Mouse Activity as an Indicator of Interestingness in Video

Gloria Zen
DISI, University of Trento
gloria.zen@unitn.it

Paloma de Juan and Yale Song
Yahoo Research
{pdjuan,yalesong}@yahoo-inc.com

Alejandro Jaimes
Ai Cure
aj27@caa.columbia.dot.edu

ABSTRACT

Automatic detection of interesting moments in video has many real-world applications such as video summarization and efficient online video browsing. In this paper, we present a lightweight and scalable solution to this problem based on user mouse activity while watching video. Unlike previous approaches that analyze video content to infer the interestingness, we leverage the implicit user feedback obtained from thousands of online video watching sessions. This makes our method computationally efficient and scalable to billions of videos. Most importantly, our approach can handle a variety of video genres because we make no assumption on what constitutes interestingness: we let the crowd tell us through their mouse activity. By analyzing 106,212 user sessions collected from a popular online video website, we show that mouse activity is highly indicative of interestingness, and that our approach has competitive performance to several state-of-the-art methods.

1. INTRODUCTION

Over the last five years, the average time US adults spend per day watching online video has grown from 21 minutes to 76 minutes, with an average annual growth rate of 38.4%.[1] Prediction estimates reveal that, by 2019, online video will be responsible for four-fifths of global Internet traffic.[2] Now more than ever, it is crucial to have an automatic method able to analyze video content and produce various types of metadata, such as time-stamped tags and highlights. This information can help improve user experience by generating high quality video previews [26] and efficient online video browsing [6], and bring more revenue to the service providers by, e.g., inserting relevant ads within a video [25]

In this paper, we investigate the usefulness of mouse activity in detecting interesting moments in videos by analyzing user data collected from thousands of online video watching sessions. Our work is motivated by recent works that analyze user feedback while watching online video [8, 22, 29, 34, 39] and those that analyze mouse cursor activities [1, 7, 15, 16, 23, 31]. Previous research has mainly focused on using mouse activity to estimate gaze density while browsing websites [7, 23], performing web search [15, 16, 31], and reading news articles [1]. We contribute to this line of research by showing that this implicit form of user feedback is highly indicative of interesting moments in video. To the best of our knowledge, our work is the first to study the relationship between mouse activity and user attention while watching online video.

There are four main advantages of using mouse activity for our task: It is efficient, scalable, non-intrusive, and generalizable. Most existing approaches to detecting interesting moments from video are content-based, extracting expensive audio and visual features

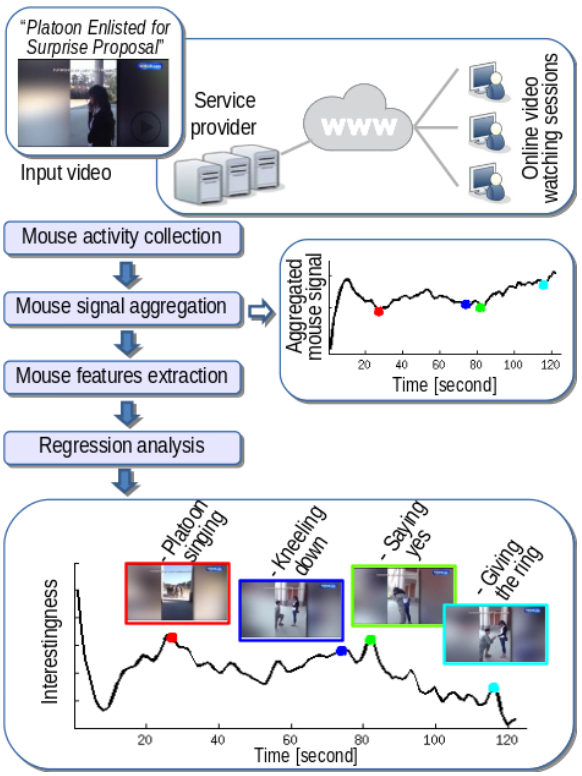

Figure 1: An overview of our approach. Shown here are the main phases of our approach for a video from our dataset, "*Platoon enlisted for surprise proposal*", for which we collected mouse activity from over 2,600 online video watching sessions. We derive several mouse features from the aggregated mouse signal, then perform regression analysis to estimate the degree of interestingness for each second.

[1]eMarketer, April 2015: "US Adults Spend 5.5 Hours with Video Content Each Day."
[2]Reelseo, June 2015: "By 2019, 80% of the world's internet traffic will be video."

ICMR'16, June 06-09, 2016, New York, NY, USA

DOI: http://dx.doi.org/10.1145/2911996.2912005

and using advanced machine learning techniques to estimate the level of interestingness [11, 13, 17, 36, 37]. Compared to those approaches, our method is computationally efficient because the mouse activity signal is low-dimensional and fast to compute. Also, our method analyzes mouse activity per video, and thus is scalable to billions of online videos by processing them in parallel. Further, our signal collection is passive and non-intrusive (mouse events are collected using JavaScript), so users can remain in their natural setting and their behavior is not affected by the data collection process, a crucial factor in designing online experiments. Finally, and most importantly, unlike most previous works [22, 24, 30, 37, 38], our approach can be applied regardless of a video genre, since we make no assumptions on what makes a video interesting.

To study the relationship between mouse activity and interestingness at a large scale, we collected mouse events that happened within the web browser from 106,212 video watching sessions on a popular online video website. We also collected a total of 1,800 responses from Amazon Mechanical Turk (AMT) to generate the timeline annotation of interestingness levels for these videos. We then extracted several features from the mouse activity signals and performed a regression analysis with the degree of interestingness obtained via crowdsourcing as the target variable. We evaluated the effectiveness of our approach by automatically detecting interesting moments in video and comparing this to the output of some of the recent content-based approaches [11, 13]. Our results show that mouse activity is more effective at predicting interesting moments than methods based on visual features, suggesting that mouse activity is indicative of interestingness in video.

