Spectators' Synchronization Detection based on Manifold Representation of Physiological Signals: Application to Movie Highlights Detection

Muszynski, Michal; Kostoulas, Theodoros; Chanel, Guillaume; Lombardo, Patrizia; Pun, Thierry

Spectators' Synchronization Detection based on Manifold Representation of Physiological Signals: Application to Movie Highlights Detection

Michal Muszynski
Computer Vision and Multimedia Laboratory
University of Geneva
Switzerland
michal.muszynski@unige.ch

Theodoros Kostoulas
Computer Vision and Multimedia Laboratory & Swiss Center for Affective Sciences
University of Geneva
Switzerland
theodoros.kostoulas@unige.ch

Guillaume Chanel
Swiss Center for Affective Sciences & Computer Vision and Multimedia Laboratory
University of Geneva
Switzerland
guillaume.chanel@unige.ch

Patrizia Lombardo
Department of Modern French & Swiss Center for Affective Sciences
University of Geneva
Switzerland
patrizia.lombardo@unige.ch

Thierry Pun
Computer Vision and Multimedia Laboratory & Swiss Center for Affective Sciences
University of Geneva
Switzerland
thierry.pun@unige.ch

ABSTRACT

Detection of highlights in movies is a challenge for the affective understanding and implicit tagging of films. Under the hypothesis that synchronization of the reaction of spectators indicates such highlights, we define a synchronization measure between spectators that is capable of extracting movie highlights. The intuitive idea of our approach is to define (a) a parameterization of one spectator's physiological data on a manifold; (b) the synchronization measure between spectators as the Kolmogorov-Smirnov distance between local shape distributions of the underlying manifolds. We evaluate our approach using data collected in an experiment where the electro-dermal activity of spectators was recorded during the entire projection of a movie in a cinema. We compare our methodology with baseline synchronization measures, such as correlation, Spearman's rank correlation, mutual information, Kolmogorov-Smirnov distance. Results indicate that the proposed approach allows to accurately distinguish highlight from non-highlight scenes.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications—*signal processing*; I.5.2 [**Pattern Recognition**]: Design Methodology—*pattern analysis*

Keywords

Synchronization; Physiological Signals; Affective Computing; Time Delay Embedding; Manifold Learning; Dimensionality Reduction; Diffusion Maps; Highlights Detection

1. INTRODUCTION

Detection and recognition of highlights in movies are difficult tasks due to the large variability of actions, scenes and sequences that strongly depend on the film genre. Adopting a supervised approach to highlights detection needs a large number of manually annotated samples [13]. Moreover, the movie annotation process is very tedious in itself and often biased by personal preferences of the annotators.

In the last decade, researchers have attempted to match physiological responses to the affective state of viewers and the appearance of highlights in films [4]. Measuring physiological responses to the movie content can provide insight into the viewers' aesthetic experiences, and can help better understand emotions elicited from the particular scenes [13].

In the field of affective computing, much research on emotion recognition in response to multimedia content has been carried out using EEG signals, peripheral physiological signals and facial expressions [9], [11]. In the area of highlights detection, various models have involved physiological measurements of a single viewer when watching of media, such as movies [7], music excerpts [14], and sport events [6]. Another approach to highlights detection which was based on the analysis of several spectators' physiological signals has been proposed in [4]. The authors used a physiological index of social interaction to determine general highlights of videos. That method allowed the detection of highlights that are relevant to the majority of viewers. Those experiments allow the evaluation of the feedback of one spectator, but they cannot take into account interactions among a group

of viewers.

Strong interactions among viewers can occur while watching spectacular types of highlights e.g. special effects. Furthermore, viewers can feel strong empathy with the movie characters, and have similar reactions to a certain movie content e.g. dramatic events. In these cases, we expect that similar physiological reactions of viewers are evoked. Under these assumptions, we propose that a level of physiological synchronization can be considered as a reliable indicator of highlights appearance in movies. Using multiple viewers' physiological recordings allows us to alleviate the impact of their personal preferences. Contrarily to methods which rely on emotion assessment, the proposed approach can be applied directly on multi-person physiological signals, and does not require the tedious annotation of movies (however, the expert's annotation is used only to evaluate the method in our study).

