Exploiting Facial Expressions for Affective Video Summarisation

Hideo Joho
Dept. of Computing Science
University of Glasgow
hideo@dcs.gla.ac.uk

Joemon M. Jose
Dept. of Computing Science
University of Glasgow
jj@dcs.gla.ac.uk

Roberto Valenti
Intelligent Systems Lab
University of Amsterdam
r.valenti@uva.nl

Nicu Sebe
Dept. of Computer Science
University of Trento
sebe@disi.unitn.it

ABSTRACT

This paper presents an approach to affective video summarisation based on the facial expressions (FX) of viewers. A facial expression recognition system was deployed to capture a viewer's face and his/her expressions. The user's facial expressions were analysed to infer personalised affective scenes from videos. We proposed two models, pronounced level and expression's change rate, to generate affective summaries using the FX data. Our result suggested that FX can be a promising source to exploit for affective video summaries that can be tailored to individual preferences.

Categories and Subject Descriptors

H.5.1 [**Information interfaces and presentation (e.g., HCI)**]: Multimedia Information Systems—*video*

General Terms

Experimentation, Human Factors

Keywords

Facial expressions, affective summarisation

1. INTRODUCTION

The explosion of multimedia contents and the need for effective access have resulted in the development of a number of video summarisation techniques. Video summaries are needed in many situations. For example, it is useful for getting a gist of video content. Summaries can also support the end-user's decision-making to view the entire video (e.g., films) or not.

Money and Agius [10] categorise video summaries based on three dimensions: content type (feature based, object based, event based, and perception based), personalisation (personalised, generic), and interactivity (interactive, static). Techniques such as shot boundary detection and keyframe extraction are the basis of the feature based summaries which have been extensively investigated [6]. This type of summaries is not designed to consider semantics of video contents. The summaries investigated in evaluation forums such as TRECVID [12] tend to be object based or event based summaries. Such a summary consists of unique scenes of an object such as "antique car" or an object in the context of an event "red hot air balloon ascending". These types of summaries are designed to present a gist of contents based on the main objects and events within a video. However, the feature based and object/event based approaches tend to suffer from the semantic gap problem in interactive use.

Recently, there has been a growing interest in perception based summaries. These look at a higher level of abstraction than the other types of summaries by exploiting viewer's affective state, perceived excitement, and attention found within or caused by video contents [9]. Perception based approaches are designed to overcome the semantic gap problem in summarisation by finding affective scenes in videos. Another prospect of the perception based summarisation is the application of creating the personalised summaries. Since the affective scenes in videos are subjective, and hence, can vary across viewers, personalised summaries that are tailored to one's preference can be generated from the same video. However, this area has not been fully exploited, and existing techniques to generate perception based summaries are expensive. For example, they requires manual annotations [17] or several physiological sensors [9, 11] to capture people's affective state.

This paper presents an approach to affective video summarisation based on viewer's facial expressions. Our approach automatically captures and analyses facial expressions using a conventional webcam. The captured facial expressions are then used for determining affective highlights of videos. The contributions of this paper are as follows. First, we show that personal highlights of videos vary across viewers. This justifies the development of personalised affective summarisation. Second, we propose affective summarisation models based on facial expression analysis. Finally, we evaluate the effectiveness of our approach using video clips in several genres (e.g., film, drama, advertisement, etc.)

CIVR '09, July 8-10, 2009 Santorini, GR
Copyright Âľ 2009 ACM 978-1-60558-480-5/09/07... $5.00

and compare it with content-based techniques.

The rest of this paper is structured as follow. We first review the work on affective video summarisation. Then we present the facial expression recognition system. The details of features based on facial expressions and content analysis are then discussed, followed by the experimental design and results. We finally conclude the paper with the highlight of our findings and future work.

2. AFFECTIVE VIDEO SUMMARISATION

Money and Agius [10] provide a taxonomy of video summaries and their generation techniques based on an extensive literature survey. We use their taxonomy to discuss existing work on video summarisation and relate our work to them.

The first aspect of their framework is the information sources analysed for summarisation. *Internal* summarisation techniques analyse internal information from video streams produced during the production stage of video contents. More specifically, they tend to use low-level image, audio, and text features of videos. *External* summarisation techniques analyse external information which can be obtained from the process of capturing, producing, or viewing videos. External summarisation techniques are further divided into *User-based information* and *Contextual information* sources. User-based information typically includes people's behaviour during the interaction with video contents. This also includes people's preference information. The user-based information can be obtained in an *obtrusive* way using explicit feedback or *unobtrusive* way using various sensors. While unobtrusive methods are generally preferred, they tend to be noisy and limited in the level of details [10]. An example of the contextual information is geographical footprints of videos using a GPS facility equipped with a video camera.

