# Bridging Video Content and Comments: Synchronized Video Description with Temporal Summarization of Crowdsourced Time-Sync Comments

**Linli Xu, Chao Zhang**
School of Computer Science and Technology
University of Science and Technology of China
linlixu@ustc.edu.cn, matgdog@mail.ustc.edu.cn

## Abstract

With the rapid growth of online sharing media, we are facing a huge collection of videos. In the meantime, due to the volume and complexity of video data, it can be tedious and time consuming to index or annotate videos. In this paper, we propose to generate temporal descriptions of videos by exploiting the information of crowdsourced time-sync comments which are receiving increasing popularity on many video sharing websites. In this framework, representative and interesting comments of a video are selected and highlighted along the timeline, which provide an informative description of the video in a time-sync manner. The challenge of the proposed application comes from the extremely informal and noisy nature of the comments, which are usually short sentences and on very different topics. To resolve these issues, we propose a novel temporal summarization model based on the data reconstruction principle, where representative comments are selected in order to best reconstruct the original corpus at the text level as well as the topic level while incorporating the temporal correlations of the comments. Experimental results on real-world data demonstrate the effectiveness of the proposed framework and justify the idea of exploiting crowdsourced time-sync comments as a bridge to describe videos.

## Introduction

The rapid growth of online sharing media has provided popular channels for people to watch videos of entertainment, sports, news, etc. As a consequence, we are facing an overwhelming collection of online videos which poses major challenges of effective management and annotation of videos. To address that, automatic video tagging techniques (Siersdorfer, San Pedro, and Sanderson 2009; Ulges et al. 2010) have been proposed which mainly focus on detecting concepts and associating tags of entire videos. Meanwhile, users can get further interested to know the information of video content along the playback time. Specifically, if videos are associated with time-sync textual description or annotation, users can preview the content with both thumbnails and text along the timeline, and this textual information can further enhance users' search experience with time positions.

On the other hand, automatic time-sync description of video content can be very expensive, which may rely on scene-text alignment or external information such as video transcripts or subtitles. Fortunately, the application of crowdsourced time-sync video comments which are receiving increasing popularity on many video sharing websites including Nico Nico Douga[1] in Japan, bilibili[2] and iQIYI[3] in China provides a new perspective. Time-sync video comments, also known as "bullet-screen comments", are the comments users send to express their opinions or interpretations along the playback time when watching a video. These comments are overlaid directly over the video in a synchronized manner, and users are allowed to respond to each other, which enhances the experience of participation and communication. Intuitively, crowdsourced time-sync video comments provide a valuable source of information regarding the temporal information of a video, while the task of extracting this information is still challenging: the comments are mostly in an informal and noisy form, such as misspellings, symbols of emotions and internet slang terms, while containing enormous redundancies; in addition, the comments are usually short sentences and on very different topics along the playback time.

Recently, methods have been proposed to generate temporal tags or labels based on crowdsourced time-sync video comments (Wu et al. 2014; Lv et al. 2016), which mainly focus on extracting keywords such as topics or semantic labels. On the other hand, keywords sometimes are not sufficient to describe a scene, especially when the scene includes a number of characters or depicts a complicated situation. In this context, we take a different perspective and extract representative comments rather than keywords, which are complete sentences and convey more meaningful information regarding the video content.

To achieve that, we propose a novel temporal summarization model based on the data reconstruction principle. A subset of representative comments are selected to describe the video at a playback time position along the timeline at two levels: the selected subset should be able to reconstruct the original corpus of video comments at the text level; in

[1] http://www.nicovideo.jp/

[2] http://www.bilibili.com/

[3] http://www.iqiyi.com/

the meantime, it should also recover the topics of video semantics conveyed in the comments. As a consequence, the proposed model consists of a text reconstruction component and a topic reconstruction component. In addition, we consider the temporal nature of the problem which implies that comments at adjacent playback time positions are correlated and representatives should be selected to reduce redundancy across adjacent positions. Based on this framework, the video can be described in a synchronized way, conveying relevant, important and non-redundant information of the video, which is easy to read and comprehend.

