이학석사학위논문

# *SportLight*: Statistically Principled Crowdsourcing Method for Sports Highlight Selection

## *SportLight*: 스포츠 하이라이트 추정을 위한 통계적 크라우드소싱 방법

2019 년 2 월

서울대학교 대학원

통계학과

정 지 원

# Abstract

Sports highlight selection has traditionally required expert opinions and manual labor of video editors. In recent years crowdsourcing viewers' live comments has emerged as a promising tool for automating this laborious task, overcoming the burden of extracting semantic information by computer vision. However, popular methods based on peak-finding are sensitive to noise and may produce highlights far from those selected by experts. In this work, we take a statistical approach that combines multiple hypothesis testing and trend filtering. By analyzing 29 baseball games played in the 2016 and 2017 seasons, we demonstrate that our approach properly sifts out the noise and generates result closer to expert-chosen highlights, achieving average precision higher than 0.7.

**Keywords**: Change point detection, Crowdsourcing, Fused lasso signal approximator , Multiple hypotheses testing, Sports highlight, Total variation penalty, Trend filtering, Video summary

**Student Number**: 2017-27318

# Contents

# List of Tables

# List of Figures

# Chapter 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 the live events is condensed into a series of short highlight reels, to be propagated to an even larger number of people. *Generating* highlight reels, however, 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 of a particular sport, making the whole process costly.

As the sports industry is becoming more and more capitalized, the need for automating the laborious task of highlight selection is ever increasing. One line of research in this direction is computer vision-based approaches, whose ultimate goal is to make the machine to understand the semantics of the games by analyzing the scenes [1]. 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. The cues of excitement – scenes containing such as crowd noise and player's roar – have to be curated to make "examples."

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 [2, 3]. Due to the burst of the Internet and smart devices, there has been a surge of time that people spend online. Consequently, more research on utilizing the collective input from the online crowd has sprung up in various domains [2, 3, 4, 5, 6, 7]. 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, peaks in data streams from popular social network platforms associated with live broadcasts are detected in efforts to extract exciting moments [8, 6]. 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 [6].

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, and they are highly correlated by the narrative of the game; in this respect, we focus on designing a model that successfully suppresses the possibility of false alarms

resulting from multiple hypotheses testing with a strong correlation. Because highlight events are by nature rare and abrupt, we adopt the $\ell_1$ trend filtering [9] not only to promote slowly varying trends but also to allow sparse, discontinuous changes [10].

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 dominating web service provider in Korea, which has 30 million daily visitors among the country of 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 also posts (traditionally edited) highlight reels on site. We found that the highlights selected by our method matches with the expert-curated ones with precision higher than 0.7, demonstrating that crowdsourcing can be *comparable* to the traditional method. This result is in contrast to the prior work [6], in which crowdsourced highlights have distinct features 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.

This dissertation is organized as follows. In Chapter 2, we summarize prior works on sports highlight selection and its background. Chapter 3 describes how our system works, and presents detailed illustrations of the algorithm used in the proposed system. Then, we evaluate the methodology via an experimental study based on both expertly chosen highlights and user evaluation in Chapter 4. The significance and limits of this study are discussed in Chapter 5, followed by a conclusion in Chapter 6.

The contribution of this dissertation is two-fold: (1) we demonstrate that statistically sound crowdsourcing techniques can produce sports video summa-

rization close to that generated by experts; (2) we provide a computationally inexpensive, simple-to-implement algorithm for this purpose.

# Chapter 2

# Background

## 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 [11]. These works utilize contextual cues appearing in an image frame relevant to the play, such as a player's position on a field [12] or a recognized scoreboard in a game [13]. 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 makes use of common audio cues such as the announcer's excited speech and ball-bat impact sound for a direct indicator of highlights [14, 15]. The system proposed in [16] measures the level of excitement and also displays its measured degree of interest along a time axis. Furthermore, machine learning approaches have applied audiovisual features for construction of

classifiers, such as support vector machines (SVMs) [17, 18] and hidden Markov model (HMMs) [12, 19], 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 four-hour long tennis match into three minutes [20]. It is reported that Watson collects audiovisual sources from the game and selectively translates the features into a quantified scale between 0 and 1 [21]. 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 [20].

## 2.2 Crowdsourcing-based sports highlight selection

### 2.2.1 Crowdsourcing systems

Although audiovisual sources have proven their power to extract a play's semantic information [22, 23], crowdsourcing viewers' live comments has emerged as a promising source for automating this laborious and burdensome task. Generally, audiovisual frameworks are labor intensive and computationally expensive; they not only require the gathering of numerous pre-specified scenes but also entail the 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 [24]. Another reason for supporting crowdsourcing method is that sports highlights are closely in line with viewers' emotional experiences [6]. 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 every single hit of the game.

A series of studies has demonstrated that viewers' live comments on pop-

ular social media [5, 6], online broadcasting platforms [25], or live-streaming platforms [26] 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 [5]; 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 *#EPIC-PLAY* 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 [6]. There are further studies on media interaction focusing on the applications in digital media indexing by collecting posts from online forums [25, 7]; in particular, [25] generates and analyzes sports highlights from baseball games. A more recent study combines viewer data with a traditional audiovisual feature-based approach [26].

