Enriching User Profiling with Affective Features for the Improvement of a Multimodal Recommender System

Ioannis Arapakis
Computing Science Dep.
University of Glasgow
Lilybank Gardens
Glasgow, G12 8QQ
arapakis@dcs.gla.ac.uk

Yashar Moshfeghi
Computing Science Dep.
University of Glasgow
Lilybank Gardens
Glasgow, G12 8QQ
yashar@dcs.gla.ac.uk

Hideo Joho
Computing Science Dep.
University of Glasgow
Lilybank Gardens
Glasgow, G12 8QQ
hideo@dcs.gla.ac.uk

Reede Ren
Computing Science Dep.
University of Glasgow
Lilybank Gardens
Glasgow, G12 8QQ
reede@dcs.gla.ac.uk

David Hannah
Computing Science Dep.
University of Glasgow
Lilybank Gardens
Glasgow, G12 8QQ
hannahd@dcs.gla.ac.uk

Joemon M. Jose
Computing Science Dep.
University of Glasgow
Lilybank Gardens
Glasgow, G12 8QQ
jj@dcs.gla.ac.uk

ABSTRACT

Recommender systems have been systematically applied in industry and academia to help users cope with information uncertainty. However, given the multiplicity of the preferences and needs it has been shown that no approach is suitable for all users in all situations. Thus, it is believed that an effective recommender system should incorporate a variety of techniques and features to offer valuable recommendations and enhance the search experience. In this paper we propose a novel video search interface that employs a multimodal recommender system, which can predict topical relevance. The multimodal recommender accounts for interaction data, contextual information, as well as users' affective responses, and exploits these information channels to provide meaningful recommendations of unseen videos. Our experiment shows that the multimodal interaction feature is a promising way to improve the performance of recommendation.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Relevance Feedback, Search Process*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces; I.5.1 [**Computing Methodologies**]: Pattern Recognition—*Models*

General Terms

Experimentation, Human Factors, Performance

CIVR 09 , July 8-10, 2009 Santorini, GR

Keywords

Affective feedback, facial expression analysis, multimedia retrieval, recommender systems, user profiling

1. INTRODUCTION

During the past two decades the explosive growth rate of the world wide web has led to a profound increase in the availability of online multimedia content. Considering the scale of this growth and the information overload problem that it introduces, it becomes evident that the average user often faces a challenge when attempting to locate items of interest or relevance online. To address these issues different techniques have been developed that aim to improve some of the facets of information retrieval, such as indexing, searching and filtering. However, while they often provide a varying level of support, they are not always tailored to the specific user and the specific situation.

In recent years recommender systems have emerged as a potential solution to the problem of information overload. Recommender systems are a personalized information filtering technology [6], designed to assist users in locating items of interest, by providing useful recommendations. They have been applied successfully in a number of different applications, to improve the quality of the web services. Such examples include Amazon.com, for recommending books, CDs and other products [9], MovieLens, for recommending movies [10], and VERSIFI Technologies, for recommending news articles [2]. Recommender systems adopt various profiling techniques to collect information about the interaction history, which they eventually integrate into user profiles. The data retained inside the user profiles are regarded as indicative of the users' preferences and interests, and, usually, refer to information such as age, gender, place of birth, preferences, needs, etc. Based on the internal form of representation of the user information, the latter profiles can be categorized into single-faceted and multi-faceted.

User profiling consists of three stages, namely: (i) relevance feedback, (ii) feature selection, and (iii) updating of profile. The feedback cycle is a a necessary practice, since users are sometimes guided by a vague information need,

which they cannot easily express in terms of keywords, or relate to unseen information items [7]. Therefore, the value of relevance assessments lies in the progressive disambiguation of that need and it is usually achieved through the application of different feedback techniques. These techniques range from explicit to implicit and help determine the relevance of the retrieved items. However, they often do so by determining relevance with respect to the cognitive and situational levels of interaction, failing to acknowledge the importance of intentions, motivations and feelings in cognition and decision-making [5, 11].

In this work, we propose a novel video search interface that applies real-time facial expression analysis to aggregate information on users' affective behaviour. We, furthermore, present a way of processing that information to determine the degree of relevance of perused videos and, eventually, enrich the user profiles with it. The value of our interface lies in the combination of different modules (facial expression recognition, recommender system, etc.), the integration of sensory data and, moreover, the application of information fusion. Finally, we contribute to the exploration of the role of emotions in the search process, by highlighting some of the factors that can influence users' affective behaviour. Overall, we examined the following research hypotheses:

***H*$_1$:** Users' affective responses are not consistent across different types of stimuli (search process, the viewed content).

***H*$_2$:** The integration of affective features, deriving from automatic facial expression analysis, to user profiling can improve the performance of a recommender system.

2. RECOMMENDATION MODEL

In the multimodal recommendation model, each instance of a user interest is represented within the profile as a vector. The user profile is updated based on the information received on a document *d(f_d)*, during the feedback cycle, which is implemented using one of the relevance feedback models described next.

