




SportLight: statistically principled crowdsourcing method for sports highlight selection

Jiwon Jung¹ · Seyong Ha² · Won Son³ · Joonhwan Lee⁴ · Joong-Ho Won¹ 

Received: 24 September 2020 / Accepted: 3 May 2021
© Korean Statistical Society 2021

Abstract

Sports highlight selection has traditionally required expert opinions and manual labor of video editors. To automate this laborious task, crowdsourcing viewers' live comments has recently emerged as a promising tool, which can remove the burden of extracting semantic information by computer vision. However, popular crowdsourcing methods based on peak-finding are sensitive to noise and may produce deviant highlights from the expert choice. To increase the accuracy of automated selection of sports highlight, we introduce a statistically sound crowdsourcing method, SportLight. In this work, we take a statistical approach that combines multiple hypothesis testing and ℓ_1 -trend filtering (fused lasso), supported by a computationally inexpensive algorithm. By analyzing 29 baseball games played in the 2016 and 2017 seasons, we demonstrate that our approach properly reduces the risk of false alarm and generates the results closer to expert-chosen highlights than that of the peak-finding method.

Keywords Sports highlighting · Video summarization · Crowdsourcing · Fused lasso · Human factors

1 Introduction

Large sports events such as the postseason in a professional baseball league are exciting experiences that attract a large number of fans. The excitement of these live events is condensed into a series of short highlight reels to be propagated to an even larger number of people. *Generating* highlight reels, on the other hand, is not an exciting task. Video editors have to spend countless hours watching, selecting, and assembling clips of impactful plays. This highly repetitive and labor-intensive task also requires a high level of knowledge in the target sport, making the whole process costly.

✉ Joong-Ho Won
wonj@stats.snu.ac.kr

Extended author information available on the last page of the article

As the sports industry is becoming more and more capitalized, the need for automating the costly task of highlight selection is ever increasing. One strand of research in this direction is computer vision-based approaches, the ultimate goal of which is to make the machine to interpret the semantics of the games by analyzing the scenes (D’Orazio & Leo, 2010). With recent advances in artificial intelligence, these efforts appear to bear fruit. In 2017, IBM debuted its Watson supercomputer at the U.S. Open tennis tournament for generating and posting a highlight reel on Facebook within two minutes after each match. However, building such an expert system is still very costly. Beside the supercomputing requirement, those systems need to be trained by a human with a huge amount of data. It is reported that Watson was “taught” using the footages from the Masters golf and the Wimbledon tennis tournaments. Cues of excitement – scenes containing such as crowd noise and players’ roar – have to be curated to make “examples” for machine learning.

Another, less explored, avenue of research is to take advantage of the power of crowdsourcing, or outsourcing certain tasks from computers to a collection of human workers where human labor is more efficient and reliable than that of computers (Von Ahn & Dabbish, 2004; Von Ahn et al., 2008). Due to the burst of the Internet and mobile devices, people spend a surge of time online. Consequently, more research on utilizing the collective input from the online crowd has sprung up in various domains (Bernstein et al. 2010; Marcus et al., 2011; Ha et al., 2013b; Tang & Boring, 2012; Von Ahn & Dabbish, 2004; Von Ahn et al., 2008). A common observation in these domains is that when an interesting event occurs, the rate of online activities increases in almost real-time. In sports highlight selection, thus, peaks in data streams from popular social network platforms associated with live broadcasts are detected in efforts to extract exciting moments (Hannon et al., 2011; Tang & Boring, 2012). However, online streams are inherently noisy; hence methods relying on local peak detection may be unstable and result in selections not comparable to professionally-edited highlight reels (Tang & Boring, 2012).

In this work, we propose SPORTLIGHT, a statistically-based method for sports highlight selection. Our approach is based on the idea that every moment of a game can be binary-classified into either highlight (1) or non-highlight (0) state. These state variables are unobservable but can be estimated by statistical hypothesis testing. There are as many hypotheses as the moments, which are highly correlated with the narrative of the game; in this respect, we focus on designing a method that successfully suppresses the possibility of false alarms resulting from multiple hypotheses testing with a strong correlation.

We evaluate our method by using a dataset of 29 postseason games of the Korean Baseball Organization (KBO) league in 2016 and 2017, drawing on live streaming video broadcasts from Naver, a dominant web service provider in South Korea, which attracts 30 million daily visitors out of its 50 million population. Baseball is one of the most popular professional sports in Korea, and roughly 400 thousand people per day watch the games online on Naver, which posts (traditionally edited) highlight reels on site. We found that the highlights selected by our method match with the expert-curated ones with precision around 0.7 for moderate number of highlights, which demonstrates that this cost efficient crowdsourcing can be *comparable* to the traditional method. This result is in contrast to the prior work (Tang

& Boring, 2012), in which crowdsourced highlights have features distinct from the reels from news corporations. This suggests that SPORTLIGHT successfully captures the semantics of the game, or plays that were ultimately meaningful to the outcome of the game.

The contribution of this paper is two-fold:

1. we demonstrate that statistically sound crowdsourcing techniques can produce sports video summarization close to that generated by experts.
2. we provide a computationally inexpensive, simple-to-implement algorithm for this purpose.

The remainder of the paper discusses related work and background of sports highlight selection, shares our system design and algorithm, and evaluates our method in both quantitative and qualitative perspective.

2 Background

Sports highlight selection tasks can be categorized into two types of systems based on its source of data: audiovisual feature-based systems and crowdsourcing-based systems. In this part, we introduce early works on each type of sports highlight selection, followed by a brief description of the KBO league as a target sports for crowdsourcing system.

2.1 Audiovisual feature-based sports highlight selection

In sports video analysis, prior studies have used the audiovisual features of video formats to extract the scenes of interest automatically (Shih, 2018). These works utilize contextual cues appearing in an image frame relevant to the play, such as the player's position on a field (Assfalg et al., 2003) or a recognized scoreboard in a game (Babaguchi et al., 2004; Metulini, 2017). Such visual cues serve as a useful source for generating sports highlights; they are incorporated and employed to identify the important moments of a game via computer vision techniques. Alongside a vision-based approach, research on audio event detection utilizes common audio cues such as the announcer's excited speech and ball-bat impact sound to identify a direct indicator of highlights (Xiong et al., 2003, 2004). The system proposed in Bettadapura et al. (2016) measures the level of excitement and displays its measured degree of interest along a time axis. Furthermore, machine learning or statistical approaches have applied audiovisual features for construction of classifiers, such as support vector machines (SVMs) (Liu et al., 2009; Rui et al., 2000) and hidden Markov model (HMMs) (Assfalg et al., 2002, 2003), to identify and estimate the state of the moment.

In conjunction with the aforementioned research, IBM has developed a highlight-suggestion system with an artificial intelligence tool, Watson. Featuring both critical and entertaining moment, Watson summarizes a 4-h long tennis match into 3 min

(Kapetanakis, 2018). It is reported that Watson collects audiovisual sources from the game and selectively translates the features into quantified scales between 0 and 1 (Thompson, 2010). The degree of excitement at each moment is measured based on Watson's scoring system, which assigns relative scores to multiple categories of indicators, crowd cheering and player gestures (Kapetanakis, 2018).

