Multimedia Tools and Applications manuscript No.
(will be inserted by the editor)

Looking at the Viewer: Analysing Facial Activities to Detect Personal Highlights of Multimedia Contents

Hideo Joho · Jacopo Staiano · Nicu Sebe · Joemon M. Jose

The final publication is available at `www.springerlink.com`

Abstract This paper presents an approach to detect personal highlights in videos based on the analysis of facial activities of the viewer. Our facial activity analysis was based on the motion vectors tracked on twelve key points in the human face. In our approach, the magnitude of the motion vectors represented a degree of a viewer's affective reaction to video contents. We examined 80 facial activity videos recorded for 10 participants, each watching eight video clips in various genres. The experimental results suggest that useful motion vectors to detect personal highlights varied significantly across viewers. However, it was suggested that the activity in the upper part of face tended to be more indicative of personal highlights than the activity in the lower part.

Keywords Facial Activity · Facial Expression · Affective Summarization

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

Dr. H. Joho
Graduate School of Library, Information and Media Studies, University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550 Ibaraki, Japan
E-mail: hideo@slis.tsukuba.ac.jp

Mr. J. Staiano
Department of Information Engineering and Computer Science, University of Trento
Via Sommarive 14, 38100 Povo, Trento, Italy
E-mail: jacopostaiano@gmail.com

Prof. N. Sebe
Department of Information Engineering and Computer Science, University of Trento
Via Sommarive 14, 38100 Povo, Trento, Italy
E-mail: sebe@disi.unitn.it

Prof. J. M. Jose
Department of Computing Science, University of Glasgow
Sir Alwyn Williams Building, Glasgow G12 8QQ UK
E-mail: jj@dcs.gla.ac.uk

needed in many situations. For example, such a summary could be useful for getting a gist of the video content. Summaries can also support the end-user's decision-making to view the entire video (e.g., films) or not. The results of such decision making can then be used for modelling the user preference. This also suggests that the techniques developed for video summarisation can be related to a task of user profiling and/or personal recommendation of unseen videos.

Money and Agius [Money and Agius(2008b)] 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 [Hanjalic et al(2008)Hanjalic, Lienhart, Ma, and Smith]. This type of summaries is not designed to consider semantics of video contents. The summaries investigated in evaluation forums such as TRECVID [Over et al(2007)Over, Smeaton, and Kelly] 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 of such summaries.

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 [Money and Agius(2008a)], [Joho et al(2009)Joho, Jose, Valenti, and Sebe]. 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 require manual annotations [Tjondronegoro et al(2004)Tjondronegoro, Chen, and Pham] or several physiological sensors [Money and Agius(2008a), Mooney et al(2006)Mooney, Scully, Jones, and Smeaton] to capture people's affective state.

In our previous work, we have developed an approach to affective video summarisation based on viewer's facial expressions [Joho et al(2009)Joho, Jose, Valenti, and Sebe]. Our approach automatically captured and analysed facial expressions using a conventional webcam. The captured facial expressions were then used for determining affective highlights of videos. During the course of this work, we came to realise the importance of understanding the behaviour of facial activity to provide an effective summarisation. Therefore, we decided to carry out further investigation on the motion of human face, which were the basic features for the classification of facial expressions in our previous work. In particular, we were interested in investigating which part of the human face was indicative of people's affective reaction to video contents, and thus, detection of personal highlights.

In this paper, we distinguish the terms *facial activity* and *facial expression*. The facial activity refers to the movement of specific points (i.e., motion vectors) in human

face, while the facial expression refers to a category of human emotion inferred from a classification of multiple motion vectors. The two concepts will be explained in more detail later.

The rest of this paper is structured as follow. We first review the work on affective video analysis and summarisation. Then we briefly present the facial expression recognition system. The data collection method and evaluation measurement are then described, followed by the results of analysis and discussion. We conclude the paper by discussing some directions of future work.

2 Affective video analysis and summarisation

Annotation according to affective or emotional categories of video is a relatively young domain, gaining more and more importance [Hanjalic and Xu(2005)], [Moncrieff et al(2001)Moncrieff, Dorai, and Venkatesh], [Xu et al(2005)Xu, Chia, and Jin], [Calvo and D'Mello(2010)], [Kang(2002)], [Wang and Cheong(2006)], [Hanjalic(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.