In [3] the authors developed a multimodal approach, were the creation and maintenance of profiles takes place automatically, through the exploitation of user feedback. The user profile is represented internally as a set of vectors. Each vector consists of key-value pairs, where the key can be a term and the value is the weight of the term for that particular vector. The number, size and elements of the vectors of the document d can be modified after each feedback cycle (f_d) using the following formula:

$$P_{act} = (1 - \lambda) \times P_{act} + \lambda \times (f_d) \times v_d$$

where λ is the learning factor, P_{act} is the active vector, (v_d) is the document vector and (f_d) represents user feedback. P_{act} is determined based on the inner product of the vectors in the user profile and the document vector. The closest vector to the document vector becomes the active one. The similarity measure between the active vector and the document vector should be bigger than a certain threshold γ. The similarity is calculated using the cosine similarity measure:

$$v_d = (t_1, t_2, t_3, \ldots, t_n)$$
$$v_u = (t'_1, t'_2, t'_3, \ldots, t'_n)$$
$$cos(v_d, v_u) = \sum(t_i \cdot t'_i) \;/\; \sqrt{(\sum t_i^2 \cdot \sum t'^2_i)}$$

If none of the interests in the profile are bigger than γ and the feedback is positive, the article vector is added to the user profile as a new interest. Otherwise, the user feedback is ignored. Each time that an interest becomes the P_{act}, its strength is increased or decreased according to the received feedback (if the feedback is positive the strength will be increased, if not the strength will decrease). If the strength of an interest becomes lower than a threshold it will be eventually removed from the profile. In addition, the interest with the higher strength has higher priority, compared to the others, to be the P_{act}. To keep the user profile optimum, after applying the formula to the P_{act}, all other interests in the profile are applied to the formula for possible merging. Each interest is then examined using the P_{act}. If the similarity between two interests is found greater, the two vectors are merged. In that case v_d becomes the second interest. The merging process takes place only once after each feedback cycle, for efficiency reasons. For additional details, the reader is referred to [3] and [4].

3. EXPERIMENTAL METHODOLOGY

Even though physiological response patterns and affective behavior are observable, there are no objective methods of measuring the subjective experience. Very often the emotional experience is captured using a combination of think-aloud protocols and forced-choice or free-response reports, and in some cases it is decomposed and examined through the application of a multi-modal analysis [8]. The most common approaches in emotion analysis have been the discrete-categories and dimensional approach.

In the first approach, emotions are classified into six or more basic categories, such as happiness, sadness, anger, fear, disgust and surprise. The term "basic" is primarily used to denote elements that can be combined to form more complex or compound emotions. However, emotions can be often experienced as mixed or blended, making sometimes the classification into a limited number of categories too restrictive. In the dimensional approach, emotions are characterized in terms of a multi-dimensional affect space. The most popular dimensions are those of arousal and valence. Valence is used to represent the pleasantness of the stimuli along a bipolar continuum, between a positive and a negative pole, while arousal is used to indicate the intensity of the emotion. This dimensional taxonomy of emotions treats all emotion categories as varying quantitatively from one another and represents their relationships as distances within the affect space.

In this work we employed eMotion [15], a facial expression recognition system of reasonably robust performance and accuracy across all individuals that applies the first approach (see section §3.3.5). Our primary goal was to relate users' facial expressions, which are regarded as indicative of their affective state, to the relevance of the perused videos and fuse that information with interaction data to enrich user profiling. We acknowledge that facial expression recognition should not be confused with human emotion recognition.

3.1 Design

This study used a repeated-measures design. There were three independent variables, namely: task domain (with two levels:"learning" and "entertainment"), task scope (with two levels: "broad" and "focused") and recommendation system (with two levels: "RS1: baseline" and "RS2: multimodal").

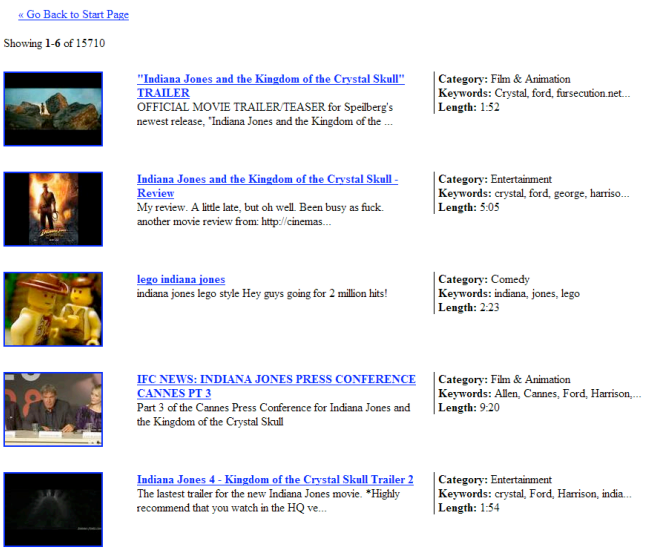

Figure 1: The results page (second layer)