Many different synchronization measures have been employed lately to process physiological signals. One such possible synchronization measure is for instance the Pearson correlation coefficient (Spearman's rank correlation coefficient) that is able to quantify linear correlations between pairs of signals. Another approach to synchronization originates from information theory. Signals can be regarded as a collection of random variables which represents the evolution of a system over time. In this context, a basic similarity measure is mutual information or Kolmogorov - Smirnov distance that can be used as a similarity measure between distributions of signals.

In the area of emotion recognition, feature vectors of images and videos have been modeled recently as points lying on some Riemannian manifold to retrieve intrinsic structure of data [10]. This approach cannot be used to investigate a family of manifolds as is the case for highlights detection where processing of several viewers' physiological signals takes place.

2. TIME DELAY EMBEDDING AND DIFFUSION MAPS

Time-delay coordinate embedding has been used in the analysis of dynamical systems [12]. This method embeds a scalar time series into an m-dimensional space to reconstruct the trajectory of a system. For each time series $\{x_i\}, i = 1, 2, 3, ..., n$ a representation of the delay-coordinate embedding can be expressed as the following vector X_i which consists of m components

$$X_i = [x_i, x_{i+j}, x_{i+2j}, .., x_{i+(m-1)j}], \tag{1}$$

where j is the index delay and m is the embedding dimension.

We assume that the high-dimensional representation of physiological signals X_i is controlled by a low-dimensional process that corresponds to a physiological response to the stimulus. Applying diffusion maps to time-delay coordinate embedding, we provide a new low dimensional parameterization that still captures physiological activity. When diffusion maps are used [5], an affinity metric $K(x_i, x_j)$ is defined between pairs of physiological samples x_i and x_j based on their representation in time-delay coordinate X_i and X_j, respectively. Then, we consider only a collection $\mathcal{M}$ of k samples x_i to define the following kernel

$$K(x_i, x_j) = e^{\frac{-||X_i - X_j||}{\epsilon}}, \tag{2}$$

where ϵ is the scale parameter of the affinity metric and $k < n$. Now, note that $i, j = 1, 2, 3, ..., k$. We can look at the collection $\mathcal{M}$ as nodes of an undirected symmetric graph, where two nodes x_i and x_j are connected by an edge with the affinity weight $K(x_i, x_j)$. We pursue the construction of a Markov chain on the graph nodes by normalizing the kernel $K(\cdot, \cdot)$. Let K be the kernel matrix, and let $P = D^{-1}K$ be the corresponding transition matrix, where D is a diagonal matrix with elements $D_{ii} = \sum_{j=1}^{k} K(x_i, x_j)$. In sequence, we can calculate P_t analogues to P. Now, $P(x_i, x_j)$ is the probability of transition in a single step from node x_i to node x_j. Similarly, we define $P_t(x_i, x_j)$ as the transition probability in t steps from node x_i to node x_j. The idea is that the transition probability between two nodes can reflect the local geometry of the data. This leads us to a definition of the diffusion distance $D_t(x_i, x_j)$ between pairs of samples, expressed by [5]:

$$D_t(x_i, x_j) = \sqrt{\sum_{l=1}^{k} (P(x_i, x_l) - P(x_j, x_l))^2 w(x_l)}, \tag{3}$$

where $w(x_l)$ is a normalization weight. Intuitively, two points are similar when many short paths with large weights connect them. It is proven that the diffusion distance $D_t(x_i, x_j)$ can be computed using the eigenvalues $\{\lambda_i\}$, that tend to 0 and have a modulus strictly less than 1, and the corresponding eigenvectors $\{\varphi_i\}$ of the transition matrix P [5]. Let $\Phi_t(x_i)$ for some $t \geq 0$ be the diffusion maps of time series samples $\{x_i\}, i = 1, 2, 3, ..., k$ into Euclidean space $\mathbb{R}^s$ that is defined by

$$\Phi_t(x_i) = [\lambda_1^{2t}\varphi_1(x_i), ..., \lambda_s^{2t}\varphi_s(x_i)], \tag{4}$$

where $s \in \{1, 2, ..., k-1\}$ is the new space dimensionality.

