USING VIEWER'S FACIAL EXPRESSION AND HEART RATE FOR SPORTS VIDEO HIGHLIGHTS DETECTION

ABSTRACT

Viewer interests, evoked by video content, can potentially identify the highlights of the video. This paper explores the use of facial expressions (*FE*) and heart rate (*HR*) of viewers captured using camera and non-strapped sensor for identifying interesting video segments. The data from ten subjects with three videos showed that these signals are viewer dependent and not synchronized with the video contents. To address this issue, new algorithms are proposed to effectively combine *FE* and *HR* signals for identifying the time when viewer interest is potentially high. The results show that, compared with subjective annotation and match report highlights, 'non-neutral' *FE* and 'relatively higher and faster' *HR* is able to capture 60%-80% of *goal*, *foul*, and *shot-on-goal* soccer video events. *FE* is found to be more indicative than *HR* of viewer's interests, but the fusion of these two modalities outperforms each of them.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human information processing.

General Terms

Algorithms, Reliability, Experimentation.

Keywords

Facial expression, Heart rate, Video segmentation, Viewer interest, Sports video highlight.

1. INTRODUCTION

Rapid growth of video data increases the need for indexing interesting video segments for supporting user-centric retrieval. Current content indexing approaches either focus on internal content (e.g., audio-visual features) or viewer's response to automatically detect 'interest evoking' segments. In particular, facial expressions and physiological responses have been found to carry information about viewer's interest in video contents [8]. Physiological signals, such as heart rate, manifest internal responses that can only be measured via biosensors. In contrast, facial expression can be observed externally and only requires a standard video camera, therefore can be more easily deployed in real-life scenarios [4; 16]. However, it can be expected that a combination of the two modalities will complement each other. Few studies have attempted to combine them, but for capturing viewer's responses to different stimuli (i.e., image) and context (i.e., topic searching) [1; 7].

The main contribution of this paper is to demonstrate the benefits of using sports videos, which have well-established structure (e.g. play-break), for exploring the temporal mapping between events, continuous heart rate, and discrete facial expression. Combination and comparison between heart rate and facial expression signals have been investigated as well. The experimental setting is more practical, compared to most of the current work, as it uses non-strapped sensors. The structure of the paper is as follows. Section 2 discusses the related work. Section 3 describes the experimental protocol, while Section 4 describes the data analysis algorithms. Section 5 and 6 describes the experimentation and analysis of results. Section 7 and 8 outlines the discussions and future work.

2. RELATED WORK

Heart rate reactions alongside valence and arousal have been used for identifying viewer's interest in movie clips [2]. Studies found that correlations between physiological features and self-assessed evaluation scores vary across users. Other physiological signals, such as galvanic skin resistance, electromyogram, and respiration pattern are also sensitive to viewer's response and have been used along with heart rate to measure interests in movie clips, music videos, and static images [12; 13]. However, current sensors for these signals require strapped components, therefore limiting the practicality for use in natural or day-to-day settings.

Facial expression is another important modality for understanding viewers' interest in response to video stimuli. Facial expression and activity features extracted from head and face regions have been used for personal highlights detection from movie, drama and TV commercial clips [5; 6]. Head oriented features (i.e., head roll, head position) can be aggregated with facial features to measure low/high engagement levels [4]. Such systems can be used for home video editing and viewer experience modelling [10; 11].

Current literature has not explored the correlations between facial expression and heart rate in response to sports video, particularly for highlights detection purposes. The main challenge is due to the time delays between an interesting event and the corresponding viewer's response. Moreover, an event may not be interesting for every person (i.e. subjectivity of viewers). This paper represents the first attempt to address all of these challenges.

3. EXPERIMENTAL PROTOCOL

3.1 Video Stimuli and Annotation

The study used three soccer matches from UEFA Euro 2012 (Spain vs Italy), English Premier League 2013-14 (Manchester City vs Tottenham), and Champion's League 2014-15 (Real Madrid vs Basel). It was ensured that none of the subjects had previously seen these matches based on a questionnaire. These three video clips (i.e., *vid1*, *vid2*, *vid3* with a frame rate of 25 frame-per-second) were trimmed to the first 20 minutes, 15 minutes and 17 minutes but was manually confirmed to contain three types of events, including *goal*, *shot-on-goal*, and *foul*. Full-length matches were not used to maintain subject's concentration and patience.

To develop the ground truth (of interesting events), two methods were adopted. First is extracting events based on the match report highlights (*HL*) collected from soccer websites (ESPN.com, UEFA.com, SkySports.com). Second is identifying specific segments based on subjects' manual annotation (*MA*) during data collection. The scoping of a highlight segment was specified by a sequence of play-break frames [15]. In total, the ground truth contained 30 sets of manually annotated segments from 10 subjects for 3 video clips. To reduce the impact of subjectivity, the final set of *MA* segments were selected from those manually annotated by more than 5, out of 10, subjects (> 50% agreement).

Table 1 summarizes the ground truth data. In total, 3 goals, 14 shots-on-goal, and 5 fouls were included in highlight segments (*HL*). Manually annotated segments (*MA*) consisted of 3 goals, 17 shots-on-goal, and no foul.

Table 1. Annotation of each type of ground truth segments (HL, MA) from three video clips. It shows the number of soccer events annotated against each type

Soccer Events	Vid 1		Vid 2		Vid 3	
	HL	MA	HL	MA	HL	MA
Goal	1	1	1	1	1	1
Shot-on-goal	5	8	4	4	5	5
Foul	0	0	4	0	1	0

3.2 Participants and Apparatus

Ten subjects (age: mean = 26.4, standard deviation = 3.2) were recruited through a pre-questionnaire for obtaining information on their support for soccer, favourite soccer team, last seen soccer match. Subjects were in good health and had no visual impairment. All of them were familiar with video watching.