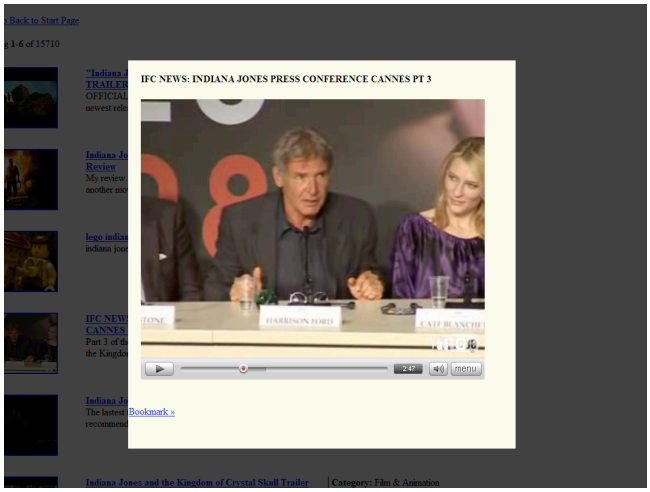

Figure 2: Browsing a video (third layer)

The task domain levels were controlled by assigning topics with the appropriate context, while the task scope levels were controlled by introducing either well-defined or less explicit relevancy criteria. The recommendation system levels were manipulated by employing a different user profiling technique. In the baseline version of our system the profiling technique integrates information that derives only from user actions (meta-data & click-throughs). The multimodal version, however, integrates affective information (users' facial expressions), on top of the interaction data that is being captured. The dependent variables were the system's performance, as it was perceived by the participants, as well as their emotional experience with respect to the search process and the viewed content.

3.2 Participants

Twenty-four participants of mixed ethnicity and educational background (3 Ph.D. students, 12 MSc students, 4 BSc students and 4 other) applied for the study through a campus-wide ad. They were all familiar with the English language (4 native, 12 advanced, 3 intermediate and 4 beginner speakers). Of the 24, 13 were male and 11 were female. All participants were between the ages of 19 and 37 and free from any obvious physical or sensory impairment.

3.3 Apparatus

For our experiment we used two desktop computers equipped with conventional keyboard and mouse. The first computer acted as the server, which hosted the recommender system, the Support Vector Machine (SVM) model, the facial expression recognition system (eMotion) and the video recording software. The second computer acted as the client and was used to provide access to the search interface. In addition, participants' desktop actions were logged using a custom-made script, which recorded information such as starting, finishing and elapsed times for interactions, and click-throughs. A "Live! Cam Optia AF" web camera, with a 2.0 megapixels sensor, and a "Logicool Qcam", with a 1.3 megapixels sensor, were also mounted on top of the client's screen. The cameras were used for recording the participants' expressions, as well as real-time facial expression analysis.

3.3.1 *Questionnaires*

The participants completed an Entry Questionnaire at the beginning of the study, which gathered background and demographic information, as well as previous experience with multimedia and online searching. The information obtained from it was used to characterize subjects, but not in subsequent analysis. A Post-Search Questionnaire was also administered after each task, to elicit participants' viewpoint on certain aspects of the search process, which was an adaptation of the Geneva Appraisal Questionnaire (GAQ) [12]. Its purpose is to assess, as much as possible, through recall and verbal report the results of a participant's appraisal process in the case of an emotional episode. All of the questions included in the questionnaire were forced-choice type, with the exception of a single question that requested a written description. This description asked for the event that elicited the emotional episode, in addition to details with respect to what has happened and the consequences it had for the participant. Out of the 35 questions of GAQ we used only 8 (4-6, 25, 29 and 32-34) that were relevant to the context of our study. Finally, an Exit Questionnaire was introduced at the end of the study.

3.3.2 *Search Tasks*

We formulated a set of search tasks that differed in their domain and scope. All topics were manually performed, prior to the experiment, to ensure the availability of relevant videos in the YouTube collection. The criteria for selecting the search tasks were that there should be enough relevant documents for each topic, to allow for the accumulation of sufficient data for the user profiles. We presented them using the structural framework of the simulated information need situations [4]. By doing so, we believe that we facilitated a better understanding of the task and increased the participants' motivation. For every search task the participants had the possibility of selecting, among a predefined list of options, the sub-topic of their choice.

3.3.3 *Search Interface*

For the completion of the search tasks we used a customized video search interface, which worked on top of the YouTube search engine and was designed to resemble its basic layout while retaining a minimum number of graphical elements. Each result is represented by a thumbnail, a short description and some meta-information (such as category, associated keywords and duration). The interface uses the YouTube API to retrieve video clips from the YouTube collection and presents them in their original order (Figure 1). The participants had the option to either perform a focused search, by formulating and submitting a query, or by looking into the available video categories for relevant clips.

The architecture of the video search interface consists of three different layers. The first layer is dedicated to support any interaction that occur at the early stages of searching, such as query formulation and search execution. It is at this layer where the participants could submit their queries or explore the predefined video categories, as mentioned above. Any output generated by that interaction (whether originating from a user query or the selection of a video category) is presented in the second layer. From there, the participants could easily select and preview any of the retrieved clips. The content of a clip is shown on a separate panel, in the foreground, which corresponds to the third layer of our system (Figure 2). The main reason behind this layered architecture was to isolate the viewed content from all possible distractions that reside on the desktop screen; therefore, establishing reasonable ground truth that allowed us to relate the recorded facial expressions to the source of stimuli (in our case, the perused video clip). Upon viewing the clip, the participants had to explicitly indicate: (i) the degree of relevancy of the video, and (ii) the emotional impact of the video content.