2.2 Crowdsourcing-based sports highlight selection

Although audiovisual sources have proven their power to extract a play's semantic information (Chao et al., 2005; Qian et al., 2012), crowdsourcing viewers' live comments has emerged as a promising source for automating this laborious and burdensome task. In general, audiovisual frameworks are labor intensive and computationally expensive; they not only require the collection of numerous pre-specified scenes but also entail computational cost of dealing with audiovisual cues. Based on the contributions of a large audience, however, we can overcome the burden of an image or audio processing (Quinn & Bederson, 2011). Another advantage of crowdsourcing method is that sports highlights are closely in line with viewers' emotional experiences (Tang & Boring, 2012). If a player makes a theatrical catch that saves his team from a loss, baseball fans are more likely to regard that moment as a highlight than usual hits of the game.

2.2.1 Crowdsourcing systems

A series of studies has demonstrated that viewers' live comments on popular social media (Marcus et al., 2011; Tang & Boring, 2012), online broadcasting platforms (Ha et al. 2013a), or live-streaming platforms (Chu & Chou, 2017) have compelling potential for sports highlight detection. These crowdsourced highlights present the extent to which fans were excited or upset at a particular moment.

The *TwitInfo* system discovers prominent peaks in the rate of incoming tweets of Twitter (Marcus et al., 2011); the system extracts main scenes of soccer events by providing a user interface which tracks the real-time occurrence of the targeted words, such as "football" or "premier league." The *#EPICPLAY* system is also built on this approach, which captures the occurrence of exciting events in American football games in a live broadcast by separating the incoming stream of microblogs into home and away records (Tang & Boring, 2012). There are further studies on media interaction focusing on the applications in digital media indexing by collecting posts from online forums (Ha et al., 2013a, b); in particular, (Ha et al., 2013a) generate and analyze sports highlights from baseball games. A more recent study combines viewer data with a traditional audiovisual feature-based approach (Chu & Chou, 2017).

2.2.2 The TwinInfo peak-finding algorithm

The key component of the crowdsourcing methods based on *TwitInfo* (Ha et al. 2013a; Marcus et al., 2011; Tang & Boring, 2012) is a so-called peak-finding

algorithm, which detects abnormal increases in the number of postings by scanning a large stream of data. While there are many deterministic peak finding algorithms based on brute-force search or divide-and-conquer (Cormen et al., 2009), this algorithm is statistical and based on the following outlier detection criterion. Given the observed stream of count data C_1, C_2, \dots, C_n , when a new count C_{n+1} is observed, the algorithm classifies the point as a peak if

$$\frac{C_{n+1} - \hat{\mu}_n}{\hat{\sigma}_n} > \tau, \quad (1)$$

for some $\tau > 0$, where $\hat{\mu}_n$ and $\hat{\sigma}_n$ are the estimated historical mean and deviation. Past data are exponentially weighted in estimation:

$$\hat{\sigma}_{n+1} = \alpha |C_{n+1} - \hat{\mu}_n| + (1 - \alpha) \hat{\sigma}_n \quad (2)$$

$$\hat{\mu}_{n+1} = \alpha C_{n+1} + (1 - \alpha) \hat{\mu}_n \quad (3)$$

An abrupt change in the temporal signal is identified with a peak when a newly entered datum exceedingly deviates from the historical mean. Once a peak is detected, the algorithm continues hill-climbing until it returns to a value that is less than or equal to the level at which it started.

Broadly speaking, the binary classification used in the peak finding algorithm resembles statistical hypothesis testing; τ serves as the critical value of the rejection region from a statistical perspective. For normally distributed data, the above formula becomes Student's t-test if sample standard deviation was used instead of the exponentially weighted absolute deviation. In effect, size of τ controls the number of peaks detected. A practical choice of the quantities τ and α is proposed in Marcus et al. (2011), as well as in Tang and Boring (2012), Ha et al. (2013a).

In fact, this algorithm is inspired by the algorithm for computing retransmission time-out (RTO) in the transmission control protocol (TCP) (Paxson et al., 2011). The mean absolute deviation is used for two reasons; it provides more conservative measurements, and is also easier to compute (Jacobson, 1988). This choice is reasonable and appropriate in TCP's context because RTO is defined in order to determine an outlier packet that takes unusually long to transmit. A conservative choice is necessary to ensure network stability.

2.3 The KBO league

Baseball is a turn-based sport in which for each of the nine innings, two teams alternate their turns for offense and defense. The offending team is allowed three outs while batting the ball pitched by the defending team. The goal is to advance bases by hitting the ball to a safe place in the field. When a batter returns to the home base, a run is scored. The game is in play when the pitcher throws the ball until the ball is caught by one of the nine defending players.

In Korea, a country of 50 million population, the KBO League is the most popular professional sport which attracts 6 million fans to the stadiums annually. To serve

this huge fanbase, every game is broadcast live nationally, in various means including the streaming videos from Naver, which alone possesses 400,000 viewers per day on average. To make the watching experience social, Naver provides a platform that encourages the viewers to post live comments on the game, specific to each of the home and away teams. In this respect, this platform resembles the *#EPICPLAY* (Tang & Boring, 2012).

For the purpose of selecting crowdsourced highlight, baseball has a number of useful features. First, its pitch-by-pitch nature clearly defines the beginning and end of a play (as opposed to soccer or basketball, a sport with continuous actions). Second, a game is in-play for only a small fraction of the its entire duration [it is often quoted that, “a baseball fan will see 17 min and 58 s of action over the course of a three-hour game” (Moyer, 2013)]. Third, there is a large number of fans who are willing to actively comment on social networking platforms during live broadcasts of games. Note that these features are very similar to those of American football, which is analyzed by Tang and Boring (2012).

3 The SportLight system design and algorithm

We divide our discussion of *SPORTLIGHT* into two parts: outline of the system components and description for how our selection algorithm is modeled.

3.1 Components of the SportLight system

After retrieving online streamers’ activity on a social broadcasting platform, our system selects a series of sports highlights with a cost-efficient algorithm.

3.1.1 Naver’s social broadcasting platform

The broadcasting platform serviced by Naver is illustrated in panel 1 in Fig. 1. The live video stream of the game in play is shown in the middle window. The scoreboard (top), batter/pitcher information for each of the home (left) and away (right) teams are also shown. Below the game information, there is a comment window where the viewers can join one of the teams to comment on the game. Comments from the home fans appear with the team logo on the left, whereas those from the away fans appear on the right with the user id and the timestamp. Although only a few comments are displayed, the entire comment history is stored in the platform with the game statistics for many years even after the game. This information can be easily retrieved in the javascript object notation (JSON) format.

