Early Prediction In Collective Intelligence On Video Users' Activity

Markos Avlonitis, Ioannis Karydis, Spyros Sioutas

Dept. of Informatics, Ionian University, Greece

{*avlon, karydis, sioutas*}*@ionio.gr*

Abstract

The huge volume of available video content calls for methods that offer insight to the content without necessitating burdensome users' extra effort or being applicable to specific types or conditions. Based on experimentation on collective users' interactions on a controlled user-experiment, this work analyses the results collected following the argument that bell-shaped reference patterns are shown to significantly correlate with scenes of interest for each video, as designated by the viewers. Though, in order to ensure the correlation of results to bell-shaped reference patterns, aggregation of a number of users' interactions is required. In order to overcome such an impediment and adhere to a real-case cold start scenario, we propose a stochastic transformation of the aggregated users' interaction signal into a space defined by its correlation to the bell-shaped reference patterns that is shown to offer significant amelioration as to the percentage of users' interaction required in order to achieve comparable results to the original users' interaction space. Moreover, to ensure further the realistic character of the proposed scenario, given an amount of already collected users' interaction, the interaction of new users' is shown to be predictable using neural network time series prediction and modeling methods. The results received indicate increased accuracy on how one can predict the most important scenes from low quantity early data of users' interactions as well as future interaction of unique users. In practice, the proposed techniques might improve both navigation within videos on the web as well as video search results with personalised video thumbnails.

Keywords: collective intelligence, early prediction, scene popularity, stochastic patterns, neural networks

1. Introduction

Nowadays, video content consumption and creation is easier than ever. On one side, widespread penetration of fast and highly interactive internet allows for an ever increasing number of users enjoying video content while on the other hand affordable storage as well as high-quality capturing devices have made creating such content an ubiquitous process with unprecedentedly high demand. The most popular web streaming video content service, YouTube [16], serves more than 1 billion unique users per month, while storing 100 hours of video every minute [17]. Accordingly, being able to make sense of the available content in a computerised manner, that is, being able to extract new and interesting information that is otherwise very difficult to be done due to the sheer volume of data, is of paramount importance.

Traditional content-based methodologies for the aforementioned data mining processes examine the actual content of each video in order to extract information. Nevertheless, their performance and capabilities fall short in certain occasions and thus research has recently focused on contextual or user-based semantics[1]. Such semantics rely on a broad spectrum of interactive behavior and "social activities" users exhibit and perform in relation to video content consumption such as sharing with others, assigning comments/tags, producing replies by means of other videos or even just expressing their preference/rating on the content. Rich as these "social metadata" may be, they have also been critiqued [2] as offering extra burden in the usual content consumption process that mainly includes viewing and browsing, by necessitating extra user effort, leading to the long-tail effect as to their existence. Thus, "social metadata" aside, research [8, 11] has examined the interaction information during the core processes of video content consumption, i.e. during viewing and browsing.

The previously mentioned increased interactivity Web 2.0 offered for the consumption of video content additionally assisted, through web-oriented architectures, in exposing content providers' functionality that other applications leverage and integrate in order to provide a set of much richer applications. In order to enhance the effectiveness of these applications, there is need for extensive studies of large users' interaction data. To this end, the design and implementation of controlled users' experiments has gained

[1]For a detailed discussion about the complementary character of the two approaches see [2]

increasing attention. Indeed, controlled experiments provide sets of data of (almost) any desirable size under controlled conditions giving thus the possibility to study specific users' interaction properties for specific Web content. Accordingly, Gkonela and Chorianopoulos [8] utilising the SocialSkip platform [4, 12] collected a pioneering user-based interaction dataset by conducting a controlled experiment during video content consumption providing a clean set of data that was easier to analyse. The platform integrated custom interface videos from YouTube with querying form functionalities of the Google Docs API in order to create an environment that would allow for video content consumption as well as user querying in order to accumulate data. Using a modified version of the SocialSkip platform, Spiridonidou et al. [13] conducted a controlled user-experiment designed and implemented in order to achieve high degree of realism to the typical contemporary scenario of web video-streaming services. In this way, the collective behavior of Web users watching the video content emerged by means of characteristic patterns in their activity leading to *collective intelligence* as to the importance of video content solely from users' interactions with the video player.

The dataset introduced at [8] was based on a set of restrictive assumptions as to its generality and thus the experiment conducted in [13] adhered to the following more real-use principles:

- the viewing interface included the controls found in YouTube in order to simulate realistic user video content consumption,
- content viewing did not have any time limit, again in order to simulate realistic user video content consumption,
- the questionnaire requested free-text replies in order to ensure that users declared the scenes they thought of as most important without interference or guidance,
- the significance of scenes was not predefined, allowing for true collective intelligence.

Nevertheless, the dataset collected from the experiment conducted in [13] did not undergo thorough testing in order to show that bell-shaped reference patterns significantly correlate with user defined scenes of interest in the video content.

Moreover, the existence of the aforementioned users' interaction signal can provide the basis for new series of metrics in order to study new characteristics of users' interactions as well as describe the content on which the interaction took place. Indeed, the aggregated users' interaction signal is useful (e.g. for identification of most important scenes), but requires a certain amount of users having interacted with the content interface in order to provide for any conclusions. Accordingly, real-use scenarios that refer to content with little consumption/interaction would not be able to benefit the methodologies' capabilities. It is thus required to devise methodologies that can predict the aggregated users' interaction signal of the necessary quantity of users from a set of low quantity early data of users' interactions.

In the same direction of early interaction prediction, aiming now not at the previously mentioned aggregated users' interaction but for a single user's interaction, such methods could be valuable in numerous cases such as: to model the specific user and his/her interaction pattern, to enhance and customise associated supportive content of the actual consumed content such as advertisements or other entities of interactive environments as well as to adjust content delivery ranging from parameters of the quality of scenes predicted to be interesting to customised content thumbnailing/summarisation to customised recommendations. Accordingly, methods are required that predict a user's interaction based on other users' interaction on the same content as well as a potential small amount of the specific user's initial interaction.

1.1. Motivation and Contribution

Bearing in mind the aforementioned lack of analysis on the correlation of the aggregated users' interaction signal of the datased collected in [13] to bell-shaped reference patterns as well as the requirements of small user quantity early interaction prediction both on the aggregated users' as well as on a single user's interactivity level, the contribution of this work is summarised as follows:

- the verification of the correlation of the dataset collected by Spiridonidou et al. [13] to bell-shaped reference patterns indicating user defined scenes of interest,
- the introduction of a stochastic transformation of the aggregated users' interaction signal into a space defined by its correlation to the bell-shaped reference patterns that is shown to offer significant amelioration

as to the percentage of users' interaction required in order to achieve comparable results to the original users' interaction space,

- the application of neural network time series prediction and modeling methods for the prediction of a single user's interactivity signal.