### 2.2.2 The peak-finding algorithm

The key component of the crowdsourcing methods based on *TwitInfo* [5, 6, 25] is the so-called peak-finding algorithm, which detects abnormal increases in the number of postings by scanning a large stream of data. This algorithm is based on the following outlier detection criterion. Given the observed stream of count data $C_1, C_2, \ldots, C_n$, when a new count $C_{n+1}$ is observed, the algorithm classifies the point as a peak if

$$\frac{C_{n+1} - \hat{\mu}_n}{\hat{\sigma}_n} > \tau, \tag{2.1}$$

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:

$$\hat{\sigma}_{n+1} = \alpha|C_{n+1} - \hat{\mu}_n| + (1-\alpha)\hat{\sigma}_n \tag{2.2}$$

$$\hat{\mu}_{n+1} = \alpha C_{n+1} + (1-\alpha)\hat{\mu}_n \tag{2.3}$$

An abrupt change in the temporal signal is identified with a peak when 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; $\tau$ 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 instead of the exponentially weighted absolute deviation was used. In effect, size of $\tau$ controls the number of peaks detected. A practical choice of the quantities $\tau$ and $\alpha$ is proposed in [5], as well as in [6, 25].

In fact, this algorithm is inspired by the algorithm for computing retransmission time-out (RTO) in the transmission control protocol (TCP) [27]. The mean absolute deviation is used because of two reasons: it provides more conservative measurements, and is also easier to compute [28]. 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 from 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 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*.

For the purpose of crowdsourced highlight selection, baseball has a number of useful features: (1) 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); (2) 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 minutes and 58 seconds of action over the course of a three-hour game" [29]); (3) 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 [6].

# Chapter 3

# The SportLight System

## 3.1 Components of the SportLight system

### 3.1.1 Naver's social broadcasting platform

The broadcasting platform serviced by Naver is illustrated in Figure 3.1, panel 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 either 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 together with the game statistics, even many years after the game. This information can be easily retrieved in the javascript object notation (JSON) format.

### 3.1.2 Highlight selection

The input to the highlight selection part is the comment history of a finished game from the aforementioned platform. The output of the highlight selection part is a number of timestamp intervals classified as highlights. The number of highlights can be specified by the user. The minimum length of an interval is one minute. Panel 2 of Figure 3.1 illustrates 20 signals of selected highlights (yellow bands) of a real playoff game (held on Oct. 21, 2016) overlaid on the plot of the frequency of comment postings vs. 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). Note that an expert-curated highlight featuring 1-2-3 inning —located in the very beginning— is missed.