Figure 1 shows the close relationship between the video's storyline and the aggregated mouse activities of over 2,600 users. The video is a news story about a man proposing to his girlfriend with help from his platoon. The mouse signal (middle plot) represents the level of mouse activity aggregated over all users. We see the inverse relationship between the mouse signal (y-axis) and the highlights (shown in the bottom plot with keyframes). For instance, a high level of mouse movement is observed after about 10 seconds as the news anchor introduces the story, which we interpret as a drop of interest. The mouse movement level sharply decreases and negatively peaks at around the 30th second when the platoon starts singing a song to the woman, representing one of the most interesting moments in this video. Later, around the 80th second, the movement decreases again as the man kneels down to propose to his girlfriend – another interesting moment in this video. Towards the end of the video, around the 118th second, another peak of interest can be observed when the man gives an engagement ring to his girlfriend. As we will show later in this paper (Table 2), the aggregated mouse activity shows higher correlation to interestingness than various other content-based features.

The main contributions of our work are: (i) we investigate the relationship between mouse activity and user attention in video, which, to the best of our knowledge, has never been explored before; (ii) we propose an efficient framework to predict interestingness in video based on user mouse activity during online video watching sessions.

2. RELATED WORK

Our work is related to three lines of research. The first includes works that use mouse movement while interacting with web pages in order to infer user intention or to evaluate the interestingness of web pages. The second includes works that focus on determining interestingness in video, either based on user responses (e.g., physiological signals, click interactions with video player), or on video content analysis (e.g., visual cues). The third includes works that can benefit from estimating video timeline interestingness for building applications for users (e.g., more positive video browsing user experience) or service providers (e.g., higher efficiency on video transmission).

Mouse Movement Analysis. Mouse movement has been used as proxy for gaze location in the context of interaction with web pages. Rodden et al. [31] found that the distance between cursor and gaze position is larger along the x-axis than the y-axis of screen, and is generally shorter when the cursor is placed over the search results. They also observed four types of mouse cursor behaviors: neglecting the cursor while reading, using the cursor as a reading aid to follow text (either horizontally or vertically), and using the cursor to mark interesting results. Huang et al [15] proposed a technique to predict the gaze position from the cursor position with high accuracy. Hauger et al. [14] found a higher correlation between gaze and cursor positions when the cursor is in motion. Arapakis et al. [1] predict the outcome of online news reading experience by analyzing mouse activity while reading the article. Huang et al. [16] use mouse movement to analyze the user behavior when interacting with search engine result pages, and show that this signal is more informative than one obtained by collecting user click interactions.

User Response Analysis. The works in this category collect implicit user feedback, such as Twitter feeds [32, 38] and video player interaction patterns (e.g., play, pause, skip, etc.) [2, 20, 21, 27, 40]. Similar to our approach, this type of user responses can be collected passively and remotely, in a scalable and efficient manner. Promising results have been reported on various video categories, such as live events [32, 38], online lectures [2, 20, 21] and sport games [27, 40], suggesting a strong correlation between the *replay* action and important video segments. However, some of the user responses are mostly found in very specific video categories (e.g., the replay action is less likely to occur in news videos compared to online lectures or sport games), making this kind of approaches not particularly suited to generic video categories.

Recent works explored crowdsourcing user responses while watching video, developing efficient annotation tools [22, 33, 39] or collecting physiological signals (e.g., brain activities, facial expressions, eye blinking, head motion, heart rate) [3, 4, 8, 29]. This line of work has no category-specific assumption and thus can generalize well to many types of videos. Bao et al. [3] infer within-content ratings by collecting reactions from users watching movies on tablet devices. Similarly, Shirazi et al. [34] annotate a video timeline based on excitement information acquired using an EEG headset, and show that this information correlates well with important scenes of a video. Wu et al. [39] propose a framework for generating video abstracts based on explicit user feedback. These approaches, however, do not translate well into the online video realm, due to the cost of recruiting annotators, and the difficulty of obtaining user consent (e.g., facial expressions) or setting up expensive laboratory devices (e.g., brain signal sensors).

Compared to the two sets of user responses described above, mouse signal has the advantages of scalability (i.e., it can be collected passively from a large sets of users) and of being generalizable to a variety of video categories.

Video Content Analysis. The works in this category focus on analyzing content of video, such as audio and visual features, and use advanced machine learning techniques to detect interesting moments in video. The main challenges include defining what is interesting in a video without an a priori knowledge about its main topic. Potapov et al. [30] use category-specific prior information based on a predefined semantic taxonomy, e.g. weighing higher on scenes that contain "blowing on a candle" for the videos in a category "birthday party." Other works crawl images and videos

from the web to learn a prior information on important moments in videos on a specific topic [37, 24, 9]. While promising results have been reported, the need for building a complete taxonomy makes these approaches difficult to scale to online videos. Song et al. [36] determine the visual importance of shots based on their visual similarity to topical images, crawled from the web using query terms derived from the video title. This approach has the advantage of being category-independent, assuming that the title describes the main subject of the video. However, its performance is largely influenced by the descriptiveness of a title.

A different thread of work specifically focuses on determining a general measure of interestingness, based on emotional and psychological studies. Grabner et al. [11] focus on predicting interestingness in webcam image sequences; Gygli et al. [13] use the predicted interestingness for generating summaries of user generated videos and egocentric videos. Other work focus on detecting interestingness of images [12, 35] and videos [18].