All the current affective analysis systems try to solve the following problems [Wang and Cheong(2006)]: 1) identification of valid affective features; 2) bridging the gap between affective features and affective states; 3) establishing an affective model to take user's personality into consideration; 4) representing the affective state.

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". The videos or parts of them are classified automatically into one of these categories. Moncrieff et al. [Moncrieff et al(2001)Moncrieff, Dorai, and Venkatesh] analyze changes in sound energy of the non-literal components of the audio tracks of films and detect four sound energy events commonly used in horror films: "surprise or alarm", "apprehension or emphasis of a significant event", "surprise followed by a sustained alarm", and "building apprehension up to a climax". They find that these four sound energy events convey well established meanings through their dynamics to portray and deliver certain affect or sentiment related to the horror film genre. Kang et al. [Kang(2002)] detect emotional events such as fear, sadness and joy from videos by computing intra-scene context (shots' coherences, shot's interactions, dominant features in color and motion information) and inter-scene context (scene's relationship with other scenes). Xu et al. [Xu et al(2005)Xu, Chia, and Jin] identify video/audio segments which make audience laugh in comedy and scary segments in horror films as affective contents. They use Hidden Markov Models (HMM) based audio classification method to detect audio emotional events (AEE) such as laughing, horror sounds, etc. Then, they use the AEE as a clue to locate the corresponding video segment.

The second type of affective analysis method is called Dimensional Affective Content Analysis method, which commonly employs the Dimensional Affective Model to compute the affective state. The psychological Arousal-Valence (A-V) Affective Model is one popular Dimensional Affective Model [Dietz and Lang(1999)]. Arousal stands for

the intensity of affective experience and Valence characterizes the level of "pleasure". Hanjalic and Xu [Hanjalic and Xu(2005)] did research on affective state representation and modeling by using the A-V Affective Model. According to the A-V affective model, the affective video content can be represented as a set of points in the two-dimensional (2-D) emotion space that is characterized by the dimensions of arousal (intensity of affect) and valence (type of affect). By using the models that link the arousal and valence dimensions to low-level features extracted from video, the affective video content can be mapped onto the 2-D emotion space. Then, an affect curve (arousal and valence time curves) can be easily detected as a reliable representation of the expected transitions from one feeling to another along a video. Pleasure-Arousal- Dominance (P-A-D) model [Mehrabian(1996)] is another popular affective model. Pleasure stands for the degree of pleasantness of the emotional experience, Arousal stands for the level of activation of the emotion, and Dominance describes the level of attention or rejection of the emotion. Based on P-A-D model, Arifin et al. [Arifin and Cheung(2007)] propose to use Dynamic Bayesian Networks (DBNs) to build up a P-A-D value estimator, which estimates the P-A-D values of the video shots of the input video. Then, the video can be segmented based on the estimated P-A-D content. Different from the Arousal and Valence modeling proposed by Hanjalic and Xu [Hanjalic and Xu(2005)], this work takes the influences of former emotional events and larger emotional events into consideration.

The third type of affective analysis method is Personalized Affective Content Analysis method. The representative work is reported in [Wang and Cheong(2006)], which introduces more personalization factors into affective analysis and apply this to Music Video (MV) retrieval. First, they build a user interface and record the users' feedback in the user profile database. Each profile records MV's ID, user's descriptions about MV's Arousal and Valence (two scores describing their opinions about Arousal and Valence level). When users play MV, they can also use feedback to change their opinions on MV at any time. Based on the users' profile, two Support Vector Regression (SVR) models (Arousal model and Valence model) are trained to fit the user's affective descriptions. Finally, the affective features extracted from MV are fed into the trained models to get the personalized affective states. The authors also provide a novel Affective Visualization interface for efficient and user-friendly MV retrieval. Through this interface, the user can easily log into the system, search MV based on their affective states (for example, anger, happy, sad/blue, or peaceful) and also provide his/her feedback on each MV.