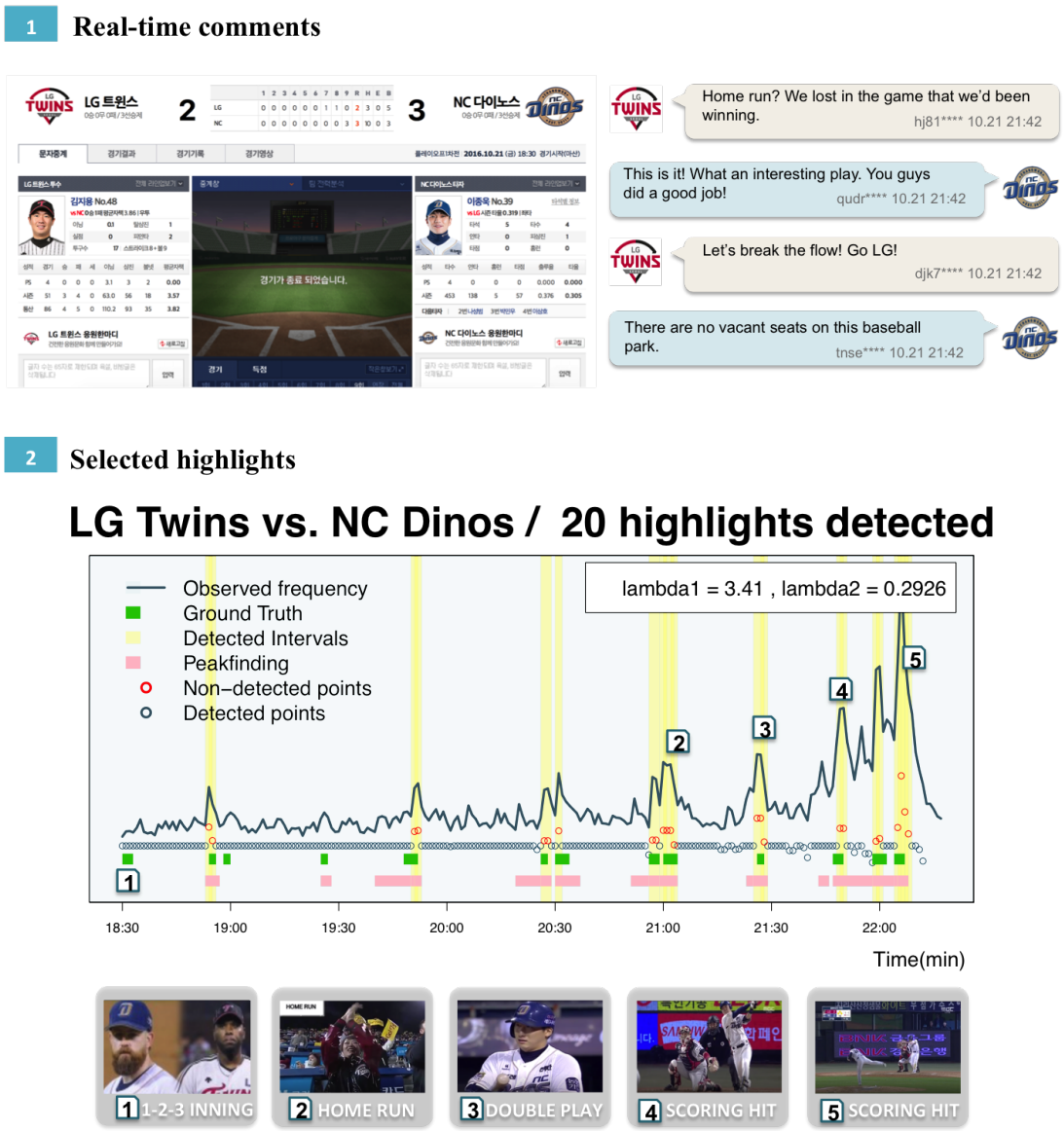


Figure 3.1 Outline of the SportLight system. Texts in panel 1 are translated from the original scripts in Korean.

## 3.2 Highlight selection algorithm

### 3.2.1 Data

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. It appears natural 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: 1) some viewers post meaningless comments like "hahaha" constantly, regardless of the game's progress; 2) 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.

### 3.2.2 Statistical model

Statistically, it is reasonable to model such data as a sequence of Poisson random variables. In other words, at time $t$, the observed frequency $X_t$ of the comments is assumed to follow the Poisson distribution with mean $\mu_t$:

$$P(X_t = x) = e^{-\mu_t} \frac{\mu_t^x}{x!}, \quad x = 0, 1, 2, \ldots \tag{3.1}$$

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

### 3.2.3 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, $\mu_t$ varies slowly, thus can be estimated from the data in the neighborhood of $t$. At a highlight state, $\mu_t$ 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 $\mu_t$, we can proceed with testing the following statistical hypothesis:

$$H_0 : \mu_t = \mu_0(t) \quad vs. \quad H_1 : \mu_t > \mu_0(t), \tag{3.2}$$

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 the $p$-value is close to 1, the test is in favor of the null hypothesis $H_0$, or the normal state. If it is close to 0, the test is in favor of the alternative hypothesis $H_1$, or the highlight state. As we assume a Poisson distribution (3.1), we conduct the Poisson test for testing (3.2).

### 3.2.4 $\ell_1$-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 *SportLight*, where we need to test more than 100 hypotheses. Although there are procedures to adjust the inflation [30, 31], often these procedures are too conservative and suppress most of true findings. Recently, a method based on $\ell_1$ trend filtering has been proposed to adaptively adjust multiple hypothesis testing [32]. The $\ell_1$ 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, $\ell_1$ trend filtering minimizes

$$\frac{1}{2}\sum_{t=1}^{n}(z_t-\zeta_t)^2+\lambda\sum_{t=1}^{n-1}|\zeta_t-\zeta_{t+1}|, \tag{3.3}$$

for variables $\zeta_1,\ldots,\zeta_n$, where $z_1,\ldots,z_n$ are given noisy observations; $n$ is the number of intervals. 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 $\zeta_1,\ldots,\zeta_n$. 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 [33]. The positive constant $\lambda$ is a tuning parameter that controls the degree of the zeroing effect. In the above case, this means that the filtered sequence $\zeta_1,\ldots,\zeta_n$ is piecewise constant with a few jumps. Thus minimizing (3.3) fits in our statistical model.

In order to adopt the $\ell_1$ trend filtering to adjust for false alarms from multiple hypothesis testing, we first convert the $p$-values from tests (3.2) into $z$-values. If time $t$ is in the normal state, then the corresponding $z$-value $z_t$ 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 $z_t$ will be far from the neighbor. The filtered $z$-values $\zeta_t$ obtained by minimizing (3.3) suppress false alarms due to the multiple hypothesis testing while allowing occasional jumps due to highlights. In order to further promote each $\zeta_t$ to tend to zero, we add an additional penalty and minimize

$$\frac{1}{2}\sum_{t=1}^{n}(z_t-\zeta_t)^2+\lambda_1\sum_{t=1}^{n}|\zeta_t|+\lambda_2\sum_{t=1}^{n-1}|\zeta_t-\zeta_{t+1}|, \tag{3.4}$$

with a pair of tuning parameters $(\lambda_1,\lambda_2)$.

The above discussion suggests that if we plot $\zeta_t$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 $\zeta_t$ may reflect distinct degree of interest. In panel 2 of Figure 3.1, those distinct levels constitute 20 highlight intervals. Additionally, selected highlights from three games are displayed in Figure 3.2.

### 3.2.5 Tuning parameter selection

The tuning parameters $\lambda_1$ and $\lambda_2$ for problem (3.4) are chosen so that the number of detected highlights match the desired number. A large value of $\lambda_1$ results in many zeros among $\zeta_1, \ldots, \zeta_n$, whereas $\lambda_2$ controls the size of jumps. There is an efficient algorithm for solving (3.4): for a fixed value of $\lambda_1$, $\zeta_t$s that minimize (3.4) for every possible values of $\lambda_2$ are computed at once [34]. Thus for each $\lambda_1$, we can find $\lambda_2$ that gives the desired number of highlights. We then select the $\lambda_1$ that exhibits the largest average of positive $\zeta_t$'s per minute, which is a reasonable choice in that sports highlights represent the climaxes of the game in which the viewers are most interested. Now we can summarize the procedure discussed above in Algorithm 1.