# Related Work

## Analysis of Time-sync Video Comments

Time-sync comments (TSC) provide a new source of information regarding the video and have received growing research interests. In (Wu and Ito 2014), correlation between emotional comments and popularity of a video is analyzed in a statistical way, while a shot boundary detection method is proposed in (Xian et al. 2015) to extract highlight shots based on time-sync comments. Efforts have also been devoted to associate comments with video content along the timeline. In (Lv et al. 2016), time-sync comments are first represented with semantic vectors, then a video splitting framework is designed to extract and label meaningful segments based on mapping the semantic vectors to pre-defined labels in a supervised way. However, this model relies on a large amount of human-labeled video segments and predefined emotional tags to train, which limits its applicability to more general scenarios. Another piece of related work is proposed in (Wu et al. 2014), where a temporal and personalized topic model is designed to select keywords in time-sync comments as the tags of a video shot. On the other hand, considering the informal nature of time-sync comments, complete comments essentially provide more effective time-sync description of a video than fragmentary keywords. In this paper, we extract representative time-sync comments for video description based on a novel temporal summarization model.

## Extractive Document Summarization

Extractive document summarization has drawn a lot of attention recently, most of which assign salient scores to sentences of a document and generate the summary with top-ranked sentences. Among them, LexRank (Erkan and Radev 2004) takes the similarity of sentences into consideration and computes the salient scores based on graph-based lexical centrality; while the algorithm proposed in (Wan and Yang 2008) incorporates the cluster-level information into the process of sentence ranking. Meanwhile, some efforts have been devoted to selecting sentences without salient scores. In (Gong and Liu 2001), singular value decomposition is used to select sentences with high rank; the algorithm proposed in (Wang et al. 2008) clusters sentences using symmetric non-negative matrix factorization and selects sentences in each cluster for summarization. More recently, a data reconstruction method is proposed in (He et al. 2012) where sentences that can best reconstruct the original document are selected. Nevertheless, all the above methods focus on the task of summarization on static documents while ignoring the temporal information in text, which is important in our problem.

Apart from traditional document summarization methods, a number of algorithms (Hu and Liu 2004; Inouye and Kalita 2011; Hu and Liu 2006; Yan et al. 2011; Shou et al. 2013; Chen et al. 2015) are designed to summarize the massive collection of tweets, reviews and news. However, to the best of our knowledge, there is no prior work on summarization of time-sync video comments from the new interactive feature on many video sharing websites. In this paper, we propose a novel temporal summarization framework based on the data reconstruction principle and select meaningful and representative comments along the timeline as the temporal description of a video.

# Proposed Framework

In this section, we propose a novel framework of temporal summarization which minimizes the reconstruction error from the text level and the topic level simultaneously in temporal segments.

## Summarization Based on Joint Text and Topic Reconstruction

We first start with summarization based on *text*-level data reconstruction. Given a set of comments $C = [\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_n]$ where $\mathbf{c}_i \in \mathbb{R}^d$ is a term-frequency vector weighted with TF-IDF scores for the $i$-th comment and $d$ is the number of terms, from the perspective of text-level reconstruction, we want to find an optimal subset of $m$ representative comments $X \subseteq C$ with $m < n$ to best reconstruct the comments in $C$. The selected subset $X$ can be denoted with an indicator vector $\boldsymbol{\beta} \in \{0,1\}^n$ such that the $j$-th comment will be selected when $\beta_j = 1$. A given comment $\mathbf{c}_i$ can then be reconstructed by a non-negative linear combination of selected comments:

$$\mathbf{c}_i = \sum_{j=1}^{n} \mathbf{c}_j \beta_j a_{ij} \tag{1}$$

where $\mathbf{a}_i \geq 0$ is a non-negative vector of length $n$ which represents the column coefficient vector of the linear reconstruction for $\mathbf{c}_i$. The non-negativity in $\mathbf{a}_i$ only allows additive combination of comments, which implicitly minimizes the redundant information (He et al. 2012).

The selection of comments $\boldsymbol{\beta}$ and the coefficients $\{\mathbf{a}_i\}$ can then be learned by minimizing the overall reconstruction error, which can be formulated as

$$\begin{aligned} \min_{A,\boldsymbol{\beta}} \mathcal{L}(A,\boldsymbol{\beta}) &= \sum_{i=1}^{n} ||\mathbf{c}_i - C \text{ diag}(\boldsymbol{\beta}) \ \mathbf{a}_i||^2 \\ &= \|C - C\text{diag}(\boldsymbol{\beta})A^\top\|_F^2 \end{aligned} \tag{2}$$

$$\text{s.t.} \quad \boldsymbol{\beta} \in \{0,1\}^n, \quad \sum_{i=1}^{n} \beta_i = m, \quad A \geq 0$$

where $\|\cdot\|$ is the $\ell_2$ norm of a vector, $||\cdot||_F$ is the Frobenius norm of a matrix and $A = [\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_n]^\top$.