Improving User Experience. Recently, a growing interest has been shown on methods for improving user experience by providing easier access to video content [22] or by allowing more efficient video browsing [5, 6, 19]. Kamvar et al. [19] propose a method for browsing multimedia collection of videos on mobile devices. The instructional videos are segmented into informational chunks (manually or automatically via a video segmentation algorithm) and a list of relevance segments is produced through a keyword query. Chen et al. consider seeking (i.e., jumping to a new position of the video) as a special form of browsing [5, 6]. They propose a smart video streaming approach to avoid user early departure due to excessive buffering time and to reduce the bandwidth wastage (i.e., downloaded video content that is never watched) [6]. Their analysis leads to a user behavior model in which a user transitions through a random number of short views before a longer view. A method for detecting the interesting segments of a video would be highly beneficial for these approaches. For example, providing the most relevant information first may allow the user to reduce the time needed to find the desired information, thus improving the experience and reducing the risk of early departure.

3. FROM MOUSE ACTIVITY TO INTERESTINGNESS

Our method collects mouse activities from online video watching sessions and extracts several features from the aggregated mouse signals. We then perform a regression analysis with the level of interestingness as the target variable. This section describes our method in detail. In the next section we present the collected dataset.

3.1 Hypothesis on Mouse Activity

We formulate the following hypothesis on user mouse activity: When a user is focused on the video content, no mouse movement will occur (but the inverse may not hold). On the other hand, when a user starts to lose focus on the video, mouse movement is the first indication of the loss of attention. Note that having no mouse movement does not guarantee that a user is focused on the video (e.g., in case when the user has switched to another tab, or minimized the window). We consider those edge cases when collecting mouse activity signals, which we describe below.

3.2 Mouse Activity Collection

To record mouse activities of a user within a video page, we set up a script that runs in the background on the browser and records the position of the mouse every 100 ms throughout the video session. A session starts when the user enters the video page, either by typing an URL in the address bar, or by clicking on a link to another page. A session can also start if the video is part of a playlist. In that case, the session starts automatically when the previous video finishes and the new one is automatically loaded. The session ends when the user leaves the video page either by closing the tab or the browser, or navigating away from the page (i.e., loading a different page in the same tab).

The mouse coordinates are buffered and sent back to the server, along with the playback status (play or pause) of the video player at each instant. This data is processed as follows:

- If a video is playing, the current mouse position is compared to the previous one: (i) if the position has changed, the event is registered as “mouse movement;" (ii) otherwise, the event is registered as “no mouse movement."
- If the video is not playing at that instant, the data is discarded.

It is important to note that mouse movement can only be tracked within the limits of the browser. Any mouse movement occurring outside the page where the video is displayed is not detected. As a consequence, if the video is playing and the mouse is moving beyond the limits of the browser, this can be misinterpreted as “no mouse movement."

To handle this case, we process two additional browser events: “blur" and “focus." The former occurs when the page loses focus, i.e., if the user switches to a different tab or window (by clicking on it or using keyword shortcuts), or minimizes the window; the latter occurs when the page gains focus again. When the page where the video is playing loses focus, the playback status is not affected (i.e., video keeps playing), but the mouse movement can not be tracked anymore.

As we do not know whether the video player is still visible to the user or not, we mark the whole period between the “blur" and the “focus" events as “mouse movement." If the user does not return to the page within the video session (i.e., there is no “focus" event), the “blur" event is interpreted as the end of the session. For each user session, our procedure leads to the collection of a binary vector, where each bin encodes the mouse activity observed during a time interval of 100 ms. We set the value to 1 for the event “mouse movement” and to 0 for the event “no mouse movement.”

The collected binary signal is process through three steps: First, in order to make it easier to align and aggregate mouse signals for all users watching the same video, we rescale the signal to a 1 second temporal resolution. This length is assumed to be small enough to keep a fine granularity that allows us to clearly separate the highlights from the subsequent attention decays. Second, we binarize the signal in order to filter out small mouse movements. Finally, we smooth the obtained binary signal using a gaussian filter, in order to take into account short temporal delays and eventual sparsity of data. To this aim, we use a temporal window of length 3 seconds.

3.3 Mouse Signal Aggregation

We define the *user mouse signal* as $x_i \in \mathbb{R}^L$ associated with the i-th user session for a specific video, where L is the video length (expressed in seconds). This signal is obtained by -1 padding the smoothed mouse signal previously obtained, thus setting to -1 the user mouse signal for the time intervals ℓ during which no mouse activity was recorded (e.g. the user session was terminated).

By combining mouse signals x_i over the N users who watched the video, we obtain a matrix $\mathscr{M} \in \mathbb{R}^{N \times L}$. Based on this matrix, the *aggregated mouse signal* $\mu \in \mathbb{R}^L$ is computed as:

$$\mu = \left\{ \frac{1}{N^\ell} \sum_{j=1}^{N^\ell} y_j^\ell \right\}, \quad \ell = 1, \cdots, L \tag{1}$$

where $N^\ell = |y^\ell|$ and $y^\ell = \{\mathscr{M}(i,\ell) \neq -1\}\, i = 1 \ldots N$. In practice, for each time interval ℓ the collected mouse signal μ indicates the percentage of mouse events observed, computed over the current active user sessions N^ℓ.

3.4 Mouse Feature Extraction

Intuitively, frames with the same value of the aggregated mouse signal μ can indicate different levels of interestingness, depending on whether they are located on a descending or an ascending slope, or on slopes with different degrees of steepness. We define a set of mouse features designed to capture the shape and the temporal variations of the aggregated mouse signal.

