Crowdsourcing for Affective Annotation of Video : Development of a Viewer-reported Boredom Corpus

Soleymani, Mohammad; Larson, Martha

Crowdsourcing for Affective Annotation of Video: Development of a Viewer-reported Boredom Corpus

Mohammad Soleymani
Computer Vision and Multimedia Laboratory
University of Geneva
mohammad.soleymani@unige.ch

Martha Larson
Multimedia Information Retrieval Lab
Delft University of Technology
m.a.larson@tudelft.nl

ABSTRACT
Predictions of viewer affective response to video are an important source of information that can be used to enhance the performance of multimedia retrieval and recommendation systems. The development of algorithms for robust prediction of viewer affective response requires corpora accompanied by appropriate ground truth. We report on the development a new corpus to be used to evaluate algorithms for prediction of viewer-reported boredom. We make use of crowdsourcing in order to address two shortcomings of previous affective video corpora: small number of annotators and gap between annotators and target viewer group. We describe the design of the Mechanical Turk setup that we used to generate the affective annotations for the corpus. We discuss specific issues that arose and how we resolve them and then present an analysis of the annotations collected. The paper closes with a list of recommended practices for the collection of self-reported affective annotations using crowdsourcing techniques and an outlook on future work.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms
Algorithms, Measurement, Design, Experimentation,

Keywords
Affective computing, multimedia benchmarking, internet video

1. INTRODUCTION
Developing video processing algorithms capable of predicting viewer boredom requires suitable corpora for development and testing. This paper reports on the development of the MediaEval 2010 Affect Task Corpus for boredom prediction of Internet video. Standard limitations on viewer affective response annotation are overcome by making use of crowdsourcing. Using Mechanical Turk (MTurk), we rapidly gather self-reported boredom scores from a large user group that is demographically diverse and also represents our target population (Internet video viewers).

Ultimately, our boredom-prediction algorithms will be used to improve multimedia retrieval and recommendation. Relatively little research has investigated topic-independent factors that contribute to the relevance of multimedia content to the user information need. In the area of text-based retrieval, incorporation of quality information has been used to improve results, as, for example in [14]. Our larger goal is to extend such techniques to multimedia information retrieval and recommendation.

Workshop on Crowdsourcing for Search Evaluation
SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

We focus on viewer affective response, and in particular on boredom, as a reflection of perceived video quality. We are also interested in variation of affective response among viewers that will help us to develop recommendation and retrieval systems that incorporate information on personal preference.

Our starting point is a set of specifications that our corpus was required to fulfill. The annotation process needed to control as much as possible for extraneous effects, such as reaction of the annotators to the topic of the video, tiredness or underlying mood of the annotators. We wanted to have a relatively large number of annotators for each video, but also a certain number of annotators who annotated the whole collection. We wanted to avoid violating copyright law in order to be able to license our corpus for public use the MediaEval 2010 benchmark. Finally, we had limited resources to invest in corpus development. After a short section on to related work, this paper describes the MediaEval 2010 Affect Task and then the MTurk task that was used to annotate the Affect Task Corpus. We discuss how we fulfilled the specifications of the corpus and met other challenges arising along the way. Finally, we present an analysis of the collected annotations and we distill our experience into a list of recommendations for using crowdsourcing for viewer affective response annotation.

2. RELATED WORK
There are two notable efforts by psychologists to create standard affective video corpora for emotional studies [8][6]. In both studies, movie excerpts extracted from Hollywood movies were used. Because only the time codes of the excerpts and their description are published, the datasets are difficult to re-use. Moreover, use of copyrighted video material depends on the regulations of individual countries. In general, it cannot be shared between researchers or shown to the public for purposes of conducting experiments, gathering annotations or demonstrating systems.

