing time did not have significant effects on the value [F $(2,15) = 0.172$, $p = 0.843$]; to the third question (How sufficient do you think the given viewing time was?), the mean evaluations were 2.500 (Std. Deviation = 0.674), 2.917 (Std. Deviation = 0.669), and 3.250 (Std. Deviation = 0.622) corresponding to the short, medium, and long viewing time, with two-way ANOVA indicating that viewing time did have a significant effect on the value [F $(2,33) = 3.631$, $p < 0.05$], whereas affective annotation [F $(1,34) = 0.238$, $p = 0.629$] and affective annotation × viewing time [F $(2,33) = 0.060$, $p = 0.942$] did not. These results suggested that participants thought that affective annotation was useful for them in comprehending the video clip and that the affective annotation of the video clip was considered accurate. The assigned viewing time, however, was not considered sufficient even at its longest.

3.2.5. Operation sequences

The operations of all participants were recorded by a screen capture tool. All records were reviewed after the experiment to analyze the strategies used by participants in comprehending the video clip given a limited viewing time. The video clip used in the second experiment was a stage drama in which there were four characters: Qiang; Qiang's father; Shuan; and Shuan's father. The numbers of happy, sad, fearful, angry, and neutral emotional states of these characters were (13, 0, 2, 2, 11), (8, 3, 0, 2, 16), (7, 4, 2, 0, 10) and (3, 3, 1, 3, 10), respectively. The length of the entire video clip was approximately 36 min.

Sequence 5 was the operation sequence of a participant (referred to as Participant 5), who was assigned 18 min (medium viewing time) to watch the video clip with affective annotation. In the experiment, Participant 5 took the presence of a character as a clue to watch the video clip and browsed all presence blocks of the character from left to right. The whole watching procedure consisted of four browsing procedures, with each one corresponding to a specific character. If an encountered scene had already been watched, Participant 5 dragged the slide to jump forward. After four browsing procedures, Participant 5 spent the remaining time reviewing blocks where different emotions were densely distributed. In the post experiment test, Participant 5 answered eight questions correctly. For the convenience of further discussion, this browsing strategy is referred to as *role-based strategy*.

Sequence 6 was the operation sequence of a participant (referred to as Participant 6), who was assigned 18 min (medium viewing time) to watch the video clip with affective annotation. Although the condition was the same as Participant 5, the sequence of Participant 6 was radically different. In the experiment Participant 6 browsed the video clip from left to right and he always dragged the slide forward. After dragging the slide to a new position, Participant 6 switched between different characters to get corresponding emotional state information. When dragging the slide, Participant 6 used small incremental jumps forward and did not utilize the rims of different colour blocks. In the post experiment test, Participant 6 answered five questions correctly. For the convenience of further discussion, this browsing strategy is referred to as *beginning to end strategy*.

After reviewing the records for all of the participants, it was found that it was difficult for the participants with short viewing time to form a clear viewing strategy. Most of them (10 out of 12) randomly dragged the slide and the operations could hardly be interpreted as strategies at all. In contrast, most participants in the medium and long viewing time groups (21 out of 24) appeared to employ a coherent strategy (either rolebased or beginning to end strategies). Table 1 illustrates the mean test scores of the participants within the medium and long viewing time groups who employed viewing strategies. It can be seen that under all four conditions (2 viewing time periods × 2 strategies), the mean test scores increased with the help of affective annotation. Additionally, without affective annotation the participants using the beginning to end strategy obtained higher mean test scores than those using the role-based strategy, while with affective annotation the results were reverse. These results suggested that the participants using the role-based strategy got more help from affective annotation. Table 1 also shows that except the group with affective annotation and medium viewing time, most participants selected a strategy through which can obtain a higher test score.

From the review of the participant records, it can be seen that employing a role-based strategy increased the workload of participants, for example Participant 5 browsed the video clip four times with different colour bars of these characters. Without affective annotation, the participants that employed a role-based strategy had to equally browse all scenes, and the workload of browsing with different colour bars forced them to spend little time on each scene. This might explain why the mean test score was lowest when a role-based strategy was employed without affective annotation. Since important scenes in the video clip often corresponded with different emotional states of the main characters, affective annotation enabled participants to quickly locate the important scenes and spend their limited time watching these scenes. This capability could be the reason why the mean test score were highest when a role-based strategy was employed in conjunction with affective annotation.

Table 1
Mean test score over different experimental conditions and strategies

Affective annotation	Viewing time	Strategy	Mean test score	Std. Deviation	Number of participants
Without	Medium	Role-based	5	0	2
		Beginning to end	6	1	3
	Long	Role-based	5	0	1
		Beginning to end	6	1.41	4
With	Medium	Role-based	8	0	1
		Beginning to end	6.4	1.29	4
	Long	Role-based	8.25	0.96	4
		Beginning to end	8	0	2

In the interface design of EmoPlayer, only one colour bar is employed to show a character's emotional distribution along timeline, and the conjunction between the emotional states of different characters cannot be explicitly displayed. From the operation sequence of Participant 6, it can be seen that with the beginning to end strategy, participants switched between different characters to get corresponding emotional state information, which suggested that the relationships between different characters were important for participants to understand the video clip. Based on this finding, the interface design of EmoPlayer could be leveraged by supporting concurrently showing multiple colour bars in terms of the emotional distributions of different characters.