The rest of the paper is organised as follows: Section 2 describes related work and Section 3 details the user-based experiment this work is based on as conducted by Spiridonidou et al. [13]. Next, Section 4 presents the methods used and proposed in order to achieve the contributions of this work, while Section 5 details experimentation on all three key contributions of the work. Finally, the paper is concluded in Section 6.

2. Related Work

Research concerning video summarisation and, more generally, important scene selection in videos has mostly been based on content-based methodologies[2]. Nevertheless, as previously mentioned, such content-based methods often fail to capture high-level semantics that adhere to non-specialist users' navigation to videos [8].

In addition to video content, research has also been carried on the users' actions concerning their viewing and searching processes. Yu et al. [18] proposed that users unintentionally show their understanding of the video content through their interaction with the viewing system. Their developed algorithm, *ShotRank*, is computed through a link analysis algorithm that utilises the voting of users on the subjective significance and "interestingness" of each shot. Moreover, in addition to user browsing log mining, *ShotRank* is also taking into consideration low-level content video analysis.

In their work [14], Syeda-Mahmood and Ponceleon, presented *MediaMiner*, a client-server-based media playing and data-mining system aiming at tracking video browsing behavior of users in order to generate fast video previews. In *MediaMiner*, users' interaction with video is recorded at the client side while gathered information is returned to the server for continuous learning and estimation of browsing states. Modeling users' states transition, while browsing through videos, is done with a Hidden Markov Model. *MediaMiner* features common video-browsing interaction buttons (e.g. play and pause)

[2]Interested readers can refer to [7] for an extensive survey.

as well as random seek into the video via a slider bar, fast/slow forward and fast/slow backward.

Gkonela and Chorianopoulos [8], presented a user-centric approach, titled *VideoSkip*, wherein by analysis of implicit users' interactions with a web video player (e.g. pause, play, thirty-seconds skip or rewind) semantic information about the events within a video are inferred. Using the simple heuristic concerning the local maxima identification on the accumulated information collected from user-activity, *VideoSkip* has been able to effectively detect the same video-events, as indicated by ground-truth manually annotated by the author of the videos.

The work of Karydis et al. [10], in contrast to the hybrid solution proposed to [18], solely relies on user interaction with the player in order to identify high semantic value video intervals. Contrary to the work in [14], the approach in [10], utilises a differentiated methodology than a Hidden Markov Model that does not necessarily require the assumption that the state of "interestingness" of a user is a function of the previous state of the user. Moreover, the approach proposed therein examines the information received from each button of the application separately, offering thus greater flexibility to the event identification, that the approach adopted in [8].

The work of Spiridonidou et al. [13] extended the work done in [8] by reporting on the design, execution and results of a controlled user-experiment wherein participants were requested to view a video and identify their opinion on the importance of the scenes viewed in a realistic web-based video content viewing scenario, based on the interface of prominent video web-streaming provider. In contrast to the set of restrictive assumptions of the experimentation done in [8], the work of Spiridonidou et al. adhered to more real-use principles such as use of viewing interface controls found in YouTube, no use of time limit in content viewing and free-text replies to the questionnaire of scene importance to users.

3. User-based Experiment

This Section details the experiment conducted under the principles discussed in Section 1.1.

3.1. Participants

Being delivered through a web application, the experiment did not require the physical presence of the subjects and thus allowed users to undertake it at

their own time and location of preference. The dissemination phase included informing the users of the experiment's URL link through emails as well as facebook messages.

Following the received link, users were requested to voluntarily participate in the experiment the duration of which would not exceed six and a half minutes to watch the video and then a few minutes answer the questions. Users were also given simple and concise instructions for both parts of the experiment through the web interface.

The participants that undertook the experiment were 103 in number, 36 of who were male and 67 female. The age range varied from 17 to 35 years old and all of them were interested in cookery and The Greek Guide Association. All participants had experience in using the internet and video streaming services such as YouTube.

3.2. Materials

The experiment, based on the open-source SocialSkip platform, was designed according to the principles discussed in Section 1.

The chosen video for the experiment was required to be interesting to a lot of people so as to motivate to be viewed by a satisfying number of participants while additionally was required to be interesting to the selected participants, again to motivate users' interactions on the areas they deemed important.

Thus a set of criteria were identified for the video content selection process. Initially, an important criterion on the selection of the content was the duration of the video that according to the creators' initial programming of the Socialskip should not exceed 10 minutes to ensure that will be watched it until its end. Another key criterion accoding to [4] was the video's structure as the less structured it was, the more important the postdata for the future viewers would be. Thus, the video should be without montage, that is, without replay, slow motion or pause from the director. Moreover, the video should not have areas of increased importance too close to one another. Accordingly, a video about "Hot Wine" [3] was selected that lasts 6 minutes and 24 seconds and was captured by The Greek Guide Association in Komotini, Greece.

[3] http://www.youtube.com/watch?v=EY2Vq4WE-5k

3.3. Procedure

Before viewing a video, participants were offered a set of instructions as to the process of the experiment, stating that:

1. You may view this video as many times required but after selecting to reply to the questions associated, video viewing will not be possible. Thus, you are advised to memorise the parts of the video that you found interesting in order to describe these later,
2. You can use the slider while watching the video in any way you want (back, forward, pause),
3. When sure you have memorised the parts you found interesting press the link to move to the questions phase.

After the participants had selected to move to the questions phase they were given the following instructions: "*Describe the scenes that you consider important in the video (indicative number of scenes A, B, C, D, E). Write your description after you have given the letter of the scene. (ex. A, the scene where the cooking instructions are given)*". Intuitively, the instructions were made in this way in order to gather from each user the most important scenes for them, while receiving unguided postdata.

During the video's content presentation, the slider was available just under the viewing area allowing the user to randomly navigate through the content at will. A "pause" button was also available for the participants' convenience that freezed the video on the current scene it was pressed. The existence of both these buttons was dictated by the requirement to ensure realistic and familiar usage to the users. Furthermore, under the viewing area, the participants could see the total duration of the video as well as the relative current viewing video time.