The research in the field of multimedia content analysis for affective understanding of videos lacks significant user studies and only relies on the feedback from limited number of participants [5][10][13]. Multimedia corpora with affective annotations make it possible to investigate interesting research questions and develop useful algorithms, but are time-consuming to generate. The number of participants contributing annotations is a significant factor that limits their usefulness. We describe the 2009 Affect Task in the VideoCLEF (now called MediaEval) benchmark [5] as an example of such a case. The 2009 Affect Task involved narrative peak detection – automatic identification of points within a video at which users experience a heightened sense of dramatic tension. Narrative peak detection is related to highlights detection in sports videos cf. [2], but cannot rely on the presence of audience reaction (the roar of the crowd) in the video. The 2009 Affect Task corpus contains 45 eight-minute videos that are

documentaries on the visual arts, hosted by a well-known Amsterdam professor, Henk van Os. Three assessors watched each video in its entirety and marked the start and end points of the segments that they identified to be the top three narrative peaks. The annotation is necessary time-consuming. In order to understand peaks against the background of their narrative context, it is necessary to watch the video as a whole. Generally, annotating the videos took 2-3 times the run-time of the video. In the 45 videos, there were only 22 peaks that all three assessors identified as among the top three. Although the agreement might have been higher, had we examined a longer top-N list, the annotations generated strongly suggest that there is a personal component determining where viewers perceive narrative peaks. In order to gain a deeper understanding of this component it would be necessary to have more than 3 assessors watch the entire video set. Moreover, assessors reported a familiarity effect. Their sensitivity to narrative peaks developed the more of Prof. van Os' material that they watched. The familiarity effect seemed to be related to a better understanding and appreciate the narrator's style, e.g., sense of humor. More annotations are necessary in order to understand better how affective response changes or develops with familiarity.

To our knowledge there has been only one effort to gather online affective annotations with a large set of participants [11]. During that study more than 1300 annotations from 40 volunteer participants were gathered for 155 video excerpts extracted from Hollywood movies. Although the number of participants is among the largest population size in its kind, the dataset is not redistributable due to the copyright violations issues. The participants who usually volunteer to participate in academic studies are from a certain age group and limited geographical locations or cultural background. In the current dataset, both copyright problems and population size and diversity are addressed.

3. AFFECT PREDICTION TASK

The MediaEval 2010 Affect Task involves automatically predicting the level of user boredom for a video. The Affect Task is running in 2010 within the MediaEval benchmarking initiative [5], which offers tasks to the multimedia research community that help consolidate and synchronize research effort and concentrate it on forward-looking, challenging research areas. Research groups build systems that predict affect and test them on the Affect Task Corpus. For the purpose of the Affect Task and related research, we adopt a fairly simply definition of boredom. We take boredom to be related to the viewer's sense of keeping focus of attention and to be related to the apparent passage of time [4]. We understand boredom to be a negative feeling associated with viewer perceptions of the viewer-perceived quality (viewer appeal) of the video being low.

We are interested in studying two aspects of viewer-reported boredom. First, the 2010 Affect Task corpus will be used to investigate universal aspects of user boredom. On the Internet, certain videos emerge as being more popular than others (as reflected by views, links or viewer-contributed ratings). This popularity can be taken as a reflection of an underlying consensus of an inherent quality of the video, i.e., in some sense it is "worth watching." If this quality is at least in part related to the video content, then we believe that it is worthwhile investigating the extent to which it can be predicted using automatic methods. We know that Internet videos differ not only in subject material, but also with respect to other factors. Among the factors that influence the creation process of a video are: novelty of videographic style, resources available, production skill of the film maker and amount of care invested in planning and realization.

Second, the corpus will be used to investigate personal variation. Affective reaction to video content differs widely from viewer to viewer. We are interested in determining if it is possible to build user-specific models for prediction of self-reported boredom. Additionally, we would like to investigate whether affective reaction can be modeled at a level between the universal and the personal. In other words, we would like to determine whether predictive models can be built for certain groups of users.