A less-confined environment was established with a closed and quiet room (with no access of other people) involving a 21.5 inches Dell monitor for showing video stimuli, a video camera and strapless heart rate sensor (*Mio Alpha*) for recording facial expression and heart rate, and a comfortable revolving chair (shown in figure 1). Facial recording involved only natural daylight rather than artificial or extra lighting. Distances from subjects to the monitor and the video camera were 80-90 cm and 100-120 cm respectively.

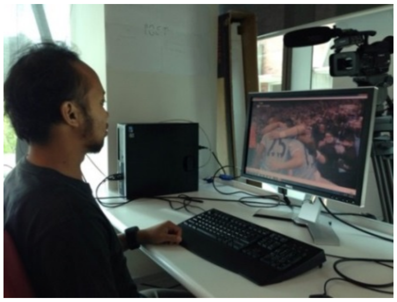


Figure 1. A subject watching soccer stimuli wearing a heart rate sensor on his left hand while his facial expression being recorded by a video camera

3.3 Recording of Facial Expression (FE) and Heart Rate (HR)

Each subject was put through three separate trials including identical steps, where each trial involves simultaneous recording of facial expression and heart rate data during watching video stimuli. Prior to each trial, a short demonstration and a preliminary training were used to familiarise each subject with the data collection procedure. Consecutive trials had a break for two weeks. The sampling frequencies of facial video and time stamped (in bit-per-minute) heart rate data were 25 frame-per-seconds (fps) and 1/3 Hz. Heart rate data was stored through an iOS application (*Digifit*) which requires 5-8 seconds for calibration. A resting time of 5 minutes was given to each subject before data collection commences, for making his/her heart rate steady. Each trial of data collection involves the following steps.

(1) *Recording*: Subject's facial video and heart rate were recorded simultaneously during watching a soccer clip. Facial video and heart rate recording were triggered using a remote controller and an iPhone.

(2) *Manual Annotation*: Each subject was asked to manually annotate each segment he/she felt interest into from the seen soccer clip, using the 'capture' option of *KMPlayer*. Annotation includes a set of starting and ending frames for each annotated segment.

(3) *Rating*: Each subject was asked to watch and rate the match report highlights in a scale from 0 to 1, where '1' indicates 'interesting' and '0' indicates 'non-interesting' segments.

(4) *Post-questionnaire*: A list of the players was presented so that each subject could tick his/her preferred player(s) and team from the seen clip. This information was obtained through a paper-based questionnaire.

Collection of facial expression and heart rate data from 10 subjects in response to 3 video clips produced 30 sets of facial expression data, 30 sets of heart rate data, and 30 sets of subjective manual annotation data. Following section describes the processing techniques of this data.

4. DATA ANALYSIS

An illustration of our system framework is provided in figure 2. Facial expression data are processed with an off-the-shelf system which provides frame-by-frame intensity scores for three emotion categories (i.e., positive, negative, and neutral). Scores of these three categories are processed separately (as functions of time) to identify consecutive frames with relatively high intensity scores, as forms of video segments. A number of temporal features are extracted from the pre-processed heart rate data. We then combine these feature values to obtain a collection of fused feature values as a function of time. Consecutive instances with high fused (feature) values are identified similarly as video segments. A segment-wise fusion method was then applied to combine segments identified from facial expression and heart rate data. More details on data processing are described in following sub-sections.

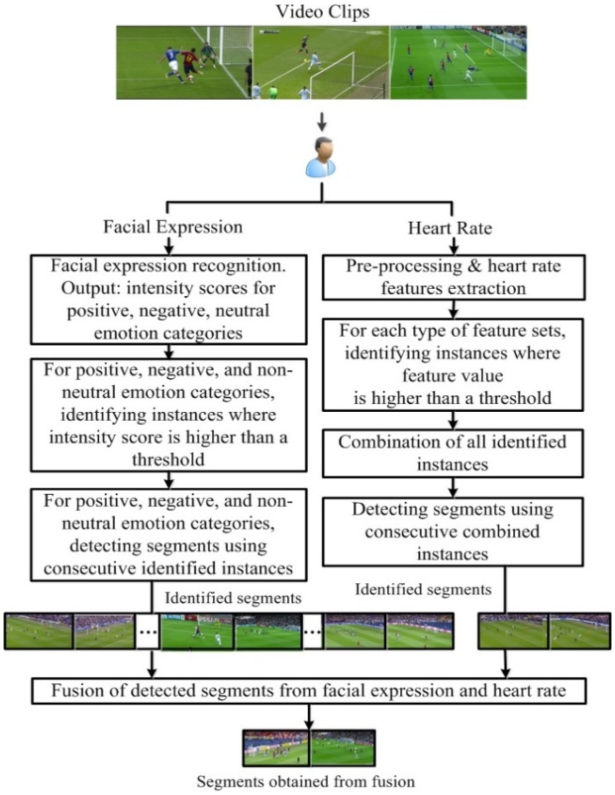

Video Clips
Facial Expression
Heart Rate
Facial expression recognition. Output: intensity scores for positive, negative, neutral emotion categories
For positive, negative, and non-neutral emotion categories, identifying instances where intensity score is higher than a threshold
For positive, negative, and non-neutral emotion categories, detecting segments using consecutive identified instances
Pre-processing & heart rate features extraction
For each type of feature sets, identifying instances where feature value is higher than a threshold
Combination of all identified instances
Detecting segments using consecutive combined instances
Identified segments
Identified segments
Fusion of detected segments from facial expression and heart rate
Segments obtained from fusion

Figure 2. System framework

4.1 Recognition of Viewer Facial Expression

We used the facial expression recognition (FER) system described in [17] for facial expression classification which provides frame-by-frame outputs and is capable of handling facial movement and pose variation. Training data of this FER system includes field-based data from TV, news, and web. This system detects the face region from a facial video with the Viola-Jones face detector and extracts as well as tracks facial landmark points using a multi-view active shape model (ASM) tracker. Scale-invariant feature transform (SIFT) features and distances between the 53 interior facial points are used as texture and geometric features. SIFT features are dimensionally reduced by the minimum redundancy maximum relevance (mRMR) algorithm and fused with geometric features. A support vector machine (SVM) classifier then predicts the fused features into one of three categorical emotion classes (i.e., positive, negative, neutral). The output of the classification phase is frame-by-frame probability (i.e., emotion intensity) scores for each of positive (*POS*), negative (*NEG*), and neutral (*NEU*) emotion categories, where each of these scores varies between [0, 1].