It has been shown that the diffusion distance between samples x_i and x_j equals the Euclidean distance in the diffusion maps space that is expressed as follows [5]

$$D_t(x_i, x_j) = ||\Phi_t(x_i) - \Phi_t(x_j)||. \tag{5}$$

3. LOCAL SHAPE DISTRIBUTION OF MANIFOLD REPRESENTATION

In this paragraph we present a geometric framework which computes the amount of synchronization between a pair of physiological signals. The concept is to measure the similarity between local shapes of reconstructed signal manifolds. Firstly, in order to capture the unique local geometric properties of a signal manifold, we introduce the local shape cumulative distribution function $F_{x_i}^{\sigma}(\delta)$ of pairwise diffusion distances for each sample x_i and its delay samples $x_i, x_{i+1}, ..., x_{i+\sigma}$ denoted by

$$F_{x_i}^{\sigma}(\delta) = \int 1_{\tilde{D}_t(x_i, x_{i+q}) \leq \delta} \mathrm{d}\mu, \tag{6}$$

where $q \in \{1, \sigma\}$, μ is a counting measure and $1_{\tilde{D}_t(x_i, x_{i+q})}$ is an indicator function with respect to a delay sample on manifolds. $\tilde{D}_t(\cdot, \cdot)$ is the cosine distance in the diffusion maps space that can be derived from the Euclidean dot product.

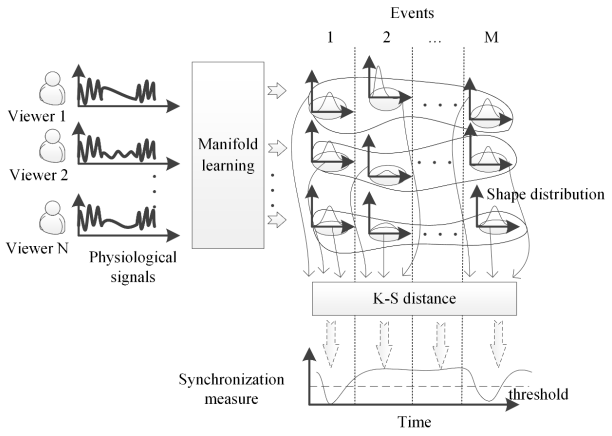


Figure 1: Overview of the proposed approach to highlights detection. In this fictitious example, all viewers 1,2,...,N are synchronized for event 1 and M in the movie since the manifolds are similar. This is not the case for event 2.

In our case, it is advantageous to use normalized the local shape distribution

$$\mathcal{F}^{\sigma}_{x_i}(\delta) = \frac{F^{\sigma}_{x_i}(\delta)}{F^{\sigma}_{x_i}(\infty)}. \quad (7)$$

For two time series $\{x_i\}$ and $\{y_i\}$, the synchronization measure is reduced to computing the Kolmogorov - Smirnov distance between two local shape distributions of their manifold representations for each time step i that is shown in Figure 1., and expressed as follows

$$S_{\sigma}(x_i, y_i) = \max_{\delta} |\mathcal{F}^{\sigma}_{x_i}(\delta) - \mathcal{F}^{\sigma}_{y_i}(\delta)|. \quad (8)$$

If two signals are the same $S_{\sigma}(x_i, y_i)$ is equal to 0. When the number of signals is more than 2, the overall synchronization can be obtained by averaging synchronization values of all possible non-overlapping pairs of signals.

4. EXPERIMENTS AND DISCUSSION

The synchronization measure was applied to physiological signals recorded during the watching of the movie (Taxi Driver, 1976) in a real cinema (Grütli cinema, Geneva) where viewers were wearing electro-dermal activity sensors. Our goal is to test if physiological signals synchrony can be used to detect the movie highlights defined by a cinema critic. In the present study, we utilize 12 skin conductance signals out of 30 recorded signals, and their sampling frequency is 10 Hz [8]. These signals are segmented in overlapping windows with time step and window length equal 0.5s and 5s, respectively. For diffusion map we set up s equals 3 based on values of eigenvalues of the transition matrix P, and for estimation of the local shape distribution we use 50 nearest samples ($\sigma = 50$) in time.