Figure 1 shows a snapshot of the application's interface as shown to users coupled with the necessary informational instructions in order to ensure the understanding of the required actions on the users' side.

4. Proposed Methodology

This Section presents the methods used and proposed in order to achieve the contributions of this work, namely the users' activity modeling, the early aggregated signal prediction as well as the early unique signal prediction methods.

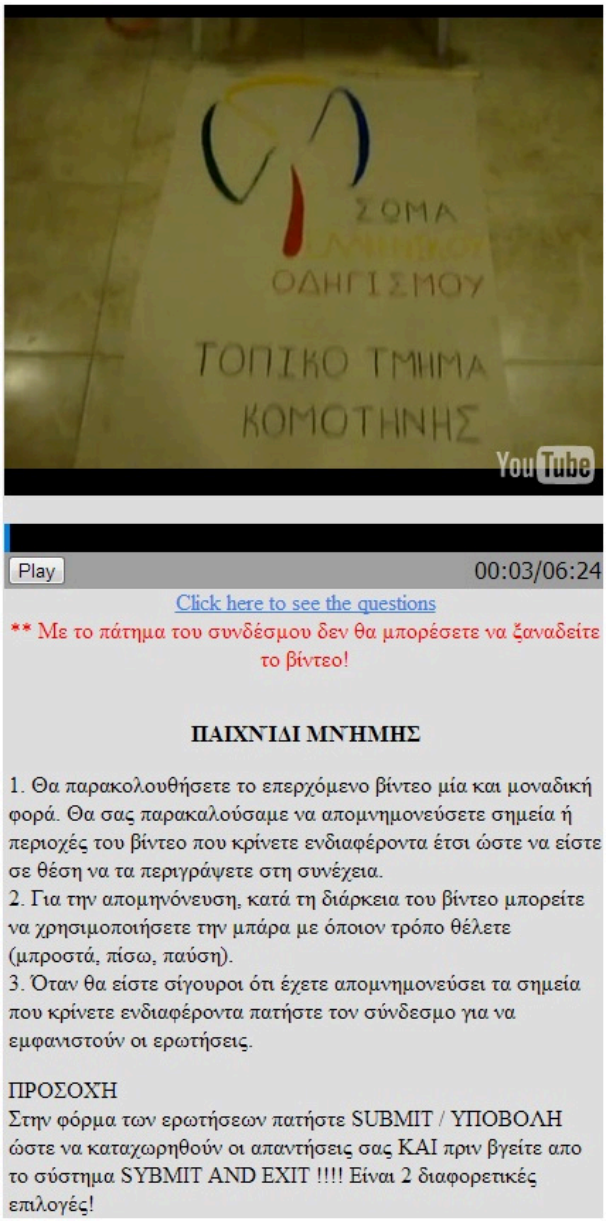

Figure 1: Snapshot of the application's interface as shown to users.

4.1. User Activity Modeling

Following the methodology presented by Karydis et al. [10] for the user activity modeling, in order to extract pattern characteristics from the users'

interaction signal, as provided through the interface of the video content provider, three distinct stages (as shown in Table 1), are used.

Stage	User activity signal processing
1	Smoothness procedure
2	Determination of users' activity aggregates
3	Estimation of pattern characteristics

Table 1: Overview of the user activity modeling and analysis.

In the first stage, a simple process is used in order to average out user activity noise in the corresponding distribution. In the context of probability theory, noise removal can be treated with the notion of the moving average [15].

Accordingly, from a curve $S^{exp}(t)$ a new smoother curve $S_T^{exp}(t)$ may be obtained as shown in Equation 1,

$$S_T^{exp}(t) = \frac{1}{T} \int_{t-T/2}^{t+T/2} S^{exp}(t')dt' \tag{1}$$

where T denotes the averaging "window" in time. The larger the averaging window T, the smoother the curve will be. The procedure of noise removal of the experimentally recording distribution is of crucial importance for the following reasons: first, in order to reveal patterns of the corresponding signals (regions of high user's activity), and second in order to estimate local maxima of the corresponding patterns. It must be noted that the optimum size of the averaging window T is entirely defined from the variability of the initial signal. Indeed, T should be large enough in order to average out random fluctuations of the users' activities and small enough in order to avoid distortion of the bell-like localised shape of the users' signal which will in turn show the area of high user activity.

In the second stage, aggregates of users' activity are estimated by means of an arbitrary bell-like reference pattern, following the intuition that there is high probability that user's actions are concentrated at a specific time interval (the center of the bell) while this probability tends to zero quite symmetrically while moving away from this interval.

In the third step, the estimation of the pattern characteristics takes place by application of two different methodologies, a stochastic and a pattern matching:

- In the stochastic approach, as described in [10], the estimation of the exact locations can be done via the estimation of the generalised local maxima. The term generalised local maxima refers to the center of the corresponding bell-like area of the average signal, as the nature of the original signal under examination may cause more than one peaks at the top of the bell due to the micro-fluctuation. It was thus shown that this is possible by estimating the well known correlation coefficient $r(x, y)$ between the two signals (time series), that is, the average experimental signal and the reference bell-like time signal.

- In the pattern matching approach, as also described in [10], the distance of the reference bell-shaped pattern to the accumulated user interaction signal is measured using 3 different distance measures. Initially, a Scaling and Shifting (translation) invariant Distance (SSD) measure (Equation 2), is adopted from [5]. Accordingly, for two time series x and y, the distance $\hat{d}_{SSD}(x, y)$ between the series is:

$$\hat{d}_{SSD}(x, y) = min_{a,q} \frac{\|x - \alpha y_{(q)}\|}{\|x\|} \tag{2}$$

  where $y_{(q)}$ is the result of shifting the signal y by q time units, and $\|\cdot\|$ is the l_2 norm. In this case, and for simplicity, the shifting procedure is done by employing a window the size of which is empirically calculated to minimise the distance, while the scaling coefficient α is adjusted through the maximum signal value in the window context.