Global features. Some videos may *globally* have more active mouse movements than other videos; one level of mouse activity in a video should not be treated as equal in another video. In order to obtain a global representation of mouse movement signals that does not depend on the type of a video, we convert all instances of the aggregated mouse signal from each video into their percentile ranks within that video.

Local features. In order to capture the local pattern of a signal (e.g., ascending, flat, or descending slope), we concatenate the values within a local temporal window of 3 seconds of length.

Local variation features. We consider the variation of μ in its local range, by computing the distance between the values of μ within $[\ell-1, \ell+2]$. We consider first-order and second-order variations. For the latter, we take as input the first-order local variation instead of the original signal μ.

Spectral Features. The mouse signal temporal variations can have different time periods. To take into account the duration of local variation of μ, we compute the Discrete Cosine Transform (DCT) [28] for the local range of a 3 second temporal window.

3.5 Interestingness Estimation via Regression Analysis

The final interestingness score is computed by combining scores obtained by the mouse features presented in the previous section.

In order to determine the importance of each feature for predicting interestingness, we follow the same procedure presented by Grabner et al [11], who focus on predicting interestingness based on a combination of various cues extracted from visual content. As in [11], features are firstly normalized to their mean and variance, and then mapped into the interval [0,1] using a sigmoid function $s = \frac{1}{(1+exp(-az+b))}$.

Secondly, the final interestingness score is obtained as a linear combination of the considered features. This is formalized as: $\hat{s} = w^T \hat{s}$, where $\hat{s} = [s^{(f1)}; ...; s^{(fN)}]$ and $\{s^{(f1)}, ..., s^{(fN)}\}$ is the set of considered features. The optimal values of the parameters of the sigmoid function (a,b) and the weights w are estimated through a regression analysis using least square minimization and leave-one-video-out cross validation.

4. MOUSE ACTIVITY DATASET

4.1 Collection Setting

We deployed a *bucket* in an online video website, Yahoo Screen[3], where the mouse movement tracking script is integrated into the regular data logging process. Having a bucket is a typical setup in large companies; it guarantees that only a small percentage of users, chosen at random, will be exposed to the script. The script collects user mouse movements non-intrusively. We recorded mouse activity only within the limits of the browser and for the specific video website; all the recorded events were completely anonymized. We recorded the mouse activities for all videos watched in the bucket during a collection period of few weeks.

The final Mouse Activity Dataset consists of a total of 45 videos and 106,212 user sessions. The videos collected belong to a variety of categories such as News, Interviews, Comedy, Slideshow, User Generated, etc., and cover a variety of topics such as "food," "nature," "fashion" or "science." Also, the videos have the following properties: (1) the video is in English, and understanding its content does not require domain-specific or prior knowledge; (2) the video is no longer than 5 minutes; (3) the mouse activity is collected from more than 100 users and the associated inter-user agreement is positive (Cronbach's alpha $\alpha_u > 0$). The first two criteria garantee that the collection of human judgment can be fairly conducted with the same AMT task developed for our study. The third criterion ensures that the collected mouse signal for the video is significant enough, considering the instrisic noisy nature of the individual mouse signal. A discussion regarding the effect of N and α_u on the validity of mouse signal as an indicator of interestingness in video is reported in the result section. We provide details of our dataset (i.e., title, length, category, number of users from whom we collected mouse activity, Cronbach's alpha values, etc.) in the supplementary material online.[4]

4.2 Inter-User Agreement

We measure the agreement of the collected mouse signals for each video using Cronbach's alpha coefficient α. Our hypothesis is that when this agreement is sufficiently high, the aggregated mouse signal is not simply noise and we can rely on this signal to discover interesting moments in video.

Since the number of users N^ℓ who have not stopped playing the video decreases as the video advances, we compute α_ℓ at each instant of the video ℓ by considering only the user sessions that are not finished by time ℓ. In detail, we compute α_ℓ over the matrix $\mathscr{M}_\ell = \mathscr{M}(i,\bar{\ell})$, where $\bar{\ell} = \{1, ..., \ell\}$ and i is the index set of the user sessions N^ℓ. The final inter-user agreement is computed as $\alpha_u = \text{median}(\alpha_2, \cdots, \alpha_L)$.

4.3 Annotation Settings

We set up an annotation task that consists of watching a video and providing an explicit response regarding the interestingness of the given content. The feedback was provided through two buttons – *thumb up* and *thumb down* – next to the video player. The participants were asked to click either button whenever they thought something interesting or uninteresting was being shown. Every time the worker clicked one of the two buttons, an event *up* or *down* was recorded. The video was played from the beginning to the end without interruption. The workers were not allowed to pause or skip the video, but they were allowed to replay it from the beginning at any time, if they needed so.

By default, each annotation m was initialized as a zero vector of length L. Each element covered a time interval of 1 second. We set $m_\ell = 1$ or $m_\ell = -1$ every time an *interesting* or *uninteresting* event was recorded within the time interval ℓ.

For quality control, we asked the participants to answer a questionnaire after having finished the video. In order to ensure they had genuinely watched the video while performing the task, we asked two objective fact-check questions about the content of video. Also, in order to ensure that the participants used their best criteria to per-

[3]As of January 2016, Yahoo Screen has been sunset. All video content is available on digital magazine properties instead.

[4]https://disi.unitn.it/∼zen/data/mouse-track-dataset.pdf

form the task, we asked them to provide at least three keywords or keyphrases explaining their choices. We collected a total of 1,800 annotations, 40 per video. The workers were free to choose how many videos to annotate (between 1 and 45), but were not allowed to annotate the same video more than once.