Annotation of the movie content was performed offline by an experienced movie critic, who annotated the movie based on the following five types of highlights [1], [2]. The so-called "Form-highlights" are:
- *H1*: Spectacular (technical choice, special effects);
- *H2*: Subtle (use of camera, lighting, music).
The so-called "Content-highlights" are:
- *H3*: Character development (characters' emotions and responses to dramatic events);
- *H4*: Dialogue (motivation of actions and tensions among characters);
- *H5*: Theme development (unusual close up, urban theme).

We apply our synchronization measure to determine scenes containing one particular type of highlights among (*H1*, *H2*, *H3*, *H4*, *H5*), as opposed to scenes without highlight. If the overall measure of the synchronization among all spectators (the average Kolmogorov-Smirnov distance between two local shape distributions) at time step i is lower than a threshold we assign this sample to a highlight scene.

We compare our methodology (*shape distribution dist.*) with baseline synchronization measures such as *correlation*, Spearman's rank correlation (*Spearman's correlation*), *mutual information* and Kolmogorov-Smirnov distance (*K-S distance*) that are applied to each signal window. The receiver operating characteristic (ROC) curves and the areas under the ROC curves (AUC) are depicted in Figure 2. and Table 1., respectively.

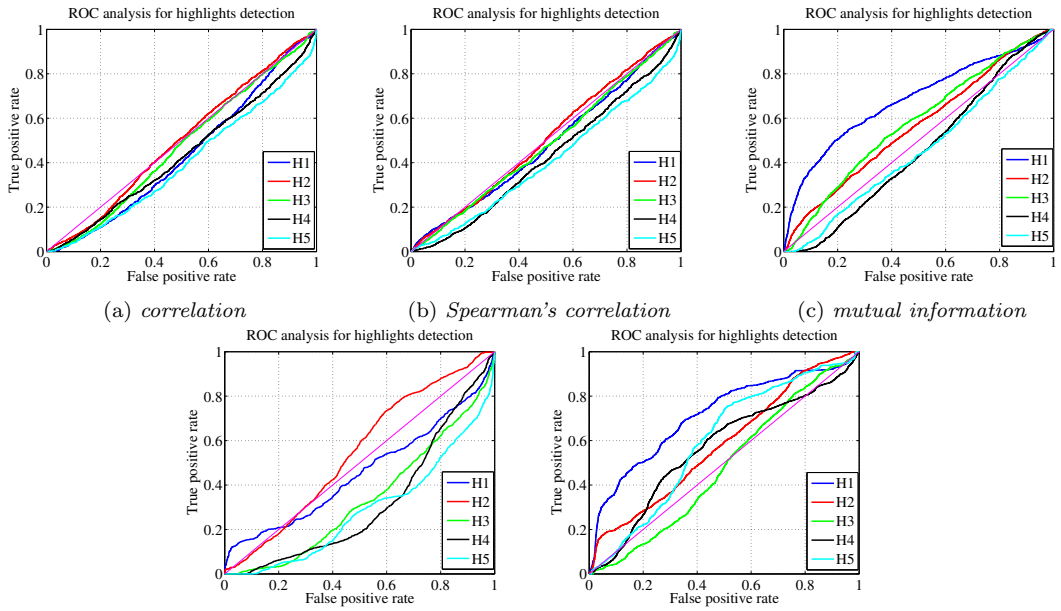
Table 1: Area under curve (AUC) for each highlight type ($H1$, $H2$, $H3$, $H4$, $H5$), and each different synchronization measure

Measure \ Highlights	*H1*	*H2*	*H3*	*H4*	*H5*
correlation	0.43	0.50	0.47	0.43	0.40
Spearman's correlation	0.48	0.50	0.48	0.42	0.41
mutual information	0.68	0.56	**0.58**	0.46	0.46
K-S distance	0.46	0.55	0.33	0.32	0.29
shape distribution dist.	**0.71**	**0.58**	0.48	**0.57**	**0.60**

The proposed methodology (*shape distribution dist.*) has significantly the highest performance for highlights *H1* which corresponds to AUC equal to 0.71 (Bradley test [3], $\alpha = 0.05$). This can be justified by the nature of the corresponding events, where it is expected to elicit strong physiological reactions from the spectators. In this case, our approach is capable of effectively discovering the similarity of skin conductance peaks because of its ability to explore the intrinsic structure of the data. On the other hand, the rest of the methods fail to detect synchronization among the spectators, and thus are not useful for the identification of movie highlights, for the given movie.

