ELVIS: Entertainment-Led Video Summaries

ARTHUR G. MONEY
Brunel University, UK
and
HARRY AGIUS*
Brunel University, UK

Video summaries present the user with a condensed and succinct representation of the content of a video stream. Usually this is achieved by attaching degrees of importance to low-level image, audio and text features. However, video content elicits strong and measurable physiological responses in the user, which are potentially rich indicators of what video content is memorable to or emotionally engaging for an individual user. This paper proposes a technique which exploits such physiological responses to a given video stream by a given user to produce Entertainment-Led VIdeo Summaries (ELVIS). ELVIS is made up of five analysis phases which correspond to the analyses of five physiological response measures: electro-dermal response (EDR), heart rate (HR), blood volume pulse (BVP), respiration rate (RR), and respiration amplitude (RA). Through these analyses, the temporal locations of the most entertaining video sub-segments, as they occur within the video stream as a whole, are automatically identified. We also demonstrate how ELVIS can be integrated into video summary applications by presenting a media player that utilises ELVIS. The effectiveness of the ELVIS technique is verified through a statistical analysis of data collected during a set of user trials. Our analysis shows that ELVIS is significantly more effective than random selection in identifying the most entertaining video sub-segments for content in the comedy, horror/comedy, and horror genres.

Categories and Subject Descriptors: H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems – Video; H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing – Abstracting Methods; H.1.2 [Models and Principles]: User/Machine Systems

General Terms: Experimentation, Human Factors

Additional Key Words and Phrases: video summarisation, video content, semantics, personalisation, physiological response, affect, emotion

1. INTRODUCTION

The amount of digital video that is available to us is growing on a daily basis. As a consequence, users need assistance in accessing this content more efficiently and effectively [Furini and Ghini 2006]. Video summarisation research responds to this need by developing video summarisation techniques that condense full length video streams through the identification and abstraction of the most entertaining content within those streams. The video summaries that arise are abbreviated surrogates of the original semantic content from the video [Barbieri et al. 2003], which can subsequently be integrated into a range of applications, such as interactive browsing and searching systems, thereby offering the user an indispensable means of managing and effectively accessing digital video content [Lew et al. 2006; Li et al. 2006].

A number of video summarisation techniques have been presented in the research literature. Previously [Money and Agius 2008], we have surveyed the research literature and identified three types of techniques that can be used to generate video summaries: *internal, external* and *hybrid video summarisation techniques*. Figure 1 shows these

Authors' addresses: Brunel University, School of Information Systems, Computing and Information Systems, St John’s, Uxbridge, UB8 3PH. * denotes corresponding author

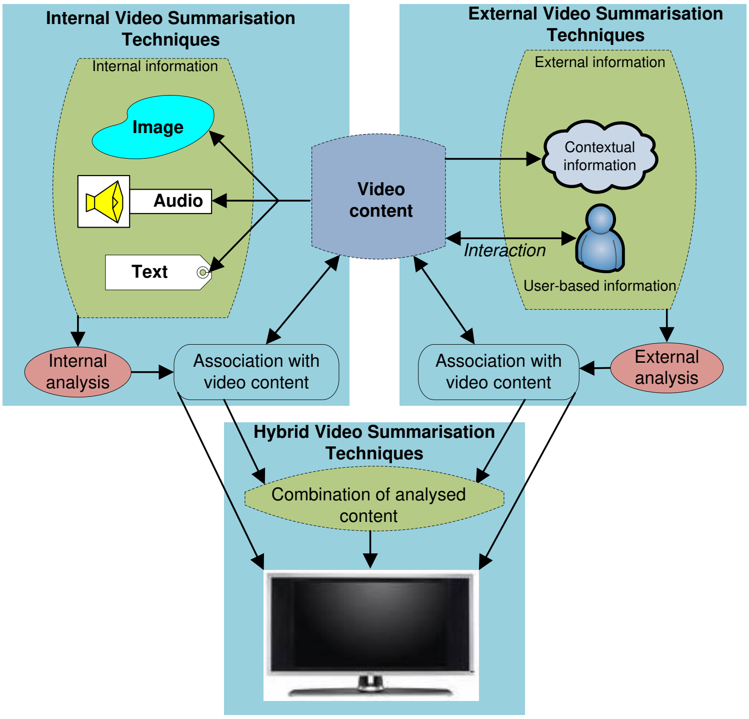

Internal Video Summarisation Techniques
Internal information
Image
Audio
Text
Internal analysis
Association with video content
Video content
External Video Summarisation Techniques
External information
Contextual information
Interaction
User-based information
External analysis
Association with video content
Hybrid Video Summarisation Techniques
Combination of analysed content

Figure 1: Internal, external and hybrid summarisation techniques

video summarisation techniques and how they relate to video content. Another useful survey of contemporary video summarisation techniques is provided by Truong and Venkatesh [2007].

The majority of video summarisation research in the last 20 years has focused on developing *internal video summarisation techniques*, which summarise video by analysing low-level features that are present (only) within the video stream such as colour, shape, object motion, speech, or on-screen text [Money and Agius 2008]. For example, Shipman et al. [2007] analyse low-level audio features to identify applause, cheering, excited speech, normal speech and music to plot a level of importance curve to represent the summarised content of the video. Damnjanovic et al. [2007] segment video by analysing the image stream. Initially scenes are identified by means of shot change detection, and then the level of motion activity measured and equated with level of importance. Jung et al. [2007] summarise television dramas. They apply theory from narrative theory in the form of a narrative abstraction model (NAM) and use it to map semantic structure onto the low level image and audio features embedded within video content. Hanjalic [2003; 2005] combines analysis of motion activity and cut density with a sound energy measure to probabilistically infer the affect related content of a video such as perceived levels of user excitement or arousal.

External video summarisation techniques tend to achieve more personalised levels of summarisation by collecting and analysing information external to the video stream, notably *contextual information*, such as the time and location in which video was recorded [de Silva et al. 2005], and/or *user-based information*, such as a user's descriptions of video content and/or browsing and viewing activity. For example, Jaimes et al. [2002] employ a high-level semantic analysis of basic manual annotations created by users, in combination with a manually supervised learning algorithm that derives a user's preference for particular content events based on their prior expressions of importance. Takahashi et al. [2005] summarise baseball videos using manual annotations;

summaries contain information such as player information, key event information (e.g. ‘plays of the ball’), and information about the extent to which the user enjoyed specific events. The annotations are temporally linked to the original video to indicate important events and individual players.

Lastly, *hybrid video summarisation techniques* use a combination of internal and external information, thereby employing both internal and external summarisation techniques, based on the premise that external techniques can compliment internal techniques by providing additional levels of detail to reduce semantic ambiguity. For example, in one hybrid technique [Aizawa et al. 2004], a video camera (worn by a user) captures video content, whilst contextual information, such as location, speed and acceleration, is captured from a worn GPS unit. Spoken voice annotations can be provided by the user. Low-level analysis on the video’s audio track is also carried out to identify interesting segments within the captured video content. Rui et al. [1999] produce video summaries automatically by using a training set of manually annotated videos which are then propagated to unannotated video content by matching against the similarity of internal video stream information. Babaguchi et al. [2001] obtain detailed contextual information sourced from sports websites about a particular soccer game and combine this with an analysis of image features. Web-based information is then associated with the video stream of that game and used to summarise the key events that occur within the game. Another example of soccer video summarisation is presented by Xu et al. [2006] who use webcast text to achieve real time event detection of live soccer coverage, where events such as goal, shot and save are identified from the text.

Despite many promising efforts, internal video summarisation techniques struggle to overcome the challenge of the semantic gap [Smeulders et al. 2000], which is the disparity between the semantics that can be abstracted by analysing low-level features and the semantics that the user associates with and primarily uses to remember the content of a video. This is primarily because contextual and user-based information is not incorporated into the analysis process at any stage. As a result, internal techniques are not able to produce personalised video summaries (summaries that represent the most significant content to an individual use), despite increasingly expectant users requiring more personalised video summaries that are in-step with their individual tastes and preferences [de Silva, et al. 2005; Lew, et al. 2006]. In light of the challenges faced by internal summarisation techniques, external and hybrid techniques are receiving more attention. This is because they are more likely to produce video summaries that are personally relevant to individual users due to the fact that they incorporate user-based and contextual information into the summarisation process. In particular, external video summarisation techniques show much promise, since they produce video summaries based purely on external information. External techniques have the added advantage of potentially being integrated into hybrid techniques if this is desirable. At present, however, there are only a small number of external techniques presented in the research literature. Furthermore, existing external techniques face challenges of their own; for example, user-based information is often obtained in the form of manual annotations from the user, which is impractical due to the time and conscious effort required. Consequently, new external information sources are needed that can be used to develop personalised video summaries but minimise the demands put on the user in terms of time and conscious mental effort.

User physiological response is one external information source since it can be captured directly from the user while requiring no conscious effort from them. Physiological response is yet to be fully incorporated into existing video summarisation techniques, however. Consequently, this paper proposes ELVIS (Entertainment-Led Video Summaries), an external video summarisation technique that identifies sub-segments (in terms of their temporal location) within a given video segment for inclusion within a video summary (which may be an entire video stream), based on real-time user physiological responses. The remainder of this paper is structured as follows. Section 2 reviews the potential of using physiological response data as video summarisation information and demonstrates the role of ELVIS within this context. In Section 3, the ELVIS technique is presented, which is used to process the users' physiological responses to video content, and identify temporally the video sub-segments for inclusion in each individual user's personalised video summaries. To demonstrate its application, the section also presents a media player that utilises ELVIS. In Section 4, the design, implementation and results of user trials to verify the effectiveness of ELVIS for three different video genres (comedy, horror/comedy, and horror) are described. Section 5 concludes the paper by discussing implications and future research directions.

2. EXTERNAL VIDEO SUMMARIES USING PHYSIOLOGICAL RESPONSE

Video content elicits strong physiological responses in the user [Brown et al. 1977; Detenber et al. 1998; Lang et al. 1999]. For example, Detenber et al. [1998] found that moving images (video) elicits higer levels of arousal, compared with still images. In addition to motion, video content (particularly that which is professionally produced) heightens arousal by using sounds and music to heighten the emotions within the user. In the words of Ian Maitland, an Emmy award-winning director [cited in Picard 1995]: "A film is simply a series of emotions strung together with a plot ... It's the filmmaker's job to create moods in such a realistic manner that the audience will experience those same emotions enacted on the screen, and thus feel part of the experience." It is therefore understandable that researchers have used video as the tool of choice to elicit user emotional response in a range of contexts, including broadcasting research [Detenber, et al. 1998; Lang, et al. 1999], psychophysiological studies [Piferi et al. 2000; Simons et al. 2000], and psychological studies [de Wied et al. 1997; Morrone-Strupinsky and Depue 2004].