3.1.2 Highlight selection results

The input to the highlight selection part is the history of the number of viewers who commented on the aforementioned platform during the game. The output of the highlight selection part is a number of timestamp intervals classified as highlights. The

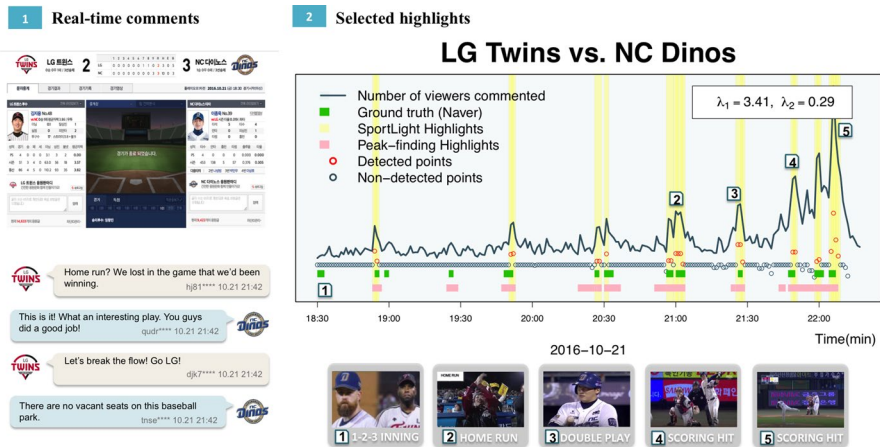


Fig. 1 Outline of the SPORTLIGHT system. (1) Naver's social broadcasting platform; texts in panel 1 are translated from the original scripts in Korean. (2) Highlights generated by SPORTLIGHT (yellow) in comparison with the peak-finding algorithm (pink) and Naver's selection (green) are displayed; tuning parameters are also displayed in panel 2. Examples of selected results are displayed on the bottom

number of highlights can be specified by the user. The minimum length of an interval is one minute. Panel 2 in Fig. 1 illustrates 20 signals of selected highlights (yellow bands) of a real playoff game on October 21, 2016 overlaid on the plot of the frequency of comment postings versus time since game started. The selected intervals contain the events of home run, double play, and scoring hits, all of which are generally considered significant plays. These selected intervals (highlighted in yellow) are shown with timings of the highlight reels curated by Naver (green boxes) and those generated by the peak-finding algorithm (pink boxes); the Naver-curated highlights have an average length of 2 min. Note that an expert-curated highlight in the very beginning is missed due to the small number of initial viewers.

3.2 Highlight selection algorithm

3.2.1 Modeling the number of viewers commented

As described, the rate of online activities, i.e., the number of viewers who post comments during a unit time interval, on the social broadcasting platform is the key quantity of the SPORTLIGHT system; we set the unit time interval as one minute.

In fact, it is common to use the number of comments per unit time interval as the basis for highlight selection algorithm. Instead, we chose the number of viewers who post comments during a unit time interval for the following reasons. For instance, some viewers post meaningless comments like “hahaha” constantly, regardless of the game's progress. In addition, some viewers split a comment into several posts, producing an excess rate. These phenomena stems from that the commenting platform is closer to an instant messaging system than a microblogging service such as Twitter.

Statistically, it is reasonable to model such data as a sequence of Poisson random variables. In other words, at time t , the observed number of viewers who comment, X_t , is assumed to follow the Poisson distribution with mean μ_t :

$$P(X_t = x) = e^{-\mu_t} \frac{\mu_t^x}{x!}, \quad x = 0, 1, 2, \dots \quad (4)$$

The mean number of viewers who comment per unit time μ_t varies with the time index t , in order to reflect the dynamics of the game and fan responses. However, the trajectory of μ_t as a function of t is unspecified and needs to be estimated from the data.

3.2.2 Multiple hypothesis testing

We then conceptually dichotomize the unit time intervals into normal (non-highlight) and highlight states. By nature, normal states will be dominant over the course of the game. In this case, μ_t varies slowly, thus can be estimated from the data in the neighborhood of t . At a highlight state, μ_t may abruptly change from the nearby normal states and is difficult to estimate from the neighborhood. Because our ultimate goal is to classify each time interval t into either highlight or normal state rather than to accurately estimate μ_t , we can proceed with testing the following statistical hypothesis:

$$H_0 : \mu_t = \mu_0(t) \quad \text{vs.} \quad H_1 : \mu_t > \mu_0(t), \quad (5)$$

for each t , where $\mu_0(t)$ is the mean counts at time t , which can be estimated from the neighborhood if H_0 is true. The result of this hypothesis testing is summarized by a p value, which is between 0 and 1. If it is close to 0 (below the significance level), the test is in favor of the alternative hypothesis H_1 , or the highlight state. Otherwise, the test does not reject the null hypothesis H_0 , or the normal state. As we assume a Poisson distribution (4), we conduct the Poisson test for testing hypothesis (5). This testing procedure is in Line 2 of Algorithm 1.

3.2.3 ℓ_1 -trend filtering

Any statistical hypothesis testing involves the risk of false alarm. This risk is inflated if we test multiple number of hypotheses simultaneously. This is the case in SPORT-LIGHT, where we need to test more than 100 hypotheses. Although there are procedures to adjust the inflation (Benjamini & Hochberg, (1995; Bland & Altman, 1995), often these procedures are too conservative and suppress most of true findings. Recently, a method based on ℓ_1 trend filtering has been proposed to adaptively detect multiple change points (Son & Lim, 2019) for serially correlated but slowly varying mean models $z_i = \mu_0^i + \varepsilon_i$ where $\varepsilon_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$, $i = 1, \dots, n$. The ℓ_1 trend filtering is a technique to estimate an (almost) slowly varying trend from a sequence of noisy observations. The slowly varying trend is allowed to change abruptly in sparse locations. In its simplest form, ℓ_1 trend filtering minimizes

$$\frac{1}{2} \sum_{i=1}^n (z_i - \zeta_i)^2 + \lambda \sum_{i=1}^{n-1} |\zeta_i - \zeta_{i+1}|, \quad (6)$$

for variables ζ_1, \dots, ζ_n , where z_1, \dots, z_n are given noisy observations. The first term is the squared error commonly found in least squares estimation, and the second term is the sum of absolute values of the differences of the sequence ζ_1, \dots, ζ_n . It is well known that minimizing squared error together with a sum of absolute values tends to make the summands in the second sum exactly zero (Tibshirani et al., 2005). The positive constant λ is a tuning parameter that controls the degree of the zeroing effect. In the above case, this means that the filtered sequence ζ_1, \dots, ζ_n is piecewise constant with a few jumps. The theory of Son and Lim (2019) states that, for an interval of constant mean values, ζ_i 's are close to the sample mean, and the corresponding sample standard deviation is smaller than that of individual observations z_i 's. Therefore, we can expect a lower false positive rate than individual hypothesis tests. Thus minimizing (6) fits in our statistical model.

