An Affect-Based Multimodal Video Recommendation System

Arturas KAKLAUSKAS, Renaldas GUDAUSKAS, Matas KOZLOVAS, Lina PECIURE, Natalija LEPKOVA, Justas CERKAUSKAS, Audrius BANAITIS

Vilnius Gediminas Technical University,
Sauletekio al. 11, Vilnius, LT-10223, Lithuania,
arturas.kaklauskas@vgtu.lt (*Corresponding author*)

Abstract: People watching a video can almost always suppress their speech but they cannot suppress their body language and manage their physiological and behavioral parameters. Affects/emotions, sensory processing, actions/motor behavior and motivation link to the limbic system responsible for instinctive and instantaneous human reactions to their environment or to other people. Limbic reactions are immediate, sure, time-tested and occur among all people. Such reactions are highly spontaneous and reflect the video viewer's real feelings and desires, rather than deliberately calculated ones. The limbic system also links to emotions, usually conveyed by facial expressions and movements of legs, arms and/or other body parts. All physiological and behavioral parameters require consideration to determine a video viewer's emotions and wishes. This is the reason an Affect-based multimodal video recommendation system (ARTIST), developed by the authors of the article, is very suitable. The ARTIST was developed and fine-tuned during the course of conducting the TEMPUS project "Reformation of the Curricula on Built Environment in the Eastern Neighbouring Area". ARTIST can analyze the facial expressions and physiological parameters of a viewer while watching a video. An analysis of a video viewer's facial expressions and physiological parameters leads to better control over alternative sequences of film clips for a video clips. It can even prompt ending the video, if nothing suitable for the viewer is available in the database. This system can consider a viewer's emotions (happy, sad, angry, surprised, scared, disgusted and neutral) and choose rational video clips in real time. The analysis of a video viewer's facial expressions and physiological parameters can indicate possible offers to viewers for video clips they prefer at the moment.

Keywords: facial expressions; physiological video retrieval; affect-based, multimodal, video recommendation system; TEMPUS CENEAST project.

1. Introduction

Those watching a film rarely speak, but they cannot suppress their body language. Body language is linked to the limbic system responsible for instinctive and instantaneous human reactions to their environment or other people. Such reactions are highly spontaneous and reflect the person's real feelings and desires, rather than calculated ones. The limbic system is also linked to emotions, usually conveyed through facial expressions and movements of legs, arms or other body parts. All this should be considered in determining the viewer's emotions and wishes. The deliberate control of one's body tends to look unnatural: movements fall behind utterances and hardly look genuine.

As someone is watching a film, affective systems can analyse the viewer's gestures, movements, touches, posture, and face and eye expressions. Such observations offer extra information on the person's character, emotions and reactions. The monitoring of a film viewer's facial expressions leads to better control over the sequence of the film's alternative video clips, or can even prompt to end the film if nothing that might suit the viewer is available in the database.

The system can consider the viewer's emotions — happy, sad, angry, surprised, scared, disgusted and a neutral state — and choose rational video clips in real time. With the analysis of the body language, viewers can be offered the video clips they prefer at the moment. Such systems are akin to people with high emotional intelligence. Persons well aware and perceptive of own feelings are better at analysing and spotting those of others and can take a deeper, more meaningful approach to the world and people around them.

Currently videos are indicated as really big data. Venter & Stein (2012), for instance, believe that today video images and image sequences comprise about 80 percent of all corporate and public unstructured big data. Recent advances in multimedia technology have led to tremendous increases in the available volume of video data, thereby creating a major requirement for efficient systems to manage such huge data volumes (Mehmood et al. 2015). With the fast proliferation of multimedia and video display devices, searching and watching videos on the

Internet has become an indispensable part of our daily lives. Many video-sharing web sites offer the service of searching and recommending videos from an exponentially growing repository of videos uploaded by individual users (Niu et al. 2015).

The term big data, when it refers to videos, often defines the exponential growth and availability of videos. The enormous supply of videos, with their numbers growing daily, and the ability of users to choose any video they need make the use of video content and prescriptive analytics a necessity. Different methods and technologies have been proposed globally to handle this task. Venter & Stein (2012), for instance, believe that, as growth of unstructured data increases, analytical systems must assimilate and interpret images and videos as well as they interpret structured data, such as texts and numbers. Prescriptive analytics leverages the emergence of big data and computational and scientific advances in the fields of statistics, mathematics, operations research, business rules and machine learning (Venter, A. Stein 2012). The explosion of user-generated, untagged multimedia data in recent years, generates a strong need for efficient search and retrieval of this data. The predominant method for content-based tagging is through slow, labor-intensive manual annotation. Consequently, automatic tagging is currently a subject of intensive research. However, it is clear that the process will not be fully automated in the foreseeable future (Koelstra and Patras 2013).