Both internal information or external information have been exploited for affective video summarisation. The examples of internal information are Hanjalic and Xu [7] and Chan and Jones [1]. Hanjalic and Xu [7] introduced a method for modelling and representing affective video content as arousal and valence time curves. These two time curves can be also combined to form the affect curve, which was considered as a reliable representation of the sequential transitions from one affect state to another, along a video, as experienced by the user. A prototype system for affect-based indexing and retrieval of films, which is based on audio feature extraction, is also presented in Chan and Jones [1]. By analyzing all the audio data (speech, music, special effects and silence), the authors extracted the continuum of arousal and valence within the time dimension and used it to develop an affect annotation scheme.

The external information is often obtained by physiological sensors. For example, Mooney et al. [11] performed a preliminary study of the role of viewer's physiological states in an attempt to improve data indexing for search and within the search process itself. Participants' physiological responses to emotional stimuli were recorded using a range of biometric measurements, such as galvanic skin response (GSR), skin temperature, and other. The study provides some initial evidence that supports the use of biometrics as the user-based external information. Soleymani, et al. [13] proposed a method for affective ranking of movie scenes, which takes into account both user emotions as well as video content. User emotion behaviour was inferred based on evidence gathered from the measurements of five peripheral physiological signals (galvanic skin response, electromyogram, blood pressure, respiration pattern and skin temperature), as well as self-assessments. In addition, the movie scenes were analysed using various video and audio features, which portrayed significant events within those scenes.

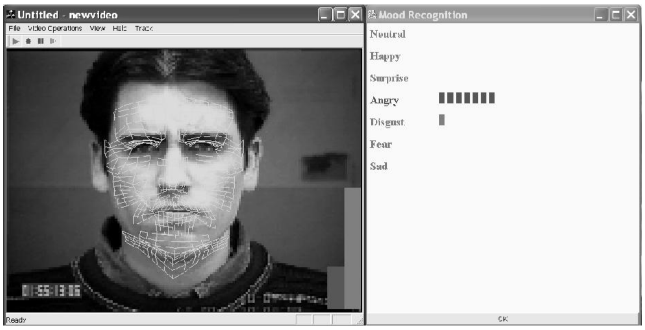

Figure 1: A snap shot of our realtime facial expression recognition system. On the left side is a wireframe model overlaid on a face being tracked. On the right side the correct expression, Angry, is detected.

The approach investigated in this paper can be seen as an *external* summarisation technique using *user-based* information. More specifically, we exploited viewer's facial expression while watching videos to find affective scenes for summarisation. Our information source (i.e., facial expression) was obtained in an *unobtrusive* way. This has a potential to make our approach simpler, more practical, and more feasible when compared to other approaches which exploited physiological signals of viewers. For example, in Money and Agius [9], subjects were wrapped by a sensor belt around their chest, a watch-type device was put around a wrist, and other signals were captured from several finger tips, and finally, their arm was rested on a cushion on the table. On the other hand, our approach required only a conventional web camera with which most recent PCs and laptops are equipped.

The next sections describe our system and the method to generate affective summaries by exploiting viewer's facial expressions.

3. FACIAL EXPRESSION RECOGNITION SYSTEM

Our real time facial expression recognition system is composed of a face tracking algorithm which outputs a vector of motion features of certain regions of the face. The features are fed as inputs to a Bayesian network classifier. We briefly describe these components in the following sections. A snap shot of the system, with the face tracking and recognition result is shown in Figure 1.

3.1 Face and facial feature tracking

The face tracking technique used in our system is an improved version of the system developed by Tao and Huang [15] called the piecewise Bezier volume deformation (PBVD) tracker. Our face tracker uses a model-based approach where an explicit 3D wireframe model of the face is constructed (see Fig. 1). A generic face model is warped to fit the detected facial features. The face model consists of 16 surface patches embedded in Bezier volumes. The surface patches defined this way are guaranteed to be continuous and smooth.

Once the model is constructed and fitted, head motion and local deformations of the facial features such as the eyebrows, eyelids, and mouth can be tracked. First the 2D image motions are measured using template matching between frames at different resolutions. Image templates from the previous frame and from the very first frame are both used for more robust tracking. The measured 2D image motions are modelled as projections of the true 3D motions onto the image plane. From the 2D motions of many points on the mesh, the 3D motion can be estimated by solving an overdetermined system of equations of the projective motions in the least squared sense.

The recovered motions are represented in terms of magnitudes of some predefined motion of various facial features. Each feature motion corresponds to a simple deformation on the face, defined in terms of the Bezier volume control parameters. We refer to these motions vectors as Motion-Units (MUs). Note that they are similar but not equivalent to Ekman's AU's [5] and are numeric in nature, representing not only the activation of a facial region, but also the direction and intensity of the motion.

The MU's are used as the basic features for the classification scheme described in the next section.