In order to adopt the ℓ_1 trend filtering, or more precisely the method of Son and Lim (2019), to adjust for false alarms from multiple hypothesis testing, we first convert the p -values from tests (5) into z values (Line 3 of Algorithm 1) in order to meet the normality assumption.¹ If time i is in the normal state, then the corresponding z value z_i will follow the standard Gaussian distribution with zero mean and unit variance and stay close to neighboring z values. If it is a highlight, then z_i will be far from the neighbor. The filtered z -values ζ_i obtained by minimizing (6) suppress false alarms due to the multiple hypothesis testing while allowing occasional jumps due to highlights. In order to further promote each ζ_i to tend to zero, we add an additional penalty and minimize

$$\frac{1}{2} \sum_{i=1}^n (z_i - \zeta_i)^2 + \lambda_1 \sum_{i=1}^n |\zeta_i| + \lambda_2 \sum_{i=1}^{n-1} |\zeta_i - \zeta_{i+1}|, \quad (7)$$

with a pair of tuning parameters (λ_1, λ_2) . This is stated in Line 10 of Algorithm 1.

3.2.4 Highlight detection

The above discussion suggests that if we plot ζ_i 's, it would look like a step function in which most plateaus are at the zero level. We select a block of a constant positive level as a highlight. This is due to that distinct levels of ζ_i may reflect distinct degree of interest (Line 11 of Algorithm 1). In panel 2 of Fig. 1, those distinct levels constitute 20 highlight intervals. Additionally, selected highlights from three games are displayed in Fig. 2.

¹ Adding the ℓ_1 penalties to the Poisson log-likelihood function may be plausible, but the theory of Son and Lim (2019) only supports normal models, and our empirical experience favors the transformation approach.

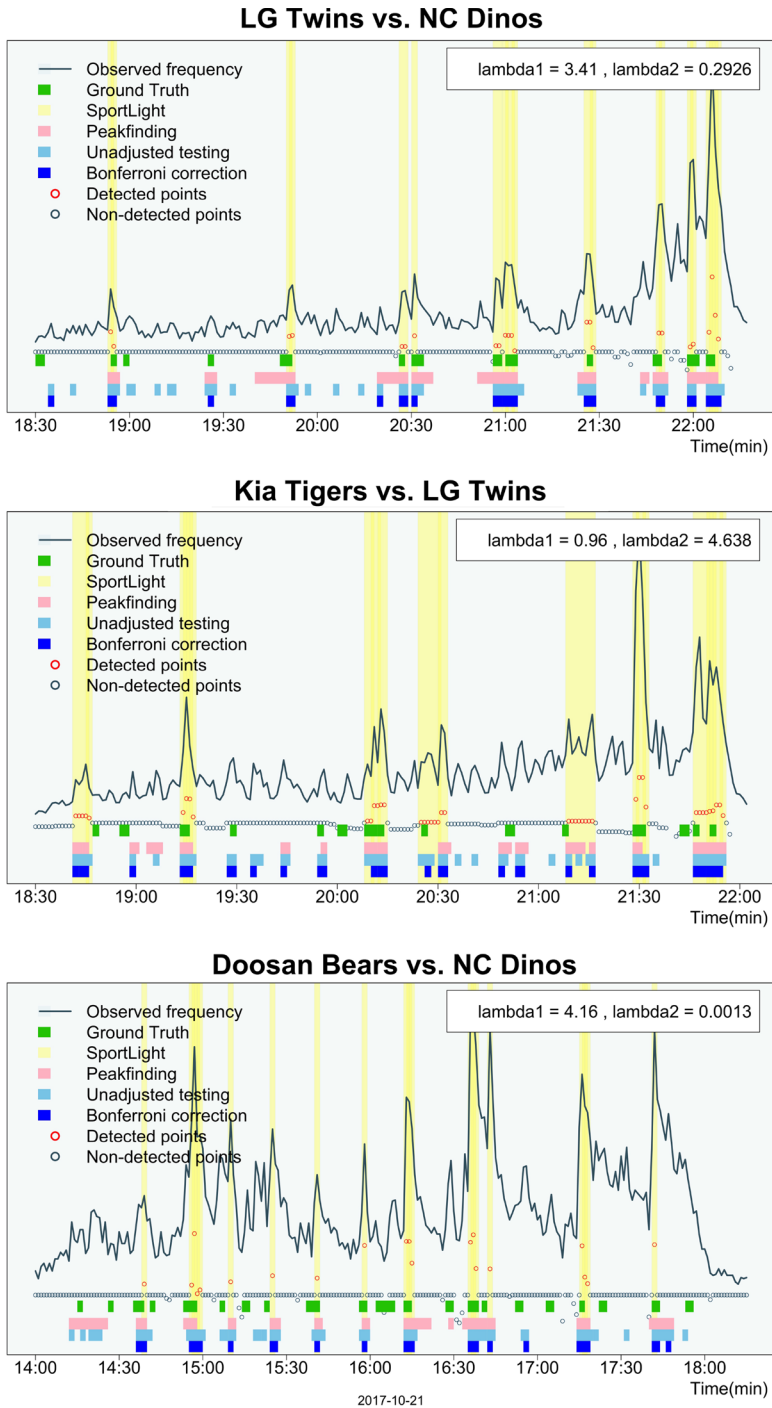


Fig. 2 Highlights retrieved by SPORTLIGHT and other methods for three selected games

3.2.5 Tuning parameter selection

The tuning parameters λ_1 and λ_2 for problem (7) are chosen so that the number of detected highlights match the desired number (Line 8 of Algorithm 1). A large value of λ_1 results in many zeros among ζ_1, \dots, ζ_n , whereas λ_2 controls the size of jumps. There is an efficient algorithm for solving (7): for a fixed value of λ_1 , ζ_i s that minimize (7) for every possible values of λ_2 are computed at once (Hoeffling, 2010). Thus for each λ_1 , we can find λ_2 that gives the desired number of highlights. We then select the λ_1 that exhibits the largest average of positive ζ_i 's per minute (Line 16 of Algorithm 1), which is a reasonable choice in that sports highlights represent the climaxes of the game in which the viewers are most interested.

3.2.6 Algorithm summary

The whole procedure discussed above is summarized in Algorithm 1. We first enter the loop in lines 1–4, which converts retrieved count data into the normal scale by conducting multiple hypothesis testing. (In conducting an individual test (4), the baseline mean $\mu_0(t)$ at time t was estimated in a moving-average fashion, with a window size of 21 centered at t .) In the following loop in lines 5–15, we perform ℓ_1 -trend filtering procedure to obtain the solution path. Line 16 selects a tuned solution out of multiple solution sets which returns desired number of highlights. Then, we store the resulting interval (line 17).