Recently annotation according to an affective or emotional video category has been gaining ground (Joho et al. 2011, Hanjalic et. al. 2008, Moncrieff et al. 2001, Calvo & D'Mello 2010, Wang & Cheong 2006).

The main objective is to make the recommendation personalized and situation sensitive. If the affective content of a video is detected, it will be very easy to build an intelligent video recommendation system, which can recommend videos to users based on users' current emotion and interest. For example, when the user is sad, the system will automatically recommend happy movies to him/her; when the user is tired, the system may suggest a relaxing movie (Joho et al. 2011). In general, there are three kinds of popular affective analysis methods. Categorical affective content analysis methods usually define a few basic affective groups and discrete emotions, for example, "happy", "sadness" and "fear". Then classify video/audio to these predefined groups. The second type of affective analysis method is called Dimensional affective content analysis method (for example, the psychological Arousal-Valence (A-V) Affective Model), which commonly employs the Dimensional Affective Model to compute affective state. The third type of affective analysis method is Personalized affective content analysis method (Lu et al. 2011).

Video classification and recommendation based on affective analysis of viewers are aimed at finding interesting and suitable videos for users by using different metadata. Metadata of videos are of two types: (i) non-affective (such as genre, director, actors, etc.) and (ii) affective (expected feeling or emotion). The methods focused on affective video analysis can be divided into two categories according to the method of generation of Affective Metadata (AM) [5]: (i) explicit (asking the user to point out an affective label for the observed video clip), and (ii) implicit (detecting the user's affective response or analyzing the affective video component element automatically) (Niu et al. 2015).

A *personalized* search is the fundamental goal of video content or *prescriptive analytics* aiming to tailor the integration of data, information and knowledge about a user beyond the explicit query precisely to that person's tasks. Niu et al. (2015) point out that the issue of finding videos suited to a user's personal preferences or measuring the similarity between videos poses various challenges.

Video content and *prescriptive analytics* are also gaining ground in the *Internet of Things*. The Affective Tutoring System for Built Environment Management (Kaklauskas et al. 2015), for instance, can track in a student's computer when and where the student was most productive and share that information with the lecturer's computer. How might different stakeholders benefit from this concept as well? Could, for instance, a student video analytics system interact with the computers of lecturers in a university? Would a lecturer wish to know, whether a student taking an examination in the classroom, in close proximity, is cheating? Obviously, a lecturer might want such information, but a student would want to conceal such a fact. Privacy issues continue to

be a major worry in the future. Partial distribution of biometrics could be valuable — and perhaps even essential — to enter a venue.

Researchers worldwide are working on video retrieval and recommendation systems that employ only unimodal affective analysis. However, multimodal video retrieval and recommendation systems are also under development aiming to overcome the limitations of unimodal systems. Many researchers and practitioners, for instance, combine affective video analysis with physiological information and data to analyze videos: Soleymani et al. (2011) use galvanic skin response (GSR), electromyography (EMG), blood pressure, breathing rate and skin temperature; Money & Agius (2008) use GSR, breathing rate, blood volume pulse feedback and heart rate and Koelstra et al. (2013) use electroencephalogram (EEG) and peripheral physiological signals. A brief review of the above-mentioned systems follows.

Viewer's attention is based on multiple sensory perceptions, i.e., aural and visual, as well as the viewer's neuronal signals (Mehmood et al. 2015). Facial expression is one of several modes of nonverbal communication. The message value of various modes may differ depending on context and may be congruent or discrepant with each other. An interesting research topic is the integration of facial expression analysis with that of gesture, prosody, and speech. Combining facial features with acoustic features would help to separate the effects of facial actions due to facial expression and those due to speech related movements (Fox et al. 2003).

Lately interdisciplinary studies (Ringeval et al. 2015, Grafsgaard et al. 2014) have been aiming to develop methods, tools, devices and analytical techniques designed for reliable real-time analysis of emotions from different modalities (physiological signals, audio, and video) and decision making (Filip et al. 2008, 2009, 2014). Achievement of such an aim involves use of physiological sensors, physiological measures, facial expression analysis systems, self-report measures and other tools.