4.2 Identifying Interesting Video Segments from Viewer FE

This work identifies three types of video segments using intensity scores obtained for three emotion categories from the FER system. The output of the FER system was considered as a signal containing three (i.e., *POS*, *NEG*, *NEU*) channels of continuous frame-based intensity scores (from 0 to 1). For example, intensity scores for i^{th} frame can be represented as $P_i^{\{POS,NEG,NEU\}}$. For any instance, summation of these three scores is equal to 1 and thus, effectively two-dimensional continuous representations (time (in frame number) vs intensity scores) are obtained for facial expression. This was important for obtaining a mapping between facial expressions and watched video clips. In the proposed method, a non-neutral (*NNEU*) score is assigned by using the maximum of positive or negative scores instead of their sum. The following procedure (algorithm 1) describes how video frames with high positive intensity scores are separated using a threshold and used for identifying interesting video segments. We used analogous procedures for identifying interesting video segments using negative and non-neutral intensity scores.

Intensity score of each missing frame is computed using linear interpolation between the intensity scores of two neighbouring frames. For noise removal, a low-pass-filter with a $\gamma = 5$ seconds window was applied. It was found that any window less than 5 seconds was not suitable to remove the rapid change in frame-by-frame scores. Our goal was to identify segments (collection of frames) with relatively high positive intensity scores. For this purpose, we computed a threshold, $Thres_1$ from mean and standard deviation of the positive intensity scores. For identifying consecutive frames, a frame-by-frame labelling was utilised where each frame was labelled as '1' if corresponding positive score is greater than the other scores as well as $Tresh_1$ (labelled as '0' otherwise). Each set of consecutive frames, labelled as '1', was considered as a preliminary interesting video segment. In such way, we identified a set of preliminary interesting video segments $S = \{s_1, s_2,\ldots\}$ where each segment s_i contains a number of frames labelled as '1'. Semantics of soccer match events are unlikely to vary by lesser than one second and based on this assumption any two sequential segments from S were merged together if distance between them (in number of frames) was less than 1 second (i.e., 25 frames). Consider the new segments were $S' = \{s'_1, s'_2, ..\}$. To remove significantly smaller segments, another threshold, $Thres_2$ was computed based on mean and standard deviation of the lengths (in number of frames) of all segments in S' and any segment in S' with length lesser than $Thres_2$ was discarded. The remaining segments were considered as the final interesting video segments, which are $S^{POS} = \{s_1, s_2,\ldots.\}$, where s_1 contains the starting and ending frame numbers of first identified video segment as indices.

Algorithm 1. Video segmentation using *POS* intensity scores

```
Input: FE intensity scores: $P_i^{\{POS,NEG,NEU\}}$ for each of $m$ frames
Output: Identified video segments, $S^{POS} = \{s_1, s_2, \ldots\}$
Compute intensity values for missing frames using linear interpolation;
Smooth intensity values using a moving average low-pass-filter with $\gamma$ seconds window and output is $P_i^{'\{POS,NEG,NEU\}}$;
$Thres_1 \leftarrow \mu\,(P^{POS'}) + \sigma\,(P^{POS'})$ ;
Label, $L \leftarrow Ø$;
For $i = 1$ to $m$
    If ( $P_i^{POS'} > P_i^{NEG'}$ ) & ( $P_i^{POS'} > P_i^{NEU'}$ ) & ( $P_i^{POS'} > Thres_1$)
        $L_i \leftarrow 1$;
    Else
        $L_i \leftarrow 0$;
Find consecutive segments, $S = \{s_1, \ldots, s_p\}$ labelled as '1';
Merge any sequential segment with a distance less than 1 second (25 frames) and the output is a new set of segments $S' = \{s'_1, s'_2,\ldots\}$;
$\mathrm{Thres}_2 \leftarrow \mu$(lengths of all $s'_i$ in $S'$) - $\sigma$(lengths of all $s'_i$ in $S'$);
Discard any segment $s'_i$ with length < $\mathrm{Thres}_2$ and consider the remaining segments as final segments, $S^{POS} = \{s_1, s_2,\ldots\}$;
```

4.3 Heart Rate Feature Extraction

Heart rate data was obtained with time, T in seconds. Consider, $HRate = \{hr_1, hr_2, \ldots., hr_n\}$ is the heart rate signal for any particular subject which contains n heart rate samples in bits-per-minutes. The following procedure (algorithm 2) describes how features are extracted from *HRate*. Prior work have used derivative of heart rate (rate-of-change) and spectral features such as energy from low frequency band, maximum approximate heart rate to understand cognitive and emotional reaction of human [3; 12-14]. We used similar features in this study. For removing the rapid fluctuation and smoothing, we used a low-pass-filter over *HRate* data containing a sliding window of $\beta = 9$ seconds (over 4 *HRate* samples). A separate sliding window was used to obtain the temporal features from smoothed *HRate* data, which were rate-of-change, gradient, variance, local maxima, and energy from low frequency band (0.04-0.15 Hz). The length of the sliding window was kept 9 seconds since smaller window (including less than 4 samples) might not capture the subtle changes in temporal features.

We computed rate-of-changes of *HRate* samples using the total absolute difference between neighbouring samples and the difference of time stamped with first and last samples within the sliding window. Standard MATLAB functions were used for computing local maxima, gradient, and variance. Function used for gradient returns a set of directional values for increasing order of the input signal and hence, averages of the absolute directional values were used as gradient features. Because of the resolution of our collected heart rate data, we computed energy feature only from low frequency band using a finite impulse response (FIR) band-pass-filter. Five extracted features were parallel to each other and scaled with time, T, in seconds.