Directly relevant to our present work, Money and Agius [Money and Agius(2008b)] 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 in an *unobtrusive* way

using various sensors. While unobtrusive methods are generally preferred, they tend to be noisy and limited in the level of details [Money and Agius(2008b)]. An example of the contextual information is the geographical footprints of videos using a GPS facility equipped with a video camera.

Both internal or external information have been exploited for affective video summarisation. The examples of internal information are Hanjalic and Xu [Hanjalic and Xu(2005)] (discussed above) and Chan and Jones [Chan and Jones(2005)]. Chan and Jones [Chan and Jones(2005)] present a prototype system for affect-based indexing and retrieval of films, which is based on audio feature extraction. 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. [Mooney et al(2006)Mooney, Scully, Jones, and Smeaton] 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. [Soleymani et al(2008)Soleymani, Chanel, Kierkels, and Pun] 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.

The approach investigated in this paper belongs to the group of Categorical Affective Analysis and 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 [Money and Agius(2008a)], 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. The system has been described in detail in [Sebe et al(2005)Sebe, Cohen, Cozman, and Huang] and for completeness

Fig. 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.

we briefly describe the components of the system 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 [Tao and Huang(1998)] 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 [Ekman and Friesen(1978)] and are numeric in nature, representing not only the activation of a facial region, but also the direction and intensity of the motion.

The 12 MUs used in the face tracker are shown in Fig. 2. As you can see, the first six vectors are roughly located in the lower part of human face while the other six vectors are located in the upper part of the face. We will denote the 12 MUs as $MU1$, $MU2$,

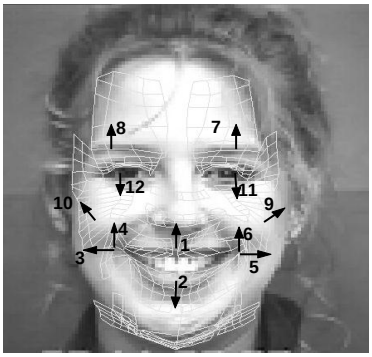


Fig. 2: The facial motion measurements.

..., $MU12$ in this paper. The MUs 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 was suggested by Chen et al. [Chen(2000)] and [Cohen et al(2003)Cohen, Sebe, Chen, Garg, and Huang], 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. Consequently, we further leveraged on two properties of semi-supervised learning: (1) the unlabeled data can indicate incorrect structure through degradation of classification performance, and (2) the 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 [Cohen et al(2004)Cohen, Sebe, Cozman, Cirelo, and Huang] 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 adopted a strategy of searching through the space of all structures in an efficient manner while avoiding local maxima. As there is no simple closed-form expression that relates structure with classification error, it is 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. The so-

lution followed in our system is the stochastic structure search (SSS) algorithm [Cohen et al(2004)Cohen, Sebe, Cozman, Cirelo, and Huang].

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

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

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

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 used 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 defined 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 defined the acceptance probability of a candidate structure, S^{new}, to replace a previous structure, S^t as follows:

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

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 [Madigan and York(1995)]. T is used as a temperature factor in the acceptance probability.

Roughly speaking, T close to 1 allows acceptance of more structures with higher probability of error than previous structures while T close to 0 mostly allows acceptance of structures that improve probability of error. Additionally, 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, finds 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 added 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 [Cohen et al(2004)Cohen, Sebe, Cozman, Cirelo, and Huang]).

Please note that we decided to use this particular tracker due to its proven robustness and its ability to cope with non-frontal faces (up to 30% in head pose

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 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 a 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.

change). There were several other alternatives, mostly based on AAM (see for example [Sung et al(2008)Sung, Kanade, and Kim] or [Cheon and Kim(2009)]) but these systems require training and have difficulties in coping with the situations that were not present in the training set.

4 Analysis

This section presents the analysis of facial activity for detecting personal highlights of video contents.

4.1 Participants and video clips

Ten people, all employees in the same software development company (holding different positions) agreed to participate in the experiment. Out of the ten, five were female and five were male. All participants were between the ages of 24 and 43, and were 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.