2. An Affect-based, Multimodal, Video Recommendation System

The Affective Multimodal Video Recommendation System for TEMPUS CENEAST Project (ARTIST) consists of the following components (see Figure 1): the Intelligent Database Management System, the Intelligent Database, the Equipment Subsystem, the Model Base Management System, the Intelligent Model Base and the User Interface. The architecture of ARTIST are briefly analysed below.

The Intelligent Database contains the Domain Database, the Universities Contacts Database,

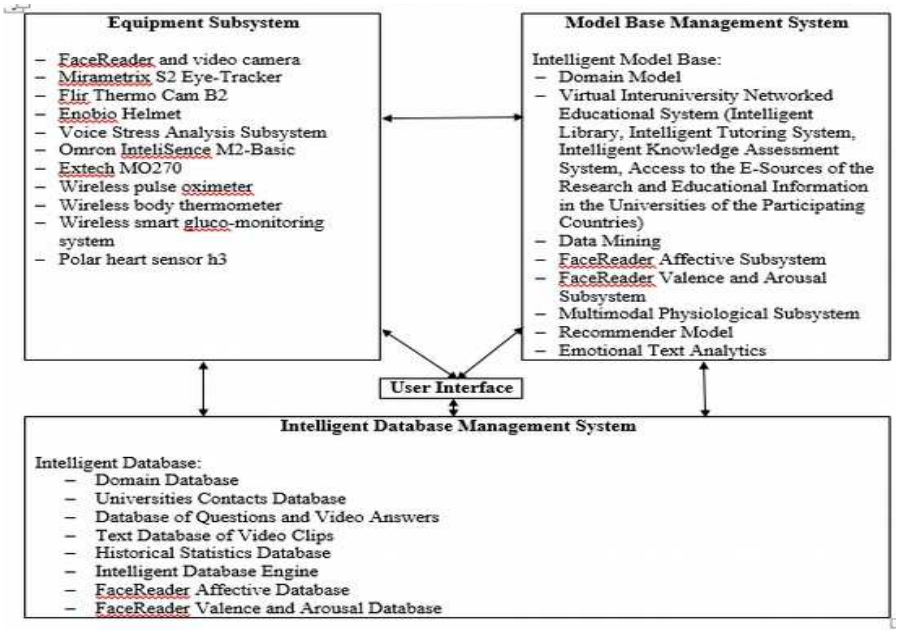

Figure 1. Architecture of the Affect-Based Multimodal Video Recommendation System for the TEMPUS CENEAST Project

the Database of Questions and Video Answers, the Text Database of Video Clips, the Historical Statistics Database, the Intelligent Database Engine, the FaceReader Affective Database, and the FaceReader Valence and Arousal Database.

The objective of the Domain Database was to create a suitable basis for education in the field of the built environment by introducing integrated, multidisciplinary BSc, MSc and PhD modules into existing programmes offered in the participating countries, or PCs (Ukraine, Russia, Belarus). The Domain Database was completed as follows:

- State-of-the-art reports on market needs; required expertise profiles in each university;
- The workshop for upgrading of BSc/specialists, MSc and PhD programmes;
- The development of a common framework for curricula based on a common philosophical and pedagogical understanding between the partner institutions;
- The development of a common approach to teaching and learning activities between PCs to ensure maximum module compatibility, while maintaining institutional and financial autonomy and accountability. Cross-partner good practice sharing, rigorous module verification and this process ensured that each partner was introduced with enhanced quality assurance and better management of teaching and student services;
- The development of the module content and teaching/learning materials suitable for the virtual interuniversity networked educational system. The new knowledge creation and dissemination was triangularised with education (input from the existing module base), innovation (new online delivery and dissemination strategies as described in the virtual interuniversity networked educational system and through industry engagement), and research (through the shared research base across institutions).
- The development of intended learning/training outcomes, assessment criteria and subject content of each module, the identification and development of learning resources (handbooks, lecture plans), the preparation of a strategy for the delivery and dissemination of the modules through the virtual inter-university networked educational system, and the development of a feedback system for continuous module content update with the engagement of the industry, students and academia to maintain the concept of lifelong learning and post-project sustainability.

Users can find more details on the available multidisciplinary BSc, MSc and PhD modules in the Universities Contacts Database; they can also use the new Virtual Interuniversity Networked Educational Centre.