Measures of user physiological response are a recognised and effective means of gaining insight into users' emotional responses to video content [Detenber, et al. 1998; Ekman et al. 1983; Gross and Levenson 1995; Lang, et al. 1999; Nasoz et al. 2003; Piferi, et al. 2000; Simons, et al. 2000; Suziki et al. 2004]. Physiological responses also provide valuable insight into real-time changes in a user's *affective state* [Allanson and Fairclough 2004; Scheirer et al. 2002], which is a generic term that refers to the user's underlying emotion, attitude, or mood at a given point in time [Simon 1982]. Affective state can be considered to be made up of two dimensions: *valence*, the level of attraction or aversion the user feels toward a specific stimulus, and *arousal*, the intensity to which an emotion elicited by a specific stimulus is felt. A range of physiological response measures have been used to infer changes in a user's affective state, the most common of which are as follows:

- *Electro-Dermal Response (EDR)* measures the electrical conductivity of the skin, which is a function of the amount of sweat produced by the eccrine glands located in the hands and feet. EDR is believed to be linearly correlated with the arousal dimension, hence the higher EDR value, the higher the arousal level and vice versa [Gomez and Danuser 2004; Gomez et al. 2004; Steinbeis et al. 2006].
- *Respiration amplitude (RA)* can be used to indicate arousal and valence levels; for example, slow deep breaths may indicate low arousal and positive valence. Shallow rapid breathing may indicate high arousal and negative valence [Frazier et al. 2004; Philippot et al. 2002].
- *Respiration rate (RR)* has been used as an indicator of arousal. An increase in the number of breaths the user takes per minute can be an indicator of increased arousal,

while lower number of breaths per minute can indicate lower levels of arousal [Gomez and Danuser 2004; Gomez, et al. 2004; Palomba and Stegagno 1993].

- *Blood Volume Pulse (BVP)* measures the extent to which blood is pumped to the body's extremities. This can serve as a measure of a user's valence: restricted blood flow to the user's extremities may indicate negative valence, while, conversely, increased blood to the extremities may indicate positive valence [Carlson 2001; Fridja 1986; Healey 2000; Picard 1997; Wang et al. 2004].
- *Heart Rate (HR)* acceleration and deceleration has also been shown to be an indicator of valence [Cacioppo et al. 1997]. Negative valence may be signified by a greater increase in HR than positive valence [Frazier, et al. 2004; Greenwald et al. 1989; Steinbeis, et al. 2006; Van Diest et al. 2001; Winton et al. 1984].

Due to advances in sensor technology, user physiological responses can be measured in real time, requiring no conscious input from the user [Money and Agius 2005]; for example, via wireless wearable sensors such as the SenseWear armband from BodyMedia. As outlined in the previous section, some internal video summarisation techniques [Hanjalic 2003; Hanjalic 2005] have been developed that summarise video streams relating to their affective content, and consequently, some multimedia metadata standards now allow for some limited affective description [Agius et al. 2006; McIntyre and Göcke 2007]. With regards to processing and evaluating physiological response data for the production of personalised video summaries, to the best of our knowledge there appears to be no research to date.

Consequently, through the development of the ELVIS technique, we explore whether users' physiological responses may serve as a suitable external source of information for producing individually personalised affective video summaries. In order to substantiate the value of physiological response as a usable external information source in its own right, ELVIS has been developed as an external video summarisation technique and therefore it also has the potential to be integrated into a hybrid technique if this were desirable. Consequently, summarisation is achieved by analysis of external information only, which in this case is user physiological response data relating to the video the user has viewed. User physiological responses are likely to be most significant during the segments of a video stream that have most relevance to that user, since these will tend to be the segments that have the most impact and are the most memorable; hence, it is these segments that are the foremost candidates for inclusion within a summarised version of the video stream. Figure 2 shows how physiological responses are used within the ELVIS technique and how the output of ELVIS can be used to playback personalised affective video summaries.

Initially, the user views the full video stream while physiological responses are captured and measured. ELVIS then processes this data and produces an internal representation of the video content in terms of the significance of physiological response. In step with the duration of the video summary requested by the user, ELVIS identifies the temporal locations of the most significant physiological responses. It should be noted that currently it is infeasible to map physiological responses onto a full range of specific emotions [Scheirer, et al. 2002], hence we do not aim to summarise and label the discrete emotions that occur within a video. Therefore, the ELVIS technique prescribes the use of the above five measures of physiological response (EDR, HR, BVP, RR and RA) and assumes that each of these measures have equal levels of importance in representing the user's response. ELVIS assumes that the higher the number of individual significant physiological responses to a video sub-segment, the higher the likelihood that this is indicative of an entertaining video sub-segment. The temporal locations of the significant responses identified by ELVIS can then be associated with the viewed video content by a media playback application (in this example, the ELVIS Media Player) and the video summary can then be played back to the user. The result is a personalised video summary that incorporates the video segments that elicited the most significant physiological responses in the user during viewing.

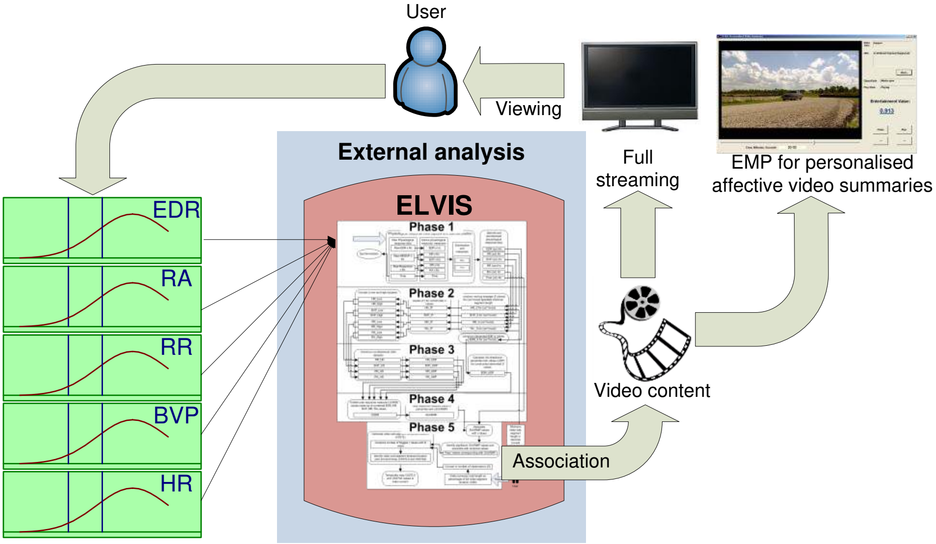


Figure 2: Physiological responses for an external video summarisation technique.

Data relating to user physiological response is typically recorded in time series format, i.e. the data is a continuous data source recorded in temporal order. Retaining the time evolving perspective, when analysing this data, gives valuable insight into the granular shifts that occur in the user’s affective state, which may not otherwise be measurable by other methods [Kramer 1991]. This format is particularly well suited to video, when considering the similarly time evolving nature of video content. However, processing of physiological response data is a non-trivial task [Picard 1997], and hence a major step towards developing video summaries, based on user physiological responses, is developing appropriate effective techniques to process this data. In the following section, we propose one such technique.

3. THE ELVIS TECHNIQUE

The ELVIS (Entertainment-Led Video Summarisation) technique processes, analyses, and evaluates user physiological responses, subsequently identifying the most entertaining video sub-segments (VSSs) within a full length video segment (VS) for inclusion in a video summary. We use ‘entertaining’ to signify content that has evoked strong and measurable physiological responses in the user. In this section, a brief summary of the ELVIS technique is presented and then a more detailed formal description is given. The ELVIS Media Player (EMP) is then presented, which serves as an example of how video summaries can be played back based on the information output by ELVIS.

3.1. ELVIS overview

Based on the premise that ELVIS has the main aim of processing user physiological responses to a given VS, so that the most entertaining VSSs from within a full length VS can be identified and included within a video summary, the five phases of processing, analysis, and evaluation are carried out. Figure 3 is an overview of the five phases that make up the ELVIS technique.