In the meantime, time-sync comments usually contain enormous noise in text, including internet slang terms and misspellings, which brings a challenge to summarization at the text level. To tackle this issue, we further exploit semantic similarity between comments from the *topic* perspective to facilitate the process of comment summarization. To achieve that, each comment can be represented as a topic distribution based on topic decomposition via Latent Dirichlet Allocation (Arora and Ravindran 2008). Compared with the text-level representation of weighted term-frequency, the topic-level representation contains more semantic information in the topic probability space.

To generate a summary that is able to reconstruct the original corpus of video comments at the text level and recover the topics of video semantics conveyed in the comments simultaneously, we propose a framework of Summarization based on Joint Text and Topic Reconstruction (SJTTR). Specifically, for a set of comments, in addition to the weighted term-frequency representation $C$, we also integrate the topic information of the comments in $T = [\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_n] \in \mathbb{R}^{K\times N}$, where $K$ is the number of topics, $\mathbf{t}_i$ is the normalized topic distribution of the $i$-th comment, which can be obtained by applying Gibbs sampling in topic decomposition. Based on the text and topic representations $C$ and $T$, the framework of joint summarization can be formulated as

$$\begin{aligned}\min_{A,B,\boldsymbol{\beta}} &\mathcal{L}(A,B,\boldsymbol{\beta}) = \rho\mathcal{L}_{\mathrm{C}}(A,\boldsymbol{\beta}) + (1-\rho)\mathcal{L}_{\mathrm{T}}(B,\boldsymbol{\beta}) \\ &= \rho\|C - C\mathrm{diag}(\boldsymbol{\beta})A^\top\|_F^2 \\ &+ (1-\rho)\|T - T\mathrm{diag}(\boldsymbol{\beta})B^\top\|_F^2 \\ \text{s.t.} \quad & \boldsymbol{\beta}\in\{0,1\}^n,\ \sum_{i=1}^{n}\beta_i = m, A, B \geq 0 \end{aligned} \tag{3}$$

where $\rho$ is the trade-off parameter between text reconstruction and topic reconstruction; $B = [\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_n]^\top$ and $\mathbf{b}_i$ represents the column coefficient vector of the non-negative linear reconstruction for $\mathbf{t}_i$.

## Temporal Summarization Based on Joint Text and Topic Reconstruction

We further consider the temporal nature of the problem and extend the above joint reconstruction model (3) to generate temporal summarization of time-sync comments.

In general, the subjects of discussion in time-sync comments dynamically change with the video content, while being correlated at adjacent playback time positions. Therefore, to summarize the temporal comments, we first divide the video into sequential segments and then generate a summary for each segment which contains information not conveyed in the previous summaries to reduce the redundancy along the timeline.

To achieve that, we propose a temporal summarization framework that generates summaries of the segments in a sequential manner. Within the framework, in the $k$-th segment, to incorporate the temporal correlations among adjacent segments, we select from the comments in that segment and the summaries generated in the previous $w$ segments which can best reconstruct the comments in the $k$-th segment. The selected comments that belong to the $k$-th segment constitute the summary of the segment, while the non-negative linear reconstruction helps to reduce the redundancy within the segment as well as among the adjacent ones. The model can be formulated as:

$$\begin{aligned}\min_{A^k,B^k,\boldsymbol{\beta}^k} &\ \rho\|C^k - \hat{C}^k\mathrm{diag}(\boldsymbol{\beta}^k){A^k}^\top\|_F^2 \\ &+ (1-\rho)\|T^k - \hat{T}^k\mathrm{diag}(\boldsymbol{\beta}^k){B^k}^\top\|_F^2 \\ \text{s.t.} &\ \boldsymbol{\beta}^k\in\{0,1\}^{\hat{n}^k},\ \sum_{i=1}^{n^k}\beta_i^k = m, \quad A^k, B^k \geq 0 \\ &\ \hat{C}^k = [C^k\cup C^{\mathrm{pre}}],\ \hat{T}^k = [T^k \cup T^{\mathrm{pre}}]\end{aligned} \tag{4}$$

where the basis of text reconstruction $\hat{C}^k$ corresponds to the text representation of the reconstruction corpus including the comments in the $k$-th segment $C^k$ and the summaries of the previous $w$ segments $C^{\mathrm{pre}}$; similarly for the basis of topic reconstruction $\hat{T}^k$. $\hat{n}^k$ denotes the number of comments in the reconstruction corpus, while $n^k$ is the number of comments in the $k$-th segment. The first $m$ non-negative elements of $\boldsymbol{\beta}^k$ indicate the summary of the $k$-th segment.