The Database of Questions and Video Answers accumulates questions and video answers to the questions. To create the system of possible questions and answers, as much information is gathered about each of them as possible and possible scripts are presented as decision tree diagrams with logical relations. The system of questions and answers is created taking into account which topics are in great demand among those interested in the programme. Topics can span a wide range, but a challenge or a dilemma is an important component. *Answers were scripted and short clips prepared*. Each answer needs a short script. The scripts are the basis for short clips with rather detailed answers to the questions asked. The clips are then uploaded to a relational database.

The Text Database of Video Clips contains all texts from the video clips in Russian and English and a list of emotional key words such as “happiness” or “joy” for mapping affective video text features or representing different affective states.

The Historical Statistics Database accumulates historical statistical data: the statistical analysis of the questions users pick; the statistical analysis of watched video clips; the emotions that dominated while users watched video clips; and the valence and arousal states that dominated while users watched video clips.

The Intelligent Database Engine consists of two main parts: 1) text mining; 2) determination of the interdependencies between interest of the users under analysis and their physiological indicator. These two composite parts are briefly described next.

Text mining covers the inputting of a bag of concepts space; the selecting, processing and indexing information in accordance with the inputted bag of concepts space; formulating the results of the retrieval and finally showing them to the users. Further, after selecting, processing

and indexing teaching materials, it covers the selecting out of composite parts (chapters/sections/paragraphs) of the teaching materials under analysis and, after that, performing the multi-criteria analysis of the composite parts. This is followed by the designing of alternative variants of the selected information and performing a multi-criteria analysis of the summarized integrated alternatives of the text by which the retrieval results are then formulated. Text mining permits selecting the maximally rational text in the coverage that the user desires.

The following factors determine a rational text (see Figure 2): Citation index of papers (Scopus, ScienceDirect, Google Scholar); Citation of authors (Scopus, ScienceDirect, Google Scholar, etc.); Top 25 papers; Impact factor of journals; Popularity of a text (citation

The following factors determine a rational text:	Publication 1	Publication 2	Publication 3	Publication 4	Publication 5
Citation of papers:					
Citation of papers (Web of Science)	35	62	110	-	127
Top 25 papers	**2**	**12**	**4**	**17**	**5**
Impact factor of journals	2.305	0.75	0.681	1.275	2.797
Density of keywords (% of a text):					
aggression	2.22517176764522	1.73611111111112	1.97120841048922	2.27706526470883	1.08082279057068
violent	3.41583385384135	1.83890716374269	0.815363478884179	0.375046043599103	1.05124237735506
Citation of authors:					
Author 1					
Web of Science					
Sum of the Times Cited	802	48	1374	114	524
Sum of Times Cited without self-citations	652	47	1344	112	520
Citing Articles	643	43	1222	112	493
Citing Articles without self-citations	29	42	1025	110	490
Average Citations per Item	70	9.6	1025	11.4	1278
H-index	9	4	22	6	12
Google					
Citations	433	3956	24680	10	2756
H-index	8	4	84	-	22
i10-index	8	13	322	-	43
Author 2					
Web of Science					
Sum of the Times Cited	897	-	-	49	884
Sum of Times Cited without self-citations	830	-	-	49	834
Citing Articles	790	-	-	49	722
Citing Articles without self-citations	757	-	-	49	-
Average Citations per Item	1495	-	-	8.17	3687
H-index	18	-	-	4	8
Google					
Citations	3476	-	-	6	2995
H-index	33	-	-	-	14
i10-index	70	-	-	-	16
Author 3					
Web of Science					
Sum of the Times Cited	-	-	-	911	-
Sum of Times Cited without self-citations	-	-	-	871	-
Citing Articles	-	-	-	785	-
Citing Articles without self-citations	-	-	-	771	-
Average Citations per Item	-	-	-	35.04	-
H-index	-	-	-	14	-
Google					
Citations	-	-	-	26	-
H-index	-	-	-	-	-
i10-index	-	-	-	-	-
Citation of papers (ScienceDirect)	-	62	-	-	-
Citation of papers (Google Scholar)	-	-	-	132	-

Figure 2. Fragment of quantitative parameters explaining selections of the most rational paragraphs

index, number of readers, time spent reading); Reputation of the documents; Supporting phrases; Document name and contents; Density of keywords. Text mining can select the desired number of pages in accordance to the assigned keywords and their significances (for example, 9, 41 and 187 pages). Additionally a user can assign a number of minutes for reading the information of interest. Text mining was developed as a Web application using Microsoft Visual Studio 2010, C# as the main programming language and the MS SQL Server 2012 as a database platform.