Algorithm 1 SPORTLIGHT highlight selection

Require: (x_1, \dots, x_n) = count data of length n , N = desired number of highlight reels ($0 < N \ll n$),
 lambda1Grid = grid sequence of λ_1

```

1: for  $t = 1, \dots, n$  do
2:    $p_t \leftarrow \text{compute\_pvalue}(x_1, \dots, x_n)$  ▷ §3.2.2
3:    $z_t \leftarrow \text{convert\_to\_zvalue}(p_t)$  ▷ §3.2.2
4: end for
5: for  $i = 1, \dots$  do
6:    $\lambda_1 \leftarrow \text{lambda1Grid}[i]$ 
7:    $m \leftarrow 0$ 
8:   while  $m \neq N$  do
9:     Choose  $\lambda_2$ 
10:     $(\zeta_1, \dots, \zeta_n) \leftarrow \ell_1\text{-trendfilter}(z_1, \dots, z_n; \lambda_1, \lambda_2)$  ▷ §3.2.3
11:     $\text{intervals} \leftarrow \text{detect\_highlight}(\zeta_1, \dots, \zeta_n)$  ▷ §3.2.4
12:     $m \leftarrow \text{size}(\text{intervals})$ 
13:   end while
14:    $\text{solutions}[i] \leftarrow ((\zeta_1, \dots, \zeta_n); \lambda_1, \lambda_2)$ 
15: end for
16:  $(\text{intervals}^*; \lambda_1^*, \lambda_2^*) \leftarrow \text{tuning\_selection}(\text{solutions})$  ▷ §3.2.5
17: return( $\text{intervals}^*$ )

```

4 Evaluation

We now evaluate the performance of the SPORTLIGHT system. In order to test our system, we conduct evaluation in both quantitatively and qualitatively. To better understand the quality of SPORTLIGHT selected highlights, we measure and assess

the relevancy of the generated reels in multiple aspects; we compare our highlights with those provided by Naver, compute the sequence distance, and collect perceptive feedback. The relevancy of produced highlights is measured in *quantitative evaluation*, and perceptive feedback is described in *perceptive evaluation*.

4.1 Quantitative evaluation

For quantitative evaluation, we use the comment data from two KBO League postseasons. This dataset covers 14 and 15 playoffs from the 2016 and 2017 seasons, respectively.

4.1.1 Setup

For the creation of the ground truth, we used the expert-chosen highlight reels provided by Naver. These reels were manually collected and labeled. The relative timing of the reels was manually measured by the first author by carefully comparing them with the video of the whole game. The ticks and time intervals used in the algorithm were also based on these whole-game videos, hence the highlight reels and time intervals of comments were synchronized. There may be a concern on the precision of manual synchronization, but the error should be at most a fraction of second, and this is sufficient for analyzing baseball games. The number of highlight reels found in this platform ranges from 11 to 38 per game. We first conduct conventional receiver operating characteristic (ROC) analysis. Since hypothesis testing is based on each one-minute interval (see Sect. 3.1.2), the false positive rate (FPR) or type I error is the fraction of such intervals claimed wrongly as included in a highlight. Likewise, the true positive rate (TPR) or power is the fraction of those intervals claimed correctly as included in a highlight.

We compare the performance of the proposed SPORTLIGHT algorithm and the peak-finding algorithm, with the baseline of vanilla multiple hypothesis testing (Sect. 3.2.2) with varying nominal significance levels. The tuning parameters of SPORTLIGHT are adjusted select 10, 20, 30, and 40 highlights for each game. These choices are displayed in Fig. 3. The peak-finding algorithm requires to determine the two parameters α and τ . We fix the weight $\alpha = 0.125$ as suggested in Marcus et al. (2011), but use a different critical value τ for each game to generate the same number of highlights as SPORTLIGHT.

4.1.2 Results

The result looks promising: in the ROC curves shown in Fig. 4, that of SPORTLIGHT lies above those of the peak-finding and vanilla multiple testing within the reasonable operating range of FPR between 10 and 30%. On average, 17 manually labeled highlights are serviced per game by Naver. Thus choosing number of highlights between 10 and 20 appears to be desirable for practice. For

these numbers of highlights, SPORTLIGHT shows a clear advantage over the peak-finding algorithm in terms of smaller FPR. The average scores are presented in Table 1; the baseline multiple testing results is based on the typical nominal significance level of 5%, and its Bonferroni correction is adjusted for this level by the length of each game.

Additionally, most of the detected signals successfully match the highlights provided by the experts in some of the selected games shown in Fig. 2; about 9 out of 10 selected reels are correctly identified. To view the performance score on each game, see Table 2 for detail.

Given the observation that the initial (expert-chosen) highlights are almost always missed due to the small number of initial viewers, the relatively small TPRs are understandable. Also, TPRs can be easily misleading in case one calls a long interval a highlight. While the peak-finding method yields higher TPR than SPORTLIGHT, this is because TPR can be overestimated if excessively lengthy intervals are selected. In fact, the Naver-chosen highlights have an average length of 2 min. However, the peak-finding algorithm generates 5.3-min long reels when retrieving 10 highlights. On the other hand, SPORTLIGHT highlights are 2.9-min long on average. The box plot in Fig. 5 directly shows the difference in average length of the selected intervals between two methods. Figure 2 also visualizes the highlight intervals selected by other methods than SPORTLIGHT; tendency to select longer intervals can be easily spotted.

4.2 Perceptive evaluation

Our next step is to gather qualitative data about the feedback on evaluation of the selected reels, especially for those identified as false alarms.

4.2.1 Setup

We collected responses from five baseball fans, who were all familiar with the baseball rules. We asked them to watch the video clips generated from the SPORTLIGHT system, and obtained their feedback to understand the characteristics of the chosen highlights. Our major focus lied in the participants' opinions on the 13 false-positive clips from the five selected games to understand the property of the crowdsourcing method. These reels were selected based on the SPORTLIGHT generation for 20 highlights. Each false positive is labeled with alphabets and numbers; five games are classified as A–E, and 13 false alarms are numbered clip by clip.

The participants were asked to classify the given clips into the following three categories;

1. This clip contains play-relevant highlights, such as scoring hits, nice defence, and strike-outs.
2. This clip has play-irrelevant highlights, which contains an intriguing part of the game, such as emotional reactions of the players and the spectators.
3. This clip does not contain any interesting scenes.