In addition, with the continuously changing subjects of discussion, the comments which are closer in playback time positions should be more correlated. Therefore we impose a penalty on $\boldsymbol{\beta}^k$ in the objective according to the distance between the current segment and the previous segments in temporal positions:

$$g(\boldsymbol{\beta}^k) = {\boldsymbol{\theta}^k}^\top\boldsymbol{\beta}^k, \tag{5}$$

where $\boldsymbol{\theta}^k$ is the penalty factor of adjacent summaries, and can be calculated as following:

$$\theta_j^k = \begin{cases} e^{(k-s_j-w)/\gamma} & \text{if} \quad \mathbf{c}_j\in C^{\mathrm{pre}} \\ 1 & \text{otherwise}\end{cases} \tag{6}$$

where $s_j$ is the segment index of the $j$-th comments in $\hat{C}^k$, $\gamma$ controls the shape of exponential decay and balances the correlation between the summaries of the previous $w$ segments and the comments in the current segment. $\theta_j^k$ takes values from $[0, 1]$ since $0 \leq k - s_j \leq w$. A summary from a closer segment in time is associated with a smaller $\theta$ value which implies that comments from that summary is more likely to be selected than the comments from the current segment, encouraging redundant information to be removed from the comments in the current segment.

Temporal summarization at the $k$-th segment can thus be formulated as:

$$\begin{aligned}\min_{A^k,B^k,\boldsymbol{\beta}^k} &\ \rho\|C^k - \hat{C}^k\mathrm{diag}(\boldsymbol{\beta}^k){A^k}^\top\|_F^2 \qquad (7)\\ &+ (1-\rho)\|T^k - \hat{T}^k\mathrm{diag}(\boldsymbol{\beta}^k){B^k}^\top\|_F^2 \\ &+ \lambda\, g(\boldsymbol{\beta}^k) \\ \text{s.t.} &\ \boldsymbol{\beta}^k\in\{0,1\}^{\hat{n}^k},\ \sum_{i=1}^{n^k}\beta_i^k = m, \quad A^k, B^k \geq 0 \\ &\ \hat{C}^k = [C^k\cup C^{\mathrm{pre}}],\ \hat{T}^k = [T^k \cup T^{\mathrm{pre}}]\end{aligned}$$

where $\lambda$ is the regularization parameter controlling the temporal shrinkage of $\boldsymbol{\beta}^k$.

The problem above is still difficult to optimize due to the discreteness of $\boldsymbol{\beta}^k$. However we can relax $\boldsymbol{\beta}^k$ to be continuous (Yu et al. 2008), reformulate (7) and obtain the framework of Temporal Summarization based on Joint Text and Topic Reconstruction (T-SJTTR):

$$\begin{aligned}
&\min_{A^k,B^k,\boldsymbol{\beta}^k} \quad \mathcal{L}^k(A^k,B^k,\boldsymbol{\beta}^k)+\lambda g(\boldsymbol{\beta}^k)\\
&=\min_{A^k,B^k,\boldsymbol{\beta}^k} \rho(||C^k-\hat{C}^k{A^k}^\top||_F^2+\sum_{i=1}^{n^k}\sum_{j=1}^{\hat{n}^k}\frac{{a_{ij}^k}^2}{\beta_j^k})\\
&\quad +(1-\rho)(||T^k-\hat{T}^k{B^k}^\top||_F^2+\sum_{i=1}^{n^k}\sum_{j=1}^{\hat{n}^k}\frac{{b_{ij}^k}^2}{\beta_j^k}) \qquad (8)\\
&\quad +\lambda\, {\boldsymbol{\theta}^k}^\top\boldsymbol{\beta}^k\\
&\text{s.t.} \quad \boldsymbol{\beta}^k\geq 0, \quad A^k\geq 0, \quad B^k\geq 0\\
&\qquad \hat{C}^k=[C^k\cup C^{\text{pre}}], \ \ \hat{T}^k=[T^k\cup T^{\text{pre}}].
\end{aligned}$$