The recording of facial activity 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

Fig. 3: Recording facial expressions of a viewer (Left) watching a video clip (Right).

new to the participants. The content video and the recording of facial activity were synchronised for subsequent analysis (See Figure 3). The facial activity videos were exported to 360x240 pixels AVI format with 25 frames per second (same as the content video clips).

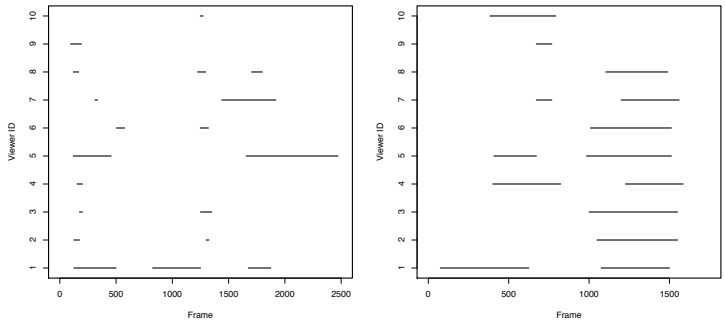
4.2 Highlight annotations

We obtained the manual annotations of highlight scenes from participants to evaluate the effectiveness of facial motion units. 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 necessary as highlights. The results of the manual annotation can be found in Figure 4, where the X-axis represents the frame number of video clips and Y-axis represented the viewer ID. Note that the frame length denoted by the X-axis varies across the video clips.

As can be seen, there was a high level of consensus as to where a highlight is present in Video.2. As summarised in Table 1, Video.2 (shown in Figure 3) 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 clip. However, such consensus did not appear to be common in most of the rest of videos. This observation is important since this suggests that people can find different parts of videos as the highlight, which is the major assumption made in this paper.

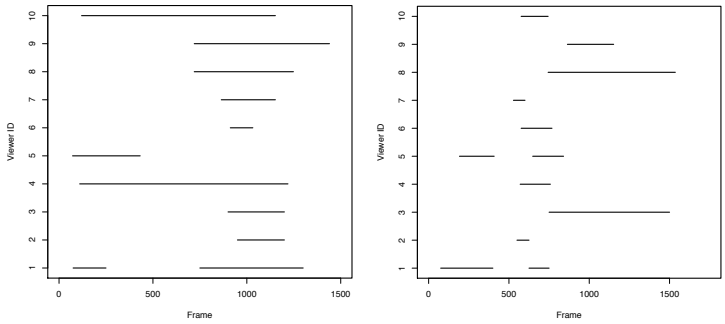
4.3 Facial features

We analysed a total of 20 facial features in this study. They included 12 motion units (denoted as MU1 to MU12), a combination of the 12 vectors (denoted as MU1-12), and 7 facial expression categories (Scared, Angry, Disgusted, Happy, Neutral, Sad, and Surprised). For each of the facial activity videos, a vector value of motion or probability of emotion categories were produced by the methods described in Section 3. We then applied a Kaiser Window process on the outputs of facial features in a similar fashion