4.4 Inter-rater Agreement

For the sake of coherece with the pre-processing applied to mouse signals, we smoothed the collected annotations using Kernel Regression and a temporal window of 3 seconds of length, and we computed Cronbach's alpha $\hat{\alpha}_r$ to assess the inter-rater agreement.

We obtained an average value of $\bar{\alpha}_r$=0.53. As there is no objective way to evaluate the correctness of a subjective annotation (i.e., it is difficult to tell whether an annotation is an outlier because it was done at random or because of a different opinion), we decided to discard those that greatly differed from the rest of the annotations based on a measure of internal consistency. For that, we filtered out up to 10 annotations for each video, by greedily leaving out the annotations that were lowering the inter-user agreement. The resulting average Cronbach's alpha of the reduced set is $\bar{\alpha}_r = 0.71$.

In the following phases of this work, we considered both the original and the reduced sets of annotations, and we did not observe significant changes in the final results. Results obtained with these two sets are reported and discussed in the Section 5.3.

4.5 Interestingness Scores

We denote as *real-valued interestingness score* the signal obtained by averaging the annotations collected via AMT. In order to evaluate the efficiency of our method in detecting interesting moments in video, we binarize the real-valued scores into "interesting" and "not interesting" and denote the obtained signal as *binarized interestingness score*. We do this by first normalizing the real-valued score distribution to have zero mean and one standard deviation, then threshold the normalized scores using value one.

5. EVALUATION

We conduct two sets of experiments to assess the validity of using mouse signal as an indicator of interestingness in video.

In the first experiment, we select videos based on different numbers of user sessions N and Cronbach's alpha α_u. This analysis aims to investigate the conditions in which the collected mouse signal can be considered a trustworthy measure of interestingness in video. In particular, we validate α as an indicator of the trustworthiness of the mouse signal, and we empirically find the optimal value of N required for collecting a trustworthy mouse signal.

In the second experiment, we validate our approach only on the set of videos for which the collected mouse signal is considered reliable. To this aim, we select videos with $\alpha > 0.7$, which is indicated in literature as the desirable minimum value of α for considering acceptable the inter-user agreement [10].

5.1 Methodology

We compare our method to recent works on visual interestingness [11] and video summarization [13, 36]. We chose the three methods as our baseline because they share certain similarities to our method – they produce frame-level scores indicating how interesting/important they are, and they do so without making any assumption on the video topic. One difference to our work is that, while our method is purely based on mouse signal, all three baselines are based on visual content of a video. Therefore, our experiments would allow us to measure the effectiveness of mouse signal in comparison to visual signal.

Table 1: Interestingness prediction results, measured in AUC at varying degree of the binarization threshold δ.

	$\delta = 5$	$\delta = 10$	$\delta = 15$	$\delta = 20$
Gygli et al. [13]	0.5141	0.5301	0.5176	0.5153
Grabner et al. [11]	0.5246	0.5377	0.5478	0.5459
Song et al. [36]	0.5741	0.5643	0.5520	0.5421
Our method	**0.6868**	**0.6728**	**0.6577**	**0.6443**

Grabner et al [11] and Gygli et al [13] define a set of visual features for predicting visual interestingness based on attention models and psychological studies. We compute visual interestingness scores by combining visual features as described in [11] and [13]. Song et al. [36] measure the importance of each video frame based on their similarity to the title of that video. We measure the frame-level interestingness by following the description in [36]. We process videos at 6 FPS in all our experiments.

We acknowledge that the two baseline methods [13, 36] are optimized to tackle video summarization, which is slightly different from our problem. Our goal is to measure the level of interestingness in each moment of a video, while the goal of video summarization is to create a concise summary of a video given a time budget (e.g., 15% of the original video length).

Below, we show results on two experiments: (i) performance in predicting the *binarized* interestingness; (ii) correlation between mouse signals and the *real-valued* interestingness scores.

Interestingness Prediction. The performance is computed in terms of Area Under the ROC Curve (AUC), where the *binarized interestingness score* is used as the ground truth for interestingness. The binarized signals is obtained by selecting the δ% top interesting moments from the *real-valued interestingness score*. We show results obtained by varying δ between 5 and 30 with a unary step. We report the results in Section 5.2.

Correlation with Interestingness. We assess the correlation between the reference *real-valued interestingness score* and the predicted interestingness scores by computing the Pearson correlation between the two signals. For the sake of comparison, both signals are normalized to zero mean and one standard deviation. We report the results in Section 5.3.

5.2 Interestingness Prediction

Table 1 and Figure 2 (a) show the performances on interestingness prediction obtained with different methods, for the videos with $\alpha > 0.7$. The performances are measured in terms of AUC and computed at varying the binarization threshold δ used for generating the *binarized interestingness score*. We observe that our method outperforms related works where the score is computed based on visual content [11, 13, 36].

Figure 2(b,c) show the performances of our method respectively by varying the value of minimum α_u and N used for selecting the video set on which our analysis is performed. Based on this observation, we can assess that both α and N are good indicators for determining the trustability of the mouse signal. In particular, we experimentally observe that the performance obtained on our dataset with $\alpha_u > 0.7$ can be reached on our dataset in the case of $N > 1,000$.