3.2 Learning the "structure" of the facial features

The use of Bayesian networks as the classifier for recognising facial expressions has been first suggested by Chen et al. [2], who used Naive Bayes (NB) classifiers and who recognised the facial expressions from the same MUs. When modelling the described facial motion features, it is very probable that the conditional independence assumption of the Naive Bayes classifier is incorrect. As such, learning the dependencies among the facial motion units could potentially improve classification performance, and could provide insights as to the "structure" of the face, in terms of strong or weak dependencies between the different regions of the face, when subjects display facial expressions.

In our approach, instead of trying to estimate the best a-posteriori probability, we try to find the structure that minimises the probability of classification error directly. The basic idea of this approach is that, since we are interested in finding a structure that performs well as a classifier, it would be natural to design an algorithm that uses classification error as the guide for structure learning. Here, we can further leverage on the properties of semi-supervised learning: we know that unlabeled data can indicate incorrect structure through degradation of classification performance, and we also know that classification performance improves with the correct structure. Thus, a structure with higher classification accuracy over another structure indicates an improvement towards finding the optimal classifier. The details of our analysis were presented in [4] and here we only briefly review the important issues that support understanding the classification component of our system.

To learn the structure using classification error, we must adopt a strategy of searching through the space of all structures in an efficient manner while avoiding local maxima. As we have no simple closed-form expression that relates structure with classification error, it would be difficult to design a gradient descent algorithm or a similar iterative method. Even if we did that, a gradient search algorithm would likely find a local minimum because of the size of the search space. Below we summarise our stochastic structure search (SSS) algorithm [4].

First we define a measure over the space of structures which we want to maximise:

Definition 1. *The* inverse error measure *for structure S' is*

$$inv_e(S') = \frac{\frac{1}{p_{S'}(\hat{c}(X) \neq C)}}{\sum_S \frac{1}{p_S(\hat{c}(X) \neq C)}}, \qquad (1)$$

where the summation is over the space of possible structures, X represents the MU's vector, C is the class space, $\hat{c}(X)$ represents the estimated class for the vector X, and $p_S(\hat{c}(\mathbf{X}) \neq C)$ is the probability of error of the best classifier learned with structure S.

We use Metropolis-Hastings sampling to generate samples from the inverse error measure, without having to ever compute it for all possible structures. For constructing the Metropolis-Hastings sampling, we define a neighbourhood of a structure as the set of directed acyclic graphs to which we can transit in the next step. Transition is done using a predefined set of possible changes to the structure; at each transition a change consists of a single edge addition, removal, or reversal. We define the acceptance probability of a candidate structure, S^{new}, to replace a previous structure, S^t as follows:

$$\min\left(1, \left(\frac{inv_e(S^{new})}{inv_e(S^t)}\right)^{1/T} \frac{q(S^t|S^{new})}{q(S^{new}|S^t)}\right) = \min\left(1, \left(\frac{p_{S^t}}{p_{S^{new}}}\right)^{1/T} \frac{N_t}{N_{new}}\right) \qquad (2)$$

where $q(S'|S)$ is the transition probability from S to S' and N_t and N_{new} are the sizes of the neighbourhoods of S^t and S^{new}, respectively; this choice corresponds to equal probability of transition to each member in the neighbourhood of a structure. This choice of neighbourhood and transition probability creates a Markov chain which is aperiodic and irreducible, thus satisfying the Markov chain Monte Carlo (MCMC) conditions [8]. T is used as a temperature factor in the acceptance probability.

Roughly speaking, T close to 1 would allow acceptance of more structures with higher probability of error than previous structures. T close to 0 mostly allows acceptance of structures that improve probability of error. A fixed T amounts to changing the distribution being sampled by the MCMC, while a decreasing T is a simulated annealing run, aimed at finding the maximum of the inverse error measures. The rate of decrease of the temperature determines the rate of convergence. Asymptotically in the number of data, a logarithmic decrease of T guarantees convergence to a global maximum with probability that tends to one.

The SSS algorithm, with a logarithmic cooling schedule T, can find a structure that is close to minimum probability of error. We estimate the classification error of a given structure using the labelled training data. Therefore, to avoid overfitting, we add a multiplicative penalty term derived from the Vapnik-Chervonenkis (VC) bound on the empirical classification error. This penalty term penalises complex classifiers thus keeping the balance between bias and variance (for more details we refer the reader to [4]).