The dataset selected for the corpus is Bill's Travel Project, a travelogue series called "My Name is Bill" created by the film maker Bill Bowles (http://www.mynameisbill.com/). The series consists of 126 videos between two to five minutes in length. This data was chosen since it represents the sort of multimedia content that has risen to prominence on the Internet. Bill's travelogue follows the format of a daily episode related to his activities and as such is comparable to "video journals" that are created by many video bloggers. We believe that results of investigations on Bill's Travel Project will extend to other video bloggers, and also perhaps to other sorts of semi-professional user generated video content. Because we are interested in aspects of the data that are independent of topic and genre, we were careful to choose data related to the same topic (travel) and genre (video blog). Further, the fact the video predominantly involves only a single speaker (Bill) helps to abstract away from personal preferences of the viewer that might be based on the gender or appearance of the central figure(s) rather than on the content of the video. The focus is kept squarely on pacing, narrative devices and manner of presentation. Finally, since the video is not Creative Commons licensed we contacted Bill, who kindly granted us permission to use it for the Affect Task. In this way, we were able to develop the corpus without concerns about copyright violations.

The relationship of the 2010 boredom prediction task to the 2009 narrative peak predication task also requires a note of explanation. We would like to investigate if there is a relationship between affective reactions within the video (i.e., their magnitude and timing) to the overall appeal of the video for users. As a result of the experience with the creation of the 2009 corpus, in 2010, we will be investigating possible "familiarity" effects in viewer-reported affective response. In other words, we are interested in whether there is a trend in viewer's reactions to Bill's videos as they grow more acquainted with his material. Specifically, we would like to know whether viewers report increasing boredom as they watch more of Bill's material or whether we find evidence a "fan of Bill" effect, namely, that they report less boredom with growing familiarity with Bill, his journey and his personal style.

The participants carrying out the Affect Task in MediaEval 2010 are various international research groups involved in multimedia information retrieval and affective computing research. The groups are free to design their own algorithms for automatic boredom detection and can make use of features derived from the visual channel, audio channel or speech recognition transcripts. Speech recognition transcripts were supplied with the corpus and generously donated to the benchmark by ICSI and SRI International [12]. Groups approach the tasks in multiple ways. Generally, they first formulate an idea of what properties of the video contribute to user perceived boredom and then build a model that captures these properties. In a typical model, the focus is on properties of the video related to production, for example the cutting or audio mixing, but they also include a wide range of factors.

We were also able to formulate theories about the sources of possible viewer interest in the video by interviewing Bill Bowles concerning the strategies that he makes use of as a film maker to add interest to his videos. In particular he mentioned, that he keeps shots short (< 1 minute), he varies the rhythm of the shot length, he doesn't make the videos any longer than necessary and he varies between close ups and distance shots. Finally, and perhaps presenting the biggest challenge to capture in an automatic algorithm, he attempts to continuously surprise his viewer with a novel approach to his subject material. For example, he switches his role (e.g., between observer, interviewer and commentator) and uses word play and comic devices. Bill also mentioned that how he makes the video is affected by his own mood at the time. This point is not relevant for the Affect Task, which deals with viewer-reported mood, but is an interesting vista for future work.

4. DESIGN OF CROWDSOURCING TASK

We approached the design of the MTurk task by first reading crowdsourcing literature, for example [3], searching for information the Internet on the subject of using MTurk and reflecting on our past experience collecting annotations online. We decided for a two-step approach. The first step was the pilot that consisting of a single micro-task (HIT) involving one video would be used for the purposes of recruiting and screening MTurk users (referred to as "workers"). The second step was the *main task* and involved a series of 125 micro-tasks, one for each of the remaining videos in the collection. We discuss each step in turn.