5.3 Correlation with Interestingness

Table 2 shows the correlation of different mouse features and methods with *real-valued interestingness scores*. In general, we observe that the correlations is low across different methods (below 0.3). We speculate that this is indicative of how challenging

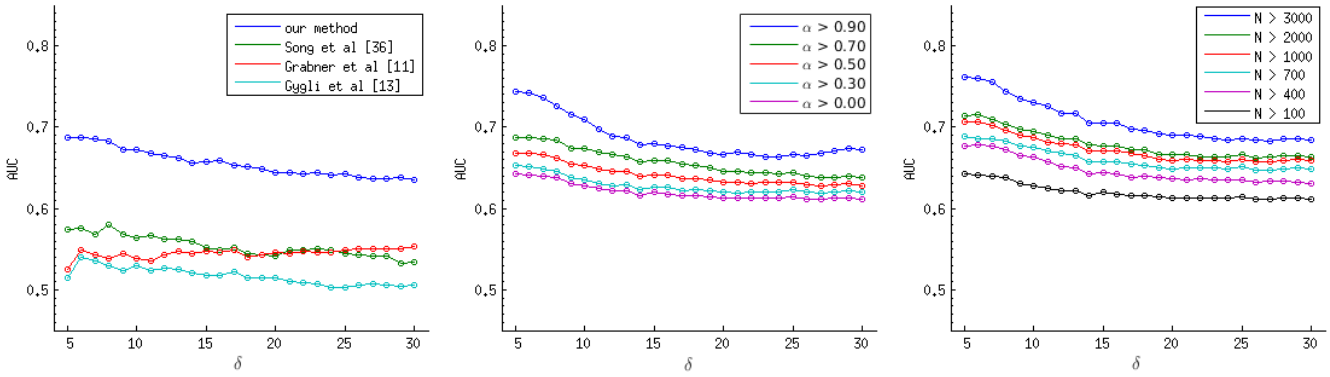

Figure 2: Overall performance on the prediction of interestingness, measured in AUC at varying δ: (a) our method compared with related works and our method when only videos with a minimum value of (b) N and (c) α are considered. The y-axis limits for (a,b,c) are magnified between 0.45 and 0.85 for sake of clarity.

Table 2: Pearson correlation between *real-valued* interestingness scores and prediction scores obtained using: (top) different mouse features and (bottom) different methods. We report results on two versions of our dataset (all videos and $\alpha_u > 0.7$)

	all videos ($\alpha_u > 0$)	$\alpha_u > 0.7$
mouse	0.250	0.283
local	0.259	0.292
local variation (1st order)	0.136	0.158
local variation (2nd order)	0.007	0.098
spectral	0.259	0.289
global	0.259	0.303
Gygli et al. [13]	0.057	0.043
Grabner et al. [11]	0.106	0.133
Song et al. [36]	0.021	0.038
Our method	**0.268**	**0.301**

the problem is in general. Indeed, videos were retrieved without making any assumption on the video format or content.

Still, just by considering mouse features (Table 2(top)), we obtain a higher correlation with respect to the recent methods based on visual features (Tab. 2(bottom)). From Table 2, we can also observe the benefit of combining mouse features, leading to a higher correlation. Also, if we compare the results obtained by considering all videos from the dataset and only those with $\alpha > 0.7$, reported respectively in the left and right column of Table 2, we observe that the increase in r for the mouse-based signal is more significant with respect to the variation of r obtained for the other methods.

Table 3 reports the two sets of Pearson correlation values computed based on the *real-valued interestingness score* obtained by using (i) the original set of AMT annotations and (ii) the reduced set, which is derived from original AMT annotations by eliminating the outliers. It can be noted that the difference of the results between the two sets is not significant, around 0.01.

5.4 Qualitative Analysis

Figure 4 shows the mouse movement and the interestingness score collected for the video "*Huge Sea Turtle Crashes Wedding*". Figure 3 shows the thumbnail frames uniformly extracted every 2 s from this video. By comparing Figure 3 and Figure 4, it is possible to observe that the *mouse signal* and the *real-valued interestingness score* exhibit a similar trend, with an increasing interestingness in the transitions from segments where the news anchor is shown, to those depicting the subject of the news (e.g., bride and turtle being shown around the 15th, 25th, 59th and 89th seconds), and vice versa (e.g., news anchor appearing around the 3rd, 21st and 83rd seconds). Interestingly, a value mismatch between the two signal can be observed when a screenshot of the Reddit Post is displayed (i.e., around the 33rd and the 47th second): while a low interesting score has been assigned by the annotators, a peak in the mouse signal can be observed. A similar situation of mismatch was observed in other videos when a screenshot or a map was shown.

Table 3: Pearson correlation between *real-valued* interestingness scores and prediction scores. The real-valued interestingness score is obtained using (left) the original set and (right) the reduced set of AMT annotations.

	original AMT set	reduced AMT set
Gygli et al. [13]	0.056	0.057
Grabner et al. [11]	0.103	0.106
Song et al. [36]	0.012	0.021
Our method	0.236	0.250

Also, it is interesting to notice that the parts of the video for which we observe a drop of interestingness are those where the newcaster is speaking and where the pictures of the celebrities are shown. This may be apparently be in contrast with attention based model theories that indicate the presence of faces or celebrities as a salient cues. Feedback provided by the AMT workers explain this choice, as shown in Table 4. This table reports some of the keywords and keyphrases used by the AMT workers to describe the moments indicated as interesting or uninteresting for one of the videos in the dataset. Some of the negative keyphrases include "*the anchor*", "*host trying to be funny*" or "*lame jokes with the celebrity photos*". This example shows us that, while this kind of subtleties may be difficult to catch with content-based method analysis, directly sensing from the users what is interesting helps disambiguate this kind of situations.