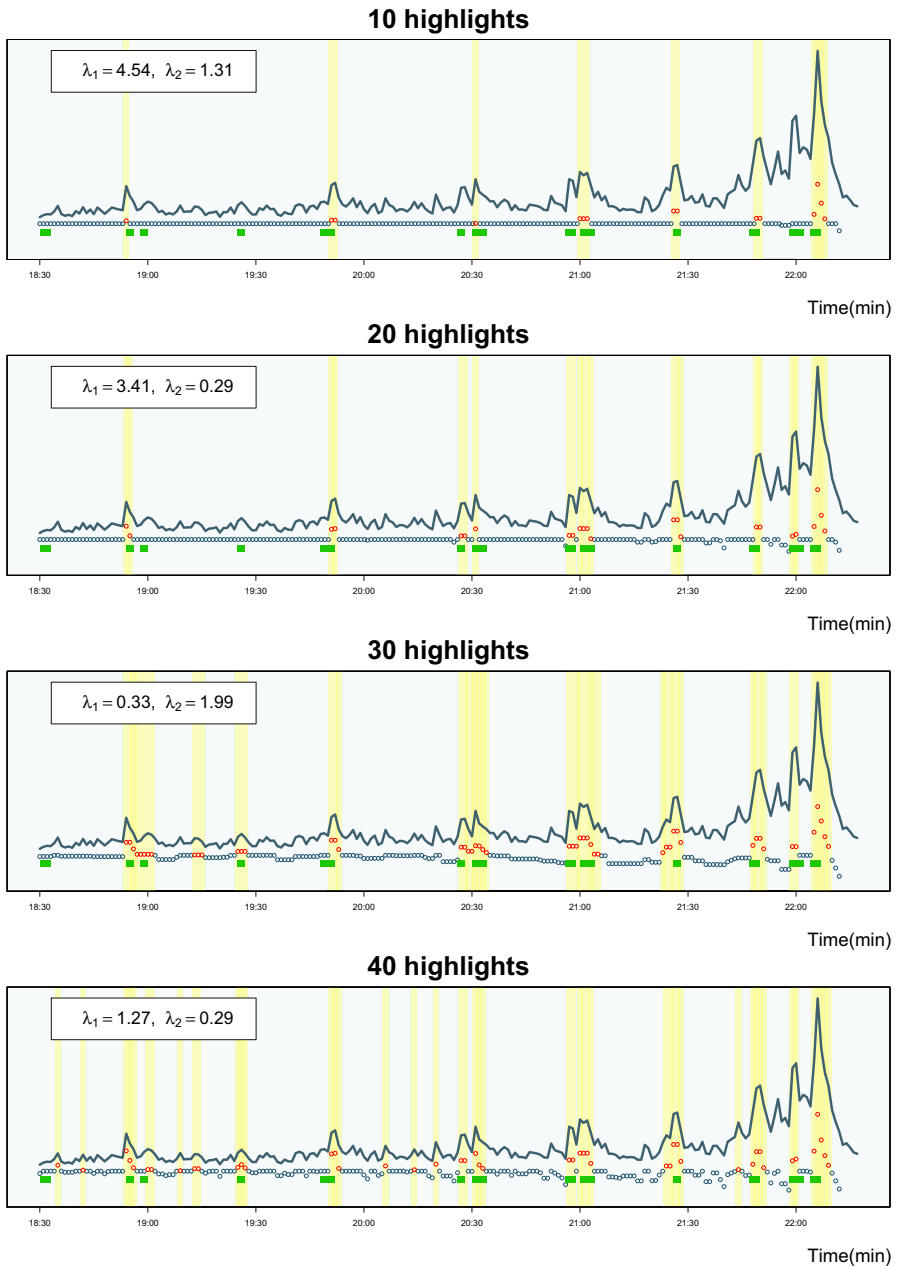


Fig. 3 Increasing number of highlights retrieved by SPORTLIGHT for a selected game

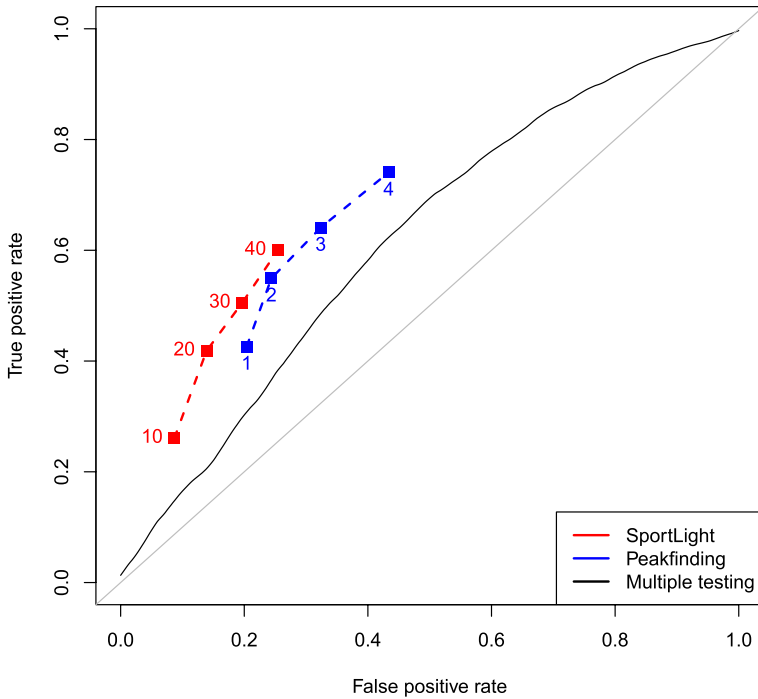


Fig. 4 Receiver operating characteristic of SportLight and other methods

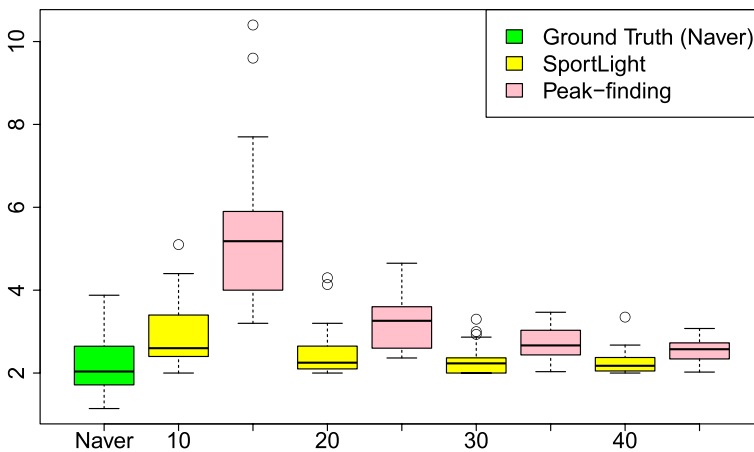
Table 1 Average performance of the 29 post-season games (% , rounded) per number of declared highlights ranging from 10 to 40

Metric	Algorithm	Number of highlights			
		10	20	30	40
FPR	SportLight	8.690	13.90	19.65	25.42
	Peak-finding	20.39	24.38	32.37	43.37
	Multiple testing	31.13			
	Bonferroni	15.71			
TPR	SportLight	26.08	41.84	50.49	60.01
	Peak-finding	42.62	54.94	64.10	74.06
	Multiple testing	44.38			
	Bonferroni	20.90			

They were asked to choose at least one category since each clip possibly contains more than one event. Therefore, the participants voted for Type 1 or Type 2 if they thought the given part of the clip of the reel was interesting or worthy of view, or voted for Type 3 if they did not find anything special or intriguing in the clip.

Table 2 Performance of three selected games (% , rounded) per number of declared highlights ranging from 10 to 40

Game (YYYY-MM-DD)	Metric	Number of highlights			
		10	20	30	40
Kia Tigers vs. LG Twins (2016-10-11)	FPR	8.750	17.50	17.50	25.63
	TPR	45.28	54.72	58.49	75.47
LG Twins vs. NC Dinos (2016-10-21)	FPR	15.38	25.64	29.06	32.48
	TPR	26.19	47.62	47.62	61.90
Doosan Bears vs. NC Dinos (2017-10-21)	FPR	2.959	7.101	12.43	23.67
	TPR	27.59	36.78	40.23	55.17

**Fig. 5** Box plot of average highlight-length in minute

They were also encouraged to give reasons for their choice on the corresponding scene. After gathering participants' opinions, we counted the number of categorized responses per clip. For each category, the number of votes ranges from 0 to 5 in the integer scale per clip based on the number of responses. In addition to the perceptive categorization, we also asked the participants to give an overall open-ended feedback on the selected highlights.