4.1 Pilot

The pilot contained three components corresponding to qualities that we required of our recruits. The first section contained questions about the personal background (age, gender, cultural background). We made judicious use of MTurk's ability to block workers from certain countries in order to maintain the overall balance. The second section contained questions about viewing habits: we asked the workers if they were regular viewers of Internet video. The third section tested their seriousness by asking them to watch the video, select a word that reflected their mood at the moment and also write a summary. The summary constituted a "verifiable" question, recommended by [3]. The summary offered several possibilities for verification. Its length and whether it contained well-formulated sentences gave us an indication of the level of care that the worker devoted to the HIT. Also, the descriptive content indicated to us whether the worker had watched the entire video, or merely the beginning. We also checked seriousness by ensuring that workers did not complete the HIT faster than the run-time length time of the video. A final question enquired if they were interested in further HITs of the same sort. We were interested in deflecting the attention of the worker away from the main goal of the task, i.e., collecting affective annotations. For this reason we placed the summary box prominently in the HIT. We also believe it was an effective distracter since it was the element of the HIT that was the longest and most intellectually challenging to answer.

4.2 Main Task

We chose the workers for the main task from the participants of the pilot by considering the quality of their description and choosing a diverse group of respondents. The qualification was only granted to the participants who answered all the questions completely. We invited workers to do the main study by sending them an invitation e-mail invitation via their ID number on the MTurk platform. The e-mail informed the users that we had assigned them our MTurk qualification. Use of a qualification serves to limit those workers that carry out the HIT to invitees only. Each HIT in the main study consisted of three parts. In the first part, the workers were asked to specify the time of day, which gave us a rough estimate of how tired they were. Also the workers were asked to choose a mood word from a drop down list that best expressed their reaction to an imaginary word, such as those used in [7]. The mood words were *pleased*, *helpless*, *energetic*, *nervous*, *passive*, *relaxed*, and *aggressive*. These questions gave us an estimate of their underlying mood. In the second part, they were asked to watch the video and give some simple responses. They were asked to choose the word that best represented the emotion they felt while watching a video from a second list of emotion words in the drop down list. The emotion list contained Ekman six basic emotions [1], namely, *sadness*, *joy*, *anger*, *fear*, *surprise*, and *disgust*, in addition to *boredom*, *anxiety*, *neutral and amusement*, which cover the entire affective space, as defined by the conventional dimensions of valence and arousal [9]. The emotion and mood word lists contained different items, which were intended to disassociate them for the user. Next, they were asked to provide a rating specifying how boring they found the video and how much they liked the video, both on a nine point scale. Then, they were asked to estimate how long the video lasted. Here, we had to rely on their full cooperation in order not to cheat and look at the video timeline. Finally, they were asked to describe the contents of the video in one sentence. We emphasized the description of the video rather than the mood word or the rating, in order to conceal the main purpose of the HIT. Quality control of the responses was carried out by checking the description of the video and also by ensuring that the time that they took to complete the HIT was reasonable.

4.3 Issues and solutions

The most important issue with the MTurk task arose because we needed each worker to finish all 125 videos. In the invitation to the main task we named the total sum workers would earn by completing all 125 HITs as an enticement, but we also mentioned that we would only accept the HITs if they completed all 125. Approximately half of the workers we invited to do the task responded positively to this arrangement. Many wrote personal e-mails with specific questions or asking for assurances from our side that we would accept their HITs. The personal communication with the workers was a key factor in collecting the annotations. We were surprised at workers' willingness to give up their anonymity by writing us e-mails and also revealing to us their worker IDs. Many also mentioned their base location in their e-mails. This evolving openness gave us more confidence in trusting the original demographic information collected in the pilot, since by revealing their identities the workers showed themselves willing to provide us with the opportunity to verify at least some of the personal information provided in the pilot. We noticed that many workers were not willing to make the commitment to do all 125 HITs. Building trust was very important. It quickly became clear that some workers were reluctant to risk starting on the series out of fear that we would reject their hits and ruin their reputations on MTurk. Receiving the payment seemed to be secondary. We noticed that at least one person really appreciated that completing the whole series gave them a substantial goal to work for and that the sum that they earned could then be used to buy a particular book. Personal communication via e-mail was essential when the video server that we were using developed a technical problem and the videos did not load. We fielded many e-mails on

those days, and on the whole were surprised at the patience that and cooperative spirit of the workers in the face of the problems.