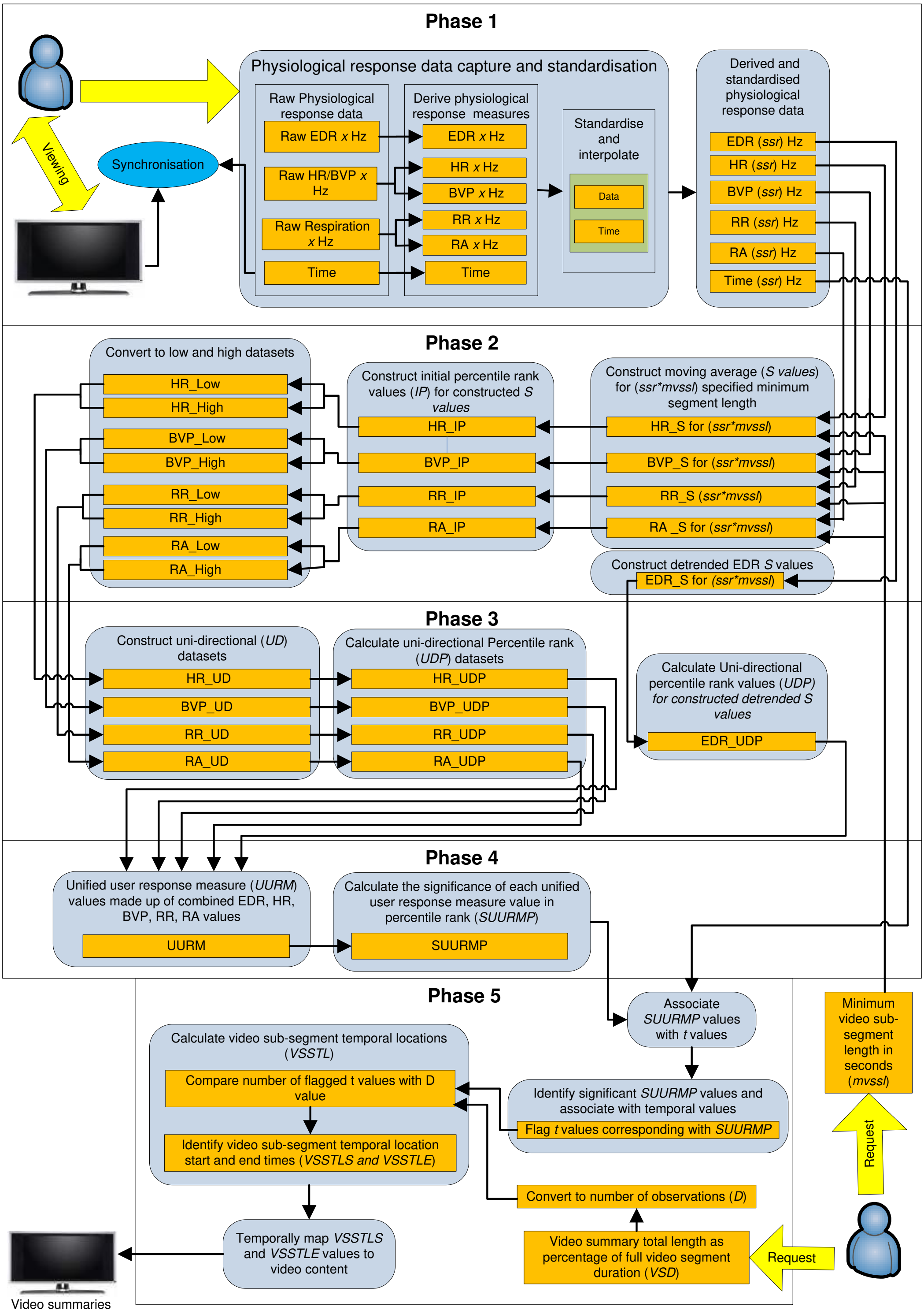

Phase 1
Viewing
Synchronisation
Physiological response data capture and standardisation
Raw Physiological response data
Raw EDR x Hz
Raw HR/BVP x Hz
Raw Respiration x Hz
Time
Derive physiological response measures
EDR x Hz
HR x Hz
BVP x Hz
RR x Hz
RA x Hz
Time
Standardise and interpolate
Data
Time
Derived and standardised physiological response data
EDR (ssr) Hz
HR (ssr) Hz
BVP (ssr) Hz
RR (ssr) Hz
RA (ssr) Hz
Time (ssr) Hz
Phase 2
Convert to low and high datasets
HR_Low
HR_High
BVP_Low
BVP_High
RR_Low
RR_High
RA_Low
RA_High
Construct initial percentile rank values (IP) for constructed S values
HR_IP
BVP_IP
RR_IP
RA_IP
Construct moving average (S values) for (ssr*mvssl) specified minimum segment length
HR_S for (ssr*mvssl)
BVP_S for (ssr*mvssl)
RR_S (ssr*mvssl)
RA _S for (ssr*mvssl)
Construct detrended EDR S values
EDR_S for (ssr*mvssl)
Phase 3
Construct uni-directional (UD) datasets
HR_UD
BVP_UD
RR_UD
RA_UD
Calculate uni-directional Percentile rank (UDP) datasets
HR_UDP
BVP_UDP
RR_UDP
RA_UDP
Calculate Uni-directional percentile rank values (UDP) for constructed detrended S values
EDR_UDP
Phase 4
Unified user response measure (UURM) values made up of combined EDR, HR, BVP, RR, RA values
UURM
Calculate the significance of each unified user response measure value in percentile rank (SUURMP)
SUURMP
Phase 5
Associate SUURMP values with t values
Minimum video sub-segment length in seconds (mvssl)
Request
Calculate video sub-segment temporal locations (VSSTL)
Compare number of flagged t values with D value
Identify video sub-segment temporal location start and end times (VSSTLS and VSSTLE)
Identify significant SUURMP values and associate with temporal values
Flag t values corresponding with SUURMP
Convert to number of observations (D)
Temporally map VSSTLS and VSSTLE values to video content
Video summary total length as percentage of full video segment duration (VSD)
Request
Video summaries

Figure 3: The ELVIS technique

In *Phase 1*, the raw physiological user response data is captured using three physiological response sensors: an EDR sensor, an HR/BVP sensor, and a respiration sensor. Five response measures (EDR, HR, BVP, RR and RA) are interpolated and standardised.

In *Phase 2*, moving window average values are calculated to produce constructed *S* (standardised) *values* for HR, BVP, RR and RA. Detrended moving window maximum deflection values are calculated for the EDR response measure. Initial percentile rank (*IP*) values are then calculated for *S values* (relating to HR, BVP, RR and RA) which are converted into bi-directional *Low* and *High* dataset values.

In *Phase 3*, *low* and *high* value datasets for each physiological response measure are combined to produce a uni-directional (*UD*) response. Uni-directional percentile rank (*UDP*) values are then calculated for each *UD* dataset; resulting in the production of *EDR_UDP*, *HR_UDP, BVP_UDP, RR_UDP* and *RA_UDP* datasets.

In *Phase 4*, *EDR_UDP, HR_UDP, BVP_UDP, RR_UDP* and *RA_UDP* are combined to produce a unified user response measure (*UURM*) dataset. The significance of unified user response measure percentile ranked (*SUURMP*) dataset is then produced from the UURM datasets.

In *Phase 5*, video sub-segment temporal location (*VSSTL)* values are derived from the values in the *SUURMP* dataset. Video sub-segment temporal location start (*VSSTLS*) and video sub-segment temporal location end (*VSSTLE*) times are based on the desired number of observations (*D*) the user has requested to be included in the video summary. The *VSSTLS* and *VSSTLE* values therefore serve as segmentation criteria that can be referenced by an external application in order to output the video summaries.

3.2. Phase 1: Standardise the raw physiological response data

The initial phase of ELVIS involves standardising the user physiological response data. This includes initial capture and temporal synchronisation of the physiological response signal with the video content, deriving appropriate response measures from the raw signal values, interpolating raw response values, standardising the sampling rate, and finally outputting completed datasets for each derived physiological response and for the time element corresponding with each physiological observation.

The five physiological measures (EDR, HR, BVP, RR and RA) are derived from the raw signal of three sensors: a skin conductance sensor for EDR, a HR/BVP sensor for HR and BVP, a respiration sensor for RA and RR. Such sensors are readily available, e.g. the ProComp Infiniti system from Thought Technologies supports a Skin Conductance Flex/Pro Sensor (SA9309M), a HR/BVP Flex/Pro Sensor (SA9308M), and a Respiration Sensor (SA9311M). At the point of capturing the user's physiological responses to video content, the data is temporally synchronised with the video content that is being viewed by the user so that physiological responses can be later mapped onto the video content. Typically, the data in its raw form is captured at various default sampling rates; for example, the ProComp Infiniti Skin Conductance and HR/BVP sensors capture raw data at 2048 Hz, whereas the Respiration Sensor captures raw data at 256 Hz. Consequently, after the data is captured, the five derived physiological response measures must be standardised and interpolated, at the standardised sampling rate (*ssr*) measured in Hz, and interpolated to enable appropriate processing and to afford comparison between response measures at a later stage. ELVIS standardises the sampling rate of the five derived physiological measures at any level up to a maximum level equal to the lowest sampling rate of the respective measures, i.e. in the above example, the maximum standardised sampling rate is 256 Hz, since this is the maximum sampling rate achieved by the respiration sensor. As demonstrated in [Money and Agius 2008], a standardisation at 8 Hz gives sufficient detail, whilst significantly reducing the necessary computation required. Hence, 8Hz is the default standardised sampling rate adopted by ELVIS.

Standardisation and interpolation of the five physiological response datasets results in all physiological observations being synchronised so that a time stamp corresponds with each physiological observation within each of the five physiological response datasets for each set of derived values (EDR, HR, BVP, RR, and RA).

3.3. Phase 2: Construct S values and convert to high-low datasets

Standardised datasets produced in Phase 1 are subjected to moving window average and detrending calculations, which then enable initial percentile rank (*IP*) datasets to be constructed. Finally, *IP* datasets can be converted into high-low datasets in order to cater for the bi-directional nature of some physiological response measures.

3.3.1. Construct S Values for HR, BVP, RR, RA and EDR

The HR, BVP, RR and RA are subjected to a moving window average calculation:

$$HR\_ \mid BVP\_ \mid RA\_ \mid RR\_ :$$

$$S_t = \frac{1}{ssr * mvssl} \sum_{i=1}^{ssr*mvssl} C_{t-(i-1)} \qquad t = ssr * mvssl, ssr * mvssl + 1, \ldots, T$$

(1)

where *ssr* is the number of Hz of the standardised physiological response data, *mvssl* is the minimum VSS length requested by the user (seconds), S_t represents the constructed HR, BVP, RR, and RA values respectively, and *t* is the indicator which identifies each point in time from which the moving average is calculated. S_t values are calculated for the full duration of user response values corresponding to the observed video content, and C_{t-i} is the actual value of the raw signal of a physiological unit at time *t-i*.

Equation 1 derives the duration of the moving window from the value provided by the user, which represents the minimum video sub-segment length (*mvssl*) the user requires to be included in the video summary. The *mvssl* value is then multiplied by the standardised sampling rate (*ssr*), i.e. the rate at which the physiological sensors have been standardised to capture response data, which determines the number of observations that must be included in each cycle of the moving average calculation. Since it is likely that the user may have their own preference as to the duration of the VSSs that are to be included in the video summary, ELVIS accommodates this by synchronising the moving average values to match the user's requested minimum segment length.

The EDR dataset is treated differently to the HR, BVP, RR and RA datasets since its baseline varies significantly during the course of an experimental session, as observed by Scheirer et al. [2002]. However, similar to the HR, BVP, RR and RA datasets, EDR_ S_t values are also calculated to reflect the user's requested minimum VSS length, where *ssr*mvssl* determines the duration of the moving window. The EDR measure can be used to identify increased eccrine gland activity, which, as discussed in Section 2, has been directly correlated with levels of arousal. However, unlike the other four physiological response measures used by ELVIS, the rate at which the EDR signal falls is often influenced by whether there has been a recent rise in the EDR signal. Typically, the EDR signal falls sharply after a rise, with the half recovery time after a rise ranging between 3 and 10 seconds [iWorx 2006]. The EDR signal will continue to fall until it reaches a baseline, unless there is another period of activation, at which point it will rise again. The rate at which the EDR signal falls cannot typically be used as a direct inference of the level to which the user is not aroused. It is only appropriate to measure the increases in EDR activity, whereby the baseline can be considered as the signal value immediately before a rise. Therefore, in order to establish an effective measure of the fluctuations of the signal, and particularly periods of increased activation, a detrended EDR signal is a more accurate representation of EDR activity. Our approach, which is an adaptation of the detrending approach taken by van Reekum and Johnstone [2004], identifies the value of the signal immediately before a rise in EDR, calculated within an appropriately sized moving window. This constitutes a method for evaluating local fluctuations in the EDR signal regardless of unpredictable baseline variations. Each *EDR_S$_t$* value is therefore calculated as follows:

$$EDR\_S_t = Max(C_t, C_{t-1}, C_{t-2}, ...C_{t-((ssr*mvssl)-1)}) - Min(C_t, C_{t-1}, C_{t-2}, ..., C_{t-((ssr*mvssl)-1)})$$

$$t = ssr*mvssl, ssr*mvssl+1, ..., T$$

(2)

where *EDR_S$_t$* represents the constructed *EDR* value, for each point in time *t* from which the moving average is calculated. *C$_t$* is the actual value of the raw *EDR* signal at time *t*. Each *EDR_S$_t$* is calculated with the assumption that the point in time corresponding with $Min(C_t, C_{t-1}, C_{t-2}, ..., C_{t-((ssr*mvssl)-1)})$ occurs prior to the point in time corresponding with $Max(C_t, C_{t-1}, C_{t-2}, ..., C_{t-((ssr*mvssl)-1)})$. If not, then *EDR_S$_t$* is set to zero.