Please note that the we decided to use this particular tracker due to its proven robustness and its ability to cope

Table 1: Description of video clips

Code	Length	Description
Video.1	01:43.5	Promotion Video of a pop song. Most parts are slow scenes where a singer is walking a downtown while singing. There is a colour effect on the picture which tones the colours to green and yellow.
Video.2	01:20.0	Documentary of a man with physical impairment demonstrating day-to-day activities. Calm background music with no speech. Visually similar across the clip. A short subtitle at the beginning introducing the contents.
Video.3	01:36.4	Documentary of people with physical impairment. Scenes of dancing with a wheelchair (First half) and travelling to the river (Last half). Calm background music with no speech (Similar to Video.2). A short subtitle at the beginning introducing the contents.
Video.4	00:39.0	Comical TV commercial of a beer. Night scenes and inside scenes with background noise of insects. Speech from three people and narrator at the end. No music. Two scenes were interwoven.
Video.5	04:29.2	A car chase scene from an action film. Upbeat background music with many sound effects of siren, scratching tires, crash, etc. Speech from four people. Many fast moving short shots.
Video.6	04:48.2	Scenes from a comedy drama film. Two scenes were interwoven: a talk show with one presenter, five guests on the stage, and large audience; and scene introducing the background of the main character. Mainly speech with many short shots.
Video.7	04:43.4	An action scene at night from a Sci-Fi film. Two groups of people are shooting and fighting. Many sound effects (guns, helicopter, breaking glasses, etc.) but no background music. Some shouts and screams in fast moving shots.
Video.8	07:03.6	Scenes from a soap drama. Amateur football game scene (60%), many conversations between people (30%), driving a car (10%), etc. No background music, but noise from the audience in the football game scene.

with non-frontal faces (up to 30% in head pose change). There were several other alternatives, mostly based on AAM (see for example [14] or [3]) but these systems require training and have difficulties in coping with the situations that were not present in the training set.

4. EXPERIMENT

This section presents the experimental design of our study.

4.1 Participants and video clips

Six people, all employees in the same software development company (holding different positions) agreed to participate in this study. Out of the six, three were female and three were male. All participants were between the ages of 22 and 36, and free from any obvious physical or sensory impairment. We used eight video clips taken from the contents in different genres. The code, duration, and brief description of the video clips, are given in Table 1. All videos had 25 frames per second.

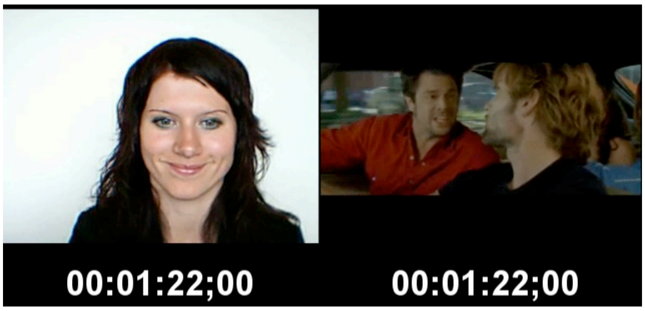

Figure 2: Recording facial expressions of a viewer (Left) watching a video clip (Right).

The recording of facial expressions was carried out in a room where a conventional video camera was set on top of a TV set. It should be noted that all video clips were new to the participants. The content video and the recording of facial expressions were synchronised for subsequent analysis (See Figure 2). The FX (facial expression) videos were exported to 360x240 pixels AVI format with 25 frames per second (same as the video clips).

4.2 FX based models

To extract the highlights of video clips based on the facial expressions, we devised three models. The first model was called a Pronounce level. This model was motivated by the observation of some facial expressions being more pronounced than the other expressions. More specifically, the categories such as Happy and Surprised were often more pronounced than the categories such as Sad and Fear. Therefore, we grouped the facial expressions of the viewers into three pronounced levels: No (Neutral), Low (Angry, Disgust, Fear, Sad), and High (Happy and Surprise). We assigned a score of 0, 0.5, and 1 to each of the levels, respectively. These scores are arbitrary and a formal study for the optimal configuration is part of our future work.

The second model was called a FX change rate. This model was designed to represent how often the detected facial expressions changed from one category to another. The frequency of change of the facial expressions can be seen as analogous to the changes that occurred in the viewers' affective state and, as a result, were treated as indicative of the level of affection of the video content. For this model, we counted the number of frames where the same FX category continued to be dominant at each frame. For example, there is a middle part (Frame 800 to 1500) where *Surprise* category continued to be the dominant expression in Figure 3. Also there was a rapid change of categories before the middle part. The change rate is low in the middle part and the rate was high in the rapid part. The length of continuing frames was normalised by the length of the entire video clip.

We applied the Kaiser Window process to the features

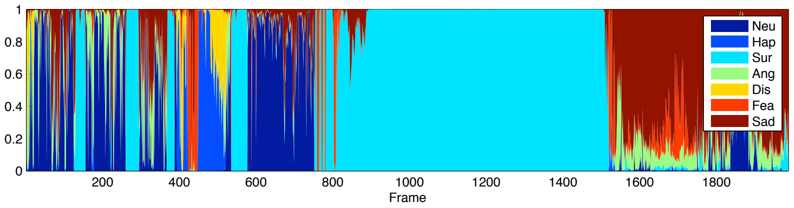

Figure 3: Visualisaton of the output of Facial Expression Recognition System (Video.2).