5. ANALYSIS OF ANNOTATIONS

Our pilot HIT was initially published for 100 workers and finished in the course of a single weekend. We re-published the HIT for more workers when we realized we needed more people in order to have an adequate number of task participants. Only workers with the HIT acceptance rate of 95% or higher were admitted to participate in the pilot HIT. In total, 169 workers completed our pilot HIT, 87.6% of which reported that they watch videos on the Internet. We took this response as confirmation that our tasks participants were close to the target audience of our research. Out of 169 workers, 105 were male and 62 were female and two did not report their gender. Their age average was 30.48 with the standard deviation of 12.39. The workers in the pilot HITs identified themselves by different cultural backgrounds from North American, Caucasian to South and East Asian. Having such a group of participants with a high diversity in their cultural background would have been difficult without using the crowdsourcing platforms. Of the 169 pilot participants, 162 had interest in carrying out similar HITs. Of the interested group, the 79 workers were determined to be qualified and assigned our task-specific qualification within MTurk. This means only 46.7% of the workers who did the pilot HIT were able to answer all the questions and had the profile we required for the main task.

In total, 32 workers have participated and also annotated more than 60 of the 125 videos in the main task HIT series. This means only 18.9% of the participants in the pilot and 39.0% of the qualified participants committed to do the main task HIT series seriously. Of this group of 32 serious participants, 18 are male and 11 are female with ages ranging from 18 to 81 (average 34.9; standard deviation 14.7).

To evaluate the quality of the annotations, the time spent for each HIT was compared to the video length. In 81.8% of the completed HITs the working duration for each HIT was longer than the video length. This means that in 18.2% of the HITs the workers did not follow the instructions. Also, their reported perception of the time is invalid. This shows the importance of having workers with the right qualifications and trustworthy pool of workers in annotation or evaluation hits. Even after the pilot task and disqualifying 60% of the first participants, 16 participants or 39.0% of our final pool did not watch at least 10% of their submitted HITs' videos completely. Rejecting those HITs reduced the number of workers who carried out more than 60 videos in the main series of HIT to 25 from which 17 are male and 8 are female ages ranging from 19 to 59 (average 33.9, standard deviation 11.8).

Three questions were asked about each video to assess the level of boredom. First, how boring the video was on nine-point scale from the most to the least boring. Second, how much the user liked the video on the nine-point scale and third how long the video was. Boredom was shown to have on average a strong negative correlation, $\rho = -0.86$ with liking scores. The time perception did not show a significant correlation for all users and it varied from 0.4 down to -0.27. Although positive correlation was expected from boredom scores and the perception of time seven participants' boredom scores have negative correlation with the time perception.

The correlation between the order of watching the videos for each participant and the boredom ratings was also examined. No positive linear correlation was found between the order and boredom score. This means that watching more videos did not increase the level of boredom and in contrary for 2 of participant it decreased their boredom. Additionally, the correlation between the video length and boredom scores was investigated. No positive correlation was found between the boredom scores and videos' duration. We can conclude that the lengthy videos are not necessarily perceived as more boring than the shorter videos.

To measure the inter-annotator agreement, the Spearman correlation between participants' pairwise boredom scores was computed. The average significant correlation coefficient was very low $\rho = 0.05$. There were even cases where the correlation coefficients were negative, which shows complete disagreement between participants. For each worker we then grouped videos into two rough categories, above and below the mean boredom score of that worker. We computed the average pair-wise Cohen's kappa for these categories and here found only slight agreement ($\kappa = 0.01$). We also compared agreement on the emotion words workers associated with viewers. Here, again Cohen's kappa indicated only slight agreement ($\kappa = 0.07$). The strong correlations suggest that it is indeed important to investigate personalized approaches to affective response prediction.