For detection of highlights *H2*, *H3*, our methodology and *mutual information* obtain significantly the best performance, respectively (Bradley test [3], $\alpha = 0.05$). The area under the ROC curve is equal to 0.58 in these cases. These results can be explained by a lack of strong synchronized reactions among all viewers to subtle contents of the movie, and character development that takes place. For highlight *H3*, this is also supported by the result of the *K-S distance* method, where we observe that the pairs of spectators are significantly low synchronized (Bradley test [3], $\alpha = 0.05$). It appears that possible single responses to these type of events cannot be well identified because of averaging synchronization over all pairs of spectators.

Furthermore, our method has significantly the highest performance (Bradley test [3], $\alpha = 0.05$) for detection of highlights *H4* and *H5* in the comparison with the baseline methods. The area under the ROC curve is equal to 0.57 and 0.60, respectively.



(a) *correlation* (b) *Spearman's correlation* (c) *mutual information*

(d) *K-S distance* (e) *shape distribution dist.*

Figure 2: ROC analysis for each highlight class detection (*H1*, *H2*, *H3*, *H4*, *H5*). Pink line corresponds to random detection.

5. CONCLUSIONS

In this work we propose a synchronization measure which is based on comparing local shapes of the manifold representation of signals. The comparison of the local shape distributions of diffusion distances on the manifolds is relatively invariant to scale (normalization in eq. 7) and topological changes of the signals. The results that we obtain on data recorded in a cinema indicate the ability of our methodology to identify some types of highlights in the movie based on synchronization of viewers' skin conductance signals.

6. ACKNOWLEDGMENTS

This work is supported by the grant from the Swiss National Science Foundation.

7. REFERENCES

[1] A. Bazin. *What is cinema?* University of California Press, 2004.

[2] D. Bordwell, K. Thompson, and J. Ashton. *Film art: an introduction*. McGraw-Hill New York, 1997.

[3] A. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[4] C. Chênes, G. Chanel, M. Soleymani, and T. Pun. Highlight detection in movie scenes through inter-users, physiological linkage. In *Social Media Retrieval*, pages 217–237, 2013.

[5] R. R. Coifman and S. Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

[6] A. Jaimes, T. Echigo, M. Teraguchi, and F. Satoh. Learning personalized video highlights from detailed mpeg-7 metadata. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, pages 133–136. IEEE, 2002.

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

[8] T. Kostoulas, G. Chanel, M. Muszynski, P. Lombardo, and T. Pun. Identifying aesthetic highlights in movies from clustering of physiological and behavioral signals. In *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*, May 2015.

[9] E. Kroupi, J.-M. Vesin, and T. Ebrahimi. Phase-amplitude coupling between eeg and eda while experiencing multimedia content. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 865–870. IEEE, 2013.

[10] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen. Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 494–501. ACM, 2014.

[11] M. Soleymani, S. Asghari-Esfeden, M. Pantic, and Y. Fu. Continuous emotion detection using eeg signals and facial expressions. In *Multimedia and Expo (ICME), 2014 IEEE International Conference on*, pages 1–6. IEEE, 2014.

[12] F. Takens. Detecting strange attractors in turbulence. *Dynamical Systems and Turbulence, Lecture Notes in Mathematics*, 898:366–381, 1981.

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

[14] J.-C. Wang, Y.-H. Yang, H.-M. Wang, and S.-K. Jeng. Modeling the affective content of music with a gaussian mixture model. *Affective Computing, IEEE Transactions on*, 6(1):56–68, 2015.