6. CONCLUSION

In this paper, we provide a novel framework for detecting interesting moments in video based on mouse activity signals. Collecting mouse activity signals is computationally efficient, non-

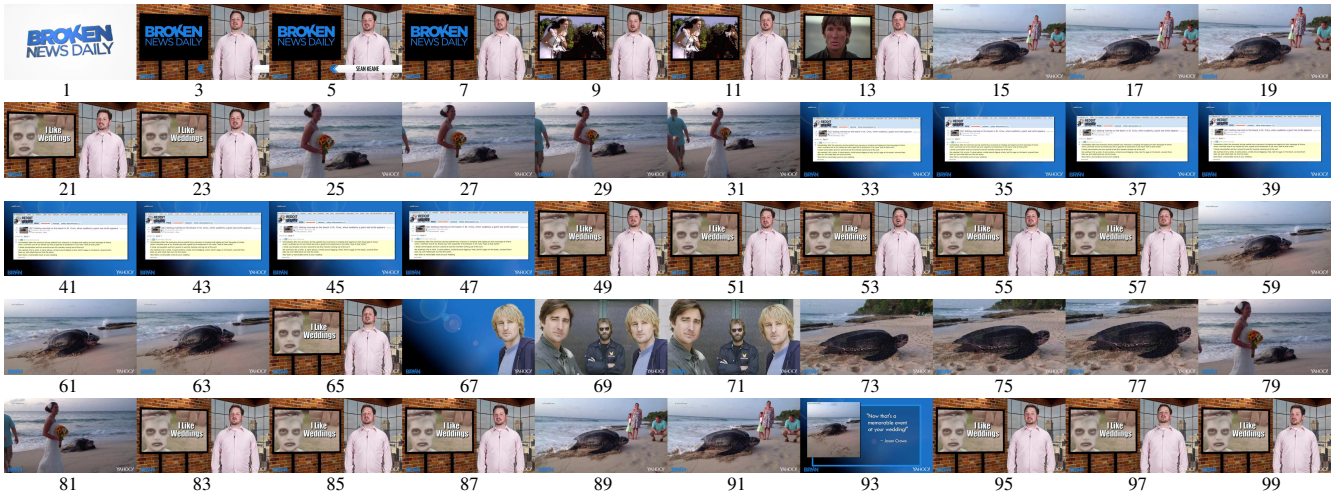

Figure 3: Thumbnail frames uniformly extracted every 2 seconds from the video "*Huge Sea Turtle Crashes Wedding*". The frames depicting the subjects of the news (e.g. bride and turtle between the 25th and the 31st second and between the 73th and 81th second) are indicated as the most interesting by the AMT workers (see Figure 4, in blue) and also correspond to peaks of the mouse signal (see Figure 4, in black).

Table 4: Sample keywords describing interesting and uninteresting moments for the video "*Huge Sea Turtle Crashes Wedding*", as indicated by the AMT workers.

	Motivations
`Interesting moments`	photos of the turtle, wedding, bride, pics of bride and groom, cute animal, unique, surprising, funny, memorable, unexpected, turtle with bride and groom, Jason's comment on Reddit, wedding ceremony on the beach
`Not interesting moments`	host trying to be funny, Reddit post photograph, references not to turtle and wedding, cheesy, the anchor, repetitive, unnecessary, lame jokes with celebrity photos, the newscater's jokes were terrible, too much text, zombie

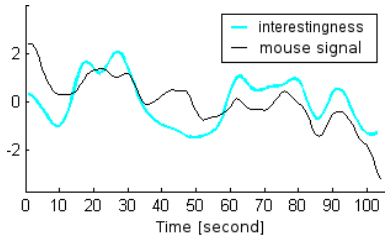

Figure 4: Scores of interestingness based on human judgments (in blue) and mouse signal (in black) for the video "*Huge Sea Turtle Crashes Wedding*". For sake of comparison, both signals are normalized to zero mean and one standard deviation.

invasive, and scalable to thousand or even millions of users. By analyzing over 100,000 online video watching sessions, we showed that mouse movement can be considered a predictive cue for the interestingness in video, yielding a performance that is significantly higher than the current state-of-the-art computer vision techniques.

As an empirical value, we found that 0.7 is the minimal desirable value of inter-user agreement over the mouse signals collected from users, for which the aggregated mouse signal can be considered significant. We also shown that the number of users watching the video is a good indicator of the mouse signal trustability for predicting interestingness in video. We empirically found that similar performances to the case of $\alpha_u > 0.7$ can be achieved when at least 1,000 users sessions are collected for each video. Considering the growing number of users who watch videos online (so called "cord cutters"), we believe this is a reasonable number for a popular video service provider. This is shown by recent statistics, indicating that the average number of views per video for different categories varies between 2,300 and 9,000.[5]

Over the last few years, the time people spend on watching online video has increased dramatically. Now more than ever, it is important to invest in developing intelligent systems able to predict how users will react to the content. Our work contributes to this line of research, and especially on detecting interesting moments in video. Our method leverages the implicit form of user feedback obtained from the mouse activity, collected from thousands of online users, in order to detect interestingness in video. We showed that our method is computationally efficient and scalable to billions of videos. We also show that our approach can handle a variety of video genres, as we make no assumption on what constitutes interestingness.

Our work has implications on video summarization, which can help improve user experience by generating high quality video previews and efficient video buffering, and also on video recommendation, which would bring more revenue to the service providers.

In the future, we plan to perform a deeper analysis on mouse activities when the user attention is lost (e.g., the user starts interacting with the video player controls, scrolling down to browse related videos, etc.). Also, more analysis is needed to understand what types of features (e.g., visual cues, sound, objects depicted in the video, plot development) drive attention, and in which specific context (e.g., reading text from the screen requires attention, but it probably does not constitute a highlight).

[5]http://www.reelseo.com/average-youtube-views

7. REFERENCES

[1] I. Arapakis, M. Lalmas, and G. Valkanas. Understanding within-content engagement through pattern analysis of mouse gestures. In *CIKM*, 2014.