3.3.4 SVM Model

We trained a two-layer hierarchical SVM model to discriminate between two categories of videos (relevant, irrelevant), by analysing facial expression data. We trained our models using a radial basis function (RBF) kernel, which, among the basic four SVM kernels (linear, polynomial, radial basis function, sigmoid), was considered as a reasonable first choice. The RBF kernel can nonlinearly map data into a higher dimensional space, unlike the linear kernel that can be applied successfully only when the relation between the class labels and the attributes is nonlinear [16]. In addition, the sigmoid kernel behaves like an RBF, for a certain set of parameters [14]. Moreover, the RBF kernel is preferable, since it encounters less numerical difficulties and has a limited number of hyperparameters.

The ground truth was obtained by classifying relevant vs. irrelevant expressions in the annotated data set we acquired from [1]. We are aware that the data we used for training purposes derived from a document retrieval experiment and was, therefore, not portraying very accurately the conditions that were encountered in our video retrieval tasks. However, it was the only available annotated data set we could employ in our study, at that point.

Our model consists of 10 weak classifiers, each trained on a different instance of the training set. The whole training set was predicted once, and the output of each weak classifier (support vectors) was used to train the meta-classifier. To deal with the imbalanced set we trained the SVM model by randomly sampling half of the relevant key-frames and one fifth of the irrelevant key-frames. Furthermore, each time a key-frame appeared in both categories it was labeled as relevant. This hierarchical framework improved the original accuracy from 78% to 89%.

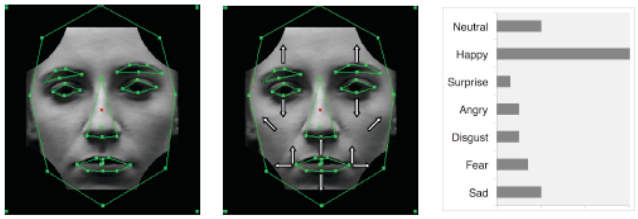

Figure 3: eMotion

3.3.5 Facial Expression Recognition System

Recent findings indicate that emotions are primarily communicated through facial expressions and other facial cues (smiles, chuckles, frowns, etc.) that are regarded as an essential aspect of social interaction. Affective information conveyed through the visual channel can be crucial to human judgement and offer valuable insights to the observer. Automatic systems are an alternative approach to facial expression analysis that can achieve performance which is comparable to that of trained human recognition [13]. In this study we applied real-time facial expression analysis using the feature-based system described in [15]. The process takes place as follows: initially, eMotion detects certain facial landmark features (such as eyebrows, the corners of the mouth, etc.) and constructs a 3-dimentional wireframe model of the face, consisting of a number of surface patches wrapped around it (Figure 3). After the construction of the model, head motion or any other facial deformation can be tracked and measured in terms of motion-units (MU's). The intensity and category of an emotion can then be determined indirectly by the presence and degree of changes in all facial regions associated with it.

Every time a clip is perused eMotion applies facial expression analysis, for every key-frame captured by the camera during that time-period. It then communicates to a predefined port the results of the emotion classification, along with the corresponding motion units, as a stream of sensory data. Our system then retrieves the data stream and forwards it to the SVM model and, depending on the outcome of the classification, it classifies the video as either relevant or irrelevant. In the first case, the recommender system will attempt to retrieve more similar results, using the meta-information of the perused video clip.

3.4 Procedure

The user study was carried out the following way: The formal meeting with the participants took place in the office of the researcher. At the beginning of each session the participant was informed about the conditions of the experiment, both verbally and through a Consent Form, and then had to complete an Entry Questionnaire. A brief training followed, which explained the basic functions of the search interface environment and the terms of interaction. Also, to ensure that the participant's face would be visible to the camera at all times, we encouraged them to keep a proper posture by indicating health and safety measures.

Every participant completed two search tasks in total.

Table 1: Average rating of recommended videos

	Baseline	Multimodal	Total
Overall	1.7 (1.4)	**2.0** (1.5)	1.8 (1.4)
Domain: Learning	1.8 (1.4)	**2.3** (1.6)	2.0 (1.5)
Domain: Entertain.	1.6 (1.3)	1.7 (1.3)	1.6 (1.3)
Scope: Broad	1.8 (1.5)	**2.2** (1.6)	2.0 (1.6)
Scope: Focus	1.6 (1.2)	1.8 (1.4)	1.7 (1.3)

Bold: Statistically significant at $p \leq .05$.