**Algorithm 1** SportLight highlight selection

**Require:** $(x_1, \ldots, x_n)$=count data of length $n$, $N$ = desired number of highlight reels ($0 < N << n$), lambda1Grid = grid sequence of $\lambda_1$

```
for t = 1, ... n do
    p_t ←compute_pvalue(x_1, ..., x_n)                ▷ Chapter 3.2.3: test (3.2)
    z_t ←convert_to_zvalue(p_t)                       ▷ Chapter 3.2.4: input of (3.4)
end for
for i = 1, ... do
    λ_1 ←lambda1Grid[i]
    m ← 0
    while m ≠ N do
        Choose λ_2
        (ζ_1, ..., ζ_n) ← ℓ_1_trendfilter(z_1, ..., z_n; λ_1, λ_2)   ▷ solve (3.4)
        intervals ←detect_highlight(ζ_1, ..., ζ_n)   ▷ Chapter 3.1.2, 3.2.4
        m ←size(intervals)
    end while
    solutions[i]← ((ζ_1, ..., ζ_n); λ_1, λ_2)         ▷ solution path
end for
(intervals*; λ_1*, λ_2*) ←tuning_selection(solutions)  ▷ Chapter 3.2.5
return (intervals*; λ_1*, λ_2*)
```

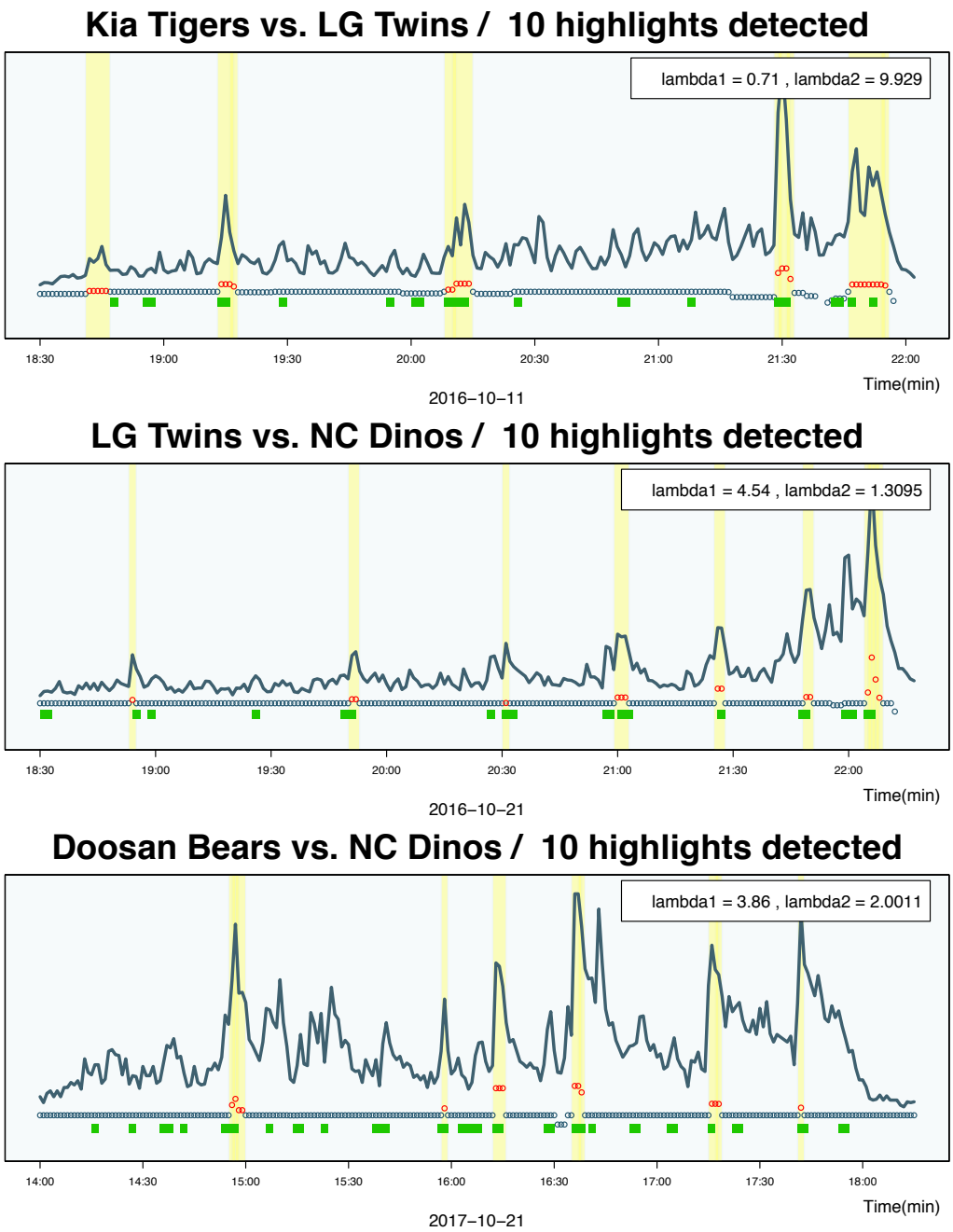


Figure 3.2 10 highlights retrieved by the SportLight for three selected games.