To summarize the above discussion, the effects of affective annotation on the comprehension of a video clip in a limited viewing time period in this experiment appear to be influenced by: (i) how much time is assigned; and (ii) what browsing strategy is employed.

3.3. Summary

The purpose of the experiments outlined above was to evaluate the efficiency of affective annotation through two representative tasks, one locating a specific scene, and the other comprehending a video clip given a limited viewing time. From the results of the experiments, it can be concluded that affective annotation can improve the speed of locating a specific scene within a video clip, and these improvements become more evident when users are not familiar with the clip. If the assigned viewing time is of a long duration, affective annotation can aid users in comprehending an entire video clip. Additionally, the browsing strategies employed by users, as well as viewing time, appear to be influential in the utilization of affective annotation. Note that our work has some limitations and the results described in this paper are preliminary. For example, each experiment only used a single video clip. Only the efficiency of affective annotation had been measured. In Experiment 1, the target scenes only related to limited emotional states. In Experiment 2, only a part of comprehension (i.e. the understanding of the specific details of the video clip) was tested.

4. Conclusions and future work

With the advances of multimedia annotation, interfaces for interacting with multimedia are receiving more and more attention. This paper presents the design and implementation of EmoPlayer, which can play video clips with affective annotations, showing each character's emotional distribution along a video clip timeline through a colour bar based interface. Additionally, two experiments were conducted to evaluate the efficiency of affective annotation. The results of these experiments indicate that affective annotation can be effective in both improving the speed of locating a specific target scene and in comprehending a video clip in a limited viewing time.

Our study has broader implications for multimedia annotation, affective computing, and HCI researches. Currently, multimedia annotation is mainly used to provide content retrieval service. Although affective annotation of text documents and a corresponding interface to visualize the emotional states have already been studied, we present an innovative application of the affective annotations of video clips. The results of our study indicate that the ability of humans, mapping between different scenes and emotions, can be utilized by media players to improve human performance under certain given tasks. Extending these findings, this ability, or other emotion related abilities of humans, might be further utilized by affective computing systems to improve human performance. To media player interface design, our work shows that augmenting some semantic information of the playing clip on the interfaces of media players can improve human performance, and the colour bar based interface is an effective method to show time related information.

Although some emotion researches in HCI have already been carried out, for example detecting and responding to human emotional expression, adding affective interventions in human–computer interaction, and structuring and representing emotions for use in digital systems etc., our work broadens this research field, and the emotions encoded in video clips are utilized to improve human performance through carefully designed interface. Additionally, our study sheds light on several potential factors that might affect the efficiency of such an interface design, for example the conjunction between content and emotion, the familiarity of content, time constraint, and using strategy etc.

Since the experimental conditions of this study were tightly constrained, the next step of our research involving affective annotation will be to evaluate their efficiency under alternative conditions, such as employing multiple colour bars, different emotion categorizations, different mappings between emotions and colours, different video clips, and different tasks. The results of this current research also suggest that showing information of the presence of characters in a video clip through the media player interface can improve the human performance in searching and comprehending video clips. Thus, further experiments will be conducted to study this performance in comparison with traditional media players.

Acknowledgements

This work was funded by the Natural Science Foundation of China (No. 60703040), the Natural Science Foundation of Zhejiang Province, China (No. Y104199), and the Scientific Research Fund of Zhejiang Provincial Education Department, China (No. 20051020). We thank all the participants in the study and the reviewers who provided very valuable feedback on an earlier version of this paper.

References

Abowd, G.D., Gauger, M., Lachenmann, A., 2003. The family video archive: an annotation and browsing environment for home movies. In: Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval, pp. 1–8.

Barbieri, M., Mekenkamp, G., Ceccarelli, M., Nesvadba, J., 2001. The color browser: a content driven linear video browsing tool. In: Proceedings of IEEE International Conference on Multimedia and Expo, pp. 627–630.

Bottoni, P., Civica, R., Levialdi, S., Orso, L., Panizzi, E., Trinchese, R., 2004. MADCOW: a multimedia digital annotation system. In: Proceedings of the Working Conference on Advanced Visual Interfaces, pp. 55–62.

Bulterman, D.C.A., 2003. Using SMIL to encode interactive, peer-level multimedia annotations. In: Proceedings of the 2003 ACM Symposium on Document Engineering, pp. 32–41.