3.3.2. Calculate initial percentile rank (IP) values for HR_S, BVP_S, RR_S and RA_S datasets

In order to standardise and normalise the *S values* for HR, BVP, RR, and RA so as to establish a measure of the significance of each value within the sample, initial percentile rank (*IP)* values are calculated for each of the four derived physiological measures, respectively. Percentile rank calculations of the constructed *S values* provide a means of ensuring that the response values for each physiological measure are directly comparable and reflect the distribution of the sample, even for skewed datasets. Therefore, percentile rank values represent the degree of significance of each physiological response value as a product of the whole dataset. ELVIS calculates percentile rank values for every *S value* in each dataset. This is a key enabling factor for ELVIS to combine and compare user responses and consequently to identify the most significant responses. The constructed *S values* produced in Equation 1 are subjected to the following calculation:

$$HR\_ \,|\, BVP\_ \,|\, RA\_ \,|\, RR\_ :$$

$$IP_t = \frac{n_e(S, S_t) + n_b(S, S_t)}{n_w(S) - 1}$$

(3)

where *S$_t$* represents a given $HR_{S(t)}|BVP_{S(t)}|RA_{S(t)}|RR_{S(t)}$ produced by Equation 1, *S* is the whole sample, $n_e(S,S_t)$ is the number of *S values* within the whole sample that are equal to *S$_t$*, $n_b(S,S_t)$ represents the number of *S values* within the whole sample that are less than *S$_t$*, and $n_w(S)$represents the total number of *S values* within the sample *S*. *IP* values are calculated for each of the physiological measures, namely HR, BVP, RA, and RR. The result of applying this calculation to the respective *S values* for each measure is a set of initial percentile rank values which represent the significance of each *S value* within the context of each physiological measure dataset.

The *IP* values for HR, BVP, RA and RR are representative of bi-directional fluctuations in user response to video content. For example, the *HR_IP$_t$* produced in Equation 3 can be used for determining significantly low heart rate values for a given threshold (*l)* or significantly high heart rate values for a given threshold (*h*) for a given HR dataset. Identifying responses representing significantly low heart rates could be achieved by $HR\_IP_t \leq l$, while significantly high heart rates could be identified by $HR\_IP_t \geq h$. A similar principle can be applied to the BVP_IP_t, RA_IP_t and RR_IP_t values.

In Phase 3, a single measure of user response to video content is constructed by combining all physiological measures to represent the overall user responses to video content. In preparation for this, HR, BVP, RA and RR *IP* values are split into *High* and *Low* datasets. The Low values are considered as *IP* values < 0.50 and High values represent values ≥ 0.50. Values in the Low datasets are then inverted (1 – Low value) so that their significance is expressed on a similar scale to the High values, i.e. on a scale of 0 to 1, with 1 being most significant and 0 being least significant. The following calculation is carried out on *IP* values in order to prepare Low and High values to be combined to form a uni-directional representation of the user responses in Phase 3. Low values are calculated as follows:

```
If HR_ | BVP_ | RA_ | RR_ : $IP_t < 0.50$
        Then HR_ | BVP_ | RA_ | RR_ : $Low_t = 1 -$ HR_ | BVP_ | RA_ | RR_ : $IP_t$
ElseIf HR_ | BVP_ | RA_ | RR_ : $IP_t \geq 0.50$
Then HR_ | BVP_ | RA_ | RR_ : $Low\_{}_t = 0$
```

High values are calculated as follows:

```
If HR_ | BVP_ | RA_ | RR_ : $IP_t < 0.50$
        Then HR_ | BVP_ | RA_ | RR_ : $High_t = 0$
ElseIf HR_ | BVP_ | RA_ | RR_ : $IP_t \geq 0.50$
Then HR_ | BVP_ | RA_ | RR_ : $High_t =$ No Change
```

3.4. Phase 3: Construct uni-directional percentile rank (UDP) datasets for each physiological response measure

In this phase the High-Low datasets produced in Phase 2 are combined to produce unidirectional datasets for each physiological response measure. These are then used as the basis for calculating uni-directional percentile rank (*UDP*) values, which represent standardised representations of each physiological response measure.

3.4.1. Construct uni-directional datasets

The Low and High value calculations carried out in Phase 2 produce two datasets for physiological measures HR, BVP, RA, and RR. Each of the datasets convert low HR, BVP, RA, and RR values so that they appear as significant responses on the same scale as high HR, BVP, RA, and RR values. Low and High HR, BVP, RA, and RR values are then combined respectively to produce significance of response values irrespective of the direction of the response measure. Combination of respective values results in one combined HR, BVP, RA, and RR dataset respectively for each physiological response measure. Combined values are calculated as follows:

$$HR\_ \mid BVP\_ \mid RA\_ \mid RR\_ :$$
$$UD_t = HR \mid BVP \mid RA \mid RR : Low_t + \ HR \mid BVP \mid RA \mid RR : High_t \qquad (4)$$

3.4.2. Calculate uni-directional percentile rank datasets

In order to facilitate comparison of values, it is necessary to re-standardise the respective measures. This is achieved by constructing percentile rank values for the respective *UD* datasets, like so:

$$HR\_|BVP\_|RA\_|RR\_:$$

$$UDP_t = \frac{n_e(UD,UD_t) + n_b(UD,UD_t)}{n_w(UD) - 1}$$

(5)

where UD_t represents a given $HR|BVP|RA|RR{:}UD_t$ value produced by Equation 4, *UD* is the whole sample, $n_e(UD,UD_t)$ is the number of *UD* values within the whole sample that are equal to UD_t, $n_b(UD,UD_t)$ represents the number of *UD* values within the whole sample that are less than UD_t, and $n_w(UD)-1$ represents the total number of *UD* values within the sample *UD* less one case. UDP_t, represents uni-directional percentile rank values which are calculated from the constructed UD_t values for HR, BVP, RR and RA calculated from point in time *t*. *UDP* values are calculated for each of the physiological measures: HR, BVP, RA, and RR respectively. The result is a uni-directional percentile ranked dataset for the measure. The higher the *UDP* value, the more significant the response value is considered to be.

Since the EDR measure is naturally presented as a uni-directional measure, representing increased levels of activation of eccrine glands, Equation 2 processes the EDR signal to identify increases in EDR activity. *EDR_S* values are converted to *UDP* values by applying the percentile rank calculation directly to the *EDR_S* values produced in Equation 2. *EDR_UDP* values for EDR are calculated as follows:

$$EDR\_UDP_t = \frac{n_e(EDR\_S, EDR\_S_t) + n_b(EDR\_S, EDR\_S_t)}{n_w(EDR\_S) - 1}$$

(6)

where *EDR_S* is the whole EDR sample, $n_e(EDR\_S, EDR\_S_t)$ is the number of *EDR_S* values within the whole sample that are equal to EDR_S_t, $n_e(EDR\_S, EDR\_S_t)$ represents the number of *EDR_S* values within the whole sample that are less than EDR_S_t, and $n_w(EDR\_S)-1$ represents the total number of *EDR_S* values within the sample *EDR_S* less one case. EDR_UDP_t are uni-directional percentile rank values calculated from the constructed EDR_S_t values calculated from point in time *t*.

Note that the *EDR_UDP* values bypass the stage of being converted to UD values, since they naturally occur in uni-directional format, whereas the HR, BVP, RA, and RR values are subjected to uni-directional conversion, before the percentile rank calculation can be carried out.

3.5. Phase 4: Construct the significance of unified user response measure percentile ranked (SUURMP) dataset

Since all physiological response measures are in standardised UDP format, these are combined to produce a unified user response measure (*UURM*) which forms the basis for constructing a significance of unified user response measure in percentile rank form (*SUURMP*) dataset.

3.5.1. Calculate UURM values

Since all physiological measures are now uni-directional percentile rank values, *EDR_UDP, HR_UDP, BVP_UDP, RA_UDP* and *RR_UDP* values are combined, representing a dataset of unified user response measure (*UURM*) values. Combining physiological measures *UURM* values are calculated as follows:

$$UURM_t = EDR\_UDP_t + HR\_UDP_t + BVP\_UDP_t + RR\_UDP_t + RA\_UDP_t$$

(7)

where EDR_UDP_t ,HR_UDP_t ,BVP_UDP_t ,RR_UDP_t , RA_UDP_t are as defined in Equations 5 and 6.

3.5.2. Calculate SUURMP from UURM values

A final percentile rank calculation is then applied to each *UURM* dataset, which standardises the responses and assigns each *UURM* value a significance rating between 0 and 1. Each *UURM* value is allocated a unique percentile rank value, which reflects the significance of each unified user response measure in percentile rank format (*SUURMP*). Each *SUURMP* value is calculated as a calculation of the total number of *UURM* values that are equal to or less than the specified *UURM* value in the dataset. *UURM* are calculated to a default of five decimal places, providing the potential for 99999 unique values, which, for an 8Hz signal, is sufficient to uniquely represent user responses to more than 208 minutes of video content. *SUURMP* values are calculated using a similar calculation as Equations 5 and 6, but substituting *UURM* values as the comparison and full sample values, as follows:

$$SUURMP_t = \frac{n_e(UURM, UURM_t) + n_b(UURM, UURM_t)}{n_w(UURM) - 1}$$

(8)

where *UURM* is the whole sample, $n_e(UURM, UURM_t)$is the number of *UURM* values within the whole sample that are equal to $UURM_t$, , $n_b(UURM, UURM_t)$ represents the number of *UURM* values within the whole sample that are less than $UURM_t$, and $n_w(UURM)$*-1* represents the total number of *UURM* values within the sample *UURM* less one case. $SUURMP_t$ are the significance of each unified user response measure in percentile rank format calculated from the constructed $UURM_t$ values calculated from point in time *t*.

3.6. Phase 5: Identify segments for inclusion in video summary

In this final phase, the *SUURMP* dataset produced in Phase 4 is used as a means of temporally identifying segments of video that should be included in the final video summary. *SUURMP* values are processed according to the percentage of the video segment duration (*VSD*) that the user requires the video summary to be. These values represent a series of video sub-segment temporal location start (*VSSTLS*) and video sub-segment temporal location end (*VSSTLE*) values, which represent the temporal location of video sub-segments that are to be included in the final video summary.

3.6.1. User requested video summary duration

In order to temporally identify VSSs for inclusion in the final video summary, the required video summary duration (*VSD*) is initially specified by the user as a percentage ranging between 1% and 100%. This is required to calculate the total number of observations (*D*) that must be selected in order to identify the correct percentage of physiological response values that correspond with the viewed video content. The total number of observations collected is determined as the product of the standardised sampling rate (*ssr*) in Hz and the full video summary duration (*VSD*) in seconds. Therefore *D* = *ssr*VSD*. As an example, consider a user who requests 30% video summary of a video that has a total run time of 100 minutes (6000 seconds); therefore VSD = 6000, so let *ssr* = 8Hz (i.e. 8 observations per second). *D* = *VSD*ssr*, which equates to 48000 observations over the course of the video; hence, *D* = 14400 (where 30% of 48000 = 14400).