The inter-dependencies between interest of the users under research and their physiological indicators are determined by Data analytics. The methods used for this purpose were Ordered Logit (regression model for ordinal dependent variables), Neural Networks and Anova (analysis of variance).

The data obtained by the analysis of spectator's face with the help of FaceReader (2014) emotions and FaceReader valence and arousal sub-systems are placed in the databases of FaceReader Affective, FaceReader Valence and Arousal.

All the data in the database is stored in tables and organized in relational database principles. It is used a typical relational Intelligent Database Management System.

The Equipment Subsystem consists of the FaceReader and video camera, Mirametrix S2 Eye-Tracker (MRS2), Flir Thermo Cam B2, Enobio Helmet (wearable and wireless electrophysiology sensor system for recording EEG), Voice Stress Analysis Subsystem, Wireless blood pressure monitor (Omron InteliSence M2-Basic), Wireless Moisture Meter (Extech MO270), Wireless pulse oximeter, Wireless body thermometer, Wireless smart glucomonitoring system and Polar heart sensor h3.

FaceReader was integrated into the smart video. FaceReader is a program for facial analysis. It can detect emotional expressions the face can manifest. It can identify six basic emotions: happy, sad, angry, surprised, scared, disgusted, and a neutral state. FaceReader also analyzes the valence, which indicates whether the person's emotional state is positive or negative, and arousal, which indicates how active the person is. FaceReader detects viewer emotions in real time. If the viewer's emotions show dissatisfaction, the current short clip is skipped and the next one played. Clips can be also skipped by the viewer.

The Intelligent Model Base consists of the following models: Domain Model; Virtual Interuniversity Networked Educational System (Intelligent Library, Intelligent Tutoring System, Intelligent Knowledge Assessment System, Access to the E-Sources of the Research and Educational Information in the Universities of the Participating Countries); Data Mining; FaceReader Affective Subsystem; FaceReader Valence and Arousal Subsystem; Multimodal Physiological Subsystem; Recommender Model; Emotional Text Analytics.

The ARTIST's Intelligent Model Base and its management subsystem use standard Microsoft Framework 4.0 and SQL Server 2008 components.

A detailed discussion of the components comprising the Intelligent Model Base appears below.

The Domain Model presents frames to the user. The Domain Model consists of the 16 modules (9 BSc/specialists, 5 MSc and 2 PhD) and 54 computer learning systems.

The Virtual Interuniversity Networked Educational Centre (see http://iti.vgtu.lt/tempus/) delivers modules from the Domain Database. In addition, this centre promotes lifelong learning in the society at large by making study materials accessible outside the traditional classroom environment to various parties within the society: from students to teachers to practitioners and policy makers. The centre ensures not only the feed-forward (information/knowledge from the centre to the beneficiaries) but also feedback (from beneficiaries to the centre). It is expected that a spiral effect will be created to drive continuous improvement of the centre. The centre comprises four major components: the Intelligent Library, the Intelligent Tutoring System, the Student Knowledge Assessment System, and the Virtual Research Environment. The centre addresses regional and national higher education priorities such as the development of international relations, enhanced quality assurance, the management of teaching and student services and triangulated knowledge creation and dissemination with education, innovation and research. Already,

training courses for the staff have been implemented. Also, 240 students were trained during the pilot project.

The *Data Mining* automatically evaluate the user's interest in watching by using Ordered Logit [regression model for ordinal dependent variables], KNN and Anova [analysis of variance]. The Data Mining aims to define user personalized learning quantitatively. The Data Mining aims to define personalized user learning quantitatively. The Data Mining gathers and analyzes the following information:

- Historical statistical data defining user interests: the statistical analysis of the questions users pick; the statistical analysis of watched video clips; the emotions that dominated while users watched video clips; and the valence and arousal states that dominated while users watched video clips.
- Biometrical information. The Model is measuring and establishing various multi-modal, physiological parameters of users (systolic and diastolic blood pressures, skin moisture and conductivity, temperature, pulse rate and changes in the eye pupil and blinking) depending on their interest in watching and establishing the reliability of these dependencies based on LOGIT, KNN and Anova methods.

The user gets a questionnaire and ticks any questions of interest (see Figure 3: http://smartvideo.vgtu.lt/web/index.php?r=site%2Fsurvey&id=49%E2%80%8B). After clicking "Save", the viewer will see a sequence of selected video clips. The system shows video clips preassigned to the questions the user has selected.