Algorithm 2. Temporal HR feature extraction

Input: HR samples, $HRate = \{hr_1, hr_2, …., hr_n\}$

Output: Temporal feature set: *RateOfChange*, *LocalMaxima*, *Energy*, *Variance*, *Gradient*

```
Smooth HRate with a moving average low-pass-filter with $\beta$ seconds window and output is $HRate' = \{hr'_1, hr'_2, …., hr'_n\}$;
For $i$ = 1 to n-1
    $diff_i \leftarrow | hr'_i - hr'_{i+1}|$;
$RateOfChange \leftarrow Ø$; $LocalMaxima \leftarrow Ø$;
$Variance \leftarrow Ø$; $Gradient \leftarrow Ø$; $Energy \leftarrow Ø$;
For $i$ = 1 to $n$-4
    $LocalMaxima_i = \max\{hr'_i, …, hr'_{i+3}\}$;
    $D_t = Time_{i+3} - Time_i$;
    $D_x = \sum_{m=i}^{i+3} diff_m$;
    $RateOfChange_i = D_x / D_t$;
    $Variance_i = \sigma^2\{hr'_i, …, hr'_{i+3}\}$;
    $Gradient_i = \mu(|\delta\{hr'_i, …, hr'_{i+3}\}/\delta\{i, …, i+3\}|)$;
Design a band-pass-filter, filter with pass bands 0.04 Hz and 0.15 Hz;
$HRate'' \leftarrow$ filter($HRate$);
For $i$ = 1 to $n$-4
    $Energy_i \leftarrow hr''_i * hr''_i$;
```

4.4 Identifying Segments from Viewer HR

Three from five extracted features (i.e., *RateOfChange*, *LocalMaxima*, and *Energy*) were selected for further processing, since *Variance* and *Gradient* were found to be highly correlated with *RateOfChange*. The following procedure (algorithm 3) describes how interesting video segments are identified from the selected heart rate features. This shows how heart rate samples with higher (than a threshold) are selected and how these consecutive selected samples are converted into segments.

Algorithm 3. Video segmentation using HR features

Input: $x = \{feat_1, feat_2, …, feat_n\}$, where x is *RateOfChange*, *LocalMaxima*, *Energy*

Output: Identified segments, $S^{HR} = \{s_1, s_2, …, s_n\}$

```
Indices, $Idx \leftarrow Ø$;
For each of $x$ = { RateOfChange, LocalMaxima, Energy}
    $Thres \leftarrow \mu(x) + \sigma(x)$;
    $I \leftarrow Ø$;
    Store index $i$ of any $feat_i$ in $I$ if $feat_i > Thres$;
    $Idx \leftarrow Idx \cup I$;
$S' \leftarrow Ø$; $j \leftarrow 1$;
Starting of first segment, $S'_{1,1} \leftarrow T(Idx_1)$;
For each index $Idx_i$ in $Idx$
    If there is a break between indices, $(Idx_{i+1} - Idx_i) > 1$
        Ending of the current segment, $S'_{j,2} \leftarrow T(Idx_i)$;
        Starting of next segment, $S'_{j,1} \leftarrow T(Idx_{i+1})$;
Merge consecutive segments with distance less than $\alpha$, $(S'_{j,2} - S'_{j+1,1} < \alpha)$ and S is the new set of segments;
For $m$ = 1 to length of $S$
    $S_{m,1} \leftarrow S_{m,1} * frm\_rate$;
    $S_{m,2} \leftarrow S_{m,2} * frm\_rate$;
$S^{HR} \leftarrow S = \{(S_{1,1}, S_{1,2}), (S_{2,1}, S_{2,2}), …\}$;
```

Our goal here was to see whether high changes in heart rate feature signify viewer interest in any manner. We computed separate threshold (i.e., *Thres*) for each selected feature set to separate the consecutive features with values higher than that threshold. As these features are time stamped, time (in seconds) was obtained from T as indices for all separated features and combined (i.e., as *Idx*). Each set of consecutive features was treated as a segment and beginning and ending of that segment was stored as time (in seconds) in S'. We then combined sequential segments with distance less than α = 3 (seconds) since it is unusual that two interesting events may occur in soccer match in an interval less than three seconds. Stored time indices were converted to frame number from second.

4.5 Fusion of Video Segments Identified from FE and HR

We fused the segments identified from non-neutral facial expressions with those identified from heart rate signals. Segments identified from positive and negative facial expression were not used in fusion since non-neutral score had already included both. All the identified segments were scaled with time (in frame number) and hence, we used frame-by-frame set union operation for fusing these two types of segments for all subjects. Consider, the resultant fused set of segments was S^{FUSED}.

5. EXPERIMENT

From the procedures described in above section, we identified five types of video segments (S^{POS}, S^{NEG}, S^{NNEU}, S^{HR}, S^{FUSED}) from facial expression and heart rate data collected from each subject in response to each video clip. We computed four statistical measurements, which are **similarity, temporal alignment (synchronicity), detection rate, and accuracy,** based on the comparison between identified video segments and the ground truth segments. Ground truth segments for our study were match report highlights and manually annotated segments (*HL*, *MA*) and the statistical measurement were done separately on each of them.

Algorithm 4 includes the steps to compute similarity score, alignment score, and detection rate. We compared each of the ground truth segments against all identified segments and computed a similarity score each time when there is any overlapping. This score was actually a Jaccard index which measures the similarity between two finite set of samples. If two sets of samples are A and B, the Jaccard index is given by equation 1.

$$J(A, B) = (A \cap B) / (A \cup B) \quad (1)$$

According to our methods, each segment was considered as a set of frame numbers and Jaccard index (i.e., *jaccard_idx*) have been calculated between two such sets of frame numbers each time. We tried to find out a particular identified segment which is most similar to a particular highlight or annotated segment. Therefore, maximum of the similarity scores was kept as the final similarity score.