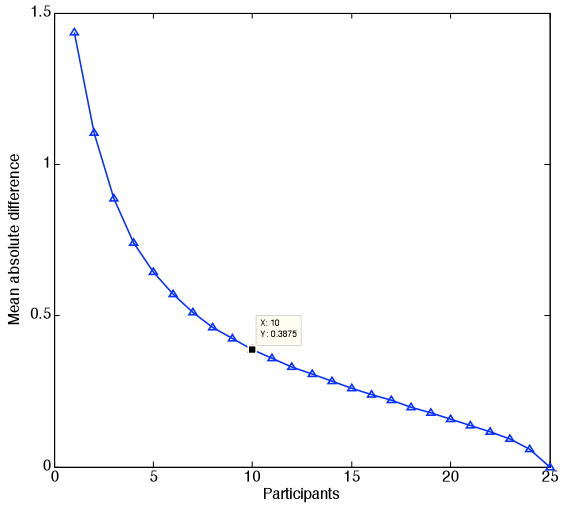

Figure 1 The mean absolute difference (on the vertical axis) versus number of participants.

One of the key questions in such studies is the number for participants for a significant result. In order to address this question, the situation of having fewer participants was simulated and the mean absolute difference with the final average was computed (see Figure 1). In this simulation, participants were randomly drawn and added to the pool of participants with the pool size of one to the maximum possible size of 25. This random simulation was performed 1000 times and the mean absolute difference between the participants' average annotations and the average scores of all 25 participants were computed. As it can be seen in Figure 1, with 10 participants the difference between the averaged scores is smaller than 5% of the possible range, $0.05 \times 8 = 0.4$. Although the gain of having more participants gets smaller after 10, in the real world applications a larger pool of annotators is always a valuable asset for information retrieval and recommendation studies.

6. BEST PRACTICES

Crowdsourcing using MTurk provided an effective means of collecting the viewer affective response annotations needed to create a corpus to be used in the development of automatic prediction of

viewer reported boredom. Our experience can be distilled into a list of recommendations that will enable the development of additional such corpora to proceed smoothly.

- The MTurk task should consist of two steps, the first identifies appropriate workers to invite and the second involves the generation of the annotations.
- For a long HIT series tasks, invite five times as many workers to do the pilot as you wish to have complete the main task.
- Expect that up to 75% of the workers you will invite will not be interested in carrying out a HIT that has the feeling of traditional "work", i.e., requires a long time commitment. In the invitation e-mail, specify a date by which they need to reply so that you can disinvite them and invite others if necessary.
- Consider breaking down long HIT series into packages and giving a small reward to the completion of individual packages in addition to a larger bonus for completing the whole series in order to prevent fatigue of the workers.
- As suggested by [3], we use multiple methods to verify that the workers are doing a good job on the question, for example, as a verifiable question and also check time.
- Include dummy questions to veil the purpose of the HIT.

Establishing trust with workers is a key factor in getting the same users to do a long HIT series. It is important to remember that they are concerned about maintaining their reputation on MTurk. Trust can be built by accepting HITs as quickly as possible and also being prompt with the bonuses. We suggest making the payment for each HIT very small and then accepting the HIT relatively indiscriminately. Workers who complete the entire series and do it well then receive the bonus. Effort invested in establishing trust accumulates since users exchange information on requesters on Turker Nation (http://www.turkernation.com/) concerning the HITs and the bonuses rewarded.