# Chapter 4

# Evaluation

## 4.1 Quantitative evaluation

We now evaluate the performance of the *SportLight* system. We use the comment data from two KBO League postseasons. This dataset covers 14 and 15 playoffs from the 2016 and 2017 seasons, respectively. For the creation of the ground truth, we used the expert-chosen clips provided by Naver. These clips are manually collected and labelled. The number of highlight reels found in this platform ranges from 11 to 38 per game.

We first evaluate conventional performance metrics, namely precision and recall, which is defined as follows:

$$
\begin{aligned}
Precision &= \frac{|\{Retrieved\ Scenes\} \cap \{Relevant\ Scenes\}|}{|\{Retrieved\ Scenes\}|} \\
Recall &= \frac{|\{Retrieved\ Scenes\} \cap \{Relevant\ Scenes\}|}{|\{Relevant\ Scenes\}|}
\end{aligned}
$$

In addition, $F_1$ scores are computed for the harmonic average of the precision and recall.

We compare the performance of the proposed algorithm and the peak-finding algorithm. The peak-finding algorithm requires to determine the two parameters $\alpha$ and $\tau$. We fix the weight $\alpha = 0.125$ as suggested in [5], but use a different critical value $\tau$ for each game to generate the same number of highlights as the SportLight. The average scores are presented in Table 4.1. We set these algorithms to select 10, 20, 30, and 40 highlights.

| Metric | Algorithm | Number of highlights | | | |
|---|---|---|---|---|---|
| | | 10 | 20 | 30 | 40 |
| Precision | SportLight | 72.13 | 67.98 | 63.22 | 61.98 |
| | Peak-finding | 74.24 | 67.11 | 60.51 | 56.18 |
| Recall | SportLight | 32.00 | 51.32 | 61.71 | 70.89 |
| | Peak-finding | 51.63 | 64.72 | 72.42 | 79.31 |
| F1 score | SportLight | 43.02 | 57.00 | 61.57 | 65.09 |
| | Peak-finding | 59.27 | 64.85 | 64.66 | 64.54 |

Table 4.1 Average performance of 29 games(%, rounded) per number of signals extracted from 10 to 40.

The result looks promising: the average precision of the proposed method reaches up to 72.14%. The precision gradually decreases if we choose to retrieve more highlights. Rather surprisingly, the cost is small: without sacrificing much in precision, increasing the desired number of highlights results in higher recall. For the maximum retrieval of 40 signals, about 62% of the chosen highlights include the ground truth intervals. On average, 17 manually labeled highlights are serviced per game by Naver. Thus choosing more than 10 highlights could be desirable for practice. In comparison to the peak-finding algorithm, the *Sport-*

*Light* appears to have an advantage in terms of precision for those number of highlights.

| **Game (YYYY-MM-DD)** | Metric | Number of highlights | | | |
|---|---|---|---|---|---|
| | | 10 | 20 | 30 | 40 |
| **Kia Tigers vs. LG Twins (2016-10-11)** | Precision | 90.00 | 85.00 | 70.00 | 72.50 |
| | Recall | 46.67 | 60.00 | 73.33 | 86.67 |
| | F1-score | 61.46 | 70.34 | 71.63 | 78.95 |
| **LG Twins vs. NC Dinos (2016-10-21)** | Precision | 100.00 | 100.00 | 90.00 | 77.50 |
| | Recall | 61.54 | 76.92 | 92.31 | 92.31 |
| | F1-score | 76.19 | 86.96 | 91.14 | 84.26 |
| **Doosan Bears vs. NC Dinos (2017-10-21)** | Precision | 100.00 | 95.00 | 83.33 | 85.00 |
| | Recall | 27.27 | 45.45 | 54.55 | 63.64 |
| | F1-score | 42.86 | 61.49 | 65.93 | 72.78 |

Table 4.2 Performance of three selected games (%, rounded) per number of signals extracted from 10 to 40.

Additionally, most of the detected signals successfully match the highlights provided by the experts in some of the selected games shown in Figure 3.2; about 9 out of 10 selected reels are correctly identified. To view the performance score on each game, see Table 4.2 for detail.

We have focused on precision because the number of highlights to select can be controlled, and recall can be easily misleading. While the peak-finding method yields higher recall than the SportLight, this is because the recall can be overestimated if excessively lengthy intervals are selected. In fact, the Naver-

chosen highlights have an average length of 2 minutes. However, the peak-finding algorithm generates 5.3-minute long reels when retrieving 10 highlights. On the other hand, SportLight highlights are 2.9-minute long on average. The box plot in Figure 4.1 directly shows the difference in average length of the selected intervals between two methods.