  The second distance measure used is the Euclidean Distance (ED) measure (Equation 3) that has been shown to be highly effective [6] in many problems, despite its simplicity:

$$d_{ED}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3}$$

  Finally, the third distance measure utilised is a Complexity-Invariant Distance (CID) measure (Equation 4) for time series as discussed by Batista et al. [3]:

$$d_{CID}(x, y) = ED(x, y) \times CF(x, y) \tag{4}$$

where the two time series x and y are of length n, $ED(x, y)$ is the Euclidean distance (Equation 3), $CF(x, y)$ is the complexity correction factor defined in Equation 5:

$$CF(x, y) = \frac{max(CE(x), CE(y))}{min(CE(x), CE(y))} \tag{5}$$

and $CE(x)$ is a complexity estimate of a time series X, calculated as shown in Equation 6:

$$CE(x) = \sqrt{\sum_{i=1}^{n-1}(x_i - x_{i+1})^2} \tag{6}$$

The aforementioned distance measures produce another time series *dist* that describes the distance of the reference bell-shaped pattern to the accumulated users' interaction signal and thus requires the identification the locations of *dist* where its value is minimal, indicating a close match of the reference bell-shaped pattern to the accumulated signal. To avoid using a simplistic global cut-off threshold the approach incorporates a local minima peak detection methodology, where a point in *dist* is considered a minimum peak if it has the minimal value, and was exceeded, to the left of the signal, by a value greater by $DELTA$, the peak detection sensitivity value.

4.2. Early Aggregated Signal Prediction

Useful as users' interaction signals may have been shown for the description of the content these take place on, the phenomenon of cold start can be a potential problem due to few or none interactions on a set of content elements that do not allow for systems to draw any safe inferences.

In the case of video content, the explosive degree of such content creation ("100 hours of video are uploaded to YouTube every minute" according to [17]) and the ability to have it ready for easy consumption through numerous streaming Web services extend the aforementioned cold-start problem, requiring thus methodologies that can predict the aggregated users' interaction signal, of the necessary quantity of users, from a set of low quantity early data of users' interactions.

Accordingly, we propose the use of a transformation of the aggregated users' interaction signal into a space defined by its correlation to the bell-shaped reference patterns. The intuition of our proposal is that as these bell-shaped reference patterns have been shown to highly correlate to the scenes of the video the users thought of as most important, such a space (hence "Stochastic space") represents a qualitative compression of the original space (hence "Direct space") of the interactions. Thus, using a representation that discards noise and focuses on the important segments of the signal will allow for greater correlation between a partial and full data signal when compared to the noisy "Direct space".

In order to derive the proposed transformation, we notice that the arbitrary bell-shaped repeating pattern may be expressed by the general relation of the Gaussian function, i.e. as shown in Equation 7

$$S^{ref}(t') = A \cdot e^{-\frac{(t'-t)^2}{B}}, \quad t' \in (t-2B, t+2B) \tag{7}$$

where t is the center of the Gaussian bell, A is its maximum height and B is a measure of its width. In detail, for each time instance we define a time integral $(t-2B, t+2B)$ where the correlation coefficient of the signals, the experimental S^{exp} and the reference S^{ref} is calculated, using the well-known relation

$$\begin{aligned} S^{cor}(t) &= Cov({S_T}^{exp}(t), S^{ref}(t)) \\ &= E[{S_T}^{exp}(t) \cdot S^{ref}(t)] - E[S^{exp}(t)]E[S^{ref}(t)] \end{aligned} \tag{8}$$

where $E(\cdot)$ is the operator for the mean value. By means of the Equation 8, the stochastic transformation is established: from $S_T^{exp}(t)$ we derive a new signal $S^{cor}(t)$ which, according to our proposal, in each time instance will isolate the noise and enhance the valuable information we search for, i.e. the existence of bell-like patterns which in turn can be mapped to represent the existence of users' interest areas.

The users' activity modeling and early aggregated signal prediction pre-processing can be represented in pseudo-code as shown in Algorithm 1. For the case of the early aggregated signal prediction, only partial users' activity is being considered.

Following the methodology previously described in Section 4.1 using the stochastic pattern for the third step we propose the representation of the

Algorithm 1 Users' activity modeling and early aggregated signal prediction algorithm

Require: Experimental time series $S_T^{exp}(t)$, upper part of Gaussian time series $S_{B,A}^{ref}(t)$ of center t, width B and height A.

1: **for** $t = 1$ **to** L **do**
2: $S_{B,A}^{ref}(t)$ {the correlation coefficient for different instance-centers}
3: **end for**
4: **return** $S_{B,A}^{ref}(t)$ {returns the transformed signal}

initial users' interaction signal from the "Direct space" to the "Stochastic space".

4.3. Early Unique Signal Prediction

Much as it is interesting to be able to have a representation of a part of the full signal in a differentiated space that exhibit increased correlation, the even partial aggregation of the users' interactions makes information about sole users non available.

Having early access to such information/prediction about the future interaction of a specific user, based on interactions of users that have already completely happened as well as a portion of the user's at hand interaction, can be invaluable for numerous tasks. Such tasks could be the modeling of the specific user and his/her interaction pattern, the amelioration and customisation of associated supportive content of the actual consumed content such as advertisements or other entities of interactive environments as well as the adjustment of content delivery ranging from parameters of the quality of scenes predicted to be interesting to customised content thumbnailing/summarisation to customised recommendations.

Accordingly, we propose the use of a Recurrent Neural Network (RNN) and especially an architectural approach of RNN with embedded memory, the "Nonlinear Autoregressive model process with eXogenous input" (NARX) [9], as shown in Equation 9

$$j(g) = f(j(g-1), \ldots, j(g-d), x(g-1), \ldots, x(g-d)) \tag{9}$$

where the next value of the dependent output signal $j(g)$ (user's interaction signal) is regressed based on d previous values of the same signal and previous values of an independent (exogenous) input signal x, which, in our case are the interactions of users that have already happened.

The NARX network is used for the prediction of a user's interaction based on other users' interaction on the same content as well as a potential small amount of the specific user's initial interaction, and is fed with the unique signal, after the smoothness procedure, of each user that provided answers to the user-based experiment as observations.

5. Experimental Evaluation

In this Section we present experimentation on all three key contributions of our work as described in Section 1.1. Initially the experimental set-up is described, then the results are presented and finally a short summarisation of the key findings is presented.

5.1. Experiment setup

The experimental setup section is divided on the three key areas of pre-processing preparations that were required for the experimental evaluation, namely the users' interaction signal pre-processing, the users' collected free-text quantisation for the definition of the ground-truth with the synthetic five element signal used in the application of neural network time series prediction and the modeling methods for the prediction of a single user's interactivity signal.

5.1.1. Interaction data processing

Initially, a simple process is used in order to average out user activity noise in the corresponding collected users' interaction signal as described in Section 4.1. Figure 2 depicts the originally received users' interaction signal, the users' activity signal as approximated to a smoothed signal using $T = 20$ and the ground truth.