To investigate the temporal synchronicity between identified and ground truth segments, we measured the incorrect positioning between these two types of segments as misalignment (i.e., *misalignment*) where smaller misalignment score means higher synchronicity. Misalignment was measured by computing the distance between each ground truth segment and its closest identified segment. An identified segment is considered closest if it is overlapped or contains a minimum distance to a ground truth segment. Distance was measured in term of number of frames and the misalignment score was converted into second. Overlapping was measured using set intersection operation. Value of *frame_rate* was kept same as the frame rate of video which is 25 fps. We computed the percentage of overlapping between each ground truth segment and identified segments, where a ground truth segment is considered as 'detected' if it overlaps any identified segment.

Detection rate, *det_rate* was computed each time when there was an overlap between identified segment and ground truth segment. For each ground truth segment, we want to see how much (in %) of that segment was overlapped with any of the identified segments.

Algorithm 4. Measuring similarity, synchronicity, and detection rate

Input: Identified segment set, $S^x = \{s_1, s_2, \ldots, s_n\}$, ground truth segment set, $Y = \{y_1, y_2, \ldots, y_m\}$, where x is *POS, NEG, NNEU, HR*; Y is *HL, MA*; and $y_i = \{f_{start}, f_{end}\}$, $s_j = \{S_{j,1}, S_{j,2}\}$
Output: Jaccard indices, *jaccard_idx*; misalignment, *misalignment*; detection rate, *det_rate*,

```
$jaccard\_idx \leftarrow \emptyset$; $misalignment \leftarrow \emptyset$; $det\_rate \leftarrow \emptyset$;
For each ground truth segment, $y_i$ in $Y$
    $jaccard\_idx_i \leftarrow 0$; $alignment \leftarrow \emptyset$;
    $seg\_annotaed \leftarrow \{f_{start}, f_{start} + 1, \ldots, f_{end}\}$;
    For each identified segment, $s_j$ in $S$
        $seg\_identified \leftarrow \{S_{j,1}, S_{j,1} + 1, \ldots, S_{j,2}\}$;
        $union \leftarrow seg\_annotaed \cup seg\_identified$;
        $intersect \leftarrow seg\_annotaed \cap seg\_identified$;
        $score \leftarrow |intersect| / |union|$;
        If there is an overlap and current score is higher, $((intersect \neq \emptyset) \& (jaccard\_idx_i < score))$
            Jaccard index, $jaccard\_idx_i \leftarrow score$;
        If there is an overlap, $(intersect \neq \emptyset)$
            Detection rate, $det\_rate_i \leftarrow |intersect| *100/ | y_i |$;
            If $s_j$ completely overlaps $y_i$ or vice versa, $((S_{j,1} > f_{start}) \& (f_{end} > S_{j,2}) | (S_{j,1} < f_{start}) \& (f_{end} < S_{j,2}))$
                $alignment_j \leftarrow 0$; break;
            If $s_j$ partially overlaps and follows $y_i$, $((S_{j,2} > f_{start}) \& (f_{start} - S_{j,1} < alignment_j))$
                $alignment_j \leftarrow (f_{start} - S_{j,1}) / frm\_rate$;
            Else if $s_j$ partially overlaps and precedes $y_i$, $((f_{end} > S_{j,1}) \& (S_{j,2} - f_{end} < alignment_j)$
                $alignment_j \leftarrow (S_{j,2} - f_{end}) / frm\_rate$ ;
        Else If $(S_{j,2} < f_{start}) \& (f_{start} - S_{j,2} < alignment_j)$
            $alignment_j \leftarrow (f_{start} - S_{j,2}) / frm\_rate$;
            Else If $(f_{end} < S_{j,1}) \& (S_{j,1} - f_{end} < alignment_j)$
                $alignment_j \leftarrow (S_{j,1} - f_{end}) / frm\_rate$ ;
    $misalignment_i \leftarrow$ min $(alignment)$;
```

We computed accuracy by measuring precision, recall, and F1 scores between identified segments and ground truth segments, as described in algorithm 5. Each time there is an overlap between each ground truth segment and any identified segment and we incremented two variables, for counting the number of ground truth (i.e., *gt_seg*) and identified segments (i.e., *id_seg*) which were detected (i.e., overlapped). From these two measures we computed number of true positive, *TP*, false positive, *FP* and false negative, *FN*.

6. ANALYSIS OF RESULTS

This section describes the findings from the results. This includes how subjects with different demography responded differently to the same soccer events. It has also shown that heart rate response can be a complementary to facial expression in measuring viewer interest. Besides, results from similarity and temporal synchronicity have shown that fusion of facial expression and heart rate performs better than each of them. Detection rate and accuracy depicts similar impression.

Algorithm 5. Measuring accuracy

Input: Identified segment set, $S^x = \{s_1, s_2, \ldots, s_n\}$, ground truth segment set, $Y = \{y_1, y_2, \ldots, y_m\}$, where x is *NNEU, HR, FUSED*; Y is *HL, MA*; and $y_i = \{f_{start}, f_{end}\}$, $s_j = \{S_{j,1}, S_{j,2}\}$
Output: Precision, *precision*; recall, *recall*; F1 score, *F*

```
$TP \leftarrow 0$; $FP \leftarrow 0$; $FN \leftarrow 0$; $gt\_seg \leftarrow 0$;
For each ground truth segment, $y_i$ in $Y$
    $id\_seg \leftarrow 0$;
    For each identified segment, $s_j$ in $S$
        If there is an overlap between $s_j$ and $y_i$
            $id\_seg \leftarrow id\_seg + 1$;
            $gt\_seg \leftarrow gt\_seg + 1$;
    If overlapping is found, $(id\_seg > 0)$
    $TP \leftarrow TP + id\_seg$;
$FP \leftarrow n - TP$;
$FN \leftarrow m - gt\_seg$;
$precision \leftarrow TP / (TP + FP)$;
$recall \leftarrow TP / (TP + FN)$;
$F \leftarrow (2*precision*recall) / (precision + recall)$;
```