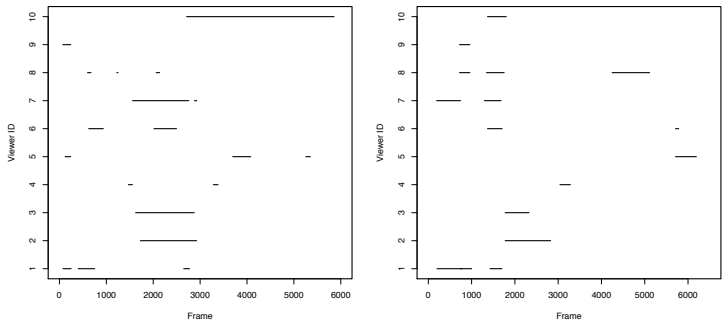
(a) Video.1

(b) Video.2



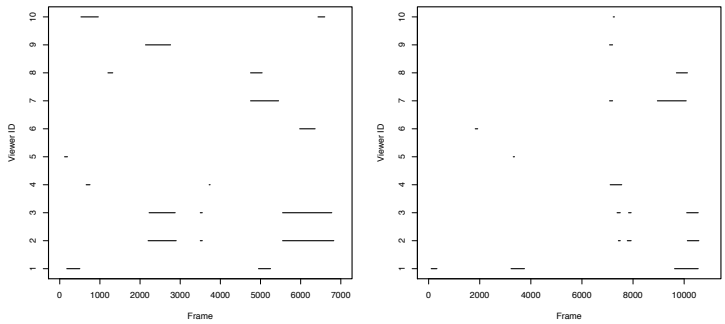
(c) Video.3

(d) Video.4



(e) Video.5

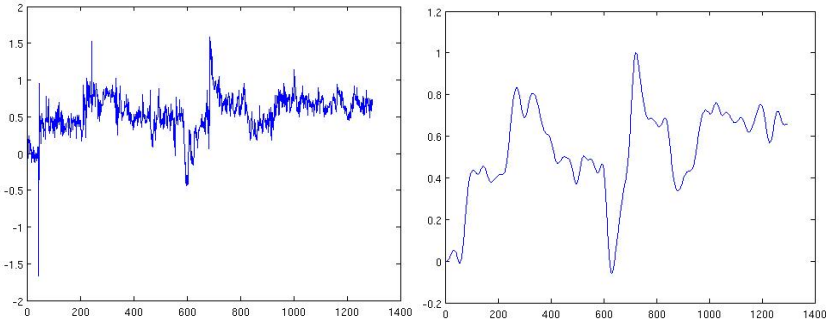
(f) Video.6



(g) Video.7

(h) Video.8

Fig. 4: Annotation of personal highlights (Video.1 to Video.8)



(a) Original Data (b) Kaiser Window

Fig. 5: Effect of Kaiser Window on MU1 feature.

to [Hanjalic and Xu(2005)]. The effect of smoothing on the original data can be found in Figure 5.

Our hypothesis was that an effective facial feature should produce a large motion or high probability of emotion category to detect personal highlights of videos. Therefore, we see this as a ranking problem where the video frames are ordered by the vector value or category probability. Consequently, we used a scoring function called Average Precision [van Rijsbergen(1979)], [Huijsmans and Sebe(2005)] to measure the effectiveness of facial features for personal highlight detection. Average Precision, $AvgP$, is one of the major performance measures in the field of Information Retrieval, and it is calculated in the following manner:

$$AvgP = \frac{\sum_{r=1}^{N} P(r)}{H} \tag{3}$$

$$P(r) = \frac{h(r)}{r} \tag{4}$$

where r is the ranking position of a frame, N is the ranking position of the lowest ranked highlight frame, $h(r)$ is the total number of highlight frame found up to the rank r, $P(r)$ is the precision at the rank r, and H is the total number of highlight frames annotated by individual participants.

5 Results and Discussion

This section reports the results of the analysis and discusses the implications of our findings on the design of personal highlights detection technologies for video contents.

5.1 Facial activity

The first analysis carried out was the performance of the 12 motion vectors to detect personal highlight scenes in video clips. The result is shown in Table 2. Motion unit IDs (MU ID) are based on the numbers shown in Figure 2. The values in the table are

Table 2: Mean Average Precision of motion vectors. Those highlighted in bold are the best performance in individual viewers.

Facial Part	MU ID	Viewer 1	2	3	4	5	6	7	8	9	10
Mouth	1	.220	.098	.166	.116	.150	.058	.097	.186	.103	.121
	2	.278	.104	.172	.220	.135	.096	.083	**.233**	.073	.111
	3	.325	.113	.171	.098	.171	.059	.138	.162	.135	.122
	4	.192	.078	.208	.075	.088	.059	.139	.103	.118	.090
	5	.175	.072	.127	.099	.095	.050	.097	.149	.055	**.131**
	6	.187	.134	**.225**	.108	.097	.068	.103	.122	.130	.109
Cheeks	9	.195	.097	.197	.148	.121	.092	.079	.145	.087	.075
	10	.325	.139	.150	**.223**	.264	.065	.094	.176	.059	.093
Eyes	7	.147	.090	.129	.091	.097	**.143**	.132	.198	**.200**	.118
	8	**.404**	.104	.130	.163	.251	.072	.088	.145	.071	.089
	11	.316	.101	.155	.077	**.337**	.094	.104	.169	.172	.078
	12	.302	**.148**	.145	.078	.207	.066	**.163**	.145	.177	.096
All	1-12	.240	.127	.123	.078	.090	.051	.102	.135	.052	.095

Table 3: Mean Average Precision of emotion categories. Those highlighted in bold are the best performance in individual viewers.