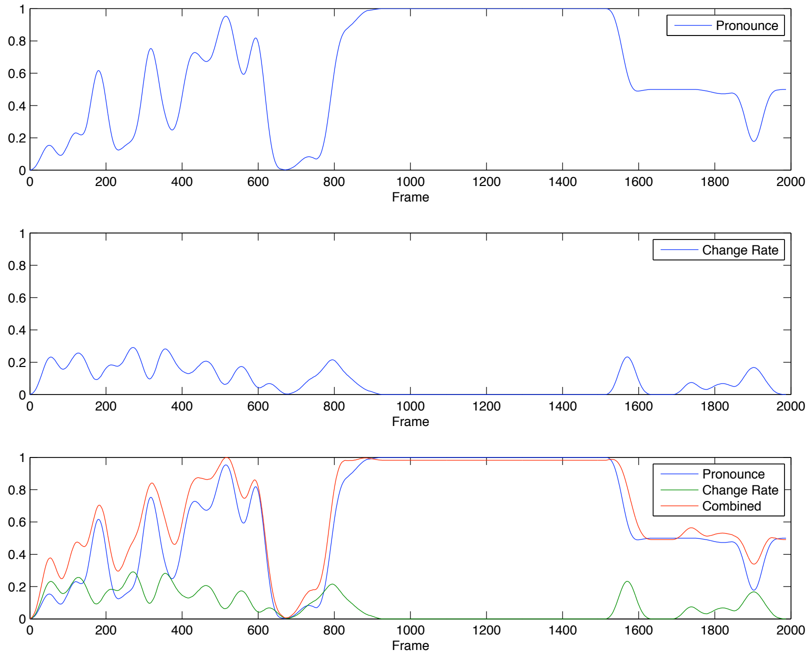

Figure 4: FX based models (Video.2). Pronounced level (Top), Change Rate (Middle), and Combination (Bottom).

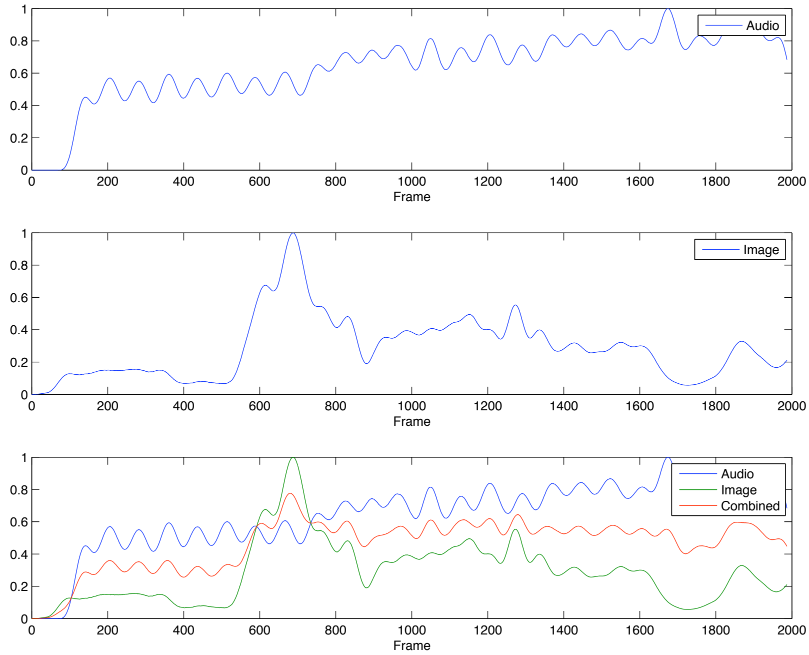

Figure 5: Content based models (Video.2). Audio (Top), Image (Middle), and Combination of two (Bottom).

for smoothing [7]. A visualisation of the facial expression models is shown in Figure 4, along with the original data from the FX recognition system in Figure 3. We used the data from Viewer.1 watching Video.2 for Figures. The top and middle charts represent the Pronounce and Change Rate models respectively, while the bottom chart shows a model that is an equally weighted combination of the two models.

As can be seen, the pronounced level appears to be considerably high during the second half of the video clip, since the detected facial expressions were classified to the surprise emotion category. The change rate was found to be higher at the beginning and end of the video clip. The combined model takes both into account with an equal weight. The combination model allows us to examine whether the two FX features are complimentary.

4.3 Content based models

We also extracted two low-level features from the original content of the video clips. The content-based features used in our study were inspired by the features examined by Hanjalic and Xu [7]. The first feature was an audio feature, root-mean-square (RMS) energy. The RMS represents the global energy of audio signals by taking the root average of the square of the amplitude. We used the MIRtoolbox[1] of Matlab to extract RMS. The second feature was a visual feature measuring a similarity between the images of two frames. There are many ways to measure a similarity between two images, and our approach was based on Karhunen-Loeve transform [16], which performed a principal component analysis on color matrix.