3.6.2. Incrementally identify the most significant SUURMP values

The *D* value is then taken as an input value for the video sub-segment temporal location (*VSSTL*) function which is applied to the *SUURMP* dataset. The *VSSTL* function incrementally steps through the *SUURMP* dataset values, starting with the highest value (representing the most significant user response value), in order to identify the temporal locations of the most significant user responses. Each *SUURMP* value is temporally associated with the video content that elicited the physiological responses from which the SUURMP values were calculated. Consequently a given $SUURMP_t$ value temporally represents the user's physiological responses to VSS for the duration of *t,…,t-(ssr*mvssl*). Therefore, a number of *t* values correspond with one $SUURMP_t$ value, thus ensuring that the calculations carried out in Equations 1 and 2 (that aggregate the raw physiological responses for the time period *ssr*mvssl)* are taken into account when identifying the time period that a selected $SUURMP_t$ corresponds with. Each *t* value in the dataset has a $SUURMP\_Flag_t$ value associated with it, which can be set to true or false (default value is false). When a $SUURMP_t$ value is selected, the associated $SUURMP\_Flag_t$ values are true, indicating that the time points corresponding with the selected $SUURMP_t$ value represent points in time mapping to the original video content that should be included in the final video summary. A count of the number of $SUURMP\_Flag_{(t)}$ values flagged as true is then performed. On each occasion an additional $SUURMP_t$ value is identified, until the number of flagged $SUURMP\_Flag_{(t)} \geq D$. This results in a number of consecutively flagged groups of *t* values, which identify the temporal locations of video sub-segments that are to be included in the video summary. Flagging of the *t* values is carried out according to the algorithm presented in Figure 4(a).

```
k = 1
For h = 1 to 99999
For t = 1 to (VSD*ssr)
                If  = k then
                        For j = 1 to (ssr*mvssl)
                                $SUURMP\_Flag_{(t+1-(j))}$ = true
                        Next
                        Count(SUURMP_Flag = true)
                        If Count ≥ D then
                                h = 99999
                        End If
                End If
        Next
k = k - 0.00001
Next
```

(a)

```
For t = 1..(VSD*ssr)
        If $SUURMP\_Flag_{(t)}$ = true then
                VSSTLS = $Time\_SUURMP_{(t)}$
                For j =1 to (VSD*ssr)
                        If  $SUURMP\_Flag_{(t+j)}$ ≠ true then
                                VSSTLE = $Time\_SUURMP_{t+(j-1)}$
                                j = (VSD*ssr)
                        End if
                Next
        End if
Next
```

(b)

Figure 4: (a) Flagging of *t* values (b) Calculating VSSTLS and VSSTLE values

3.6.3. Calculate video sub-segment temporal locations

Finally, each group of consecutively flagged *t* values is processed in order to identify the start and end points of the VSSs. The *t* value that corresponds with the first value of each flagged group represents a VSS temporal location start (*VSSTLS*) and the last *t* value in a flagged group represents the video sub-segment temporal location end (*VSSTLE*). These *VSSTLS* and *VSSTLE* values are output as a series of time stamps (seconds) which can then be fed into an external application for the rendering of the video summary. Figure 4(b) gives an example of how the *VSSTLS* and *VSSTLE* values are calculated.

3.7. Video summaries produced by ELVIS

To demonstrate how the ELVIS technique can be used to produce video summaries, the ELVIS Media Player (EMP) has been developed to playback video summaries based on the data output by ELVIS. EMP was developed using standard VB.NET tools and the

Windows Media Player 11 Software Development Kit (WMP11SDK). It plays back personalised video summaries based on the VSSs identified for inclusion in a video summary by ELVIS. EMP directly references the *VSSTLS* and *VSSTLE* values calculated by ELVIS, which are then used to identify the respective video sub-segment start and end temporal locations. EMP also references the *SUURMP* values produced by the ELVIS technique, which are presented as the *entertainment value* associated with each VSS as it is played back on-screen. The entertainment value for each VSS is the highest *SUURMP* value that was recorded for each respective video sub-segment. Entertainment values range between 0 and 1, where 1 is the highest entertainment value.

To demonstrate how *VSSTLS*, *VSSTLE* and *SUURMP* values produced by the ELVIS technique are used by the EMP, ELVIS was applied to user physiological response values for video segments from various TV shows, including a 30 minute episode of the BBC's 'Top Gear' (a very popular UK motoring show). Figure 5 plots, in temporal order, the *SUURMP* values calculated by the ELVIS technique which are used to represent the entertainment values as they unfold over the course of the video segment, which are used to produce *entertainment value curves*, and shows the EMP playing back the first six selected video sub-segments. In this example, the user requested that 40% of the original video segment should be included in the video summary. In addition, a minimum video sub-segment length of 30 seconds was specified. As can be seen, nine video sub-segments were selected by the ELVIS technique.

During playback of the VSSs, EMP displays the entertainment value. As can be seen, the entertainment value of 0.981 was observed for the 'Shopping for a Corvette' video sub-segment, which corresponds with the highest entertainment value recorded during this video sub-segment (this can be verified by examining the entertainment value curve). Similarly, the second selected video sub-segment, 'Threat of being shot beyond 79th street', had an entertainment value of 0.916 which also corresponds with the entertainment value curve. The same can be observed for all nine video sub-segments.

4. VERIFYING VIDEO SUMMARIES PRODUCED BY ELVIS

In order to empirically verify the effectiveness of ELVIS to identify the most entertaining video sub-segments for inclusion in a video summary, a set of user trials was carried out. The aim of these user trials was to verify the performance of ELVIS compared with chance (RANDOM) in matching the most entertaining video sub-segments as self-reported by individual users. Consequently the following hypotheses are posed as being of primary concern for this study:

Null Hypothesis (H_0): The ELVIS technique, on average, does no better at matching the self-reported video sub-segments than a RANDOM selection of video sub-segments.

Research Hypothesis (H_1): The ELVIS technique, on average, does better at matching self-reported video sub-segments than a RANDOM selection of video sub-segments.

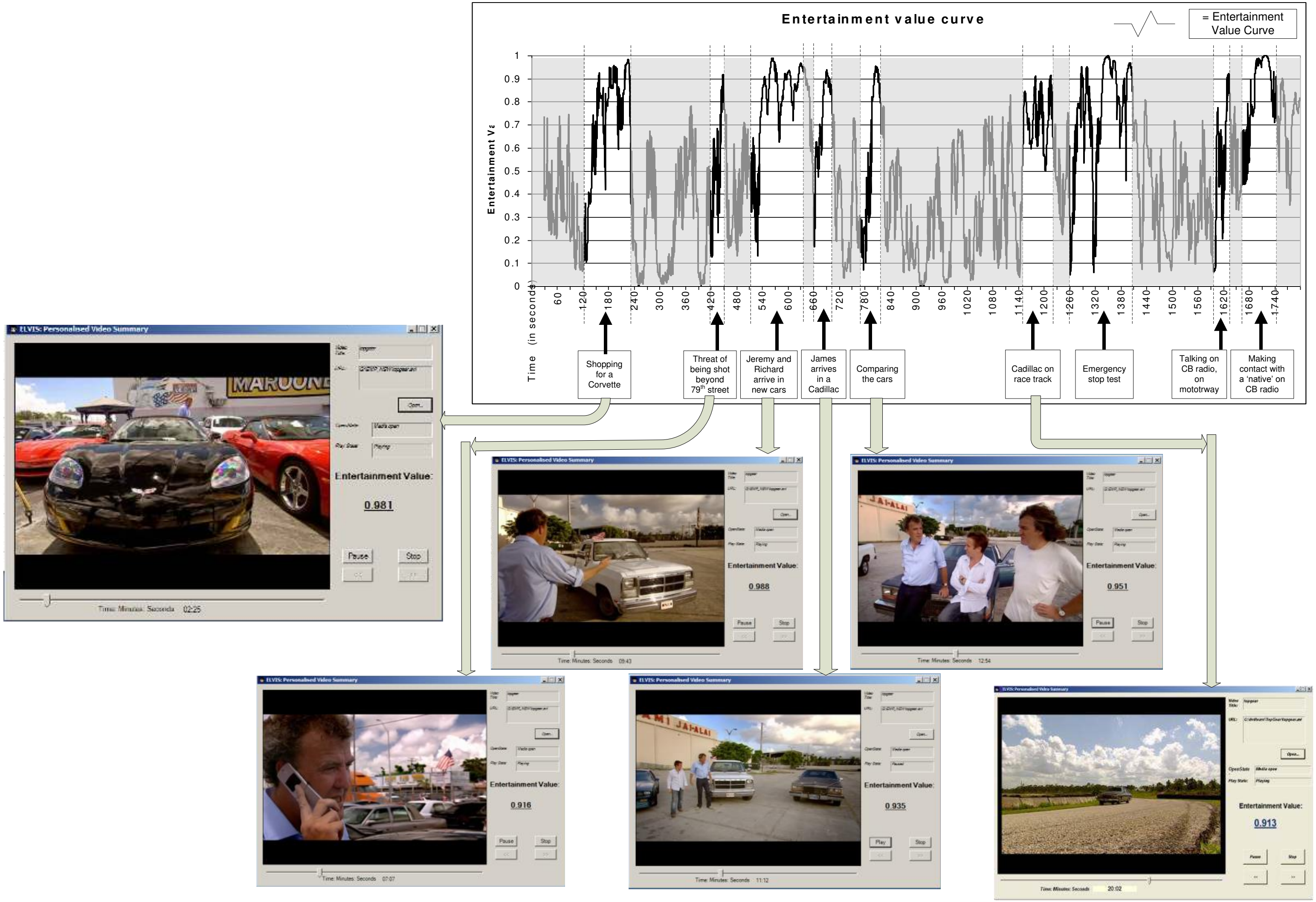


Figure 5: Graphical summary showing the entertainment value curve and the video sub-segment selections identified by ELVIS (as displayed in the EMP) for the first 30 minutes of Top Gear.

4.1. Trials procedure

In these user trials, physiological response data was collected from 60 users as they viewed video content using the ProComp Infiniti system and BioGraph software produced by Thought Technologies. Each user viewed one of three 35 minute video segments (VSs), thus each of the three VSs was viewed 20 times (20 users per VS). Based on research findings from a prior study [Money and Agius 2006] that revealed their efficacy, VSs from the Comedy, Horror/ Comedy, and Horror genres were used. These were an episode from Series 2 of Fawlty Towers entitled 'The Psychiatrist' [Spiers 1979]), Shaun of the Dead [Wright 2004], and The Others [Amenabar 2001], respectively.

After viewing the VS, the user was presented with still screenshot cards representing the video content they had just viewed. Each card, containing three screenshots (one screenshot per five seconds of video), represented 15 seconds of video content. Therefore, 140 cards were used to represent each 35 minute VS. The user was required to self-report a set of video sub-segments by selecting 42 screenshot cards out of 140 (also equalling 30% of the total duration of the viewed video segment) that they deemed to be most entertaining (note that this stage is not part of the ELVIS technique, it is for verification purposes only). The user response data collected during the trial was then processed by ELVIS, and as a result specific video sub-segments (totalling 30% of the viewed video content) for each of the users that took part in the trials were identified for inclusion within video summaries. A RANDOM selection of video sub-segments (also totalling 30%) was also identified.

The extent to which the ELVIS and RANDOM video sub-segment selections overlapped (matched) with the self-reported video sub-segment selections was then calculated. The extent to which the RANDOM selection overlapped with self-reported video VSS selections served as a baseline/control (equivalent to chance), with which the overlap percentages achieved by ELVIS could be compared. This then served as a basis against which a statistical analysis of the differences between ELVIS and RANDOM overlap percentages could be carried out, in order to establish whether ELVIS significantly outperformed RANDOM in matching the most entertaining VSSs as reported by the end user. Verifying the effectiveness of video summaries, by measuring the extent to which video sub-segments overlap with a benchmark selection of video sub-segments, is a recognised verification approach, and hence followed by a number of video summarisation studies [e.g. Babaguchi et al. 2004; Moriyama and Sakauchi 2002; Rui et al. 2000]. Figure 6 provides an overview of the process adopted to evaluate the ELVIS and RANDOM video sub-segment selections. In the next section, the statistical analysis method is presented in more detail.