6.1 Subject Dependency in Responses

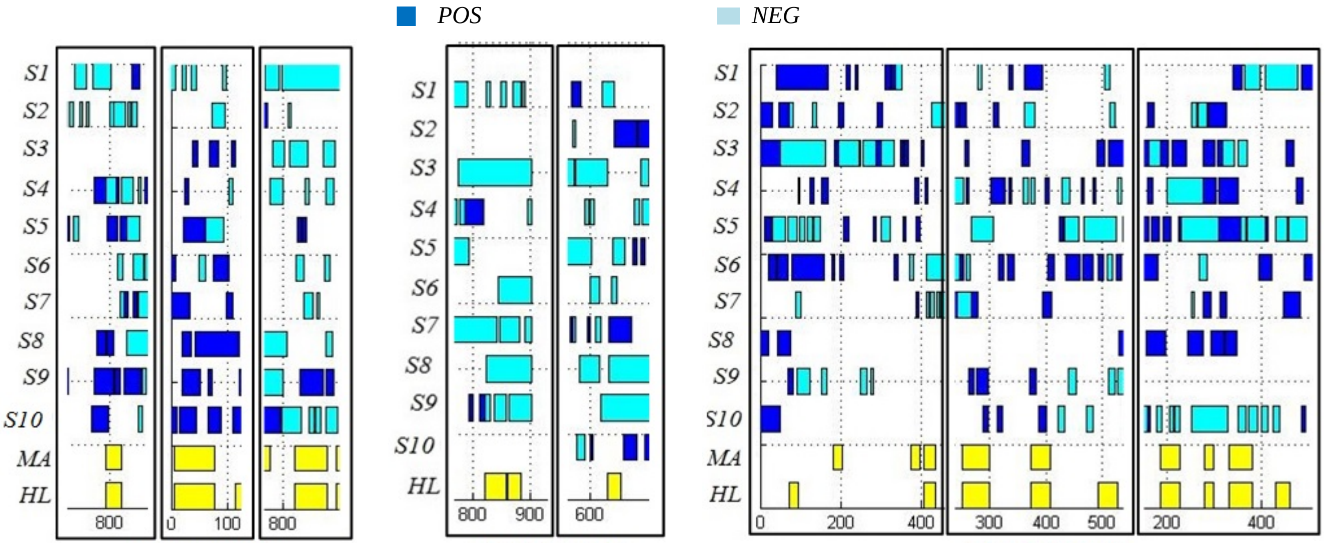
Figure 3(a) illustrates identified segments from positive and negative facial expression (*POS*, *NEG*) in response to the *goal* events from the video clips across ten subjects. Out of ten subjects three (*S1*, *S3*, *S8*) were not soccer fans. It was found that negative emotion was evoked in cases of those subjects in response to *goal* events. For other subjects, the majority of the identified segments were positive emotion evoking. Figure 3(c) illustrates similar identified segments from positive and negative facial expressions in response to *shot-on-goal* events from three video clips for all subjects. It can be seen from figure 3(c) that *shot-on-goal* events evoked both positive and negative facial expression equally where subject dependency is still evident. In the third case, majority of the *foul* events stimulated negative facial expression in all subjects regardless of soccer fans or not, as illustrated in figure 3(b).

6.2 Heart Rate as a Complementary to Facial Expression

It was found that segments identified from facial expression were not able to detect all of the ground truth segments. Some ground truth segments which were not detected by facial expression were detected by segments identified from heart rate signal which is shown in figure 4. Therefore, fusion of facial expression and heart rate can represent interest evoked in viewers better than any one of facial expression and heart rate. Following sections will show how performance is enhanced by fusing facial expression and heart rate.

6.3 Overlapping in term of Jaccard Index

Algorithm 4 describes how we computed the similarity score (Jaccard index) for each of the five types of video segments identified from each subject's response to soccer events in ground truth segments. These scores were averaged across all subjects for each type of soccer events and each type of identified segments. Figure 5 illustrates averaged Jaccard index scores for all subjects, between identified segments and events (goal, shot-on-goal, foul) of ground truth segments (i.e., *HL*, *MA*). The Jaccard index varies between 0 to 1, where 1 means 'similarity' and 0 means 'no similarity'.



(a) Goal (b) Foul (c) Shot-on-goal

Figure 3. Examples of identified segments from positive (*POS*) and negative (*NEG*) facial expression for all subjects (*S1* to *S10*) in response to goal, foul, and shot-on-goal events in three video clips (note that *Vid 1* and *MA* do not contain any foul event)

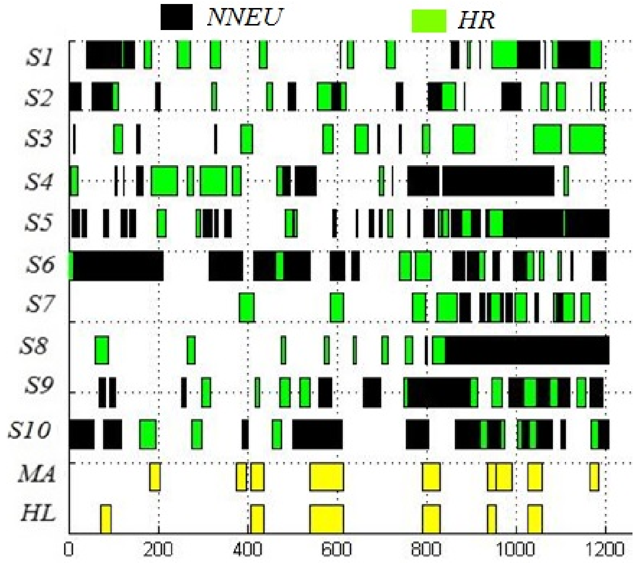


Figure 4. Identified segments from non-neutral facial expression (NNEU) and heart rate (HR) for all subjects (S1 to S10) in response to video clip 1

Results show that the Jaccard index of *POS* segments for *goal* event (0.4) is higher than *NEG* segments (0.2) and is opposite (*POS*: 0.27, *NEG*: 0.61) in case of *foul* events. *NNEU* segments include the best results from both *POS* and *NEG*. Segments identified by heart rate show lesser similarity compared to segments identified from *POS*, *NEG*, and *NNEU* scores. Segments identified from fusion of *NNEU* and *HR* show the highest similarity (Jaccard index: 0.54 and 0.53 for *goal* events) with ground truth segments.