Emotion ID	Viewer 1	2	3	4	5	6	7	8	9	10
Afraid	.228	.094	.137	.077	.145	.046	.078	**.211**	.062	.086
Angry	.225	.149	.187	.071	.174	.051	.119	.144	.049	.089
Disgusted	**.336**	.144	**.264**	.077	.147	.042	**.127**	.127	**.195**	.091
Happy	.238	.101	.208	.086	.099	.050	.107	.170	.041	.069
Neutral	.256	**.224**	.233	.104	**.234**	**.183**	.086	.152	.143	**.307**
Sad	.296	.134	.210	.100	.122	.044	.107	.161	.077	.182
Surprised	.258	.138	.126	**.110**	.179	.049	.069	.160	.078	.061
Best MV	.404	.148	.225	.223	.337	.143	.163	.233	.200	.131

the mean of average precision of all video clips. The bottom row of the table shows the performance of a feature which combined the magnitude of 12 motion vectors. To highlight the effect of facial parts, the features are divided into four parts: Mouth, Cheeks, Eyes, and All. We consider the mouth as the lower part of human face and the cheeks and eyes as the upper part of the face.

There are several observations from the result. First of all, the most useful features to detect personal highlights significantly vary across the viewers. This suggests that people's facial activity to react to their highlight scenes can be indeed very different. Second, relatively speaking, the motion units in the upper part of human face appear to be more indicative of personal highlights than the lower part. Although the best performing features varied across viewers, seven out of ten were based on the upper part of human face, which included eyes (MU7, 8, 11, and 12) and cheeks (MU9 and 10). Of those, the MUs around the eyes had the largest number of best performing cases. This suggests that the effectiveness of motion units across the 12 points are not equal, and a greater level of attention to the upper part of human face might allow us to capture individual preferences. Finally, the performance of `MU1-12` suggests that a simple addition of all motion vectors was not sufficient for accurate estimation of personal highlights.

5.2 Comparison to facial expression features

The second analysis compared the performance of motion vectors to that of emotion categories. The results are shown in Table 3. In the bottom row of the table are the best performing MUs from Table 2 for reference.

Unlike the performance of motion vectors, most of the best performing features in the emotion categories were based on the `Disgusted` and `Neutral` categories. However, if we compare these performance to the best MU features, we can observe that it was the Neutral feature which often outperformed the motion vector features. We speculate that the performance of Neutral category is partly due to the fact that many frames are categorised as Neutral when no particular facial activity was detected. Therefore, the Neutral category was more likely to perform better than other categories.

Overall, the comparison to the emotion categories suggests that some users can be modelled by a single point (motion unit) while others need multiple points (i.e., emotion category) to model their affective states.

5.3 On scalability

We have looked at people's facial activity to detect the personal highlights in video clips. This can be seen as a subtask of affective video summarisation based on human-centred multimodal approach [Jaimes et al(2007)Jaimes, Gatica-Perez, Sebe, and Huang]. A limitation of multimodal approach which exploits physiological aspects of human beings using various sensors is the scalability. Unlike the content-analysis approach, we can only collect the data while the users engage with multimedia contents. While our approach was using only a conventional webcam which is much less obtrusive than other approaches, the limitation still applies. In our previous work, this issue was briefly discussed as follows (Note that FX stands for facial expression in the following quote):