Our future work will concentrate on scaling up to be able to collect annotations for a larger set of videos with less intervention on our part. We now realize that for long HIT series, such as the ones necessary for a single person to annotate many videos, MTurk does not “run by itself”, but rather requires constant attention in terms of contacting workers and answering e-mail. In the future, we plan to be highly active during the initial stage of our main task to help speed up the process. In the future, we would like to develop a more complex pilot HIT that provides a more effective recruitment tool for workers. We are considering including more videos in the pilot HIT, or implementing a two-stage pilot, involving two HITs. A key factor here might be to use the MTurk API more extensively to achieve a higher level of automation. Addressing a practical problem, we would also like to work on developing a mechanism to deal elegantly with the failure of external resources. If a video fails to load, then the HIT is lost for the worker and needs to be manually reinitiated. The speed of the response depends on the amount of the reward offered. We paid viewers US $37.50 for watching 125 short videos. Paying less might have been possible. It would be worthwhile to determine if we can offer lower rewards without compromising quality. We also would like to investigate the bias introduced into the system by the fact that a certain type of personality is attracted to MTurk tasks and in particular to our Affect Task. Finally, we would like to move from boredom detection to other affective annotations. Our experiences with the MediaEval 2010 Affect Task Corpus suggest that crowdsouring is a valuable technique to collect affective annotations and we have just begun to tap its potential.

7. ACKNOWLEDGMENTS

The research leading to these results has received funding from the EC FP7 under grant agreement n° 216444 (PetaMedia Network of Excellence). The work of the first author is supported in part by the Swiss National Science Foundation.

8. REFERENCES

[1] P. Ekman et al., Universals and cultural differences in the judgments of facial expressions of emotion, *Journal of Personality and Social Psychology*, 53(4):712-717, 1987.

[2] A. Hanjalic and L.-Q. Xu. A selective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143-154, 2005.

[3] A. Kittur, E.H. Chi, and B. Suh. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the 26th Annual ACM conference on Human Factors in Computing Systems (CHI'08),* 453-456, 2008.

[4] J.D. Laird. *Feelings: The Perception of Self*. Oxford University Press, 2007.

[5] M. Larson, E. Newman, G.J.F. and Jones, G. J. F. Overview of Video-CLEF 2009: New perspectives on speech-based multimedia content enrichment. In C.Peters et al. (eds.), *Multilingual Information Access Evaluation*, *Vol. II Multimedia Experiments,* Springer, to appear 2010.

[6] P. Philippot. Inducing and assessing differentiated emotion-feeling states in the laboratory. *Cognition & Emotion* 7(2):171-193, 1993.

[7] M. Quirin, M. Kazen, and J. Kuhl, J. When nonsense sounds happy or helpless: The Implicit Positive and Negative Affect Test (IPANAT). *Journal of Personality and Social Psychology***,** 97:500-516, 2009.

[8] J. Rottenberg, R.D.Ray and J.J.Gross. Emotion elicitation using films. In A.Coan and J.J.B.Allen (eds.), *The Handbook of Emotion Elicitation and Assessment*. Oxford University Press, 2007.

[9] J. Russell and A. Mehrabian. Evidence for a 3-factor theory of emotions, *Journal of Research in Personality*, 11(3):273-294, 1977.

[10] M. Soleymani, G. Chanel, J.J.M. Kierkels, and T. Pun. Affective characterization of movie scenes based on content analysis and physiological changes. *International Journal of Semantic Computing*, 3(2):235-254, 2009.

[11] M. Soleymani, J. Davis, and T. Pun. A collaborative personalized affective video retrieval system. In *Proceedings of the International Conference on Affective Computing & Intelligent Interaction (ACII 2009),* 588-589.

[12] A. Stolcke, X. Anguera, K. Boakye**,** Ö Çetin, A. Janin, M. Magimai-Doss, C. Wooters, and J. Zheng. In Steifelhagen et al. (eds.) *The SRI-ICSI Spring 2007 Meeting and Lecture Recognition System*. LNCS vol. 4625, Springer, 450-463, 2008.

[13] H.L. Wang and L.F. Cheong Affective understanding in film, *IEEE Transactions on Circuits and Systems for Video Technology*, 16(6):689-704, 2006.

[14] X. Zhu and S. Gauch. Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '00), 288-295, 2000.