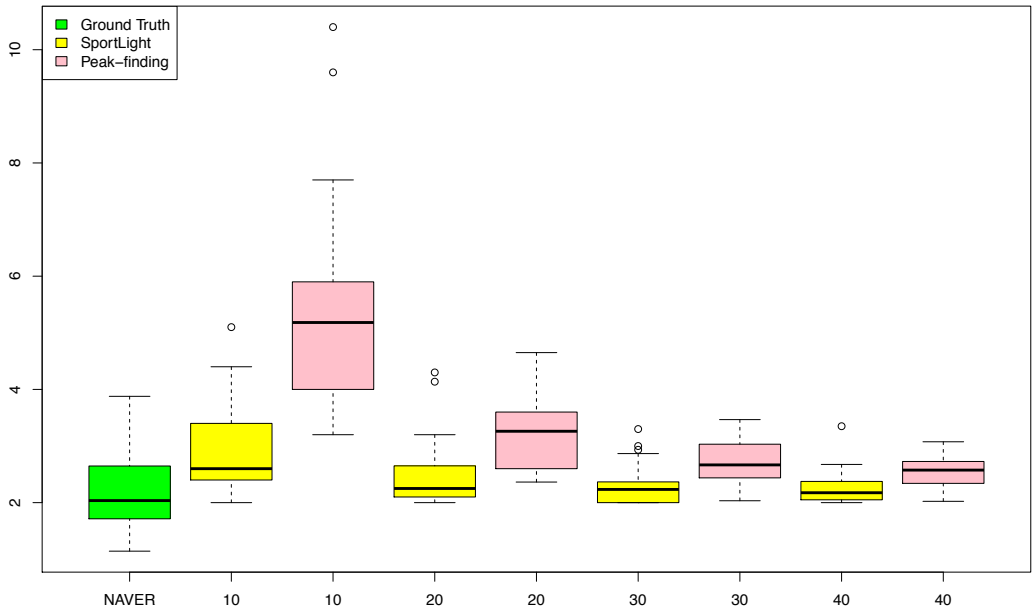


Figure 4.1 Box plot of average highlight-length in minute.

Hence, a better metric is needed for more precise assessment of the performance. We use the Hamming distance, the number of different digits in two binary sequences of the same length, to compare the performance in a scene length-adjusted fashion. We convert a game into a binary sequence in such a way that a unit (one-minute) interval corresponds to 1 if it is classified as that in a highlight, and 0 otherwise. The Hamming distances between each of the

*SportLight* and the peak-finding methods and the expert-chosen highlights are computed. Since the play times differ from game to game, we normalize the measured Hamming distance by dividing it with the total length of the game. These normalized results are shown in Table 4.3 for both *SportLight* and the peak-finding method. Along with the Hamming distance, minute-wisely compared performance is also attached including conventional metrics: precision, recall and $F_1$ score. For minute-wise evaluation, we compare each binary sequence produced by the *SportLight* and the peak-finding methods with that of ground truth.

| Metric (minute-wise comparison) | Algorithm | Number of highlights | | | |
|---|---|---|---|---|---|
| | | 10 | 20 | 30 | 40 |
| Hamming distance | SportLight | 29.20 | 30.07 | 32.51 | 34.79 |
| | Peak-finding | 33.41 | 34.34 | 38.98 | 45.59 |
| Precision | SportLight | 49.87 | 47.75 | 44.07 | 41.98 |
| | Peak-finding | 42.28 | 42.36 | 39.08 | 35.42 |
| Recall | SportLight | 24.47 | 38.95 | 47.03 | 56.33 |
| | Peak-finding | 40.07 | 51.10 | 59.72 | 69.33 |
| F1 score | SportLight | 32.01 | 41.68 | 44.54 | 47.15 |
| | Peak-finding | 40.20 | 45.49 | 46.30 | 46.21 |

Table 4.3 Average Hamming-distance (%, normalized) and average minute-wise performance (%) per number of signals extracted from 10 to 40.

In terms of minute-wise assessment, the difference becomes more obvious. It suggests that the proposed system yields reels that are closer to the manual selection; the discrepancy measured with Hamming distance is remarkably smaller than that with the peak-finding method in all of our experiment: 29%

versus 33%, respectively, for 10 highlights. In terms of minute-wise precision, the performance is much more stringently measured; we observed significant decrease in precision for both algorithms. However, our system produces higher precision in all of our experiment: 50% versus 42%, respectively, for 10 signals. This evidence supports statistical modeling on crowdsourced online activity data as a feasible approach to summarize the sports games.

## 4.2 Qualitative evaluation

Our next step is to gather the qualitative evaluation of the selected reels, especially for reels identified as false alarms. We collected responses from five baseball fans. They are all familiar with the baseball rules. We asked them to watch the provided video clips generated from the *SportLight* system. Then, we obtained feedback to understand the characteristics of the chosen highlights. Specifically, we mainly focused on the users' opinions on the 13 false-positive clips from the five selected games to understand the property of the crowdsourcing method.

The participants were asked to classify the given clips into 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.

They were asked to choose at least one category because each clip possibly contains more than one event. They were also encouraged to give reasoning for their choice on the corresponding scene. As a summary of the evaluation, 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.

Results are shown in Figure 4.2; out of 13 false positives, 8 were voted for the first category, 2.8 were for the second category, and the rest were considered to be the non-highlights or unknowns.