Carter, S., Churchill, E., Denoue, L., Helfman, J., Nelson, L., 2004. Digital graffiti: public annotation of multimedia content. In: Proceedings of CHI'04 Extended Abstracts on Human Factors in Computing Systems, pp. 1207–1210.

Chan, C.H., Jones, G.J.F., 2005. Affect-based indexing and retrieval of films. In: Proceedings of the 13th Annual ACM International Conference on Multimedia, pp. 427–430.

Christel, M.G., Warmack, A.S., 2001. The effect of text in storyboards for video navigation. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 1409–1412.

Cockton, G., 2002. From doing to being: bringing emotion into interaction. Interacting with Computers 14 (2), 89–92.

Costa, M., Correia, N., Guimarães, N., 2002. Annotations as multiple perspectives of video content. In: Proceedings of the Tenth ACM International Conference on Multimedia, pp. 283–286.

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

Craggs, R., Wood, M.M., 2004. A categorical annotation scheme for emotion in the linguistic content of dialogue. In: Proceedings of the Affective Dialogue Systems, Tutorial and Research Workshop, pp. 89–100.

Crockford, C., Agius, H., 2006. An empirical investigation into user navigation of digital video using the VCR-like control set. International Journal of Human–Computer Studies 64 (4), 340–355.

Davidoff, L.L., 1980. Introduction to Psychology, 2nd ed. McGraw-Hill, New York.

Ekman, P., 1982. Emotion in the Human Face. Cambridge University Press, Cambridge.

Fogli, D., Fresta, G., Mussio, P., 2004. On electronic annotation and its implementation. In: Proceedings of the Working Conference on Advanced Visual Interfaces, pp. 98–102.

Költringer, T., Tomitsch, M., Kappel, K., Kalbeck, D., Grechenig, T., 2005. In: Implications for designing the user experience of DVD menus, Proceedings of CHI'05 Extended Abstracts on Human Factors in Computing Systems, pp. 1565–1568.

Kwon, O.W., Chan, K., Hao, J., Lee, T.W., 2003. Emotion recognition by speech signals. In: Proceedings of the 8th European Conference on Speech Communication and Technology, pp. 125–128.

Liu, H., Lieberman, H., Selker, T., 2003a. A model of textual affect sensing using real-world knowledge. In: Proceedings of the 8th International Conference on Intelligent User Interfaces, pp. 125–132.

Liu, H., Selker, T., Lieberman, H., 2003b. Visualizing the affective structure of a text document. In: Proceedings of CHI'03 Extended Abstracts on Human Factors in Computing Systems, pp. 740–741.

Murray, I.R., Arnott, J.L., 1993. Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. Journal of the Acoustical Society of America 93 (2), 1097–1108.

Norman, D.A., 2001. DVD menu design: the failures of web design recreated yet again. Available at: http://www.useit.com/alertbox/20011209.html (last retrieved 16 October 2006).

Partala, T., Surakka, V., 2004. The effects of affective interventions in human–computer interaction. Interacting with Computers 16 (2), 295–309.

Peter, C., Herbon, A., 2006. Emotion representation and physiology assignments in digital systems. Interacting with Computers 18 (2), 139–170.

Picard, R.W., Klein, J., 2002. Computers that recognise and respond to user emotion: theoretical and practical implications. Interacting with Computers 14 (2), 141–169.

Qian, Y., Feijs, L.M.G., 2004. Exploring the potentials of combining photo annotating tasks with instant messaging fun. In: Proceedings of the 3rd International Conference on Mobile and Ubiquitous Multimedia, pp. 11–17.

Ramos, G., Balakrishnan, R., 2003. Fluid interaction techniques for the control and annotation of digital video. In: Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology, pp. 105–114.

Russell, J.A., 1991. Culture and the categorization of emotions. Psychological Bulletin 110 (3), 426–450.

Thompson, W.F., Schellenberg, E.G., Husain, G., 2004. Decoding speech prosody: do music lessons help? Emotion 4 (1), 46–64.

Tokitomo, A., Kazuhiro, N., Hiroshi, Y., 2004. Effect of facial colors on humanoids in emotion recognition using speech. In: Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication, pp. 59–64.

Valdez, P., Mehrabian, A., 1994. Effects of color on emotions. Journal of Experimental Psychology: General 123 (4), 394–409.

Wang, H.C., Ahuja, N., 2003. Facial expression decomposition. In: Proceedings of the 9th IEEE International Conference on Computer Vision, pp. 958–965.

Williams, M.A., Moss, S.A., Bradshaw, J.L., Mattingley, J.B., 2005. Look at me, I'm smiling: visual search for threatening and nonthreatening facial expressions. Visual Cognition 12 (1), 29–50.

Xia, F., Wang, H., Fu, X.L., Zhao, J.Y., 2005. An XML-based implementation of multimodal affective annotation. In: Proceedings of the 1st International Conference on Affective Computing and Intelligent Interaction, 535–541.