For each search task they were given a short cover story, which introduced them to an artificial situation, thus facilitating the formulation of better-defined relevance criteria. To negate the order effects we counterbalanced the task distribution by using a Latin Squares design. The participants were asked every time to bookmark as many relevant videos as possible and were given 15 minutes to complete the task, during which they were left unattended to work. At the end of each task the participants had to complete the first part of a Post-search Questionnaire and afterwards evaluate a set of recommended videos, which were selected using one of the recommendation strategies. After that they were asked to complete the second part of the Post-search Questionnaire.

An Exit Questionnaire was also administered at the end of each session, along with the receipt of payment. Finally, the participants were asked to sign a Payment Form, prior to receiving the participation fee of £12.

4. RESULTS

This section presents the experimental results of our study, based on 48 sessions that were carried out by 24 participants. We collected questionnaire data on three aspects of the information seeking process, namely: (i) perceived relevance of recommendations, (ii) emotional experience with respect to the search process, and (iii) emotional experience with respect to the viewed content. A 5-point Likert scale was used in all questionnaires. Questions that ask for user rating on a unipolar dimension have the positive concept corresponding to the value of 1 (on a scale of 1-5) and the negative concept corresponding to the value of 5. Questions that ask for user rating on a scale of 1-5 represent in our analysis stronger perception with high scores and weaker perception with low scores. Friedman's ANOVA and Wilcoxon Signed-Ranked test were used to establish the statistical significance ($p < .05$) of the differences observed among the four types of tasks ("Learning - Broad", "Learning - Focused", "Entertainment - Broad" and "Entertainment - Focused"). Pearson's Chi-Square test and the Dependent t-test were applied in the analysis of emotion variance (between search process and viewed content) and the performance of the recommender systems. To take an appropriate control of Type I errors we applied a Bonferroni correction, and so all effects are reported at a .0125 level of significance.

4.1 Recommender Systems

This section presents the results from the evaluation of the recommender systems performance, as it was determined by the participants' ratings. The main and interaction effect of our independent variables are examined, again with respect to participants' reported relevance. To distinguish the first half (search task) from the second half (evaluation or recommended videos) of each session, we name the former "Initial" and the latter "Recommendation". Our recommender systems are also categorised as "Baseline" and "Multimodal".

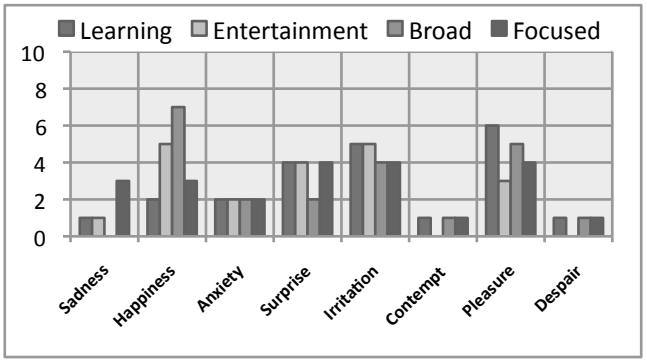

Figure 4: Reported emotions (domain & scope effect)

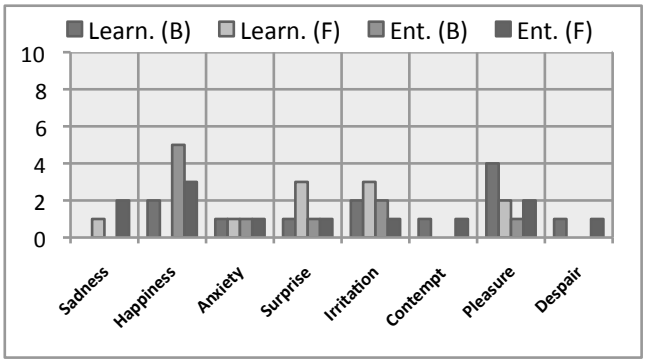

Figure 5: Reported emotions (interaction effect)

4.1.1 Main Effect

Table 1 shows the means and standard deviations (in brackets) of participants' ratings for the two recommendation systems. The second row shows the overall performance of the two systems. As can be seen, participants gave a higher rating to the videos recommended by the multimodal system when compared to the baseline system. The Mann-Whitney Test shows that the difference is significant ($W = 28791.5$, $p = 0.020$). Note that we used the independent test since participants made several ratings within individual blocks, although the experiment was a within-subject design.

4.1.2 Domain & Scope Effect

We were also interested in the effect of tasks on the performance: task domains and task scope. First, we split the rating data based on the blocks of domains or scopes. Then the Mann-Whitney Test was applied to individual blocks. The results are presented in rows 3 to 6 of Table 1. As can be seen, the difference between the two systems was significant in the "Learning" set of the task domain ($W = 7297$, $p = 0.006$), and "Broad" set of the task scope ($W = 7550.5$, $p = 0.015$).