It is quite visually evident from Figure 2 that the peaks of the smoothed aggregated users' interaction signal coincide with the time segments designated by users as most important scenes.

5.1.2. Users' postdata processing

Users' postdata from the free-text replies on the description of scenes' significance were collected, processed and based on the textual descriptions eight scenes were selected as shown in Table 2.

Of the scenes submitted by the users, only five were retained after exclusion of scenes based on users' link to the video content provider and the

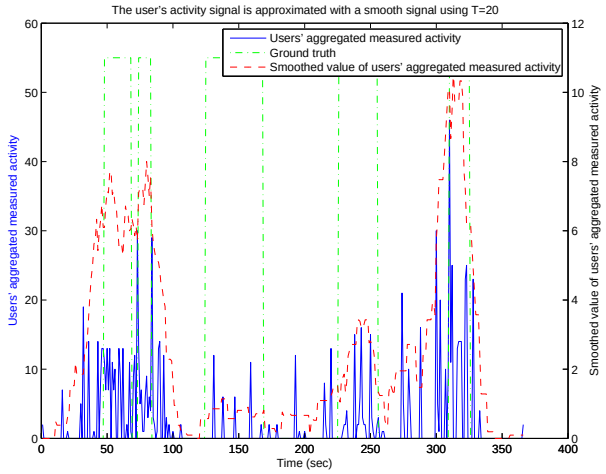

Figure 2: Users' activity signal approximated to a smoothed signal using $T = 20$ with markup of the duration of scenes deemed important by users.

#	**Scene description**	**Scene starts at**	**Scene ends at**	**Mentioned by # users**
1	Inaugural - The Greek Guide Association logo	0	4	31
2	Description of ingredients	48	68	63
3	1st recipe secret: Make it hot (60 °C)	74	83	40
4	1st part of recipe preparation	86	118	16
5	Historical information	125	168	39
6	2nd recipe secret: figs in liquor, raisin skewers on cloves	226	255	49
7	The appearance of Santa Claus	310	325	75
8	Trying the wine	330	346	6

Table 2: Quantised descriptions and time of beginning/ending of important scenes as submitted by users.

mentions each scene accumulated. Thus, scene 1 was excluded as most experiment participant users were part of the Greek Guide Association that made the video and thus their association to the first image of the inaugural scene showing the logo of the Greek Guide Association impacted on describing the scene as important without regard to the content. Moreover, scenes 4 and 8 referring to the first part of recipe preparation and the tasting of the produced the wine collected too few mentions.

Additionally, each user and thus the user's interaction signal was linked

to a five element signal wherein, based on the scenes that where retained for the experiment, each element of the signal was assigned an one unit value (1) if the user had designated the equivalent scene as important or a zero (0) value otherwise. As not all 103 participants returned free-text description on their perceived important scenes, but only 76 of them, the association of the five element signal was only performed for those users that submitted description of the scenes deemed important.

5.1.3. NARX-based prediction

In the RNN experimentation using the NARX architecture, a dynamic recurrent NN with feedback connections enclosing several layers of the network, is used containing a varying number of neurons in order to test the effect of the neuron availability. In addition, the experimentation has also included the division of the dataset into training, validation of generality, and testing subsets. In the experiment with the NARX presented herein, the evaluation of the performance was based on the quantised descriptions of important scenes as submitted by users formed into a five element signal, as described in Section 5.1.2.

The learning function used was the Levenberg-Marquardt function while the performance function was the mean squared error (MSE) performance function. For each of the parameters examined (number of neurons, division of the dataset and input/feedback delays) the resulting performance of the NARX was averaged over multiple runs due to the randomness in the division of the dataset into training, validation, and testing subsets and in order to receive high quality results.

5.2. Experiment results

In this section we present the experimental results attained by the estimation of the correlation of the users' interaction signal to bell-shaped reference patterns, the signal transformation for the early aggregated signal prediction and finally the RNN Time-series prediction for the early unique signal prediction.

5.2.1. Users' interaction signal verification

Following the methodology presented by Karydis et al. [10] for the correlation of the users' interaction signal to bell-shaped reference patterns indicating scenes of interest we experimented with both the stochastic and pattern matching approaches described therein.

As far as the stochastic approach is concerned, the analysis of the users' activity distributions was based on an exploration of several alternative averaging window sizes. Results of the proposed modeling methodology for the video content are shown in Figure 3, and in this case, the pulse width D is 20 seconds and the smoothing window T is 20 seconds. The results are depicted by means of pulses instead of the bell shapes in order to compare with the corresponding pulses of the ground-truth designated by the users of the experiment. The mapping between pulses and bells is based on the rule that the pulse width is equal to the width between the two points of the bell where the second derivative changes sign. Moreover, pulse signals were extracted from the corresponding local maxima indicating time intervals where aggregates were detected according to the definition given in Section 4.1.

For the stochastic approach, the correlation of the estimated high-interest intervals and the ground-truth as submitted by the viewers of the video was calculated at 0.862 indicating very strong correlation between the two pulses.

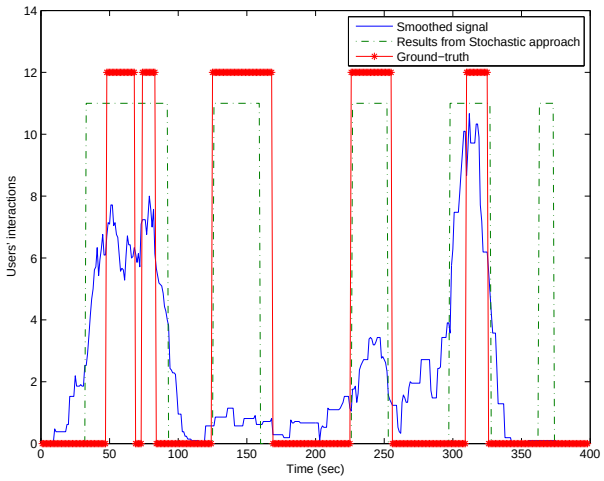

Figure 3: Cumulative users' interaction vs. time including results from stochastic approach and ground-truth.

The pattern matching approach was examined for each distance measure as described in Section 4.1. The time intervals of each video, provided by the cumulative users' descriptions of scenes' importance during the experimentation, once again, constitute the ground-truth, based on which the classifier's prediction is evaluated.