4.2.2 Results

The number of votes is reported in Table 3; out of 71 votes in total, 40 votes were given to the first category (Type 1, 56%), and 15 votes to the second category (Type 2, 21%), and the rest were considered to the third category (Type 3, 23%) or non-highlights.

Interestingly, a substantial portion (77%) of the votes were given to the first two categories (especially for Type 1 highlights), although these types of clips are not

Table 3 The number of votes on false positive clips for perceptive evaluation

False positive clip #	Type 1	Type 2	Type 3
A1	4	0	2
A2	4	1	0
A3	2	3	0
B1	0	1	4
C1	4	2	1
C2	1	0	4
C3	2	3	0
C4	5	0	0
D1	4	0	2
E1	3	1	1
E2	5	0	1
E3	4	1	1
E4	2	3	0

usually preferred by experts. These scenes include “highlights” either overlooked or regarded less important by experts: a failed bunt play (A1), a defense failing to tag the runner out (C4), or a defense for three-pitch strikeout (E2). The number of votes in Type 2 Highlights also shows a gap between the experts’ judgment and participants’ preference on sports highlights. These clips include the emotional reactions of the audience and renowned baseball managers, such as the players and fans celebrating the victory of their supporting team (A3, E4), or a close-up of a player’s face who hit a home run (C3).

Another noteworthy observation here is that none of these false positive clips belongs to one category except for one clip (C4), out of 13. This indicates that the participants had different opinions on how they categorize and interpret the scenes of interest.

5 Discussion

In this section, we discuss the implication of our statistical model and its practical applicability with qualitative feedback from our evaluation participants. In participants’ subjective comments on SPORTLIGHT’s usability, their recommendations for improvements are also reported to identify the key aspects of SPORTLIGHT. We also discuss possible extensions of SPORTLIGHT.

5.1 Call for statistical principles

The stream of real-time comments of the audience on sports fluctuates significantly as the game progress. Highlight events and the corresponding postings are abrupt, rare, and only last for a short time. Consequently, identifying the underlying state of the game from the fluctuating stream of comments is not an easy task. A naive

approach would lead to an excessive number of false alarms and unacceptable lengths of declared highlights. Given the time-varying property of real-time comments, a sound statistical analysis based on sound principles is required. In SPORTLIGHT, multiple hypotheses testing is used instead of naively filtering the comments signal. Filtering is applied after the statistical tests, reflecting the time-varying nature of the underlying states. This helps suppressing the noise. Another novel feature is to apply the (zeroth-order) ℓ_1 trend filtering. This promotes the filtering outcome to be constant in most regions except for the locations of abrupt changes. As a result, the nature of sports events is correctly encoded in the statistical model.

5.2 A “benefit” of false alarms

SPORTLIGHT is not 100% immune to false alarms. However, those a few false positives provide an interesting feature to this statistically crowdsourced method. Although many expert-curated highlights tend to choose scenes that are highly relevant to scoring events, baseball fans may have different opinions on what highlights are. From our perceptive evaluations of the false positive clips, we noticed that baseball fans may consider highlights not only as scoring-related events but also as emotionally impressive ones. For example, all the participants chose clip C4 as a highlight, and one of the participants gave the following description:

“Though the player could not tag the runner out, that defense was outstanding.” (P1)

Such highlights are hard to be curated because experts mainly care about events relevant to scoring. However, SPORTLIGHT is able to generate them since the system takes viewers’ responses into account. Also, we found that baseball fans express mixed opinions on whether and how they perceive a particular event as appealing depending on personal preferences. For instance, our participants provided different reasons for their choice of a particular highlight (C1):

“The gloomy face of the manager watching his team’s poor performance is quite impressive.” (P1)

“ I think this excellent pitching is one of the highlights because it used the strike zone cleverly to strike out the batter and close the inning.” (P2)

This feature justifies our crowdsourcing method as a plausible alternative to expert-curated highlights since the generated highlights are able to cover a wide range of baseball fans.

That crowdsourced highlight reels “are more tied to the drama and emotion of the game as experienced by fans” is reported in *#EPICPLAY* by Tang and Boring (2012). However, it is also reported that the set of reels does not closely match that of nightly sportscasts. To the contrary, SPORTLIGHT provides a reasonable trade-off between reproducing expert-curated highlights and finding emotional moments, by using a statistically principled approach.

5.3 Extensions

Real-time generation While we have analyzed the archive of the entire postseason games for two years, SPORTLIGHT can generate real-time highlight reels from alive broadcasts. The viewer comments data can be updated on-line. Thanks to the path-following algorithm (Hoefling 2010) for Algorithm 1, the desired analysis can be conducted very quickly. Thus by refreshing the algorithm at regular intervals to include fresh comments, important live events can be efficiently tracked.

Team-specific highlights SPORTLIGHT can be immediately extended to generate team-specific highlight reels, by running two instances of Algorithm 1 for both teams' count sequences. Since the set of populations are partitioned in this case and the algorithm is data-adaptive, the union of these highlights would be different from the selected highlights from the combined data. The resulting sets of highlights are likely to be more specific for the fans of each team.

Other sports SPORTLIGHT can be applied to other sports than baseball. Similarity between baseball and American football is discussed in Sect. 2.3. Naver provides essentially the same platform as baseball for the English Premier League. In fact soccer was one of the major event streams studied by Marcus et al. (2011); replacing the TCP-based peak-finding algorithm with Algorithm 1, a more statistically robust analysis of tweets is expected in the *TwitInfo* system.

Live streaming Live streaming platforms such as Twitch or YouTube Live also have a similar structure to Naver's social broadcasting platform. Streamers broadcast varying content to viewers in real-time for up to several hours. Viewers and the streamers can interact with each other via the chat rooms supported by the platform. The live streams are often curated and archived on video-sharing platform (e.g. YouTube) in order to engage with possible future viewers. Considering this apparent similarity, we may expect that streamers, most of whom running a one-man business, can benefit from SPORTLIGHT in generating highlights of their live streams.

6 Limitation

Despite the promises, there is still a room for further research regarding practical applicability.

Suggesting the need for further improvement of SPORTLIGHT's quality of chosen reels, SPORTLIGHT does not completely eliminate the need for manual intervention. Some selected results contain redundant parts of a play such as replayed events and advertisements, which indicates that removal of the unnecessary parts is required. One of the participants' comment states that redundant parts are selected without any special event between pitches are observed, pointing out usability issues regarding the length of SPORTLIGHT highlights. Additionally, some are reported as incomplete highlight in that highlight events are partly captured.

Detailed evaluation on different sets of parameters may help discover the best model choices for sports highlight selection. An appropriate choice of parameters such as tuning quantities and the size of unit interval is required to obtain more

qualified highlights. Further experiments on these choices can help optimize the algorithmic choices.

In addition, our model does not provide textual description of what a particular highlight stands for. Labeling is not available for selected highlights in our current model. The proposed system depends solely on the number of viewers who posted comments. Further study on text analysis would be interesting to conclusively link the detected signals with the content of the game in detail.