Three integrated subsystems determine whether specific selected video clips will be shown:

- FaceReader Affective Subsystem. FaceReader uses a webcam to detect the viewer's emotions (happy, sad, angry, surprised, scared, disgusted and a neutral state) every five seconds (see Figure 4). If the overall emotional state the subsystem detects is more negative than positive, it suggests

Figure 4. FaceReader results of the emotions subsystem analysis

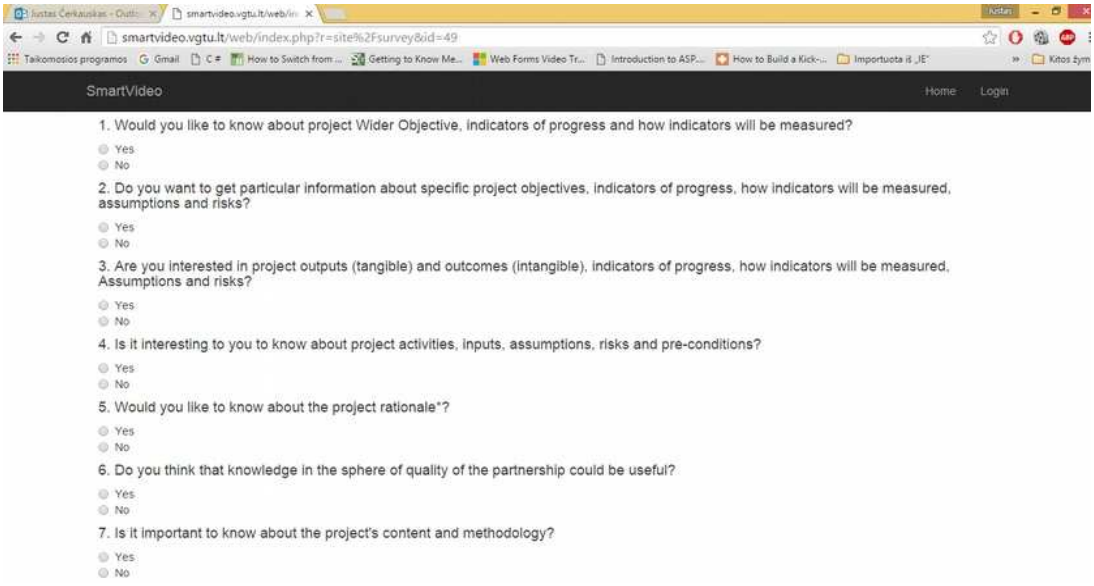

Figure 3. Personalized questionnaire

skipping to the next video clip, and vice versa — if the emotional state is more positive than negative, the subsystem suggests continuing with the current video clip.

- FaceReader Valence and Arousal Subsystem. This subsystem detects the viewer's valence and arousal levels every five seconds (see Figure 5). The valence indicates whether the user's emotional status is positive or negative, and arousal indicates how active the user is. According to Lang et al. (1993), the dimension of valence ranges from highly positive to highly negative, whereas the dimension of arousal ranges from calming or soothing to exciting or agitating. Thus, there can be events that are negative and agitating; positive and soothing; positive and exciting; etc. (Kensinger 2004). If the subsystem determines emotions with high arousal and high valence (excited, astonished, delighted, happy, pleased), it suggests continuing with the current video clip, but if the emotions are tending towards negative (for example, low valence and low arousal: miserable, depressed, bored, tired), it suggests moving to the next video clip.

Figure 5. FaceReader results of the valence and arousal subsystem analysis

- Multimodal Physiological Subsystem. It can record a user's multimodal physiological responses (heart rate, systolic and diastolic blood pressure, skin humidity, perspiration, temperature and conductance, VSA, pupil size, EEG) while the user is engaged in watching. In their previous studies (Kaklauskas et al. 2010, 2015), the authors determined dependencies linking interest and the above multimodal physiological parameters. If the subsystem determines that the video clip interests the viewer, it suggests continuing watching, if not, then it suggests skipping to the next video clip.

If two or three of the above subsystems suggest showing specific selected video clips, the system will show them; in other cases it will skip to the next video clip. The process is repeated until the last selected video clip ends or the user turns the system off manually.