We also ran the two-way ANOVA tests by using systems and task domains as independent variables. The results show that both main effects are significant but no interaction

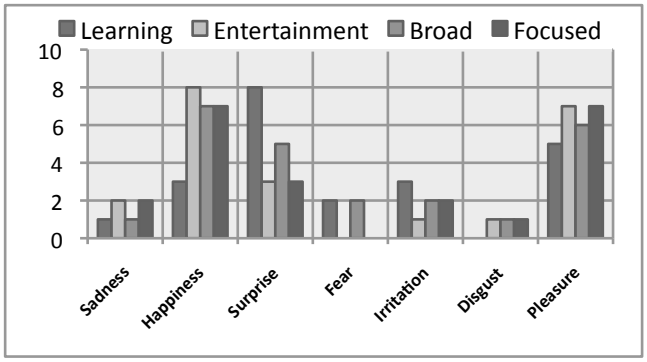

Figure 6: Reported emotions (domain & scope effect)

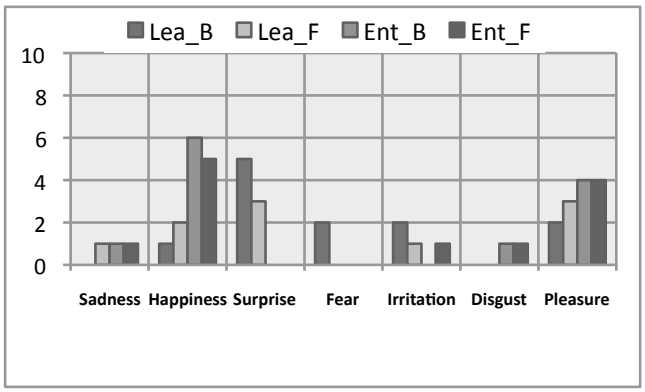

Figure 7: Reported emotions (interaction effect)

effect was found. We repeated the same test for system type and task scopes. The results were similar: significant main effects without interaction effect. Therefore, more research is needed to determine the effect of tasks on the system performance, although some supporting evidence has been found.

4.2 Emotional Experience

To evaluate the progression of the emotion patterns across the tasks we asked our participants to self-assess the emotional episodes they experienced during the study. Tables 2, 3 and 4 show a summary of some of the most important aspects of the emotional episodes, such as the perceived unpleasantness of the stimuli, the intensity of the experienced emotions, as well as the amount of effort that the participants put to control or mask their emotional expressions. Friedman's ANOVA was applied to evaluate the effect of the independent variables to participants' emotion behaviour but it did not reveal any statistically significant differences. We also examined the main effect of the task domain and task task scope on the same variables. Again, the Friedman's ANOVA test did not reveal any significant differences between any of the individual tasks.

4.2.1 Search Process

A major goal of this study was to study the occurrence of emotions during multimedia retrieval. The column chart in Figure 4 illustrates the effect of the task domain on the emotion of the participants, within the context of the search. It appears that pleasure, irritation and surprise were the most intense emotions for tasks in the learning domain, followed by happiness, anxiety, contempt, despair and sadness. The Entertainment domain follows a similar pattern, with happiness, irritation and surprise being reported as the dominant emotions, followed by pleasure, anxiety and sadness. The column chart also depicts the effect of the task, again with respect to the search process. It is evident that happiness and pleasure were the dominant emotions, for the tasks with broad scope, followed by irritation, anxiety, surprise, despair and contempt. A more balanced distribution characterizes the tasks with focused scope, which appear to have pleasure, happiness, irritation and surprise as equally dominant, followed by sadness, anxiety, despair and contempt. Pearson's Chi-Square test was applied to both domain and scope categories but it did not reveal any associated relationships between them.

The column chart in Figure 5 illustrates the distribution of the most intense emotions, with respect to the search. We can see that for the first task category (Learning - Broad) pleasure, happiness and irritation were the most intense emotions, among all other reported emotions, followed by despair, anxiety, contempt and surprise. The second task category (Learning - Focused) shows a different distribution, with irritation and surprise being reported by almost one third of the participants as the most intense emotions. Other emotions, such as pleasure, anxiety and sadness were also reported, but at a lesser rate. The third task category (Entertainment - Broad) is dominated mainly by emotions of happiness, followed by irritation, anxiety, surprise and pleasure. Finally, the fourth task category (Entertainment - Focused) has a more balanced emotional blend, with happiness being the only dominant emotion, followed by pleasure, sadness, anxiety, despair, contempt, irritation and surprise. Pearson's Chi-Square test was again applied without revealing an associated relationship between any of the emotions and the task categories.

4.2.2 Viewed Content

The column chart in Figure 6 illustrates the effect of the task domain on the emotional experiences of the participants, with respect to the viewed content. For the Learning domain we have many reports of surprise and pleasure, followed by happiness, irritation, fear and sadness at a lesser rate. The Entertainment domain shows a similar pattern, with happiness and pleasure as the dominant emotions, followed by surprise, sadness, disgust and irritation. The tasks with broad scope appear to have happiness and pleasure as the dominant emotions, followed by surprise, fear, irritation, sadness and disgust. A very similar distribution also appears in the tasks with focused scope, with pleasure, happiness being again the dominant emotions, followed by surprise, sadness, irritation and disgust. Pearson's Chi-Square test was applied but it did not reveal an associated relationship between any of the emotions for the task domain or scope.