Figures 4, 5 and 6 show the achieved precision, recall, specificity percent-

age & Matthews Correlation Coefficient (MCC) values. The values shown therein are for varying peak detection sensitivity values for each of the three distance measures, SSD, ED and CID respectively. It should be noted that the left y-axis of the Figures depicts precision, recall, F1 score, accuracy and specificity while the right y-axis depicts only the MCC value. As it can be seen in Figure 4, the SSD metric achieves an F1 score of 0.7 in a scale of $[0, 1]$, with 1 being the best value. Nevertheless, F1 score does not take the true negative rate into account. Thus, the MCC value has been additionally computed leading to a 0.54 value on a scale of $[-1, 1]$, with 1 implying a perfect prediction. The claim that the Euclidean Distance is able to perform relatively high, despite its simplicity, is once again shown in Figure 5 where ED scored an F1 score of 0.53 and an MCC value of 0.26. Finally, the CID measure, as shown in Figure 6, was outperformed by the other two measures having scored an F1 score of 0.57 and an MCC value of 0.35.

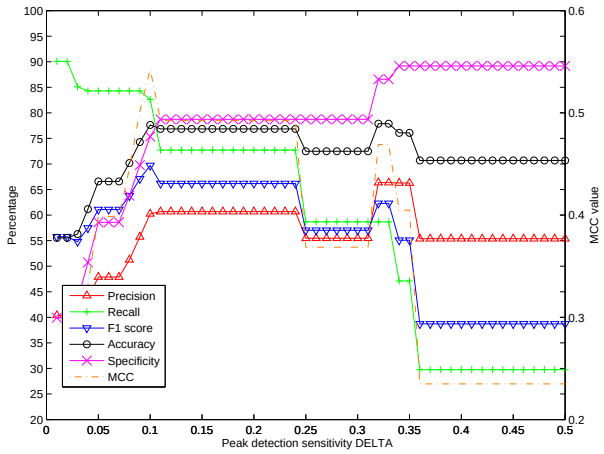

Figure 4: Pattern matching approach, SSD measure: Precision, recall, F1 score, accuracy, specificity percentage & MCC value.

5.3. Signal Transformation

In order to test the capability of the proposed (as described in Section 4.2) transformation, all three distance measures described in Section 4.1 were used. In addition to the aforementioned Scaling and Shifting (translation) invariant Distance (SSD), Euclidean Distance (ED) and Complexity-Invariant Distance (CID) measures, another well-known distance measure was used, the Dynamic Time Warping (DTW) [1].

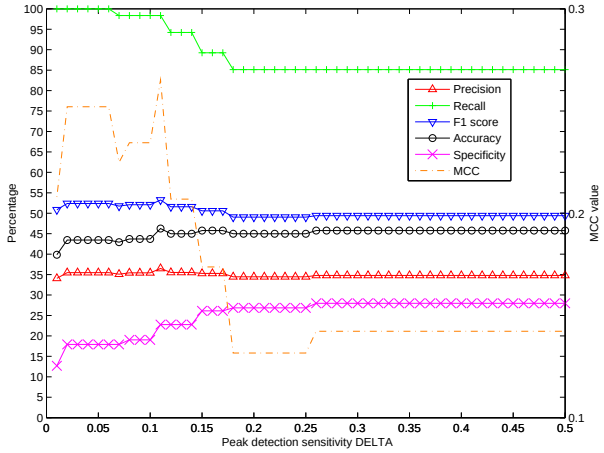

Figure 5: Pattern matching approach, ED measure: Precision, recall, F1 score, accuracy, specificity percentage & MCC value.

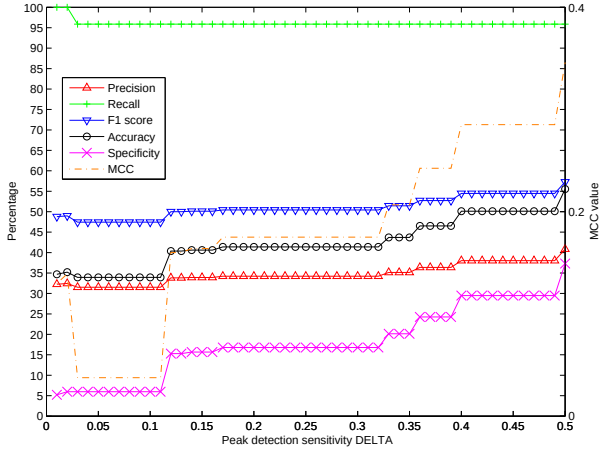

Figure 6: Pattern matching approach, CID measure: Precision, recall, F1 score, accuracy, specificity percentage & MCC value.

As it can be seen, from Figures 7, 8, 9 and 10 the distance of partial signals to the full signals in the stochastic space is in all cases less than the equivalent distance in the direct space, thus confirming the claim that conversion to the stochastic space allows for early data processing. This is quite evident in the cases of the CID and DTW measures where the aggregated signal of 20% of the users in the stochastic space is as similar as the signal of 70% of the users both in relation to the aggregated signal of all the users.

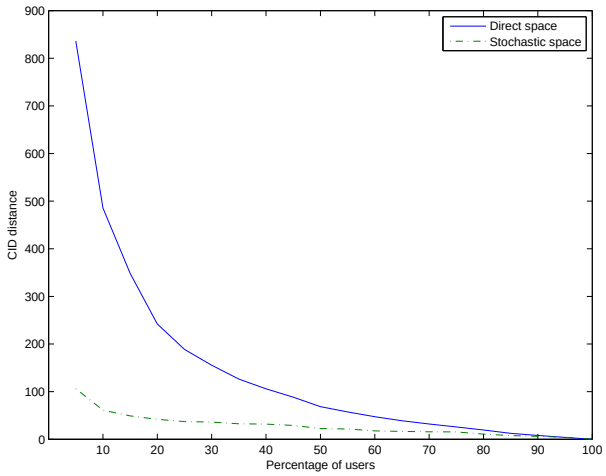

Figure 7: Distance of partial signal to the full in both spaces using the CID measure.