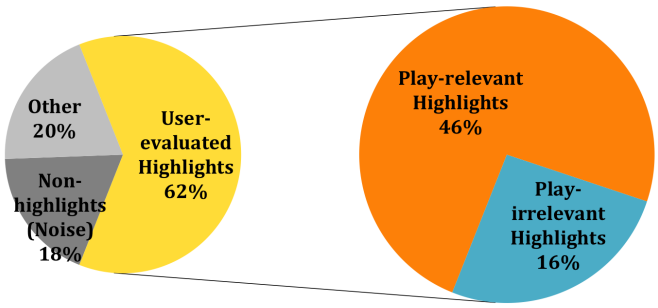


Figure 4.2 Composition of false alarms.

The result of the user evaluation of the false positives is two fold. First, the participants had different opinions on how they categorize and interpret the scenes of interest. For example, these scenes include the emotional reactions of the audience and renowned baseball managers: the players and fans celebrating the victory of their supporting team, and a close-up of the managers' frustrating faces. Some of the participants viewed these reactions as important features of a game, but the rest did not consider them as interesting. That is, baseball fans expressed mixed opinions on how they regard and perceive the importance of a given scene based on their personal taste. Second, there is a gap between the experts' judgement and the general public's view on their preference of highlight. These scenes include "highlights" either overlooked or regarded less

important by experts: an impressive defense which failed to tag the runner out, and a failed bunt play. While the experts prioritize and focus on the scoring part of a match, the participants identified some of the false alarms as worthy of notice.

Although the user evaluation showed the feasibility of SportLight for sports highlight selection, the participants also pointed out some limitations of the generated scenes: (1) some of the generated intervals are incomplete; some of the main events are partly captured; (2) some of the generated intervals are excessively retrieved; in many cases, they often contain unnecessary and irrelevant scenes such as replayed events and advertisements. Fortunately, these limitations can be improved by using a shorter unit interval.

# Chapter 5

# Discussion

We have shown that the proposed *SportLight* system successfully analyzes online activities on sports events. Largely matching the expert opinions, our method produces the most spotlighted parts of a game. Along with the success on KBO data analysis, however, our design had several limitations. First, our system is not able to provide the highlights with descriptions. We designed our system to generate highlight intervals depending solely on the number of viewers who posted comments. Hence, the system does not take the semantic content into account. In this respect, further study on text analysis would be required to conclusively link detected signals with the content of the game in detail. We also see that there still is a room for improving quality of retrieval of the signals. As described in evaluation, the generated highlights raise an issue of retrieving excess or partial highlights. Further experiments on choosing a proper size of the unit interval and tuning quantities are warranted regarding this issue.

# Chapter 6

# Conclusion

Video summarization task has accompanied arduous manual exploration of its suppliers. As a remedy for the demand, crowdsourcing models are emerging. The challenge in the crowdsourced highlight generation process, however, is to accurately sort out the significant increase in the viewers' activities. Highlight signals are rare and abrupt in a sports game; to discover the parameter change of the embedded mean signals, we proposed a system that combines multiple hypotheses testing and $\ell_1$-trend filtering, which effectively extracts the moments of viewers' interest in the presence of noise.

Technically, our main contribution lies in statistically principled application of highlight identification of the count data. By analyzing two baseball post-seasons, we demonstrated that our system successfully estimates unobserved binary game states—highlights and non-highlights. As a result, we could obtain sports highlights comparable to expert-curated ones in terms of accuracy and perceptively appealing; our model successfully generated a sequence of highlights closer to the ground truth. By applying statistically principled methods,

*SportLight* can provide much shorter and more relevant highlight reels than the peak-finding algorithm.

We hope our approach paves the way for statistically principled crowdsourcing in event detection. We expect *SportLight* can be further extended to other domains than baseball.

# Bibliography


[1] T. D'Orazio and M. Leo, "A review of vision-based systems for soccer video analysis," *Pattern Recognition*, vol. 43, no. 8, pp. 2911–2926, 2010.

[2] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 319–326, ACM, 2004.

[3] L. von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum, "Recaptcha: Human-based character recognition via web security measures," *Science*, vol. 321, no. 5895, pp. 1465–1468, 2008.

[4] M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, and K. Panovich, "Soylent: a word processor with a crowd inside," in *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pp. 313–322, ACM, 2010.

[5] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller, "Twitinfo: aggregating and visualizing microblogs for event exploration," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 227–236, ACM, 2011.

[6] A. Tang and S. Boring, "#EpicPlay: Crowd-sourcing sports video highlights," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1569–1572, ACM, 2012.

[7] S. Ha, D. Kim, and J. Lee, "Crowdsourcing as a method for indexing digital media," in *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pp. 931–936, ACM, 2013.

[8] J. Hannon, K. McCarthy, J. Lynch, and B. Smyth, "Personalized and automatic social summarization of events in video," in *Proceedings of the 16th International Conference on Intelligent User Interfaces*, pp. 335–338, ACM, 2011.

[9] S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky, "$\ell_1$ trend filtering," *SIAM Review*, vol. 51, no. 2, pp. 339–360, 2009.

[10] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical learning with sparsity: the lasso and generalizations.* CRC press, 2015.

[11] H.-C. Shih, "A survey of content-aware video analysis for sports," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 5, pp. 1212–1231, 2018.

[12] J. Assfalg, M. Bertini, C. Colombo, A. Del Bimbo, and W. Nunziati, "Semantic annotation of soccer videos: automatic highlights identification," *Computer vision and image understanding*, vol. 92, no. 2-3, pp. 285–305, 2003.