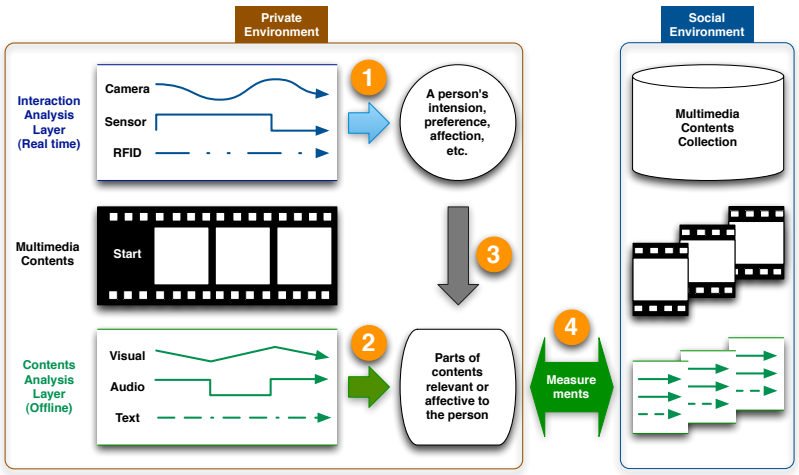


Fig. 6: A framework for the multimodal approach to multimedia personalisation

> "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." [Joho et al(2009)Joho, Jose, Valenti, and Sebe]

This section expands our view of this issue by looking at the multimodal interaction analysis of multimedia contents in a larger context, which is illustrated as a research framework in Figure 6.

The framework assumes two major environmental parts, namely, a private environment and social environment. The collection and storage of an end-user's multimodal information should be carried out in the private environment given that many of the recordings can contain sensitive data. On the other hand, the majority of multimedia contents is available online as the social environment. A key issue is to bridge these two environments in a practical way.

The framework broadly divided the analysis into two layers. One is the interaction analysis layer which includes the multimodal interaction analysis presented in this paper, or others with various sensory devices as described in Section 2. The analysis of this layer tends to be carried out in realtime when the end-user engages with multimedia contents to capture a user's affection, preference, intention, and other user profile information. Another is the content analysis layer which includes the analysis of visual, audio, and textual data extracted from multimedia contents. This layer's analysis can be done in offline to capture the characteristics of multimedia contents at various levels.

Given that the interaction analysis layer can provide a rich representation of user profile information, one way to scale the multimodal interaction approach is to map the significant parts (e.g., affective) of multimedia contents onto the representation of the content-analysis layer. Once this mapping is successfully carried out, then a measurement such as similarity measure can be done with all the other contents available in the social environment at the content-analysis layer. We do not claim that this is the only way to further our research. For example, a successful mapping of the interaction analysis layer to the content-analysis layer can be challenging. However, it is clear from the framework that there is ample room for further investigation to achieve a scalable multimodal approach to personal highlight detection and affective multimedia summarisation.

6 Conclusion and Future Work

Detecting a personal highlight of multimedia contents is a key research issue for affective multimedia analysis and summarisation. We proposed a facial activity-based approach to personal highlight detection, which required only a conventional webcam system unlike other approaches. The preliminary analysis of our approach suggested that the motion vectors in a upper part of human face were more likely to be indicative of personal highlights than the lower part of the face. We plan to develop a more sophisticated technique to detect personal highlights based on this finding in the future. We are also interested in the issue of mapping interaction analysis data to content analysis data to achieve a scalable multimodal profiling for multimedia contents.

Acknowledgements Funding was provided by the MIAUCE Project (EU IST-033715). Any opinions, findings, and conclusions described here are the authors and do not necessarily reflect those of the sponsor.

References

[Arifin and Cheung(2007)] Arifin S, Cheung P (2007) A computation method for video segmentation utilizing the pleasure-arousal-dominance emotional information. In: ACM International Conference on Multimedia

[Calvo and D'Mello(2010)] Calvo R, D'Mello S (2010) Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Transactions on Affective Computing 1(1):18–37

[Chan and Jones(2005)] Chan CH, Jones GJF (2005) Affect-based indexing and retrieval of films. In: ACM International Conference on Multimedia, pp 427–430

[Chen(2000)] Chen L (2000) Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction. PhD thesis, University of Illinois at Urbana-Champaign

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

[Cohen et al(2003)Cohen, Sebe, Chen, Garg, and Huang] Cohen I, Sebe N, Chen L, Garg A, Huang T (2003) Facial expression recognition from video sequences: Temporal and static modeling. Computer Vision and Image Understanding 91(1–2):160–187