4.2. Statistical analysis method

A statistical analysis of overlap percentages achieved by the two video sub-segment selection procedures (ELVIS and RANDOM) was carried out for each group of 20 users to evaluate the extent to which each of the respective video sub-segment selections achieved statistically significant overlaps with self-reported video sub-segments. The primary aim of this analysis was to verify the performance of ELVIS in identifying the most entertaining video sub-segments in accordance with the hypotheses outlined at the start of this section. For each of the three video segments used, the user trials can be considered as a within subjects one-way repeated measures design with the above two treatment conditions:

- ELVIS selection
- RANDOM selection

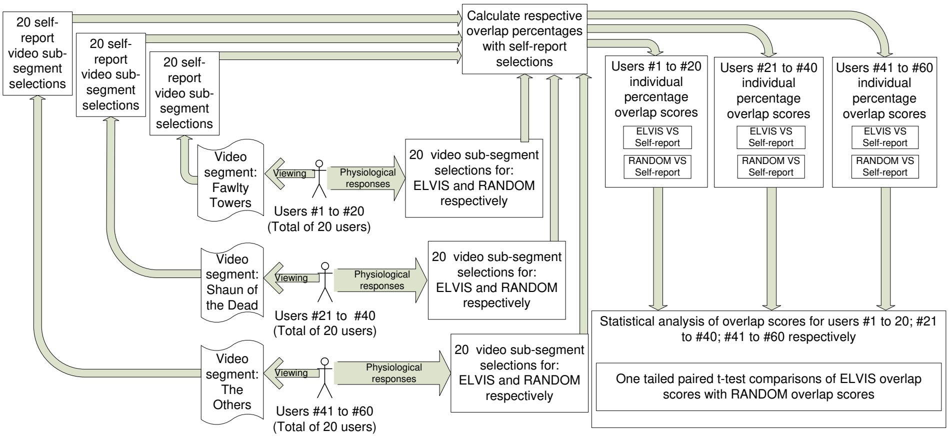

20 self-report video sub-segment selections
20 self-report video sub-segment selections
20 self-report video sub-segment selections
Calculate respective overlap percentages with self-report selections
Users #1 to #20 individual percentage overlap scores
ELVIS VS Self-report
RANDOM VS Self-report
Users #21 to #40 individual percentage overlap scores
ELVIS VS Self-report
RANDOM VS Self-report
Users #41 to #60 individual percentage overlap scores
ELVIS VS Self-report
RANDOM VS Self-report
Video segment: Fawlty Towers
Viewing
Physiological responses
20 video sub-segment selections for: ELVIS and RANDOM respectively
Users #1 to #20 (Total of 20 users)
Video segment: Shaun of the Dead
Viewing
Physiological responses
20 video sub-segment selections for: ELVIS and RANDOM respectively
Users #21 to #40 (Total of 20 users)
Video segment: The Others
Viewing
Physiological responses
20 video sub-segment selections for: ELVIS and RANDOM respectively
Users #41 to #60 (Total of 20 users)
Statistical analysis of overlap scores for users #1 to 20; #21 to #40; #41 to #60 respectively
One tailed paired t-test comparisons of ELVIS overlap scores with RANDOM overlap scores

Figure 6: Overview of trials and statistical analysis process for evaluating video sub-segment selections of ELVIS

In order to formally test hypotheses H_0 and H_1, paired t-tests are performed for overlap scores achieved by ELVIS compared to RANDOM. The P-values produced by the paired t-tests allow the performance of ELVIS to be evaluated and to establish whether ELVIS outperforms the RANDOM treatment condition to a statistically significant degree ($\alpha = 0.05$).

In addition, the effect size (Cohen's d) of the difference between respective treatment conditions as defined by Cohen [Clark-Carter 1997] is also calculated. Often, Cohen's d and the associated power analysis are carried out retrospectively. First, the hypothesis test is performed using a one-tail paired t-test. Should the null hypothesis be rejected then the effect size is estimated from the sample values. After this, the associated power for the estimated effect size is determined by looking up the corresponding d value from a table similar to Table A15.3 in [Clark-Carter 1997]. This is the approach taken in this study.

We propose that a large effect size would be necessary if ELVIS is to be of practical use to the end user. Consequently, for the repeated measures trial in this study, it can be determined that for a sample of size 20 (20 users per VS), the use of a one-tailed paired t-test performed at the 5% level of significance is capable of detecting a large effect size (d = 0.8) with power 0.96 (see [Clark-Carter 1997], Table A15.3, p.609). In practice, researchers usually perform the test at the 5% level of significance and try to achieve a power of 80% [Clark-Carter 1997]. So for this study, for each group of 20 users, a power of 96% is substantially above the level normally aspired to in practice.

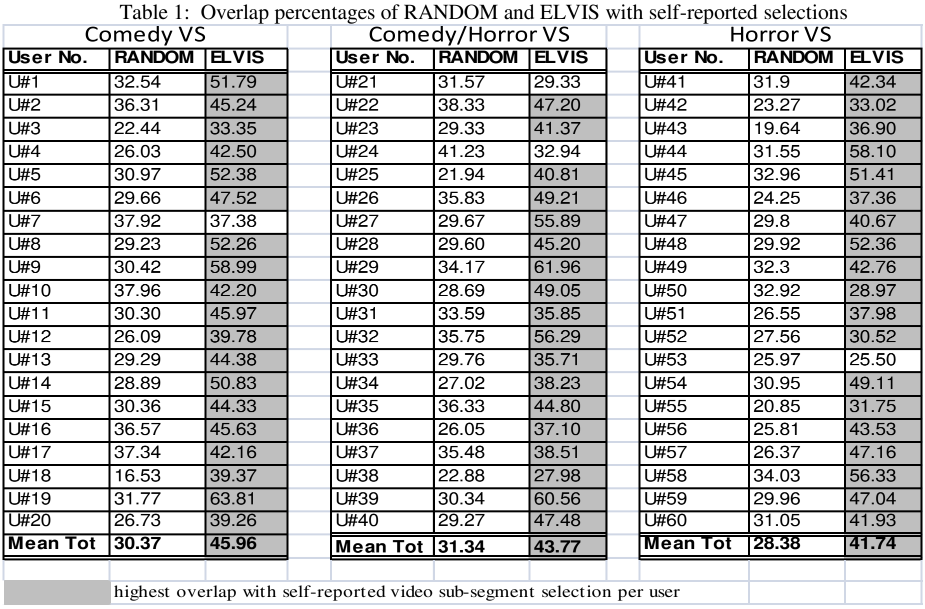
4.3. User trial results

In this section, the overlap percentages achieved for each respective user by ELVIS and RANDOM treatment conditions are first presented, followed by the results of the paired t-tests comparing ELVIS mean overlap percentages with the RANDOM treatment condition. The results for each video segment, Comedy VS, Horror/Comedy VS, and Horror VS, are presented in turn.

4.3.1. User trial results

The overlap percentages achieved by ELVIS and RANDOM treatment conditions for each individual user are presented in Table 1. These percentages show the extent to which the ELVIS and RANDOM selections overlapped with the users' self-reported selection of VSSs. The results in Table 1 are presented in three parts, relating to each of the three video content types used in the user trials: Comedy, Comedy/Horror, and Horror.

Table 1: Overlap percentages of RANDOM and ELVIS with self-reported selections



Comedy VS				Comedy/Horror VS				Horror VS		
User No.	RANDOM	ELVIS		User No.	RANDOM	ELVIS		User No.	RANDOM	ELVIS
U#1	32.54	51.79		U#21	31.57	29.33		U#41	31.9	42.34
U#2	36.31	45.24		U#22	38.33	47.20		U#42	23.27	33.02
U#3	22.44	33.35		U#23	29.33	41.37		U#43	19.64	36.90
U#4	26.03	42.50		U#24	41.23	32.94		U#44	31.55	58.10
U#5	30.97	52.38		U#25	21.94	40.81		U#45	32.96	51.41
U#6	29.66	47.52		U#26	35.83	49.21		U#46	24.25	37.36
U#7	37.92	37.38		U#27	29.67	55.89		U#47	29.8	40.67
U#8	29.23	52.26		U#28	29.60	45.20		U#48	29.92	52.36
U#9	30.42	58.99		U#29	34.17	61.96		U#49	32.3	42.76
U#10	37.96	42.20		U#30	28.69	49.05		U#50	32.92	28.97
U#11	30.30	45.97		U#31	33.59	35.85		U#51	26.55	37.98
U#12	26.09	39.78		U#32	35.75	56.29		U#52	27.56	30.52
U#13	29.29	44.38		U#33	29.76	35.71		U#53	25.97	25.50
U#14	28.89	50.83		U#34	27.02	38.23		U#54	30.95	49.11
U#15	30.36	44.33		U#35	36.33	44.80		U#55	20.85	31.75
U#16	36.57	45.63		U#36	26.05	37.10		U#56	25.81	43.53
U#17	37.34	42.16		U#37	35.48	38.51		U#57	26.37	47.16
U#18	16.53	39.37		U#38	22.88	27.98		U#58	34.03	56.33
U#19	31.77	63.81		U#39	30.34	60.56		U#59	29.96	47.04
U#20	26.73	39.26		U#40	29.27	47.48		U#60	31.05	41.93
Mean Tot	**30.37**	**45.96**		**Mean Tot**	**31.34**	**43.77**		**Mean Tot**	**28.38**	**41.74**

▨ highest overlap with self-reported video sub-segment selection per user

As can be seen, for the Comedy VS, ELVIS achieved higher overlap scores than RANDOM in 19 out of 20 cases; the only exception being U#7, for which RANDOM achieved an overlap of 37.92% compared with ELVIS which achieved 37.38%. Overall for the Comedy/Horror VS, ELVIS achieved higher overlap scores than RANDOM. The mean total of all percentage overlap scores for ELVIS at 45.96% was higher than the RANDOM mean total of 30.37%.

The Comedy/Horror VS once again resulted in ELVIS achieving higher overlap scores than RANDOM. For this type of video content, ELVIS achieved higher overlap scores for 18 out of 20 users. The only exceptions were for U#21 and U#24, where the RANDOM versus ELVIS overlap scores were 31.57% and 41.23% versus 29.33% and 32.94% respectively. The overall mean total overlap score achieved by ELVIS was 43.77%, which was higher than the RANDOM mean total overlap score of 31.34%.

The Horror VS also showed that ELVIS achieved higher overlap percentages than RANDOM for the majority of users (19 out of 20 cases). The only exception was U#53, where RANDOM achieved an overlap score of 25.97% compared with ELVIS which achieved 25.50%. Overall the mean total on the overlap percentages showed that ELVIS achieved 41.74% which was a higher score than RANDOM which achieved 28.38%.