[13] N. Babaguchi, Y. Kawai, T. Ogura, and T. Kitahashi, "Personalized abstraction of broadcasted american football video by highlight selection," *IEEE Transactions on Multimedia*, vol. 6, no. 4, pp. 575–586, 2004.

[14] Z. Xiong, R. Radhakrishnan, and A. Divakaran, "Method and system for extracting sports highlights from audio signals," Aug. 26 2004. US Patent App. 10/374,017.

[15] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang, "Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework," in *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, vol. 3, pp. III–401, IEEE, 2003.

[16] V. Bettadapura, C. Pantofaru, and I. Essa, "Leveraging contextual cues for generating basketball highlights," in *Proceedings of the 2016 ACM on Multimedia Conference*, MM '16, pp. 908–917, ACM.

[17] C. Liu, Q. Huang, S. Jiang, L. Xing, Q. Ye, and W. Gao, "A framework for flexible summarization of racquet sports video using multiple modalities," *Computer Vision and Image Understanding*, vol. 113, no. 3, pp. 415–424, 2009.

[18] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," in *Proceedings of the eighth ACM international conference on Multimedia*, pp. 105–115, ACM, 2000.

[19] J. Assfalg, M. Bertini, A. Del Bimbo, W. Nunziati, and P. Pala, "Soccer highlights detection and recognition using HMMs," in *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, vol. 1, pp. 825–828, IEEE, 2002.

[20] A. Kapetanakis, "IBM Watson: inside the 'black box'," *US Open News*, Sept. 08 2018. Accessed: 2019-01-15.

[21] C. Thompson, "What is I.B.M.'s Watson?," *The New York Times Magazine*, June 16 2010. Accessed: 2019-01-15.

[22] C.-Y. Chao, H.-C. Shih, and C.-L. Huang, "Semantics-based highlight extraction of soccer program using dbn," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 2, pp. ii–1057, IEEE, 2005.

[23] X. Qian, H. Wang, G. Liu, and X. Hou, "HMM based soccer video event detection using enhanced mid-level semantic," *Multimedia Tools and Applications*, vol. 60, no. 1, pp. 233–255, 2012.

[24] A. J. Quinn and B. B. Bederson, "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, 2011.

[25] S. Ha, D. Kim, and J. Lee, "Crowdsourcing as a method for digital media interaction," in *HCI 2013*, pp. 153–154, The HCI Society of Korea, 2013.

[26] W.-T. Chu and Y.-C. Chou, "On broadcasted game video analysis: event detection, highlight detection, and highlight forecast," vol. 76, no. 7, pp. 9735–9758.

[27] V. Paxson, M. Allman, J. Chu, and M. Sargent, "Computing TCP's retransmission timer," *RFC 6298*, 2011.

[28] V. Jacobson, "Congestion avoidance and control," in *ACM SIGCOMM computer communication review*, vol. 18, pp. 314–329, ACM, 1988.

[29] S. Moyer, "In America's pastime, baseball players pass a lot of time," *The Wall Street Journal*, July 16 2013.

[30] J. M. Bland and D. G. Altman, "Multiple significance tests: the bonferroni method," *Bmj*, vol. 310, no. 6973, p. 170, 1995.

[31] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300, 1995.

[32] W. Son and J. Lim, "Modified path algorithm of fused lasso signal approximator for consistent recovery of change points," *Journal of Statistical Planning and Inference*, vol. 200, pp. 223–238, 2019.

[33] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 1, pp. 91–108, 2005.

[34] H. Hoefling, "A path algorithm for the fused lasso signal approximator," *Journal of Computational and Graphical Statistics*, vol. 19, no. 4, pp. 984–1006, 2010.

# 국문초록

스포츠 하이라이트 장면의 선정은 전통적으로 영상 편집자의 전문가의 판단과 수작업을 필요로 한다. 최근 몇년 동안 하이라이트 선정 작업을 자동화하는 유망한 도구로써 실시간 영상 시청자들의 댓글을 크라우드소싱하는 방법이 시도되어 왔으며, 이를 통해 컴퓨터 비전을 통한 의미 정보 추출의 계산 비용을 극복할 수 있게 되었다. 그러나 기존의 최고점찾기에 기반한 알고리즘은 잡음에 민감할 뿐만 아니라 전문가가 선정한 하이라이트 영상과 질적인 괴리가 생기는 문제점이 제기된다. 따라서 이 논문에서는 다중 가설 검정과 추세 필터링을 결합한 통계적 접근 방법을 통해 기존의 방법과 비교하여 개선된 하이라이트 구간을 추정하고자 한다. 본문에서 제안한 시스템을 2016년과 2017년 포스트 시즌에 시행된 29개의 야구 경기 분석에 적용한 결과 평균 정밀도 0.7 이상을 달성하였고, 기존의 방법에 비해 잡음을 적절히 처리하며 전문가가 선정한 장면들과 더 유사한 결과를 생성한다는 것을 검증하였다.

**주요어**: 변화점 감지, 크라우드소싱 방법, FLSA 알고리즘 , 다중가설검정, 스포츠 하이라이트, 총변이 벌점함수, 추세필터, 영상 요약

**학번**: 2017-27318