INTEGRATING FACIAL EXPRESSIONS INTO USER PROFILING FOR THE IMPROVEMENT OF A MULTIMODAL RECOMMENDER SYSTEM

Ioannis Arapakis, Yashar Moshfeghi, Hideo Joho, Reede Ren, David Hannah, Joemon M. Jose

Department of Computing Science
University of Glasgow
Lilybank Gardens
Glasgow, G12 8QQ
{arapakis, yashar, hideo, reede, hannahd, jj}@dcs.gla.ac.uk

ABSTRACT

Over the years, recommender systems have been systematically applied in both industry and academia to assist users in dealing with information overload. One of the factors that determine the performance of a recommender system is user feedback, which has been traditionally communicated through the application of explicit and implicit feedback techniques. In this paper, we propose a novel video search interface that predicts the topical relevance of a video by analysing affective aspects of user behaviour. We, furthermore, present a method for incorporating such affective features into user profiling, to facilitate the generation of meaningful recommendations, of unseen videos. Our experiment shows that multimodal interaction feature is a promising way to improve the performance of recommendation.

***Index Terms*—** Affective feedback, facial expression analysis, muiltimedia retrieval, recommender systems, user profiling

1. INTRODUCTION

In recent years, recommender systems have emerged, as a potential solution to the problem of information overload. Recommender systems have been successfully applied in a number of different applications to improve the quality of their services. Such examples include Amazon.com, for recommending books, CDs and other products [1], MovieLens, for recommending movies [2], and VERSIFI Technologies, for recommending news articles [3]. Recommender systems are a personalized information filtering technology [4], designed to assist users in locating items of interest by providing useful recommendations. They often do so by applying various profiling techniques to aggregate interaction-related information, which they eventually integrate into user profiles. The data retained inside the user profiles are regarded as indicative of the users' preferences [5] and interests, and often 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 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. 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 degree 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 [6, 7].

In this work, we propose a novel video search interface that applies real-time facial expression analysis to aggregate information on the users' affective behaviour. We, furthermore, present a way of exploiting that information to classify the topical relevance of the perused videos, with the help of a Support Vector Machine (SVM), and eventually enrich the user profiles. The value of our interface lies in the combination of different modules (facial expression recognition system, recommender system, etc.), the integration of sensory data and, finally, the application of information fusion. Similar work has been published by Yeasin et. al in [8], who applied facial expression recognition to identify six universal facial expressions from video sequences, and measured levels of interest based on a 3-dimentional affect space.

Overall, we examined the following research question:

$\boldsymbol{H_1}$**:** User affective feedback, as determined from automatic facial expression analysis, can improve the performance of a recommender system when taken into account.

2. EXPERIMENTAL METHODOLOGY

Even though physiological response patterns and affective behavior are observable, there are no objective methods of measuring the subjective experience [9]. 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. The most common approaches in emotion analysis have been the discrete-categories and dimensional approach.

Discrete emotion theorists suggest the existence of six or more basic emotions (happiness, sadness, anger, fear, disgust, and surprise), which are universally displayed and recognized [10, 11]. The existence of basic emotions is supported by evidence of cross-cultural universals for facial expressions and antecedent events, as well as the presence of such states in other primates. Experiments in many countries have shown that people express and recognise basic emotions the same way [12].

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

In the dimensional approach emotions vary quantitatively and are characterized in terms of a multi-dimensional affect space [13]. 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 [14, 13]. Support for the dimensional emotion theories comes from physiological correlates, such as heart rate and skin conductance levels which that correlate with emotional stimuli.

In this work we employed eMotion, a facial expression recognition system that applies the first approach. Research indicates that emotions are primarily communicated through facial expressions rather than bodily gestures [15] and provide facial cues (smiles, chuckles, smirks, frowns, etc.) that are considered an essential aspect of our social interaction. Automatic systems are an alternative approach to facial expression analysis and have exhibited performance that is comparable (under controlled conditions) to that of trained human recognition, which reaches the accuracy of 87% [16]. eMotion applies a generic classifier that has been trained on a diverse data set, combining data from the Cohn-Kanade database. Its main advantage is its reasonable performance across all individuals, irrespectively of the variation introduced from mixed-ethnicity groups. Results of the person-dependent and person-independent tests presented in [17] support our performance-related assumptions.

2.1. 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"). 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 variable was the system's performance, as it was perceived by the users.

2.2. 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 proficient 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.

2.3. 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 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. Participants' desktop actions (URLs visited, starting, finishing and elapsed times for interactions, click-throughs) were logged using a custom-made script. A "Live! Cam Optia AF" web camera (2.0 megapixels) and a "Logicool Qcam" (1.3 megapixels) were 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.