The Recommended model, which these authors are proposing herein, is designed to accumulate data and generate feedback to users. Meanwhile the authors of the article may analyse areas that need improvement with the ARTIST's recommender model. The authors learn which parts of the questionnaire and video clips need improvements: by gathering suggestions from users and by analysing the data on users' answers to the questions. ARTIST is able to offer more interesting video clips, as an alternative, upon analyzing the biometric parameters of a user on his/her interest. Users concerned about the quality of watching may: suggest supplemental video clips, which they believe would explain individual topics or subtopics more comprehensively; suggest new topics or subtopics, as additions to the existing ones; suggest rephrasing the questions to make it more intelligible.

The administrator receives all this information together with a user's video clips watching statistical results. It is up to the administrator to use the user's suggestions in practice. The administrator considered the statistical information as well as the recommendations users provided and made use of the users' suggestions in the following ways: added the additional video suggested by the users to the Database of Questions and Video Answers; added the suggested new question and video clips answer; rephrased the question.

The Emotional Text Analytics can analyse the text from the video clips in Russian and English stored in the text database by typical emotional key words. Such text analytics determines the affective degree of text. This

analytics is then used to select video clips by their affective degree (see Figure 6).

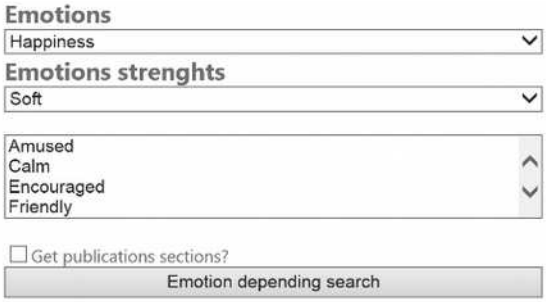

Figure 6. Fragment of the emotional text analytics

3. Discussion and the Conclusions

The smart video about the TEMPUS project 530603-TEMPUS-1-2012-1-LT-TEMPUS-JPCR (hereinafter ARTIST) highlights the possibilities the ARTIST opens. It is a personalised attempt to entice and attract prospective BSc, MSc and PhD students to the modules and the virtual interuniversity networked educational system (intelligent library, intelligent tutoring system, intelligent knowledge assessment system, access to e-sources with research and educational information available in the universities of the participating countries) developed earlier, to promote the project adequately, and to make prospective students aware of it. The plans for the next stage of the ARTIST's development involves integrating this system with other intelligent voice and IRIS analysis systems, which the authors herein have also developed. Such an integration of intelligent and physiological systems would provide even better assessments of the users and the submissions of the best video clips to them.

Acknowledgements

This research was funded largely by the project "Reformation of the Curricula on Built Environment in the Eastern Neighbouring Area" (No. 530603-TEMPUS-1-2012-1-LT-TEMPUS-JPCR) of EU programme Tempus IV (2007 – 2013), Action 1: Joint Projects (JP) (see http://www.ceneast.com/).

REFERENCES

1. CALVO, R., S. D'MELLO, **Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications**, IEEE Trans. on Affective Computing, vol. 1(1), 2010, pp. 18-37.
2. FACEREADER, **Reference Manual Version 6. Tool for Automatic Analysis of Facial Expressions.** Noldus Information Technology, 2014, p. 183.
3. FILIP, F. G., **Decision Support and Control for Large-scale Complex Systems**, Annual Reviews in Control, vol. 32(1), 2008, pp. 61-70.
4. FILIP, F. G., A. SUDUC, M. BÎZOI, **DSS in Numbers**, Technological and Economic Development of Economy, vol. 20(1), 2014, pp. 154-164.
5. FILIP, F. G., K. LEIVISKÄ, **Large-Scale Complex Systems.** Springer Handbook of Automation. 2009, pp. 619-638.
6. FOX, N., R. GROSS, P. DE CHAZAL, J. COHN, R. REILLY, **Person Identification using Multi-modal Features: Speech, Lip, and Face.** in Proc. of ACM Multimedia Workshop in Biometrics Methods and Applications (WBMA 2003), CA, 2003.
7. GRAFSGAARD, J. F., J. B. WIGGINS, K. E. BOYER, E. N., WIEBE, J. C., LESTER, **Predicting Learning and Affect from Multimodal Data Streams in Task-oriented Tutorial Dialogue.** In: Stamper, J., Pardos, Z., Mavrikis, M., McLaren, B. M. (Eds.), Proceedings of the 7th international conference on educational data mining, London, England: International Educational Data Mining Society, 2014, pp. 122-129.
8. HANJALIC, A., R. LIENHART, W. Y. MA, J. R. SMITH, **The Holy Grail of Multimedia Information Retrieval: So Close or Yet So Far Away?** Proceedings of the IEEE, vol. 96(4), 2008, pp. 541-547.
9. JOHO, H., J. STAIANO, N. SEBE, J. M. JOSE, **Looking at the Viewer: Analysing Facial Activity to Detect Personal Highlights of Multimedia Contents.** Multimedia Tools and Applications, vol. 51(2), 2011, pp. 505-523.
10. KAKLAUSKAS, A., A. KUZMINSKE, E. K. ZAVADSKAS, A. DANIUNAS, G. KAKLAUSKAS, M. SENIUT, J. RAISTENSKIS, A. SAFONOV, R. KLIUKAS, A. JUOZAPAITIS, A.