7 Conclusion

We have presented SPORTLIGHT, a crowdsourcing model which identifies the most spotlighted parts of a game by statistically analyzing online activities on sports events. By employing computationally inexpensive algorithm, our method provides much shorter and more relevant highlight reels than those of the peak-finding method. Our evaluation of SPORTLIGHT on two baseball postseasons showed that we could obtain perceptively appealing sports highlights reels comparable to expert-curated ones. This demonstrates feasibility of SPORTLIGHT for sports highlight identification. We hope our approach paves the way for statistically principled crowdsourcing in event detection.

Acknowledgements Won Son was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT) (No. 2020R1F1A1A01051039). Joong-Ho Won was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2C1007126).

References


- Assfalg, J., Bertini, M., Colombo, C., Del Bimbo, A., & Nunziati, W. (2003). Semantic annotation of soccer videos: Automatic highlights identification. *Computer Vision and Image Understanding*, 92(2–3), 285–305.
- Assfalg, J., Bertini, M., Del Bimbo, A., Nunziati, W., & Pala, P. (2002). Soccer highlights detection and recognition using HMMs. In *Proc. 2002 IEEE international conference on multimedia and expo (ICME'02)* (Vol. 1, pp. 825–828). IEEE.
- Babaguchi, N., Kawai, Y., Ogura, T., & Kitahashi, T. (2004). Personalized abstraction of broadcasted American football video by highlight selection. *IEEE Transactions on Multimedia*, 6(4), 575–586.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, 289–300.
- Bernstein, M.S., Little, G., Miller, R.C., Hartmann, B., Ackerman, M.S., Karger, D.R., Crowell, D., & Panovich, K. (2010). Soylent: A word processor with a crowd inside. In *Proceedings of the 23rd annual ACM symposium on User interface software and technology* (pp. 313–322). ACM.
- Bettadapura, V., Pantofaru, C., & Essa, I. (2016). Leveraging contextual cues for generating basketball highlights. In: *Proceedings of the 2016 ACM on Multimedia Conference, MM '16* (pp. 908–917). ACM. <https://doi.org/10.1145/2964284.2964286>
- Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: The bonferroni method. *BMJ*, 310(6973), 170.

- Chao, C.Y., Shih, H.C., & Huang, C.L. (2005). Semantics-based highlight extraction of soccer program using DBN. In *Proc. 2005 IEEE international conference on acoustics, speech, and signal processing (ICASSP'05)* (Vol. 2, p. ii-1057). IEEE.
- Chu, W. T., & Chou, Y. C. (2017). On broadcasted game video analysis: Event detection, highlight detection, and highlight forecast. *Multimedia Tools and Applications*, 76(7), 9735–9758.
- Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to algorithms*. Cambridge: MIT Press.
- D'Orazio, T., & Leo, M. (2010). A review of vision-based systems for soccer video analysis. *Pattern Recognition*, 43(8), 2911–2926.
- Ha, S., Kim, D., & Lee, J. (2013). Crowdsourcing as a method for digital media interaction. In *HCI 2013* (pp. 153–154). The HCI Society of Korea.
- Ha, S., Kim, D., & Lee, J. (2013). Crowdsourcing as a method for indexing digital media. In *CHI'13 extended abstracts on human factors in computing systems* (pp. 931–936). ACM.
- Hannon, J., McCarthy, K., Lynch, J., & Smyth, B. (2011). Personalized and automatic social summarization of events in video. In *Proceedings of the 16th international conference on intelligent user interfaces* (pp. 335–338). ACM.
- Hoefling, H. (2010). A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4), 984–1006.
- Jacobson, V. (1988). Congestion avoidance and control. In *ACM SIGCOMM computer communication review* (Vol. 18, pp. 314–329). ACM.
- Kapetanakis, A. (2018). IBM Watson: Inside the 'black box'. US Open News. Accessed 29 Apr 2019.
- Liu, C., Huang, Q., Jiang, S., Xing, L., Ye, Q., & Gao, W. (2009). A framework for flexible summarization of racquet sports video using multiple modalities. *Computer Vision and Image Understanding*, 113(3), 415–424.
- Marcus, A., Bernstein, M.S., Badar, O., Karger, D.R., Madden, S., & Miller, R.C. (2011). Twitinfo: Aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 227–236). ACM.
- Metulini, R. (2017). Filtering procedures for sensor data in basketball. *Statistica & Applicazioni*, 15(2), 133–150.
- Moyer, S. (2013). In America's pastime, baseball players pass a lot of time. *The Wall Street Journal*. <https://www.wsj.com/articles/SB10001424127887323740804578597932341903720>.
- Paxson, V., Allman, M., Chu, J., & Sargent, M. (2011). Computing TCP's retransmission timer. RFC 6298.
- Qian, X., Wang, H., Liu, G., & Hou, X. (2012). HMM based soccer video event detection using enhanced mid-level semantic. *Multimedia Tools and Applications*, 60(1), 233–255.
- Quinn, A.J., & Bederson, B.B. (2011) Human computation: A survey and taxonomy of a growing field. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1403–1412). ACM.
- Rui, Y., Gupta, A., & Acero, A. (2000). Automatically extracting highlights for TV baseball programs. In *Proceedings of the eighth ACM international conference on Multimedia* (pp. 105–115). ACM.
- Shih, H. C. (2018). A survey of content-aware video analysis for sports. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(5), 1212–1231.
- Son, W., & Lim, J. (2019). Modified path algorithm of fused lasso signal approximator for consistent recovery of change points. *Journal of Statistical Planning and Inference*, 200, 223–238.
- Tang, A., & Boring, S. (2012). #EpicPlay: Crowd-sourcing sports video highlights. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1569–1572). ACM.
- Thompson, C. (2010). What is I.B.M.'s Watson? The New York Times Magazine. <https://www.nytimes.com/2010/06/20/magazine/20Computer-t.html>. Accessed 29 Apr 2019.
- Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1), 91–108.
- Von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 319–326). ACM.
- Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895), 1465–1468.
- Xiong, Z., Radhakrishnan, R., & Divakaran, A. (2004). Method and system for extracting sports highlights from audio signals. US Patent App. 10/374,017

Xiong, Z., Radhakrishnan, R., Divakaran, A., & Huang, T.S. (2003). Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework. In *Proc. 2003 IEEE International Conference on Multimedia and Expo (ICME'03)* (Vol. 3, p. III-401). IEEE.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Jiwon Jung¹ · Seyong Ha² · Won Son³ · Joonhwan Lee⁴ · Joong-Ho Won¹ 

Jiwon Jung
jungjw1994@snu.ac.kr

Seyong Ha
seyong.ha@gmail.com

Won Son
son.won@dankook.ac.kr

Joonhwan Lee
joonhwan@snu.ac.kr

¹ Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, South Korea

² Department of Computer Science, University of Toronto, 40 St. George Street, Toronto, ON M5S 2E4, Canada

³ Department of Statistics, Dankook University, Jukjeon-ro 152, Yongin-si, Gyeonggi-do 16890, South Korea

⁴ Department of Communication, Seoul National University, Seoul, South Korea