[Cohen et al(2004)Cohen, Sebe, Cozman, Cirelo, and Huang] Cohen I, Sebe N, Cozman F, Cirelo M, Huang T (2004) Semi-supervised learning of classifiers: Theory, algorithms, and applications to human-computer interaction. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(12):1553–1567

[Dietz and Lang(1999)] Dietz R, Lang A (1999) Aefective agents: Effects of agent affect on arousal, attention, liking and learning. In: Cognitive Technology Conference

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

[Hanjalic(2006)] Hanjalic A (2006) Extracting moods from pictures and sounds: Towards truly personalized TV. IEEE Signal Processing Magazine 2(23):90–100

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

[Hanjalic et al(2008)Hanjalic, Lienhart, Ma, and Smith] Hanjalic A, Lienhart R, Ma WY, Smith JR (2008) The holy grail of multimedia information retrieval: So close or yet so far away? Proceedings of the IEEE 96(4):541–547

[Huijsmans and Sebe(2005)] Huijsmans D, Sebe N (2005) How to complete performance graphs in content-based image retrieval: Add generality and normalize scope. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(2):245–251

[Jaimes et al(2007)Jaimes, Gatica-Perez, Sebe, and Huang] Jaimes A, Gatica-Perez D, Sebe N, Huang T (2007) Human-centered computing: Towards a human revolution. IEEE Computer 5(40):30–34

[Joho et al(2009)Joho, Jose, Valenti, and Sebe] Joho H, Jose JM, Valenti R, Sebe N (2009) Exploiting facial expressions for affective video summarisation. In: ACM International Conference on Image and Video Retrieval (CIVR)

[Kang(2002)] Kang H (2002) Analysis of scene context related with emotional events. In: ACM International Conference on Multimedia

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

[Mehrabian(1996)] Mehrabian A (1996) Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Current Psychology 14(4):261–292

[Moncrieff et al(2001)Moncrieff, Dorai, and Venkatesh] Moncrieff S, Dorai C, Venkatesh S (2001) Affect computing in film through sound energy dynamics. In: ACM International Conference on Multimedia

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

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

[Mooney et al(2006)Mooney, Scully, Jones, and Smeaton] Mooney C, Scully M, Jones GJF, Smeaton AF (2006) Investigating biometric response for information retrieval applications. In: European Conference on Information Retrieval, pp 570–574

[Over et al(2007)Over, Smeaton, and Kelly] Over P, Smeaton AF, Kelly P (2007) The TRECVID 2007 BBC rushes summarization evaluation pilot. In: TVS '07: Int. workshop on TRECVID video summarization, pp 1–15

[van Rijsbergen(1979)] van Rijsbergen CJ (1979) Information Retrieval, Second Edition. Butterworths

[Sebe et al(2005)Sebe, Cohen, Cozman, and Huang] Sebe N, Cohen I, Cozman F, Huang T (2005) Learning probabilistic classifiers for human-computer interaction applications. Multimedia Systems 10(6):484–498

[Soleymani et al(2008)Soleymani, Chanel, Kierkels, and Pun] Soleymani M, Chanel G, Kierkels JJ, Pun T (2008) Affective ranking of movie scenes using physiological signals and content analysis. In: ACM workshop on Multimedia semantics, pp 32–39

[Sung et al(2008)Sung, Kanade, and Kim] Sung J, Kanade T, Kim D (2008) Pose robust face tracking by combining active appearance models and cylinder head models. International Journal of Computer Vision 80(2):260–274

[Tao and Huang(1998)] Tao H, Huang T (1998) Connected vibrations: A modal analysis approach to non-rigid motion tracking. In: IEEE Conference on Compter Vision and Pattern Recognition, pp 735–740

[Tjondronegoro et al(2004)Tjondronegoro, Chen, and Pham] Tjondronegoro D, Chen YP, Pham B (2004) Highlights for more complete sports video summarization. IEEE Multimedia 11(4):22–37

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

[Xu et al(2005)Xu, Chia, and Jin] Xu M, Chia L, Jin J (2005) Affective content analysis in comedy and horror videos by audio emotional event detection. In: IEEE International Conference on Multimedia and Expo