RADZEVICIENE, R. CERKAUSKIENE,). **Affective Tutoring System for Built Environment Management**. Computers & Education, vol. 82, 2015, pp. 202-216.

11. KAKLAUSKAS, A., E. K. ZAVADSKAS, V. PRUSKUS, A. VLASENKO, M. SENIUT, G. KAKLAUSKAS, A. MATULIAUSKAITE, V. GRIBNIAK, **Biometric and Intelligent Self-assessment of Student Progress System**, Computers & Education, vol. 2010, pp. 821-833.
12. KAKLAUSKAS, A., E. K. ZAVADSKAS, V. PRUSKUS, A. VLASENKO, L. BARTKIENE, R. PALISKIENE, L. ZEMECKYTE, V. GERSTEIN, G. DZEMYDA, G., TAMULEVICIUS, **Recommended Biometric Stress Management System**. Expert Systems with Applications, vol. 38(11), 2011, pp. 14011-14025.
13. KENSINGER, E. A. **Remembering Emotional Experiences: The Contribution of Valence and Arousal**. Reviews in the Neurosciences, vol. 15(4), 2004, pp. 241-252.
14. KOELSTRA, S., I. PATRAS, **Fusion of Facial Expressions and EEG for Implicit Affective Tagging**. Image and Vision Computing, vol. 31(2), 2013, pp. 164-174.
15. LANG, P. J., M. K. GREENWALD, M. M. BRADLEY, A. O. HAMM, **Looking at Pictures: Affective, Facial, Visceral, and Behavioural Reactions**, Psychophysiology, vol. 30, 1993, pp. 261-273.
16. LU, Y., N. SEBE, R. HYTNEN, Q. TIAN, **Personalization in Multimedia Retrieval: A Survey**. Multimedia Tools and Applications, vol. 51, 2011, pp. 247-277.
17. MEHMOOD, I., M. SAJJAD, S. RHO, S. W. BAIK, **Divide-and-Conquer based Summarization Framework for Extracting Affective Video Content.** Neurocomputing, in Press, Corrected Proof.
18. MONCRIEFF, S., C. DORAI, S. VENKATESH, **Affect Computing in Film through Sound Energy Dynamics.** In: ACM International Conference on Multimedia, 2011.
19. MONEY, A. G., H. AGIUS, **Video Summarisation: A Conceptual Framework and Wurvey of the State of the Art.** Journal of Visual Communication and Image Representation, vol. 19(2), 2008, pp. 121-143.
20. NIU, J., X. ZHAO, M. A. ABDUL AZIZ, **A Novel Affect-based Vodel of Similarity Measure of Videos**. Neurocomputing, In Press, Corrected Proof, (2015).
21. RINGEVAL, F., F. EYBEN, E. KROUPI, A. YUCE, J.-P. THIRAN, T. EBRAHIMI, D. LALANNE, B. SCHULLER, **Prediction of Asynchronous Dimensional Emotion Ratings from Audiovisual and Physiological Data.** Pattern Recognition Letters, vol. 66, 2015, pp. 22-30.
22. SOLEYMANI, M., M. PANTIC, T. PUN, **Multi-modal Emotion Recognition in Response to Videos.** Affective Computing, IEEE Transactions on, vol. 3(2), 2011, pp. 211-223.
23. SONG, M., M. YOU, N. LI, C. CHEN, **A Robust Multimodal Approach for Emotion Recognition.** Neurocomputing, vol. 71(10-12), 2008, pp. 1913-1920.
24. VENTER, F., A. STEIN, **Images & Videos: Really Big Data**. The Institute for Operations Research and the Management Sciences (INFORMS), 2012.
25. WANG, H., L. CHEONG, **Affective understanding in film.** IEEE Transactions on Circuits and Systems for Video Technology, vol. 16(6), 2006, pp. 689-704.