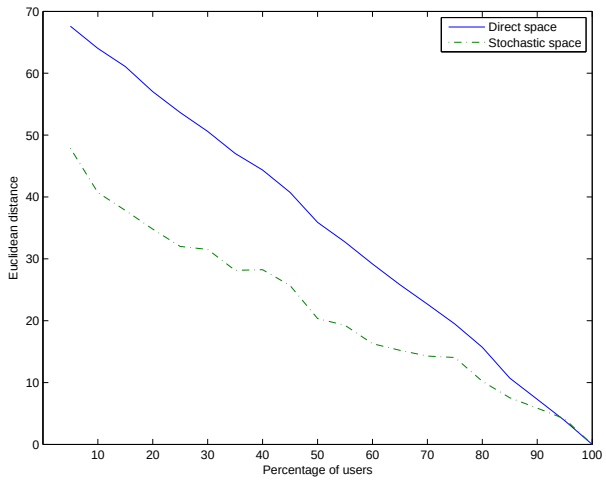

Figure 8: Distance of partial signal to the full in both spaces using the ED measure.

In other words, this experiment indicates that prediction of the aggregated users' interaction signal, at the necessary quantity of users, is a possibility. This prediction is derived from a set of low quantity early data of users' interactions with the use of a transformation of the aggregated users' interaction signal into a space defined by its correlation to the bell-shaped reference patterns, the "Stochastic space". Accordingly, these experiments confirm our intuition that the bell-shaped reference patterns represent a qualitative

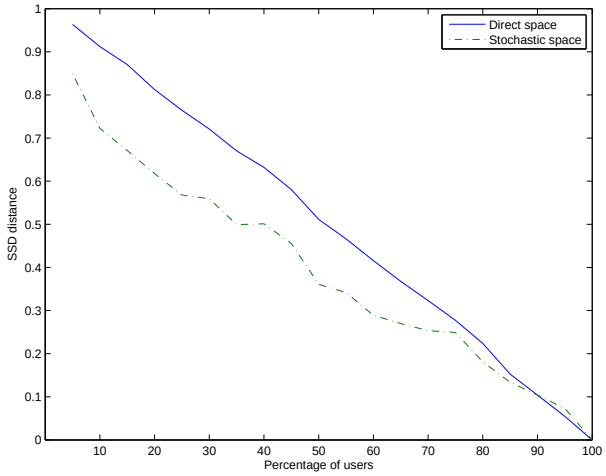

Figure 9: Distance of partial signal to the full in both spaces using the SSD measure.

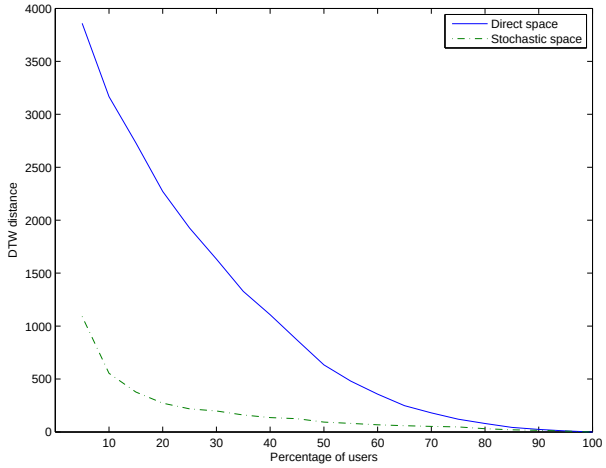

Figure 10: Distance of partial signal to the full in both spaces using the DTW measure.

compression of the "noisy" original space of the interactions.

5.4. NARX NN Time-series Prediction

In order to test the capability of the proposed (as described in Section 4.3) methodology for the prediction of a user's interaction based on other users' interaction on the same content as well as a potential small amount of the specific user's initial interaction, we experimented with the number of neurons, division of the dataset and input/feedback delays of the NARX architecture.

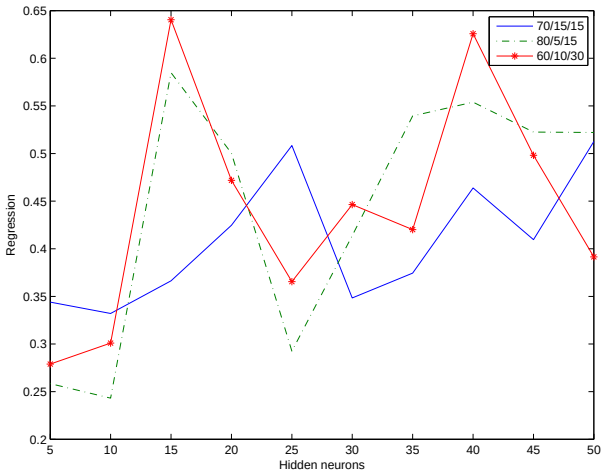

Figure 11: Regression for varying hidden neurons, input/feedback delay = 1.

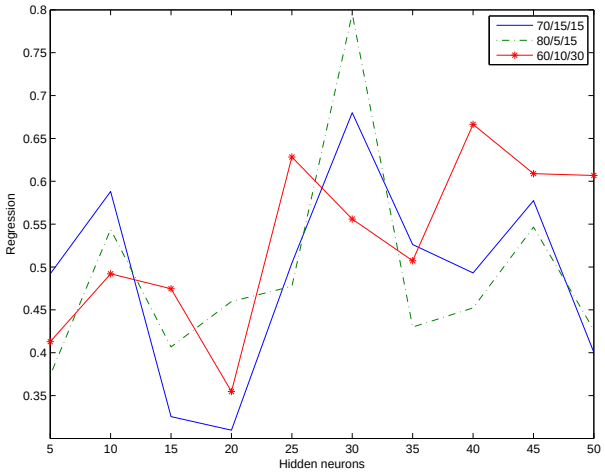

Figure 12: Regression for varying hidden neurons, input/feedback delay = 2.

Figures 11, 12 and 13 show the regression value achieved for varying hidden neurons. The regression metric measures the correlation between input and outputs of the NARX architecture. For each of the input/feedback delay values the regression attained in all experiments shows an increasing tendency for increasing number of hidden neurons, while for increasing input/feedback delay value the regression attained shows again increasing tendency reaching almost 0.8.

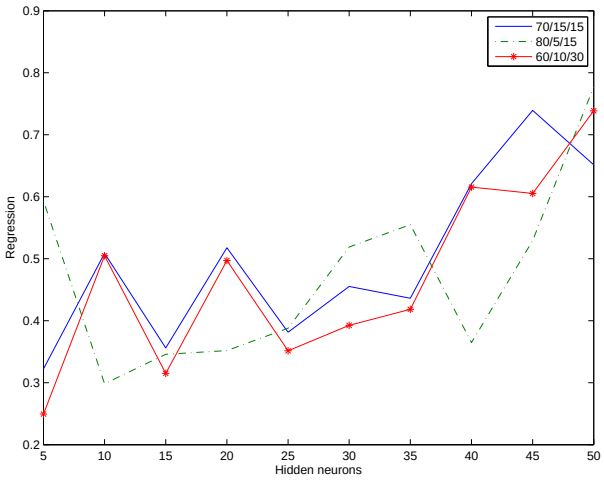

Figure 13: Regression for varying hidden neurons, input/feedback delay = 3.