When considering the mean total overlap percentages achieved for each of the three video content types, ELVIS achieved the highest score for Comedy VS (45.96%), the next highest score was achieved for the Comedy/Horror VS (43.77%) and lowest mean total overlap score was achieved for the Horror VS (41.74%). In all three cases, ELVIS achieved higher mean total overlap scores than RANDOM.

The results of the paired t-tests comparing mean overlap differences between ELVIS and RANDOM treatment conditions for the Comedy, Horror/Comedy and Horror VSs are now presented. Included in the results are measures of significance of the differences in means overlap scores, and effect size and power values. Table 2 presents the mean paired differences in percentage overlap between ELVIS and RANDOM for Comedy, Horror/Comedy and Horror VSs respectively. To assess ELVIS from a practical significance perspective, a retrospective power analysis (as described in Section 4.2) was carried out. As a result of this analysis, Table 2 presents the effect sizes achieved in the user trials, which are estimated and recorded in the column labelled "Effect size (d)", and the power (i.e. the probability of correctly accepting the alternative hypothesis) corresponding to the effect size, which is determined by looking up the power corresponding to the estimated effect size in Table A15.3 in [Clark-Carter 1997]; the latter is presented in the last column of Table 2 and labelled "Est. Power".

Table 2: Paired t- test, effect size and power for Comedy, Horror/Comedy and Horror VSs

	Paired Differences						
	Mean	Std. Dev.	t	df	Sig. (1-tailed)	Effect size (d)	Est. Power
Comedy: ELVIS - RANDOM	15.59	8.18	8.524	19	0.000	1.91	1.00
Horror/Comedy: ELVIS - RANDOM	12.43	10.06	5.529	19	0.000	1.24	1.00
Horror: ELVIS - RANDOM	13.36	7.75	7.712	19	0.000	1.72	1.00

▨ statistically significant at the 5% level

As can be seen from the "Sig. (1-tailed)" column, ELVIS performs significantly better than RANDOM at the 5% level of significance. Therefore, in statistical terms, ELVIS performed significantly better than RANDOM. The effect size achieved for the Comedy VS is large (Cohen's d suggests over 0.8 can be considered a large effect size), with the estimated power of 100%. There is therefore strong evidence of the ability of ELVIS to outperform RANDOM by a 'large' amount. Based on the results for the Comedy VS, there is strong evidence to reject the null hypothesis (H_0) and support the research hypotheses (H_1), i.e. that ELVIS achieves on average significantly higher mean percentage overlap scores compared to RANDOM.

As can be seen from results of the paired t-tests comparing mean overlap differences between ELVIS and RANDOM treatment conditions for the Horror/Comedy VS in Table 2 in the column labelled "Sig. (1-tailed)", statistically ELVIS performs significantly better than RANDOM at the 5% level of significance. The effect sizes achieved in the user trials were also large for ELVIS compared to RANDOM, with an effect size of 1.24. So, there is strong evidence of the ability of ELVIS to outperform RANDOM by a large amount. Based on the results for the Horror/Comedy VS, there appear to be significant differences in mean percentage overlap scores between the ELVIS and the RANDOM treatment condition i.e. ELVIS achieved a significantly higher mean percentage overlap score compared to RANDOM. There is therefore strong evidence to reject the null hypothesis (H_0), and support the research hypotheses (H_A), i.e. that ELVIS achieves on average significantly higher mean percentage overlap scores compared to RANDOM.

The results of the paired t-tests comparing mean overlap differences between ELVIS and RANDOM treatment conditions for Horror VS, as can be seen from the column labelled "Sig. (1-tailed)" in Table 2**,** statistically ELVIS performs significantly better than RANDOM at the 5% level of significance. The effect sizes achieved in the user trials were again large for ELVIS compared to RANDOM with an effect size of 1.72. Therefore, in this case, there is strong evidence of the ability of ELVIS to outperform the RANDOM treatment condition by a large amount. The results of this study show that there appear to be significant differences in mean percentage overlap scores between treatment conditions. Specifically, ELVIS achieved significantly higher mean percentage overlap scores compared to RANDOM, and therefore there is strong evidence to reject the null hypothesis (H_0), and support the research hypotheses (H_1).

To summarise the results findings of these trials, statistically ELVIS achieved significantly higher mean overlap percentages compared to RANDOM for each of the three video segments viewed in the user trials. Therefore, in all cases, not only was there strong evidence that the null hypothesis (H_0) could be rejected at the 5% level of significance, but in all cases, the estimated power values were close to 100%, all with a large effect size. This also provided strong evidence that there is a high probability that the research hypothesis (H_1) could be correctly accepted. In other words, there was strong evidence to show that on average ELVIS does significantly better at matching self-reported video sub-segments than a RANDOM selection. Furthermore, due to the large effect size achieved by ELVIS compared with RANDOM in every case, it is proposed that the differences in mean overlap percentage are large enough to be of practical value to the user of a video summarisation system that uses the ELVIS technique. The statistically significant results achieved by ELVIS has a number of implications for our own video summarisation research and the video summarisation research domain as a whole, which are now discussed.

5. CONCLUSIONS

Current video summarisation research has shown that although internal summarisation techniques successfully summarise video content, they are not able to produce personalised video summaries and still face the challenge of overcoming the semantic gap [Smeulders, et al. 2000]. In response to these challenges, external and hybrid techniques are receiving more attention, however, there is only a limited number of existing external video summarisation techniques presented in the literature. Consequently there is a need to identify new external information sources, and develop external video summarisation techniques that successfully use these information sources.

In this paper, we have proposed that physiological response data may potentially serve as a valuable external information source for personalised affective video summarisation. As a result, we have presented the ELVIS technique, which effectively processes and analyses physiological response data and identifies the most entertaining video sub-segments according to the user's physiological responses to video content. The Elvis Media Player (EMP) was also presented, to demonstrate how ELVIS can be used within real world video browsing and playback applications. In order to verify the effectiveness of ELVIS in identifying the most entertaining video sub-segments, a set of laboratory based user trials were carried out in which 60 users viewed one of three video segments representing content from comedy, comedy/horror and horror genres. Subsequently, one-tailed paired t-tests were carried out to compare the extent to which video sub-segments identified by ELVIS and randomly selected video sub-segments matched the most entertaining video sub-segments as self-reported by the user. ELVIS was shown to consistently overlap with self-reported video sub-segment selections at a significantly higher level of accuracy compared with a randomly selected selection. Furthermore, the level of overlap achieved by ELVIS was significant enough to achieve a large effect size which indicates that the video summaries produced by ELVIS are likely to be of practical value to the end user.

The fact that ELVIS has been shown to consistently identify the most entertaining video sub-segments for individual users across a range of video content, has numerous implications for future video summarisation research, which include the following:

- In light of the need to find new external sources of information to assist in the video summarisation process overcoming the long-standing challenge of the semantic gap [Smeulders, et al. 2000], this study demonstrates that physiological response data can be used to produce personalised video summaries. This is a valuable level of detail relating to the extent to which the individual user was 'entertained' while viewing specific sub-segments of video content that does not appear to be available via video stream based information sources.
- In light of the fact that processing physiological response data is a non-trivial task [Picard 1997], the ELVIS technique provides a valuable means of achieving video summaries based on this information in spite of the complexities posed by this type of data. As a result, the ELVIS technique serves as an example of the feasibility of physiological response data being used within the video summarisation context, and opens the door to such data being more frequently incorporated within the context of video summarisation research.
- Only a small number of external summarisation techniques exist within current video summarisation research literature [Money and Agius 2008]. As demonstrated in this study, the ELVIS technique has the potential of summarising video content as a standalone solution, as demonstrated via EMP, and thus can be considered as a valuable addition to the range of existing external video summarisation techniques.
- Existing internal and hybrid video summarisation techniques now have the potential to widen the range of summarisable semantics by incorporating the ELVIS technique into existing solutions. In turn, this would be beneficial to the user by providing a wider range of personally relevant semantics by which the user could access the content within a video.

In terms of future research, the potential hybridisation of ELVIS may be explored by integrating the ELVIS technique into existing internal video summarisation techniques. This should serve to maximise the range of semantics that can be extracted from video content. In this way, the relative strengths of ELVIS can be used to complement other existing video summarisation techniques. Furthermore, as wireless wearable sensors that stream physiological response data directly from the user to a remote home entertainment device become more of a reality, the notion of an ELVIS based system used in a home based 'living room' setting becomes feasible. With such sensors, the ELVIS technique could be used to unobtrusively generate video summaries as family members view content streamed to their set top boxes, thus creating a repository of popular, entertaining video sub-segments that family members may wish to later enjoy with each other, less immediate family members, and friends. Future research would need to address the development of software that was appropriate for use in the home by less experienced users and which could seamlessly store and organise summarised video content so that it was readily accessible at a later time.

6. REFERENCES

AGIUS, H., CROCKFORD, C. AND MONEY, A.G. 2006. Geographic video content. In *Encyclopedia of Multimedia*, B. FURHT Ed. Springer, New York, NY, USA, 257-259.

AIZAWA, K., TANCHAROEN, D., KAWASAKI, S. AND YAMASAKI, T. 2004. Efficient retrieval of life log based on context and content In *Proceedings of the 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences (CARPE '04)*, New York, NY, USA, 15 October, ACM Press, 22-31.

ALLANSON, J. AND FAIRCLOUGH, S.H. 2004. A research agenda for physiological computing. *Interacting with Computers 16*, 857-878.

AMENABAR, A. 2001. *The Others*. Miramax.

BABAGUCHI, N., KAWAI, Y. AND KITAHASHI, T. 2001. Generation of personalized abstract of sports video. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '01)*, Tokyo, Japan, 22-25 August, IEEE, 800-803.

BABAGUCHI, N., KAWAI, Y., OGURA, T. AND KITAHASHI, T. 2004. Personalized abstraction of broadcasted American football video by highlight selection. *IEEE Transactions on Multimedia 6*, 575-586.

BARBIERI, M., AGNIHOTRI, L. AND DIMITROVA, N. 2003. Video summarization: methods and landscape. In *Internet Multimedia Management Systems IV*, J.R. SMITH, S. PANCHANATHAN AND T. ZHANG Eds. SPIE, Bellingham, WA, USA, 1-13.

BROWN, W.A., CORRIVEAU, D.P. AND MONTI, P.M. 1977. Anger arousal by a motion picture: A methodological note. *American Journal of Psychiatry 134*, 930-931.

CACIOPPO, J.T., BERNTSON, G.G., KLEIN, D.J. AND POEHLMANN, K.M. 1997. The psychophysiology of emotion across the lifespan. *Annual Review of Gerontology and Geriatrics 17*, 27-74.

CARLSON, N.R. 2001. *Psychology of Behaviour*. Allyn and Bacon, Boston, MA, USA.

CLARK-CARTER, D. 1997. *Doing Quantitative Psychological Research: From Design to Report*. Psychology Press, London.

DAMNJANOVIC, U., PIATRIK, T., DJORDJEVIC, D. AND IZQUIERDO, E. 2007. Video summarisation for surveillance and news domian. In *Proceedings of the the second international conference on semantic and digital media technologies*, Genova, Italy, 5-7 December 2007, Springer-Verlag, 99-102.