Due to the non-negativity of $\boldsymbol{\theta}^k$, adding ${\boldsymbol{\theta}^k}^\top\boldsymbol{\beta}^k$ in the objective will enforce sparsity in $\boldsymbol{\beta}^k$; when $\beta_j^k=0$, the corresponding coefficients $a_{1j}^k,...,a_{nj}^k$ and $b_{1j}^k,...,b_{nj}^k$ must be 0, which implies the $j$-th comment is not selected in the $k$-th segment. Based on the solution of (8), we can generate summaries of the segments in a sequential manner, which provide a temporal description of the video.

## Optimization

In the $k$-th segment, the problem (8) is convex regarding $\boldsymbol{\beta}^k$, $A^k$ and $B^k$, which guarantees a global optimal solution. In this section, we propose an alternating optimization algorithm to solve for $\boldsymbol{\beta}^k$, $A^k$ and $B^k$.

Firstly, when fixing $A^k$ and $B^k$, we can get the analytical solution of $\boldsymbol{\beta}^k$ by setting the derivative of the objective regarding $\boldsymbol{\beta}^k$ to zero:

$$\beta_j^k=\sqrt{\frac{\rho\sum_{i=1}^{n^k}{A_{ij}^k}^2+(1-\rho)\sum_{i=1}^{n^k}{B_{ij}^k}^2}{\lambda\ \theta_j^k}}. \qquad (9)$$

To solve for $A^k$ with non-negative constraints, the update can be obtained by using the Lagrange method and minimizing the following:

$$J^k=\mathcal{L}^k(A^k,B^k,\boldsymbol{\beta}^k)+\text{tr}(U^k{A^k}^\top) \qquad (10)$$

where $U^k=[u_{ij}^k]\geq 0$ is the Lagrange multiplier for $A^k$. By analyzing the Karush-Kuhn-Tuker condition (Boyd and Vandenberghe 2004), we can get the update rule for $a_{ij}^k$:

$$a_{ij}^k\leftarrow\frac{(\hat{C^k}^\top C^k)_{ij}}{(A^k\hat{C^k}^\top C^k+A^k\text{diag}(\boldsymbol{\beta}^k)^{-1})_{ij}}a_{ij}^k. \qquad (11)$$

Similarly, the update rule for $b_{ij}^k$ is

$$b_{ij}^k\leftarrow\frac{(\hat{T^k}^\top T^k)_{ij}}{(B^k\hat{T^k}^\top T^k+B^k\text{diag}(\boldsymbol{\beta}^k)^{-1})_{ij}}b_{ij}^k. \qquad (12)$$

The overall optimization procedure is summarized in Algorithm 1. The optimization problem (8) is convex, and the alternating optimization procedure in Algorithm 1 solves for a variable by minimizing the objective value when the other two variables are fixed (Sha et al. 2007), as a consequence the objective function is non-increasing during the iterations and the algorithm will converge. Assuming the maximum numbers of iterations for step (4) and inner loop (5-9) are $K_1$ and $K_2$ respectively, the total computational cost for Algorithm 1 is $O(K_1n^k(1+K_2\hat{n^k}^2))$.

**Algorithm 1** Temporal Summarization Based on Joint Text and Topic Reconstruction (T-SJTTR)

```
Input:
    Comments of all $v$ segments with text representation: $C = [C^1, C^2, ..., C^v]$;
    Comments of all $v$ segments with topic representation: $T = [T^1, T^2, ..., T^v]$
    Parameters : $\rho$ , $\gamma$ , $\lambda$
Output: summaries $X = [X^1, X^2, ..., X^v]$ for all the segments
 1: for each $k \in [1, v]$ do
 2:     Initialize $\boldsymbol{\beta}^k$, $A^k$ and $B^k$
 3:     repeat
 4:         Update $\boldsymbol{\beta}^k$ according to equation (9)
 5:         repeat
 6:             Update $A^k$ according to equation (11)
 7:             Update $B^k$ according to equation (12)
 8:         until converge
 9:     until converge
10:     $X^k \leftarrow \{\mathbf{x}_i^k | \mathbf{x}_i^k \in C^k, \beta_i^k$ is the top-$m$ non-zero element of $\boldsymbol{\beta}_{1:n_k}^k\}$
11: end for
12: $X = [X^1, X^2, ..., X^v]$
```