[2] M. Avlonitis and K. Chorianopoulos. Video pulses: User-based modeling of interesting video segments. *Advances in Multimedia*, 2014.

[3] X. Bao, S. Fan, A. Varshavsky, K. Li, and R. Roy Choudhury. Your reactions suggest you liked the movie: Automatic content rating via reaction sensing. In *UbiComp*, 2013.

[4] P. R. Chakraborty, L. Zhang, D. Tjondronegoro, and V. Chandran. Using viewer's facial expression and heart rate for sports video highlights detection. In *ACM ICMR*, 2015.

[5] L. Chen, Y. Zhou, and D. M. Chiu. A study of user behavior in online vod services. *Computer Communications*, 46, 2014.

[6] L. Chen, Y. Zhou, and D. M. Chiu. Smart streaming for online video services. *IEEE Transactions on Multimedia*, 17(4), 2015.

[7] M. C. Chen, J. R. Anderson, and M. H. Sohn. What can a mouse cursor tell us more?: correlation of eye/mouse movements on web browsing. In *CHI*, 2001.

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

[9] W.-S. Chu, Y. Song, and A. Jaimes. Video co-summarization: Video summarization by visual co-occurrence. In *CVPR*, 2015.

[10] G. Goldstein and M. Hersen. Handbook of psychological assessment. 2000.

[11] H. Grabner, F. Nater, M. Druey, and L. Van Gool. Visual interestingness in image sequences. In *ACM Multimedia*, 2013.

[12] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Van Gool. The interestingness of images. In *ICCV*, 2013.

[13] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool. Creating summaries from user videos. In *ECCV*. 2014.

[14] D. Hauger, A. Paramythis, and S. Weibelzahl. Using browser interaction data to determine page reading behavior. In *UMAP*. 2011.

[15] J. Huang, R. White, and G. Buscher. User see, user point: gaze and cursor alignment in web search. In *CHI*, 2012.

[16] J. Huang, R. W. White, and S. Dumais. No clicks, no problem: using cursor movements to understand and improve search. In *CHI*, 2011.

[17] Y. Ito, K. M. Kitani, J. A. Bagnell, and M. Hebert. Detecting interesting events using unsupervised density ratio estimation. In *ECCV Workshops and Demonstrations*.

[18] Y.-G. Jiang, Y. Wang, R. Feng, X. Xue, Y. Zheng, and H. Yang. Understanding and predicting interestingness of videos. In *AAAI*, 2013.

[19] M. Kamvar, P. Chiu, L. Wilcox, S. Casi, and S. Lertsithichai. Minimedia surfer: browsing video segments on small displays. In *CHI*, 2004.

[20] I. Karydis, M. Avlonitis, K. Chorianopoulos, and S. Sioutas. Identifying important segments in videos: A collective intelligence approach. *International Journal on Artificial Intelligence Tools*, 23(02), 2014.

[21] J. Kim, P. J. Guo, D. T. Seaton, P. Mitros, K. Z. Gajos, and R. C. Miller. Understanding in-video dropouts and interaction peaks in online lecture videos. In *L@S*, 2014.

[22] J. Kim, P. T. Nguyen, S. Weir, P. J. Guo, R. C. Miller, and K. Z. Gajos. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *CHI*, 2014.

[23] D. Lagun and E. Agichtein. Inferring searcher attention by jointly modeling user interactions and content salience. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015.

[24] M. Mazloom, A. Habibian, D. Liu, C. G. Snoek, and S.-F. Chang. Encoding concept prototypes for video event detection and summarization. In *ACM ICMR*, 2015.

[25] T. Mei, X.-S. Hua, L. Yang, and S. Li. Videosense: towards effective online video advertising. In *ACM Multimedia*, 2007.

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

[27] D. R. Olsen and B. Moon. Video summarization based on user interaction. In *EuroITV*, 2011.

[28] A. V. Oppenheim, R. W. Schafer, J. R. Buck, et al. *Discrete-time signal processing*, volume 2. Prentice-hall Englewood Cliffs, 1989.

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

[30] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid. Category-specific video summarization. In *ECCV*, 2014.

[31] K. Rodden, X. Fu, A. Aula, and I. Spiro. Eye-mouse coordination patterns on web search results pages. In *CHI*, 2008.

[32] D. A. Shamma, L. Kennedy, and E. F. Churchill. Tweet the debates: understanding community annotation of uncollected sources. In *ACM SIGMM Workshop on Social media*, 2009.

[33] D. A. Shamma, R. Shaw, P. L. Shafton, and Y. Liu. Watch what I watch: using community activity to understand content. In *ACM Workshop on Multimedia Information Retrieval*, 2007.

[34] A. S. Shirazi, M. Funk, F. Pfleiderer, H. Glück, and A. Schmidt. Mediabrain: Annotating videos based on brain-computer interaction. In *Mensch & Computer*, 2012.

[35] M. Soleymani. The quest for visual interest. In *ACM Multimedia*, 2015.

[36] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes. TVSum: Summarizing web videos using titles. In *CVPR*, 2015.

[37] M. Sun, A. Farhadi, and S. Seitz. Ranking domain-specific highlights by analyzing edited videos. In *ECCV*, 2014.

[38] A. Tang and S. Boring. #EpicPlay: crowd-sourcing sports video highlights. In *CHI*, 2012.

[39] S.-Y. Wu, R. Thawonmas, and K.-T. Chen. Video summarization via crowdsourcing. In *CHI*, 2011.

[40] B. Yu, W.-Y. Ma, K. Nahrstedt, and H.-J. Zhang. Video summarization based on user log enhanced link analysis. In *ACM Multimedia*, 2003.