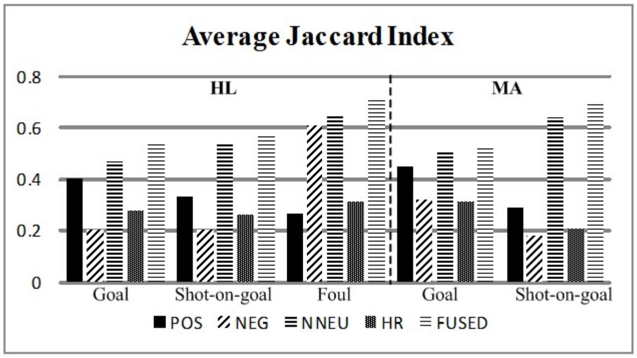


Figure 5. Averaged Jaccard index for different soccer events for video segments identified from positive, negative, non-neutral facial expression, heart rate, and fusion

6.4 Temporal Synchronicity

Algorithm 4 describes how distance (in seconds) between each ground truth segment (*HL* and *MA*) and its closest identified segment has been measured as misalignment scores. Table 2 illustrates the averaged score across all subjects for each type of ground truth segments and each type of identified segments. All ground truth segments (both which are detected and not detected) were compared during computing the misalignment scores. Segments identified from *NEG* segments are found to be mostly misaligned. Segments identified from *NNEU* scores have better temporal alignment (lower misalignment score) with both types of ground truth segments. Subjects' annotated segments have lesser misalignment with segments identified from *HR*. Segments identified from *HR* signals have more consistent alignment in relation to both highlights and subjects' annotated segments. Segments obtained from fusion were found to be mostly synchronous across all events regardless of highlights or subjects' annotated segments. The misalignment we found can be corrected by adding these computed values as offset with ground truth segments.

Table 2. Average misalignment score for different soccer events across identified segments

Average Misalignment Scores (in seconds)				
Ground truth	**Segment categories**	**Goal**	**Shot-on-goal**	**Foul**
HL	**POS**	3.3	4.57	12.72
	NEG	13.59	7.89	2.21
	NNEU	4.36	9.35	2.3
	HR	4.11	8.35	7.21
	FUSED	**0.23**	**1.15**	**1.43**
MA	**POS**	3.18	5.4	-
	NEG	43.56	21.33	-
	NNEU	7.22	10.20	-
	HR	3.39	5.32	-
	FUSED	**0.78**	**2.72**	-

6.5 Event Detection

Mixtures of common soccer events such as *goal*, *shot-on-goal*, and *foul* helped us to investigate both positive and negative evoked emotions. Figure 6 illustrates the average detection rate across all subjects for each type of ground truths and each soccer event. Results show that the majority of 'Goal' events were corresponding to extracted segments from *POS* scores (71.11%, 69.23%) while the least correspondence was found to segments from *NEG* scores (40.12%, 34.12%). A majority of 'Foul' events were corresponding to *NEG* segments (50%, 59.13%). Segments extracted from *NNEU* values outperform *POS* and *NEG* as *NNEU* contains the maximum absolute value of POS and *NEG* (see figure 6). Segments identified from *HR* show most consistent detection for all events. This is because heart rate can be affected regardless of the 'positive' or 'negative' nature of the emotional response. However, it is not accurate in terms of detection rate. Segments found from fusion were found to have higher detection rate (90%-100%) than all.

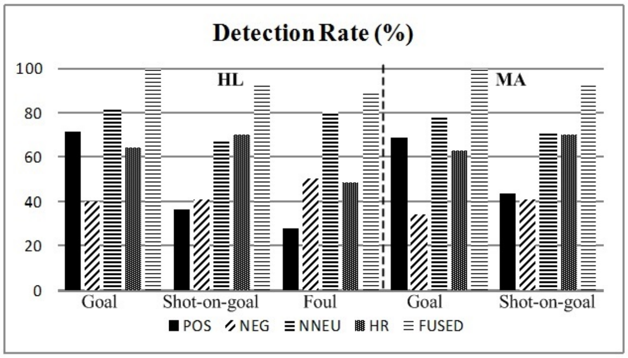


Figure 6. Average detection rate for different soccer events

6.6 Accuracy

We computed precision, recall and F1 scores by comparing identified segments (from each subject's response) with highlight and annotated segments separately and then averaged the scores across all subjects for each type of identified segments (see table 3). Results show that segments obtained through fusion outperform others (F1 scores: 0.71, 0.72). Segments identified from *POS* scores have the least accuracy (F1 score: 0.44). In overall, precision score is higher than recall score which means that number of false positive is fewer than number of false negative.