Table 1: Datasets of time-sync video comments

| Videos | #. TSCs | #. Frames |
|---|---|---|
| QPS-EP06 | 27653 | 3721 |
| QPS-EP07 | 28812 | 3856 |
| QPS-EP12 | 30307 | 4187 |
| LYB-EP07 | 51991 | 2639 |
| LYB-EP26 | 48508 | 2637 |
| LYB-EP40 | 46732 | 2637 |

## Experiments

### Datasets

To evaluate the proposed methods, we collect time-sync comments from "iQIYI"[4], which is one of the largest video websites in China. We consider two types of video data, variety shows and TV-series, which usually give rise to enthusiastic discussions and appropriate to be exploited for video description. We construct two datasets of time-sync comments on a phenomenal Chinese debate show known as "Qi Pa Shuo" and the latest TV series "Lang Ya Bang", both of which have gained wide popularity and attracted large numbers of comments. Specifically, we choose three episodes

[4] http://www.iqiyi.com/

Table 2: The average F-measures of ROUGE-1 and ROUGE-2. A bold number indicates the highest ROUGE score.

| | QPS-EP06 | | QPS-EP07 | | QPS-EP12 | | LYB-EP07 | | LYB-EP26 | | LYB-EP40 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R-1 | R-2 | R-1 | R-2 | R-1 | R-2 | R-1 | R-2 | R-1 | R-2 | R-1 | R-2 |
| Random | 0.2353 | 0.0526 | 0.2374 | 0.0583 | 0.2443 | 0.0616 | 0.2694 | 0.0583 | 0.2611 | 0.0657 | 0.2573 | 0.0547 |
| ClusterHITS | 0.2808 | 0.0985 | 0.2816 | 0.0993 | 0.3033 | 0.1045 | 0.3108 | 0.1013 | 0.3213 | 0.1167 | 0.3087 | 0.0973 |
| LexRank | 0.2787 | 0.0968 | 0.2932 | 0.1082 | 0.2908 | 0.1026 | 0.3076 | 0.0966 | 0.3142 | 0.1128 | 0.2985 | 0.0926 |
| DSDR | 0.3356 | 0.1172 | 0.3407 | 0.1218 | 0.3346 | 0.1288 | 0.3386 | 0.1207 | 0.3511 | 0.1202 | 0.3353 | 0.1198 |
| TopicDSDR | 0.2793 | 0.0977 | 0.2976 | 0.1116 | 0.3067 | 0.1102 | 0.3150 | 0.1064 | 0.3052 | 0.1097 | 0.3043 | 0.0954 |
| SJTTR | **0.3682** | **0.1375** | **0.3761** | **0.1463** | **0.3874** | **0.1404** | **0.3775** | **0.1413** | **0.3913** | **0.1475** | **0.3704** | **0.1385** |
| T-SJTTR | **0.3758** | **0.1402** | **0.3895** | **0.1525** | **0.3982** | **0.1486** | **0.3969** | **0.1517** | **0.4095** | **0.1583** | **0.3843** | **0.1447** |

from “Qi Pa Shuo” and three from “Lang Ya Bang” with maximum number of comments.

The tokenization and stemming of comments are completed with ICTCLAS[5], a Chinese natural language processing toolbox. After that, we filter out stopwords and only keep the comments with more than 3 terms. The details of the datasets after pre-processing are summarized in Table 1.

## Evaluation Metrics

We take ROUGE (Recall-Oriented Understudy for Gisting Evaluation) as our main evaluation criterion. The summarization quality is measured by counting the overlapping units, such as $n$-grams, word sequences between the candidate summaries generated by algorithms and the reference summaries created by human. To provide the reference summaries, we ask 5 students who are big fans of TV shows to watch the video segments and select the most representative comments in the discussions. The length of summary in each segment is limited to 250 words.