These results confirm the intuition behind this experiment, as for increasing input/feedback delay or, in the context of our research more values of the user at hand, the capability of the NARX to predict the future interactions of the same user is ameliorating. Nevertheless, even for minimal input/feedback delay values of unit, the regression value is well above 0.6.

5.5. Discussion

The performance evaluation results can be summarised as follows:

- As far as the correlation of the aggregated users' interaction signal to bell-shaped reference patterns is concerned, both the stochastic and pattern matching approaches indicate that the collected users' interaction from the user-based experiment by Spiridonidou et al. [13] and the respective free-text descriptions of the scenes deemed by users as important exhibit a very high correlation.

- The transformation of the aggregated users' interaction signal into a space defined by its correlation to the bell-shaped reference patterns was shown to offer significant amelioration as to the percentage of users' interaction required in order to achieve comparable results to the original users' interaction space. Thus, the claim that conversion to the stochastic space allows for early data processing is supported.

- The valuable for numerous tasks application of neural network time series prediction and modeling methods was shown to be fruitful paving the way for accessible early information about the future interaction of a specific user based on interactions of users that have already happened as well as a portion of the user's at hand interaction.

6. Conclusion

In this work, a controlled user-experiment was utilised in order to identify the segments of video content participants thought of as most important by means of registering their interactions with the video interface. The experiment was designed and implemented in order to achieve high degree of realism to the typical contemporary scenario of web video-streaming services.

Extensive results on the data collected showed high correlation of the areas (video time segments) that received more interactions from the users with their free-text submitted replies on what they thought of as important. In other words, the most important scenes according to the participants of the experiment, that form the so called "ground truth", highly coincide with the patterns emerged in users' interaction time signal.

In order to ensure the realism of the typical contemporary scenario of web video-streaming consumption, in terms of how one can predict the most important scenes from low quantity early data of users' interactions, a transformation is proposed herein that maps the users' aggregated interaction signal based on its correlation to bell-shaped reference patterns. The transformation was shown to offer significant amelioration as to the percentage of users' interaction required in order to achieve comparable results to the original users' interaction space.

In addition, in order to provide for user-oriented customisation methods on video content presentation and consumption, given an amount of collected users' interaction, we propose the use of a dynamic recurrent Neural Network for the prediction of the interaction of new users'. The approach proposed is shown to achieve high correlation between input and outputs.

The existence of a signal (here the signal counts how many times the slider was located at a specific second) that can carry the information about the most important video scenes could be a valuable tool for Web applications. Moreover, the existence of the aforementioned users' signal can provide the basis for new series of metrics in order to study new characteristics of users' interactions. Indeed, the identification of most important scenes is just a

first order video characteristic. On a higher level, one could search for second order characteristics: what is the duration of each important scene and how popular each scene is (there could be more than one popular scenes in each video but what is their gradation popularity). The above issues can be addressed in a very rigorous manner since these features can be mapped to well known functions in standard signal processing theories. These aspects will be addressed in a future work.

References

[1] Agrawal, R., Lin, K.I., Sawhney, H.S., Shim, K., 1995. Fast similarity search in the presence of noise, scaling, and translation in time-series databases, in: Proceedings of the 21th International Conference on Very Large Data Bases, pp. 490–501.

[2] Avlonitis, M., Karydis, I., Chorianopoulos, K., Sioutas, S., 2013. Treating Collective Intelligence in Online Media. CRC Press / Taylor & Francis. chapter Semantic Multimedia Analysis and Processing.

[3] Batista, G.E.A.P.A., Wang, X., Keogh, E.J., 2011. A complexity-invariant distance measure for time series, in: Proc. SIAM Conference on Data Mining, pp. 699–710.

[4] Chorianopoulos, K., Leftheriotis, I., Gkonela, C., 2011. Socialskip: pragmatic understanding within web video, in: Proceddings of the 9th international interactive conference on Interactive television, pp. 25–28.

[5] Chu, K.K.W., Wong, M.H., 1999. Fast time-series searching with scaling and shifting, in: PODS, pp. 237–248.

[6] Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., Keogh, E., 2008. Querying and mining of time series data: experimental comparison of representations and distance measures. Proc. VLDB Endowment 1, 1542–1552.

[7] Geetha, P., Narayanan, V., 2008. A survey of content-based video retrieval. Journal of Computer Science 4, 474–486.

[8] Gkonela, C., Chorianopoulos, K., 2012. VideoSkip: event detection in social web videos with an implicit user heuristic. Multimedia Tools and Applications , 1–14.

[9] Haykin, S., 1998. Neural Networks: A Comprehensive Foundation. 2nd ed., Prentice Hall PTR, Upper Saddle River, NJ, USA.

[10] Karydis, I., Avlonitis, M., Chorianopoulos, K., Sioutas, S., 2013. Identifying important segments in videos: A collective intelligence approach. International Journal on Artificial Intelligence Tools .

[11] Karydis, I., Avlonitis, M., Sioutas, S., 2012. Collective intelligence in video user's activity, in: Artificial Intelligence Applications and Innovations (2), pp. 490–499.

[12] SocialSkip, 2014. User-based video analytics. https://code.google.com/p/socialskip/.

[13] Spiridonidou, A., Karydis, I., Avlonitis, M., 2013. Mimicking real users' interactions on web videos through a controlled experiment, in: Mining Humanistic Data Workshop, pp. 60–69.

[14] Syeda-Mahmood, T., Ponceleon, D., 2001. Learning video browsing behavior and its application in the generation of video previews, in: Proc. of ACM International Conference on Multimedia, pp. 119–128.

[15] Vanmarcke, E., 1983. Random fields, analysis and synthesis. MIT Press.

[16] YouTube, 2014a. Share your videos with friends, family, and the world. http://www.youtube.com/.

[17] YouTube, 2014b. Statistics. http://www.youtube.com/t/press_statistics.

[18] Yu, B., Ma, W.Y., Nahrstedt, K., Zhang, H.J., 2003. Video summarization based on user log enhanced link analysis, in: Proc. of ACM International Conference on Multimedia, pp. 382–391.