DE SILVA, G., YAMASAKI, T. AND AIZAWA, K. 2005. Evaluation of video summarization for a large number of cameras in ubiquitous home. In *Proceedings of the 13th ACM International Conference on Multimedia*, Singapore, 6-11 November, ACM Press, 820-828

DE WIED, M., HOFFMAN, K. AND ROSKOS-EWOLDSEN, D.R. 1997. Forewarning of graphic portrayal of violence and the experience of suspenseful drama. *Cognition and Emotion 11*, 481-494.

DETENBER, B.H., SIMONS, R.F. AND BENNETT, G. 1998. Roll 'em!: The effects of picture motion on emotional responses. *Journal of Broadcasting & Electronic Media 42*, 113-127.

EKMAN, P., LEVENSON, R.W. AND FRIESEN, W.V. 1983. Autonomic nervous system activity distinguished between emotion. *Science 221*, 1208-1210.

FRAZIER, T.W., STRAUSS, M.E. AND STEINHAUER, S.R. 2004. Respiratory sinus arrhythmia as an index of emotional response in young adults. *Psychophysiology 41*, 75 - 83.

FRIDJA, N. 1986. *The Emotions*. Cambridge University Press, Cambridge.

FURINI, M. AND GHINI, V. 2006. An audio-video summarisation scheme based on audio and video analysis. In *Proceedings of the IEEE Consumer Communications and Networking Conference (CCNC '06)*, Las Vegas, NV, USA, 8-10 January, IEEE, 1209-1213.

GOMEZ, P. AND DANUSER, B. 2004. Affective and physiological responses to environmental noises and music. *International Journal of Psychophysiology 53*, 93-103.

GOMEZ, P., STAHEL, W. AND DANUSER, B. 2004. Respiratory responses during affective picture viewing. *Biological Psychology 67*, 359 - 373.

GREENWALD, M.K., COOK, E.W. AND LANG, P.J. 1989. Affective judgement and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli. *Journal of Pyschophysiology 3*, 51-64.

GROSS, J.J. AND LEVENSON, R.W. 1995. Emotion elicitation using films. *Cognition and Emotion 9*, 87-108.

HANJALIC, A. 2003. Generic approach to highlight extraction in a sport video. In *Proceedings of the IEEE International Conference on Image Processing (ICIP 2003)*, Barcelona, Spain, 14-18 September, IEEE, 1-4.

HANJALIC, A. 2005. Adaptive extraction of highlights from a sport video based on excitement modeling. *IEEE Transactions on Multimedia 7*, 1114-1122.

HEALEY, J.A. 2000. Wearable and Automotive Systems for Affect Recognition from Physiology. In *Department of Electrical Engineering and Computer Science* MIT, Cambridge, MA, USA, 158.

IWORX 2006. *Experiment 33: The Galvanic Skin Response (GSR) and Emotion*. Psychological Physiology Courseware No. <http://www.iworx.com/LabExercises/lockedexercises/LockedGSRANL.pdf>

JAIMES, A., ECHIGO, T., TERAGUCHI, M. AND SATOH, F. 2002. Learning personalized video highlights from detailed MPEG-7 metadata. In *Proceedings of the IEEE International Conference on Image Processing (ICIP 2002)*, New York, NY, USA, 22-25 September, IEEE, 133-136.

JUNG, B., SONG, J. AND LEE, Y. 2007. A narrative-based abstraction framework for story-oriented video. *ACM Transactions on Multimedia Computing, Communications and Applications 3*, 1-28.

KRAMER, A.F. 1991. Physiological metrics of mental workload: a review of recent progress. In *Multiple-Task-Performance*, D.L. DAMOS Ed. Taylor & Francis, London, 329-360.

LANG, A., BOLLS, P., POTTER, R. AND KAWAHARA, K. 1999. The effects of production pacing and arousing content on the information processing of television messages. *Journal of Broadcasting and Electronic Media 43*, 451-476.

LEW, M.S., SEBE, N., DJERABA, C. AND JAIN, R. 2006. Content-based multimedia information retrieval: state of the art and challenges. *ACM Transactions on Multimedia Computing, Communications and Applications 2*, 1-19.

LI, Y., LEE, S., YEH, C. AND KUO, C. 2006. Semantic retrieval of multimedia. *IEEE Signal Processing Magazine 23*, 79-89.

MCINTYRE, G. AND GÖCKE, R. 2007. The Composite Sensing of Affect. In *Affect and Emotion in Human-Computer Interaction. LNCS*, C. PETER AND R. BEALE Eds. Springer, Heidelberg, Germany.

MONEY, A. AND AGIUS, H. 2006. Are affective video summaries feasible? In *Emotion in HCI: Joint Proceedings of the 2005, 2006, and 2007 International Workshops at the BCS HCI Group Annual Conferences*, C. PETER, R. BEALE, E. CRANE, L. AXELROD AND G. BLYTH Eds. London, UK, 12 September 2006, Fraunhofer IRB Verlag, Stuttgart, Germany, 142-149.

MONEY, A.G. AND AGIUS, H. 2005. 'Once more, with feeling': an emotional approach to multimedia content analysis. In *Proceedings of the 9th IASTED International Conference on Internet and Multimedia Systems and Applications (IMSA 2005)*, Honolulu, Hawaii, USA, 15-17 August, ACTA Press: Anaheim, CA, USA, 436-441.

MONEY, A.G. AND AGIUS, H. 2008. Feasibility of personalized affective video summaries, Lecture Notes in Computer Science, vol. 4868. In *Affect and Emotion in Human-Computer Interaction*, C. PETER AND R. BEALE Eds. Springer-Verlag, Berlin Heidelberg, Germany.

MONEY, A.G. AND AGIUS, H. 2008. Video summarisation: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation 19*, 121-143.

MORIYAMA, T. AND SAKAUCHI, M. 2002. Video summarization based on the psychological unfolding of drama. *Systems and Computers in Japan 33*, 1122-1131.

MORRONE-STRUPINSKY, J.V. AND DEPUE, R.A. 2004. Differential relation of two distinct, film-induced positive emotional states to affiliative and agentic extraversion. *Personality and Individual Differences 36*, 1109-1126.

NASOZ, F., ALVAREZ, K., LISETTI, C.L. AND FINKELSTEIN, N. 2003. Emotion recognition from physiological signals for presence technologies. *International Journal of Cognition 6*, 1 - 32.

PALOMBA, D. AND STEGAGNO, L. 1993. Physiology, perceived emotion and memory: responding to film sequences. In *The Structure of Emotion: Psychophysiological, Cognitive, and Clinical Aspects*, N. BIRBAUMER AND A. OHMAN Eds. Hogrefe & Huber, Toronto, 158-168.

PHILIPPOT, P., CHAPELLE, C. AND BLAIRY, S. 2002. Respiratory feedback in the generation of emotion. *Cognition and Emotion 16*, 605-627.

PICARD, R.W. 1995. *Affective Computing*. MIT Media Laboratory Perceptual Computing Section Technical Report No. 321, November. <http://vismod.media.mit.edu/tech-reports/TR-321.pdf>

PICARD, R.W. 1997. *Affective Computing*. MIT Press, Cambridge, MA.

PIFERI, R.L., KLINE, K.A., YOUNGER, J. AND LAWLER, K.A. 2000. An alternative approach for achieving cardiovascular baseline: Viewing an aquatic video. *International Journal of Psychophysiology 37*, 207-217.

RUI, Y., GUPTA, A. AND ACERO, A. 2000. Automatically extracting highlights for TV Baseball programs In *Proceedings of the 8th ACM International Conference on Multimedia*, Los Angeles, CA, USA, 30 October, ACM Press, 105-115.

RUI, Y., ZHOU, S.X. AND HUANG, T.S. 1999. Efficient access to video content in a unified framework. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems (ICMCS '99)*, Florence, Italy, 7-11 June, IEEE, 735-740.

SCHEIRER, J., FERNANDEZ, P., KLEIN, J. AND PICARD, R.J. 2002. Frustrating the user on purpose: A step toward building an affective computer. *Interacting with Computers 14*, 93-118.

SHIPMAN, S., DIVAKARAN, A. AND FLYNN, M. 2007. Highlight scene detection and video summarization for PVR-enabled television systems. In *Proceedings of the IEEE International Conference on Consumer Electronics*, Hiroshima, Japan, 10-14 January 2007, IEEE, 1-2.

SIMON, H.A. 1982. Comments. In *Affect and Cognition*, C. SYDNOR AND S.T. FISKE Eds. Lawrence Erlbaum Associates, Hillsdale, NJ, 333-342.

SIMONS, R.F., DETENBER, B.H., REISS, J.E. AND SHULTS, C.W. 2000. Image motion and context: A between- and within-subject comparison. *Psychophysiology 37*, 706-710.

SMEULDERS, A.W.M., WORRING, M., SANTINI, S., GUPTA, A. AND JAIN, R. 2000. Content-Based Image Retrieval at the End of the Early Years. *IEEE Transactions on pattern analysis and machine intelligence 22*, 1349-1380.

SPIERS, B. 1979. *The Psychiatrist*. BBC Television.

STEINBEIS, N., KOELSCH, S. AND SLOBODA, J.A. 2006. The role of harmonic expectancy violations in musical emotions: evidence from subjective, physiological, and neural responses. *Journal of Cognitive Neuroscience 18*, 1380-1393.

SUZIKI, J., HIROSHI, N. AND HORI, T. 2004. Level of interest in video clips modulates event-related potentials to auditory probes. *International Journal of Psychophysiology 55*, 35-43.

TAKAHASHI, Y., NITTA, N. AND BABAGUCHI, N. 2005. Video summarization for large sports video archives. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2005)*, Amsterdam, The Netherlands, 6-8 July, IEEE, 1170-1173

TRUONG, B.T. AND VENKATESH, S. 2007. Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications 3*, 1-37.

VAN DIEST, I., WINTERS, W., DEVRIESE, S., VERCAMST, E., HAN, J.N., VAN DE WOESTIJNE, K.P. AND VAN DEN BERGH, O. 2001. Hyperventilation beyond fight/flight: respiratory responses during emotional imagery. *Psychophysiology 38*, 961 - 968.

VAN REEKUM, C.M. AND JOHNSTONE, T. 2004. Psychophysiological responses to appraisal dimensions in a computer game. *Cognition and Emotion 18*, 663-688.

WANG, H., PRENDINGER, H. AND IGARASHI, T. 2004. Communicating emotions in online chat using physiological sensors and animated text. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '04)*, Vienna, Austria, 24-29 April, ACM Press, 1171-1174.

WINTON, W.M., PUTNAM, L.E. AND KRAUSS, R.M. 1984. Facial and autonomic manifestations of the dimensional structure of emotion. *Journal of Experimental Social Psychology 20*, 195-216.

WRIGHT, E. 2004. *Shaun of the Dead*. Universal Pictures.

XU, C., WANG, J., WAN, K., LI, Y. AND DUAN, L. 2006. Live sports detection based on broadcast video and Web-casting text. In *Proceedings of the 14th ACM International Conference on Multimedia*, Santa Barbara, CA, 23-27 October, ACM Press, 221-230.