The column chart in Figure 7 shows the distribution of the most intense emotions, with respect to the viewed content. It is evident that for the first task category surprise was the most dominant, among all other reported emotions, followed by pleasure, irritation, fear and happiness. The second task category has a similar blend, with pleasure and surprise being reported by almost one third of the participants as the

Domain	Unpleasantness of stimuli		Intensity of emotion		Effort to mask emotional expressions	
	M	SD	M	SD	M	SD
Learning	2	1.0488	3.1905	1.1233	2.0476	1.0235
Entertainment	1.7619	.9952	3.3333	.9661	1.8571	1.0142

Table 2: Descriptive statistics on emotional experience (domain effect)

Scope	Unpleasantness of stimuli		Intensity of emotion		Effort to mask emotional expressions	
	M	SD	M	SD	M	SD
Broad	2	1	3.1429	1.1526	2.0952	.9952
Focused	1.619	.9207	3.1429	3.1429	1.8571	1.0623

Table 3: Descriptive statistics on emotional experience (scope effect)

Task	Unpleasantness of stimuli		Intensity of emotion		Effort to mask emotional expressions	
	M	SD	M	SD	M	SD
Learn. - Broad	2.2222	1.0929	3.3333	1.3229	2	.866
Learn. - Focused	1.7778	1.0929	3.4444	.7265	2.2222	1.3017
Entert. - Broad	1.7778	.9718	3.3333	.866	2.1111	1.1667
Entert. - Focused	1.3333	.7071	3.3333	.7071	1.5556	.8819

Table 4: Descriptive statistics on emotional experience (interaction effect)

most intense emotions. Other emotions, such as happiness, irritation and sadness were also reported at a lesser rate. The third task category is primarily dominated by happiness, which was reported by half of the participants, and followed by pleasure, sadness and disgust. Likewise, the fourth task category appears to have significantly more incidents of happiness, followed by less frequent incidents of pleasure, sadness, irritation, and disgust. Pearson's Chi-Square test was applied but it only revealed an associated relationship between the task category and the emotions of happiness ($\chi^2(1, N{=}48) = 4.181$, $p < .05$) and surprise ($\chi^2(1, N{=}48) = 5.839$, $p < .05$).

4.2.3 Search Task & Viewed Content

Pearson's Chi-Square test was applied but it did not reveal an associated relationship between any of the emotions and the emotional stimuli (search process, viewed content), for all domains (learning, entertainment), scope (broad, focused), as well as their interaction effect.

5. DISCUSSION

The post-hoc analysis of the results indicates that both recommender systems facilitated a more effective search by suggesting unseen relevant videos to the users, with the multimodal recommender achieving a higher performance. This finding suggests that the performance of profiling was enhanced by the facial expression data and validates our second research hypothesis. Both main and interaction effect analysis show that the participants gave higher ratings for the task domain "Learning" and task scope "Broad". This reveals that the multimodal system was more effective than the baseline system when the tasks involved some form of learning or when the tasks involved a wide range of videos. Overall, the multimodal recommender system was found to perform better than the baseline, which accounted only for click-through data, since the differences in the performance of the two systems were found to be statistically significant. Our findings outline the benefits of enriched profiling and the use of facial expression data, and support the design of multimodal recommender systems. However, we acknowledge the need for additional evaluation of the models, using more sophisticated training techniques and a data set that addresses the conditions of video, rather than document, retrieval.

The analysis of the emotional experience of the participants, as this was inferred from the questionnaire data, revealed that very little effort was put to mask their emotional expressions. Therefore, we can reasonably assume that overall the captured facial expressions were spontaneous and authentic. This behavior appears to be consistent across all three tasks, regardless of domain or scope of the search task. It also suggests that, from the viewpoint of the participants, the presence of the recording equipment did not affect significantly their emotional behavior. Moreover, the intensity of the experienced emotions did not vary largely across the different tasks, indicating that, on average, the participants did not experience very intense or very mild emotions, but rather of medium scale. However, this was an expected finding since the employed video collection was not expected to induce strong affective states to the users. This was a desired condition, since any extreme/strong affective states detected were more likely produced during the appraisal process of the videos' relevance to the user's information need. Similarly, the unpleasantness of the emotional stimuli was also found consistent different the different tasks, revealing a trend towards positive stimuli.

In relation to the search process, the column graphs in Figures 4 - 7 did not make evident any emerging emotion patterns (unlike the progression from positive to negative emotions from study [1]). It appears that neither the domain or the scope, or the interaction between them, has had an effect. The latter finding is supported by the results of the Pearson's Chi-Squared test, which did not indicate any associated relationship between the reported emotions and the independent parameters of the study. Similarly to the emotion patterns of the search process, those that emerged in the context of the viewed videos do not appear to vary significantly between the domain or scope. Again, this find-

ing is supported by the Pearson's Chi-Squared test results, which did not reveal any associated relationship. However, the interaction of the independent variables had an effect on two of the reported emotions. More specifically, the Pearson's Chi-Squared test showed that happiness was significantly more frequently reported in the "Entertainment - Broad" task, compared to the "Learning - Broad" task ($\chi^2(1,\ N{=}48) = 4.181,\ p < .05$), while surprise was significantly less frequently reported in the "Entertainment - Broad/Focused" tasks, compared to the "Learning - Broad" task ($\chi^2(1,\ N{=}48) = 5.839,\ p < .05$).