A visualisation of the audio and visual feature is shown in Figure 5. The top and middle charts represent the audio and image features, respectively, while the bottom charts shows a combined model of the two. As can be seen, there was a gradual increase in audio towards the end of the video clip, while a high rate of image changes occurred around Frame 700. Like the FX models, The combined model takes both into account with an equal weight.

To evaluate the performance of these models for a personalised summarisation task, we obtained manual highlight annotations from participants. The next section describes it in detail.

4.4 Highlight annotations

We obtained the manual annotations of highlight scenes from participants to evaluate the performance of the techniques. After the end of a video clip, participants were presented with a simple video annotation tool where they could select parts of video clips. Participants were allowed to annotate as many separate scenes as they found it as highlights.

Viewers annotated on average 545.8 frames as highlight of video clips (SD: 491.0, Min: 23, Max: 2025), which was just less than 22 seconds. Of 48 annotations, 24 annotations had one highlight segment, 14 had two, and 10 had three segments. The average was 1.7 segments (SD: 0.8).

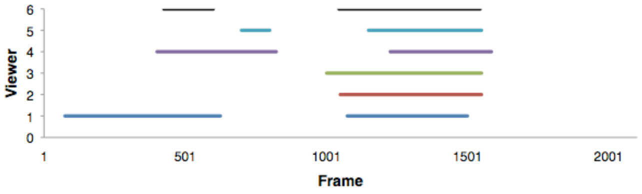

Figure 6: Manual Annotation (Video 2).

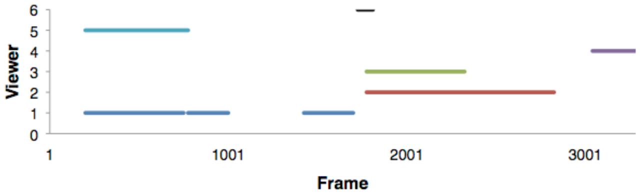

Figure 7: Manual Annotation (Video 6).

Figure 6 and Figure 7 visualise the manual annotations on the two of the video clips. As can be seen, there was a high level of consensus as to where to be a highlight of the video clips in Video.2, while people seemed to have different opinions about the highlight of Video.6. As summarised in Table 1, Video.2 (shown in Figure 6) was a documentary of people with physical impairment. In the frames between 1000 and 1500, one of the people skillfully folded a piece of paper using their feet. Most viewers selected this scene as the highlight of the video clips.

However, this was a rare case in the eight video clips. The annotation patterns were similar to Figure 7 for the rest of video clips. This observation is important since this suggests that people can find different parts of videos as the highlight. This justifies the development of personalised summarisation techniques that can be tailored to individual preferences. The use of multimodal interaction analysis offers us one way to achieve such dynamic summarisation as opposed to the content-based techniques that often produced the same summary from the same video. However, the effect of individual difference might be too large to produce an effective personalised summary. The rest of the paper compares these two approaches based on the manual annotations obtained from our viewers.

We used three measures for the evaluation of the highlighting techniques: precision represents a proportion of annotated frames in all extracted frames from a video clip; recall represents a proportion of extracted frames that are annotated in all annotated frames; finally F-score represents a weighted score of precision and recall. We weighted them equally in this study, and thus, F-score was the mean of precision and recall. We present the F-scores in our result section.

Since the length of annotated scenes differed across participants and videos, the systems were set to extract the same length of scenes as each of the annotations. For example, if a total length of an annotation of Viewer.1 was 25 seconds, the systems were set to extract a total of 25 seconds as the highlight for this viewer. In addition, we measured the performance using the summary length of 10%, 25%, 50%, and 75% of the original videos. It should be noted that the summary ratio of videos was longer than those used in TRECVid. This is because our video clips were much shorter than TRECVid collections.

5. RESULTS

We first looked at the F-score of seven FX categories (Neutral, Happy, Surprised, Angry, Disgust, Fear, and Sad) that

[1] http://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox

were detected by our facial expression recognition system. Each category was used as an independent feature. The probability of each category was plotted in the time series across the video frames. The Kaiser Window was applied for smoothing in the same fashion as other features. The highlight scenes were determined by selecting the frames with the highest value for a given length. The result is shown in Figure 8.

As can be seen, the Surprise category was found to perform better than the other categories when the summary length was set to the individual annotation length. This supports our design of the pronounce level model. However, the Happy category (another category in the high pronounce level group) was not as strong indicator of highlight as the Surprise category. The result also suggests that the Sad category performed well. This suggests that the sad category was more indicative of personal highlight than we had expected, and thus, it should be considered for further development.

The F-score improved as the length of summaries increased. This is because the recall tended to be higher when a longer summary was generated than shorter summary. As reported in Section 4.4, on the other hand, the length of annotated highlights varied across videos and across viewers.