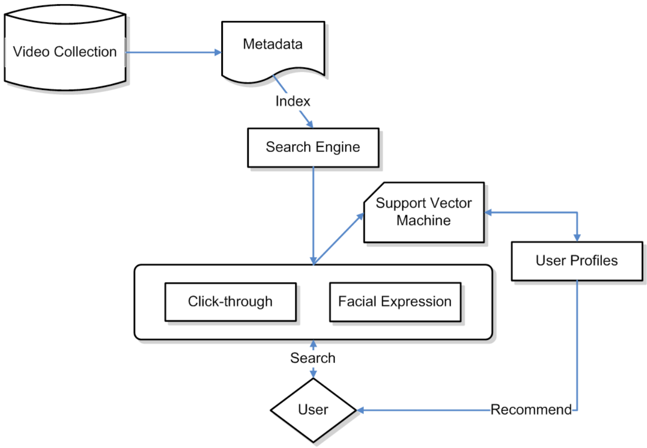

Fig. 1. System Architecture

2.3.1. Search Tasks & Interface

We formulated a set of search tasks that differed in their domain and scope. All topics were manually performed to ensure the availability of relevant videos. We presented them using the structural framework of the simulated information need situations [18]. By doing so, we believe that we facilitated a better understanding of the task and re-inforced 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. For the completion of the search tasks we used a customized video search interface, which worked on top of YouTube search engine and was designed to resemble its basic layout while retaining a minimum number of graphical elements. Each result was represented by a thumbnail, a short description and some meta-information (category, associated keywords, duration).

The architecture of the video search interface consists of three different layers. The first layer was dedicated to support any interaction that would occur at the early stages of searching, such as query formulation and search execution. Any output generated by that interaction was presented in the second layer. From there, the participants could easily select and preview any of the retrieved clips. The content of a clip was shown on a separate panel, in the foreground, which corresponds to the third layer of our system. 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 (the perused video). Upon viewing the clip the participants had to explicitly indicate its degree of relevance to the current task.

2.3.2. SVM Model

We trained a two-layer hierarchical SVM model to discriminate between two categories of videos (relevant, irrelevant), by analysing facial expression data. The ground truth was obtained by classifying relevant vs. irrelevant expressions in the annotated data set we

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: Entertainment	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 $p \leq .05$.

acquired from [19]. By isolating reading sessions, we used users' explicit feedback (bookmarking of documents) as ground truth and associated each document to a class. We are aware that the data set we used for training 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.

The model was trained using a radial basis function (RBF) kernel, which was considered as a reasonable first choice. Optimisation of the SVM parameters or feature engineering were not performed at this stage. Our model consists of 10 weak classifiers, each trained on a different instance of the training set. Each key-frame portraying the user while perusing a document is judged independently of the neighbour keyframes and is characterised as relevant or irrelevant. The whole training set was predicted once, and the output of each weak classifier was used to train the meta-classifier. This hierarchical framework improved the initial accuracy from 78% to 89%.

2.3.3. Facial Expression Recognition System

Real-time facial expression analysis was applied, using the system described in [17]. 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. 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), and, finally, classified into one of the seven detectable emotion categories. 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 pre-defined port the results of the classification, along with the corresponding motion units, as a stream of sensory data. Our system then forwards the data to the SVM model and, depending on the outcome of the classification, classifies the video as either relevant or irrelevant. In the former case, the recommender system will attempt to retrieve more similar results, using the meta-information of the perused video clip (Figure 1).

2.3.4. Multimodal Recommender System

The interests of each individual is stored in a profile, which is generated during registration time. Users' interests are dynamic in nature and can change over time. It is, therefore, important to have a system that can accommodate to such changes. Moreover, the efficiency of the system is dependent on the accuracy of captured interests, thus, it is important to break-down interaction into several phases, allowing us to develop a better understanding of these interests and their changes. The first phase of capturing user interests is during query submission. At this point the system is be able to perform recommendations, using the terms that appear in the search query. The profile is be updated each time the user formulates a new query. The second phase occurs during click-through action. We assume that this action is an implicit indicator of interest towards the selected item. The item's meta-data is used as source of information for updating the user profile. These two steps consist our baseline user-profiling technique. In the enriched user-profiling technique the captured facial expressions are treated as an additional source of implicit feedback, which is used to update the user profile. The perused item's meta-data are also used as an additional source of information, along with the positive and negative feedback obtained from the facial expression analysis. A feedback is regarded as positive if the user finds at least one frame interesting during the time he/she is watching a video.

Each of these actions has its own degree of significance. The search query is considered to be the least significant, since users often have problems expressing their initial information need. On the contrary, click-throughs are considered a more important source of feedback, since the users have the opportunity to go through the meta-data and decide whether to view the video clip or not. Finally, any feedback deriving from the facial expression analysis is regarded as the most significant, because it is generated while the users are watching the actual clip. After each feedback cycle, the user profile is updated, following the multimodal approach presented in [20].