Looking at the first research hypothesis, the post-hoc analysis did not reveal an associated relationship between the reported emotions and the emotional stimuli (search process, viewed content), for any of the domain or scope categories. Therefore, we do not have enough evidence to refute the null hypothesis ($\boldsymbol{H}_0$**:** Users' affective responses are not consistent across different types of stimuli). This, however, does not necessarily mean that the null hypothesis is true. It only suggests that there is not sufficient evidence against $\boldsymbol{H}_0$, in favor of $\boldsymbol{H}_1$.

6. CONCLUSIONS

In this work we studied the behaviour of emotions in the feedback process. We, furthermore, introduced a novel video retrieval system, which accounts for user feedback that derives from real-time facial expression analysis. We believe that this approach can facilitate and sustain a different form of relevance feedback, which accounts for the affective dimension of human-computer interaction. The value of our system lies in the combination of different modules and modalities, as well as the seamless integration of affective elements into user profiling. In addition, we have presented a way to process that information in order to determine the relevance of perused videos and generate meaningful recommendations.

Our system is realistically applicable; we have implemented it using an inexpensive web camera and a standard browser, which has been modified to communicate with a facial expression recognition system. The study's findings validate our research hypothesis that user affective feedback, as determined from automatic facial expression analysis, can improve the performance of a recommender system when taken into account. This suggests a correlation between the facial expressions produced by the users and the topical relevance of the viewed media. However, this is an ongoing work that warrants additional investigation, especially with respect to the factors that introduce noise to the facial expression analysis, the optimisation of the SVM parameters, as well as the training set, which should address the conditions of video rather than document retrieval. In addition, we intend to investigate the role of complementary emotional cues such as the head pose, upper-body gestures, etc.

7. ACKNOWLEDGMENTS

The research leading to this paper was supported by the European commission, under the contracts FP6-033715 (MIAUCE Project) and FP6-045032 (SEMEDIA project). Our thanks to Roberto Valenti and Nicu Sebe for providing us with an evaluation version of eMotion.

8. REFERENCES

[1] I. Arapakis, J. M. Jose, and P. D. Gray. Affective feedback: an investigation into the role of emotions in the information seeking process. In *SIGIR '08*, pages 395–402. ACM, 2008.

[2] D. Billsus, C. A. Brunk, C. Evans, B. Gladish, and M. Pazzani. Adaptive interfaces for ubiquitous web access. *Commun. ACM*, 45(5):34–38, 2002.

[3] U. Cetintemel, M. Franklin, and C. Giles. Self-adaptive user profiles for large-scale data delivery. *Proceedings. 16th International Conference on Data Engineering, 2000*, pages 622–633, 2000.

[4] U. Cetintemel, M. J. Franklin, and C. L. Giles. Flexible user profiles for large scale data delivery, 1999.

[5] A. R. Damasio. *Descartes Error: Emotion, Reason, and the Human Brain*. Putnam/Grosset Press, 1994.

[6] E.-H. S. Han and G. Karypis. Feature-based recommendation system. In *CIKM '05*, pages 446–452, New York, NY, USA, 2005. ACM.

[7] D. Harman. Relevance feedback revisited. In *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 1–10, New York, NY, USA, 1992. ACM.

[8] A. Jaimes and N. Sebe. Multimodal human-computer interaction: A survey. *Comput. Vis. Image Underst.*, 108(1-2):116–134, 2007.

[9] G. Linden, B. Smith, and J. York. Amazon.com recommendations: item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1):76–80, 2003.

[10] B. N. Miller, A. Istvan, S. K. Lam, J. A. Konstan, and J. Riedl. Movielens unplugged: experiences with an occasionally connected recommender system. In *IUI '03: Proceedings of the 8th international conference on Intelligent user interfaces*, pages 263–266, New York, NY, USA, 2003. ACM.

[11] H.-R. Pfister and B. Gisela. The multiplicity of emotions: A framework of emotional functions in decision making. *Judgment and Decision Making*, 3:5–17, 2008.

[12] K. R. Scherer. *Appraisal considered as a process of multi-level sequential checking.* Appraisal processes in emotion: Theory, Methods, Research. Oxford University Press, New York and Oxford, k. r. scherer, a. schorr, & t. johnstone (eds.) edition, 2001.

[13] N. Sebe, I. Cohen, and T. S. Huang. *Multimodal Emotion Recognition.* Handbook of Pattern Recognition and Computer Vision. World Scientific, 2005.

[14] H. tien Lin and C. jen Lin. A study on sigmoid kernels for svm and the training of non-psd kernels by smo-type methods. Technical report, National Taiwan University, 2003.

[15] R. Valenti, N. Sebe, and T. Gevers. Facial expression recognition: A fully integrated approach. *ICIAPW 2007*, pages 125–130, Sept. 2007.

[16] C. wei Hsu, C. chung Chang, and C. jen Lin. A practical guide to support vector classification. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, 2003.