Figure 9 shows the F-score of the FX based models (Pronounced, ChangeRate, and P+C) and content based models (Audio, Image, and A+I) for the affective video summarisation. There are several observations to be made from the results. First, the FX based models generally performed as well as or better than the content based models. The difference became more evident when the summary length increased to 75%. Second, the combination model, P+C, tended to perform better than the individual FX based models, Pronounced or ChangeRate. The combination effect appeared to be stronger when the summary length was shorter. This suggests that the two FX based features were complementary to detect affective scenes from video clips. However, the combination effect was not found in the content based models, and the Image feature tended to outperform the Audio or combination models.

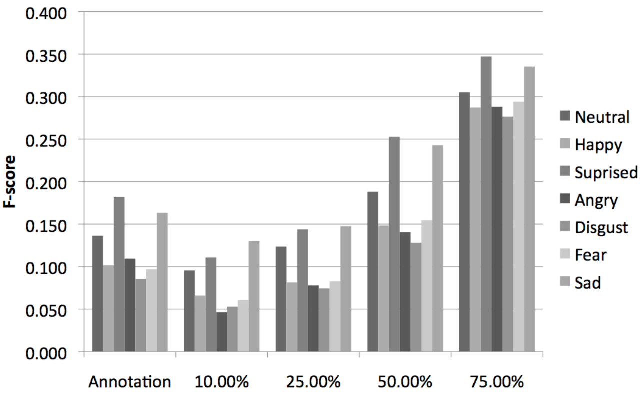

Figure 8: F-score of FX categories.

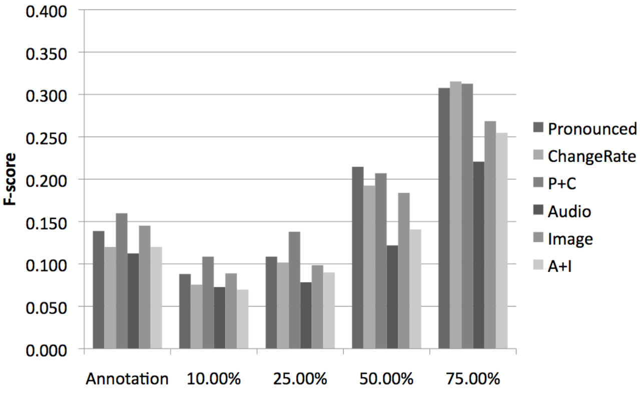

Figure 9: F-score of affective models.

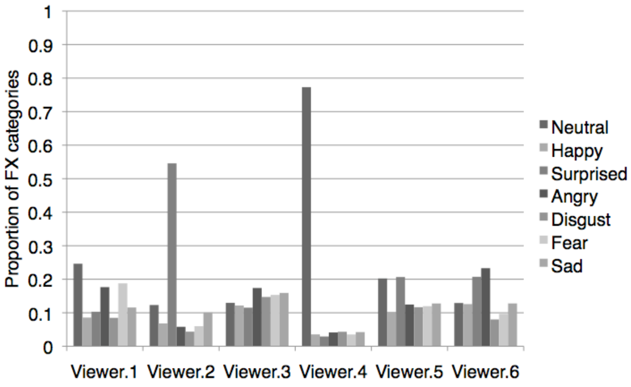

Figure 10: Distribution of facial expressions.

We were also interested in the analysis of the distribution of FX categories across the viewers and videos. Figure 10 depicts the emotion behaviour of the six viewers, across all videos. The average intensity is presented on a scale of 0 to 1, per emotion category and per viewer, as the latter information was obtained from the analysis of the viewers' facial expressions. The distribution of the emotions appears to be uniform, without revealing any dominant expressions. This suggests that our FX system detected a variety of facial expressions from viewers, with exceptions of Viewer.2 and Viewer.4. In the case of Viewer.2, the expression of surprise received a higher score (compare to the other six expressions), therefore, indicating a more intensified emotion. However, after a manual inspection of the FX video we concluded that this effect was caused by that person's large oval eyes, which resulted in a misclassification of some of the facial expressions. Viewer.4, on the other hand, retained a neutral expression throughout the study which resulted low scores of the rest of the categories, compared to other participants. In other words, Neutral was the dominant category of this viewer. However, this was not common as you can see from Figure 10.

6. DISCUSSION

In this work we explored the viewer's facial expressions as the basis of summarisation. This section discusses the implications of our study on the design and development of affective video summarisation techniques.

In Section 4.4, we analysed the manual annotation of video highlights. The result indicated that it was unlikely that a single summary could be commonly seen as the highlight of videos by viewers. While we are currently extending our analysis with a larger number of viewers, the result so far suggested that at least there were two or three distinguished parts of videos that can be seen as the highlight by various viewers. This reinforced our motivation of exploring affective summarisation techniques using user based information such as facial expressions. The user based information has the potential to generate the personalised summaries that are tailored to individual preferences, as opposed to content based techniques.