2.4. 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. For each task they were given 15 minutes, during which they were asked to bookmark as many relevant videos as possible. One task type was to search for videos that would facilitate the learning of some new skill (e.g., dancing), while another task type was to locate videos of entertainment. For each search task they were given a short cover story, which introduced them to a simulated situation, thus promoting the formulation of better-defined relevance criteria. This story also controlled the scope and domain of the search task. The participants performed one broadly defined task and one focused task. The order of task domains and scopes was rotated to reduce learning effects [21]. An Exit Questionnaire was also administered at the end of each session.

3. RESULTS & DISCUSSION

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 effects of our independent variables are examined with respect to participants' perceived degree of relevance, scaled from 1 (low) to 5 (hight). 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 ($W = 28791.5, p = 0.020$). This finding suggests that the performance of profiling was enhanced by the facial expression data. Note that we used the independent test since participants made several ratings within individual blocks, although the experiment was a within-subject design.

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 ($W = 7550.5$, $p = 0.015$). This suggests that the multimodal system was more effective than the baseline system when tasks involved some form of learning or when tasks involved a wide range of videos. 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 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. Overall, our findings outline the benefits of enriched profiling and the use of facial expression data, and support the design of multimodal recommender systems.

4. CONCLUSIONS

In this work we introduced a novel video search interface that applies real-time facial expression analysis to aggregate information about the affective state of the user. 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 components into user profiling. We have additionally 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.

Our 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. 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.

5. REFERENCES

[1] G. Linden, B. Smith, and J. York, “Amazon.com recommendations: item-to-item collaborative filtering,” *Internet Computing, IEEE*, vol. 7, no. 1, pp. 76–80, 2003.

[2] 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*, NY, USA, 2003, pp. 263–266, ACM.

[3] D. Billsus, C. A. Brunk, C. Evans, B. Gladish, and M. Pazzani, “Adaptive interfaces for ubiquitous web access,” *Commun. ACM*, vol. 45, no. 5, pp. 34–38, 2002.

[4] E. H. Han and G. Karypis, “Feature-based recommendation system,” in *CIKM ’05*, New York, NY, USA, 2005, pp. 446–452, ACM.

[5] G. Adomavicius and A. Tuzhilin, “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions,” *IEEE Trans. on Knowl. and Data Eng.*, vol. 17, no. 6, pp. 734–749, 2005.

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

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

[8] M. Yeasin, B. Bullot, and R. Sharma, “Recognition of facial expressions and measurement of levels of interest from video,” *Multimedia, IEEE Transactions on*, vol. 8, no. 3, pp. 500–508, June 2006.

[9] K. R. Scherer, “What are emotions? and how can they be measured?,” *Social Science Information*, vol. 44, no. 4, pp. 695–729, December 2005.

[10] P. Ekman and H. Oster, “Facial expressions of emotion,” *Annual Review of Psychology*, vol. 30, no. 1, pp. 527–554, 1979.

[11] P. Ekman, *Basic Emotions*, chapter 3, pp. pp. 301–320, Handbook of Cognition and Emotion. John Wiley, 1999.

[12] P. Ekman, *Unmasking the face*, MA: Malor books, Cambridge, 2003.

[13] J. A. Russell and J. H. Steiger, *The Structure in Person’s Implicit Taxonomy of Emotions*, Journal of Research in Personality, 1982.

[14] J. A. Russell and A. Mehrabian, “Evidence for a three-factor theory of emotions,” *Journal of Research in Personality*, vol. 11, no. 3, pp. 273–294, September 1977.

[15] M. Pantic and L. Rothkrantz, “Expert system for automatic analysis of facial expression,” *Image and Vision Computing Journal*, vol. 18, no. 11, pp. 881–905, July 2000.

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

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

[18] P. Borlund and P. Ingwersen, “Experimental components for the evaluation of interactive information retrieval systems,” *JOURNAL OF DOCUMENTATION*, vol. 56, no. 1, pp. 71–90, 2000.

[19] I. Arapakis, J. M. Jose, and P. G. Gray, “Affective feedback: an investigation into the role of emotions in the information seeking process,” in *SIGIR ’08*. 2008, pp. 395–402, ACM.

[20] U. Cetintemel, M. J. Franklin, and C. L. Giles, “Self-adaptive user profiles for large-scale data delivery,” in *ICDE ’00*, Washington, DC, USA, p. 622.

[21] J. Tague-Sutcliffe, “The pragmatics of information retrieval experimentation, revisited,” *Inf. Process. Manage.*, vol. 28, no. 4, pp. 467–490, 1992.