Table 3. Precision, recall, and F1 scores

Accuracy				
Ground Truth	**Segment categories**	**Precision**	**Recall**	**F1**
HL	**POS**	0.39	0.49	0.44
	NEG	0.80	0.41	0.54
	NNEU	0.72	0.59	0.65
	HR	0.45	0.53	0.47
	FUSED	**0.86**	**0.61**	**0.71**
MA	**POS**	0.61	0.53	0.57
	NEG	0.74	0.51	0.61
	NNEU	0.80	0.62	0.70
	HR	0.63	0.48	0.54
	FUSED	**0.82**	**0.64**	**0.72**

7. DISCUSSION AND LIMITATION

The goal of our study was to explore facial expression and non-invasively collected heart rate data of viewer in response to soccer stimuli, for highlight event detection. It is common understanding that heart rate response is more subtle and carries more sophisticated emotion information than facial expression. However, our findings show that facial expression performs better than heart rate in identifying viewer interest. To maintain less constrained and non-invasive setting for data collection, we used a photo diode based heart rate sensor (*Mio Alpha*) instead of strapped ECG sensor and therefore, the heart rate we collected had smaller resolution than standard ECG. According to a recent study, pulse plethysmography (PPG) based Mio Alpha has a mean absolute error of 4.43 bpm and the overall reliability score found was 77.83% against standard ECG for heart rate data collected during physical activity and exercise [9]. But our findings show that low resolution heart rate data is still useful to extract interest information of viewer with a reasonable accuracy (F1 score: 0.47, 0.54). Moreover, compared to strapped ECG sensor, the advantage of strapless heart rate sensor is that it can be easily applicable in real-life scenarios.

Viewer interest information extracted from heart rate was found different and independent of facial expression. Heart rate signals can identify interesting segments which are missed by facial expression and therefore heart rate signals can be used as a complementary to facial expression. Therefore, fusion of these two modalities (*HR* and *FE*) demonstrated better results than each of them individually.

This study used only soccer stimuli to investigate the response collected from ten subjects. The scope of soccer events included in the stimuli was limited to *goal*, *foul*, and *shot-on-goal*. Subject independent (e.g., expert's) annotation could be used in validation, besides match report highlights and subjective annotation. Use of standard ECG could be used for accurate heart rate data.

8. CONCLUSION AND FUTURE WORK

This study collects facial expression and heart rate data from ten subjects in three different trials using three stimuli. It has been shown that interesting video segments can be identified using facial expression and heart rate which complement each other. Our findings show that a useful temporal mapping can be obtained between viewer response and interesting soccer video events and in our case it was found that viewer response may not necessarily (temporally) synced with different types of key soccer events. This work could be extended using data collected from more subjects for training to improve the performance. Different stimuli other than sports video can be used. Accurate heart rate sensor could bring up more interesting findings from a similar study.

9. REFERENCES

[1] Arapakis, I., Konstas, I., and Jose, J.M., 2009. Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance. In *17th ACM international conference on Multimedia* ACM, 461-470.

[2] Bos, M.G.N., Jentgens, P., Beckers, T., and Kindt, M., 2013. Psychophysiological Response Patterns to Affective Film Stimuli. *PLoS ONE* 8, 4, e62661. DOI= http://dx.doi.org/10.1371/journal.pone.0062661.

[3] Fleureau, J., Guillotel, P., and Huynh-Thu, Q., 2012. Physiological-Based Affect Event Detector for Entertainment Video Applications. *IEEE Transactions on Affective Computing* 3, 3, 379-385.

[4] Hernandez, J., Liu, Z., Hulten, G., Debarr, D., Krum, K., and Zhang, Z., 2013. Measuring the engagement level of TV viewers. In *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013* IEEE, 1-7.

[5] Joho, H., Jose, J.M., Valenti, R., and Sebe, N., 2009. Exploiting facial expressions for affective video summarisation. In *ACM International Conference on Image and Video Retrieval* ACM, 31.

[6] Joho, H., Staiano, J., Sebe, N., and Jose, J.M., 2011. Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents. *Multimedia Tools and Applications* 51, 2, 505-523.

[7] Lang, P.J., Greenwald, M.K., Bradley, M.M., and HAMM, A.O., 1993. Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology* 30, 3, 261-273.

[8] Money, A.G. and Agius, H., 2009. Analysing user physiological responses for affective video summarisation. *Displays* 30, 2, 59-70.

[9] Parak, J. and Korhonen, I., 2014. Evaluation of wearable consumer heart rate monitors based on photopletysmography. In *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2014* IEEE, 3670-3673.

[10] Peng, W.-T., Chu, W.-T., Chang, C.-H., Chou, C.-N., Huang, W.-J., Chang, W.-Y., and Hung, Y.-P., 2011. Editing by viewing: automatic home video summarization by viewing behavior analysis. *IEEE Transactions on Multimedia* 13, 3, 539-550.

[11] Peng, W.-T., Huang, W.-J., Chu, W.-T., Chou, C.-N., Chang, W.-Y., Chang, C.-H., and Hung, Y.-P., 2009. A user experience model for home video summarization. In *Advances in Multimedia Modeling* Springer, 484-495.

[12] Soleymani, M., Chanel, G., Kierkels, J.J., and Pun, T., 2008. Affective ranking of movie scenes using physiological signals and content analysis. In *2nd ACM Workshop on Multimedia Semantics* ACM, 32-39.

[13] Soleymani, M., Chanel, G., Kierkels, J.J., and Pun, T., 2009. Affective characterization of movie scenes based on content analysis and physiological changes. *International Journal of Semantic Computing* 3, 02, 235-254.

[14] Soleymani, M., Koelstra, S., Patras, I., and Pun, T., 2011. Continuous emotion detection in response to music videos. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on* IEEE, 803-808.

[15] Tjondronegoro, D., Chen, Y.-P.P., and Pham, B., 2004. The power of play-break for automatic detection and browsing of self-consumable sport video highlights. In *6th ACM SIGMM international workshop on Multimedia information retrieval* ACM, 267-274.

[16] Wang, S., Liu, Z., Zhu, Y., He, M., Chen, X., and Ji, Q., 2014. Implicit video emotion tagging from audiences' facial expression. *Multimedia Tools and Applications*, 1-28.

[17] Zhang, L., Tjondronegoro, D., and Chandran, V., 2013. Facial expression recognition experiments with data from television broadcasts and the World Wide Web. *Image and Vision Computing*.