The results shown in Section 5 suggested that the facial expressions were a promising source for affective video summarisation. The overall performance of FX based models was comparable to content based models. However, when the summary length increased, FX based models tended to outperform content based models. This was an encouraging result given that this was our first attempt to use facial expressions for summarisation. Our study also revealed interesting characteristics of viewer's facial expressions during watching video clips. For example, we found a case where a single category of expression dominated the data. While this was not a common case, it should be considered for the development of FX based models.

Overall, however, the performance of both FX based and content based models was still not satisfactory. This simply demonstrates that personalised video summarisation is a very challenging task and requires further research. The use of manual annotations was found to be essential for the robust evaluation of the affective summarisation techniques.

Finally, it should be pointed out that physiological approaches do not tend to scale as much as content based approaches do. Therefore, we need to explore ways to leverage user based information in a practical fashion. One way might be the combination with content based approaches. For example, the highlight scenes are determined by FX based models in unobtrusive way, but the scenes were represented by low level feature models. This will allow us to generate a personalised summary for unseen videos by measure the similarity between existing FX profile and new video contents. This is just one possibility and we will continue to explore other options.

7. CONCLUSION AND FUTURE WORK

We have presented an ambitious but innovative approach to affective summarisation based on the exploitation of people's facial expressions. While they were preliminary results, there are positive outcomes in our work. First, the models based on FX features achieved a comparable performance to the content-based models. This is a promising result given that this was our first attempt in affective summarisation. Second, the combination of the FX models appeared to improve the performance over the single models. Therefore, the models seemed to be complementary, which is encouraging. Finally, the observation of the manual annotations provided us empirical evidence that justifies the development of personalised summarisation over content-based summarisation.

8. ACKNOWLEDGMENTS

Thanks to Reede Ren for initial extraction of the audio feature, Punitha Puttu Swamy for extracting the visual feature, and David Hannah for helping data analysis. Funding was provided by the MIAUCE project (Ref: IST-033715).

9. REFERENCES

[1] C. H. Chan and G. J. F. Jones. Affect-based indexing and retrieval of films. In *ACM Multimedia*, pages 427–430, 2005.

[2] L. Chen. *Joint Processing of Audio-visual Information for the Recognition of Emotional Expressions in Human-computer Interaction*. PhD thesis, University of Illinois at Urbana-Champaign, 2000.

[3] Y. Cheon and D. Kim. Natural facial expression recognition using differential-aam and manifold learning. *Pattern Recognition*, 42(7):260–274, 2009.

[4] I. Cohen, N. Sebe, F. Cozman, M. Cirelo, and T. Huang. Semi-supervised learning of classifiers: Theory, algorithms, and applications to hci. *PAMI*, 26(12):1553–1567, 2004.

[5] P. Ekman and W. Friesen. *Facial Action Coding System: Investigator's Guide*. Consulting Psychologists Press, Palo Alto, CA, 1978.

[6] A. Hanjalic, R. Lienhart, W. Y. Ma, and J. R. Smith. The holy grail of multimedia information retrieval: So close or yet so far away? *Proceedings of the IEEE*, 96(4):541–547, 2008.

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

[8] D. Madigan and J. York. Bayesian graphical models for discrete data. *International Statistical Review*, 63:215–232, 1995.

[9] A. Money and H. Agius. Feasibility of personalized affective video summaries. In *Affect and Emotion in Human-Computer Interaction*. Springer, 2008.

[10] A. Money and H. Agius. Video summarisation: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation*, 19(2):121–143, February 2008.

[11] C. Mooney, M. Scully, G. J. F. Jones, and A. F. Smeaton. Investigating biometric response for information retrieval applications. In *ECIR*, pages 570–574, 2006.

[12] P. Over, A. F. Smeaton, and P. Kelly. The TRECVID 2007 BBC rushes summarization evaluation pilot. In *TVS '07: Int. workshop on TRECVID video summarization*, pages 1–15, 2007.

[13] M. Soleymani, G. Chanel, J. J. Kierkels, and T. Pun. Affective ranking of movie scenes using physiological signals and content analysis. In *ACM workshop on Multimedia semantics*, pages 32–39, 2008.

[14] J. Sung, T. Kanade, and D. Kim. Pose robust face tracking by combining active appearance models and cylinder head models. *IJCV*, 80(2):260–274, 2008.

[15] H. Tao and T. Huang. Connected vibrations: A modal analysis approach to non-rigid motion tracking. In *CVPR*, pages 735–740, 1998.

[16] P. Tianqiang, Z. Keke, and L. Bicheng. Video abrupt transition detection based on k-l transform. In *ICIG '07: Proceedings of the Fourth International Conference on Image and Graphics*, pages 845–848, Washington, DC, USA, 2007. IEEE Computer Society.

[17] D. Tjondronegoro, Y.-P. Chen, and B. Pham. Highlights for more complete sports video summarization. *IEEE Multimedia*, 11(4):22–37, 2004.