There are several evaluation measures implemented in the ROUGE toolkit (Lin 2004). Among them, ROUGE-N is an $n$-gram recall metric which is computed as follows:

$$\mathrm{ROUGE-N} = \frac{\sum_{S \in \mathrm{Ref}} \sum_{\mathrm{gram}_n \in S} \mathrm{Count}_{\mathrm{match}}(\mathrm{gram}_n)}{\sum_{S \in \mathrm{Ref}} \sum_{\mathrm{gram}_n \in S} \mathrm{Count}(\mathrm{gram}_n)}$$

where $n$ denotes the length of the $n$-gram and $\mathrm{Ref}$ is the set of reference summaries. $\mathrm{Count}_{\mathrm{match}}(\mathrm{gram}_n)$ is the maximum number of $n$-grams co-occurring in the candidate summary and the set of reference summaries, $\mathrm{Count}(\mathrm{gram}_n)$ is the number of $n$-grams in the reference summaries. The previous implementation in the ROUGE toolkit does not support Chinese very well. In our experiments, we apply the ROUGE 2.0 toolkit[6] to evaluate the summarization quality which can also report scores for unigram, bigrams, trigrams, etc. Among the various scores, ROUGE-1 has been shown to be the most consistent with human judgement (Lin and Hovy 2003). ROUGE can generate three types of scores including recall, precision and F-measure. In our experiments, we find similar trends among the three criterions, and for simplicity we use F-measure as the representative.

## Algorithms for Comparison

As our method is totally unsupervised, we compare with the following widely used unsupervised summarization algorithms as baselines in our experimental investigation:

**Random**, the method that selects sentences randomly for each segment as the summary.

**ClusterHITS** (Wan and Yang 2008), the method that considers topic clusters as hubs and sentences as authorities. It selects sentences according to authority scores.

**LexRank** (Erkan and Radev 2004), a graph-based summarization method which selects sentences based on eigenvector centrality.

**DSDR** (He et al. 2012), a summarization model based on a non-negative linear reconstruction where sentences are selected to represent the original documents by minimizing the reconstruction error.

**TopicDSDR** (Zhang, Li, and Huang 2013), an extension of DSDR which applies topic reconstruction to summarize the documents.

The parameters for the above algorithms are chosen with line search for optimal performance in the experiments. For topic decomposition, the number of topics $K$ is set to 20.

## Overall Performance

Empirical investigation of all the algorithms is carried out on two different shows described in Table 1. For “Qi Pa Shuo”, we divide the videos into segments with a length of 300 frames, while for “Lang Ya Bang”, the length is set to 200 frames. Here, we measure the ROUGE scores for every segment and then average over all $v$ segments to evaluate the overall performance. The results are summarized in Table 2. We can observe that the proposed methods including SJTTR and T-SJTTR outperform the other algorithms with all evaluation metrics. Random selection performs the worst among all algorithms, which is not surprising. Compared to LexRank, ClusterHITS considers topics as hubs and sentences as authorities where hubs and authorities can interact with each other, which improves the quality of summarization. However, the ROUGE scores of both LexRank and ClusterHITS are lower than the methods based on data reconstruction most of the time due to the fact that they select the top ranked sentences which usually share much redundancy within segments. Meanwhile, the advantages of the proposed SJTTR and T-SJTTR methods over DSDR and TopicDSDR justify the idea of selecting sentences which reconstruct the original comments from the text perspective and topic perspective simultaneously. In addition, by exploiting the temporal nature in data, T-SJTTR improves over SJTTR with a temporal regularization to remove redundancy along the timeline.

[5] http://ictclas.nlpir.org/

[6] https://bitbucket.org/kganes2/rouge-2.0/

Table 3: Selected comments and the corresponding video plots along the timeline of a segment in “LYB-EP26”

| | | | | |
|---|---|---|---|---|
| <br>22:11 | <br>23:06 | <br>24:20 | <br>25:27 | <br>25:51 |
| Nian Nian has slanted eyebrows. | Mr. Su used to be Brother Su. How sad. | I feel sorry for both Jing Rui and Su from their conversation. | Mr. Su lost a good friend forever! | Jing Rui leaves the place where his heart broke and dreams faded away. |
| **keywords by LDA:** ‘Su’ , ‘say’ , ‘really’ , ‘Jing Rui’ , ‘leave’ , ‘eyebrow’ , ‘Nian Nian’ , ‘come’ , ‘friends’ , ‘love’ | | | | |

We further evaluate the quality of summary with varying size of segment. We take “LYB-EP26” and split the video into segments with a length of 200, 250, 300 frames respectively, ROUGE-1 is used as the evaluation criterion. Figure 1 shows the performance of various algorithms with different segment sizes. Overall, the F-measures of all the algorithms decrease when the segment size grows, which is reasonable since a larger segment contains more information to cover. In all the settings, the proposed methods outperform the baseline algorithms significantly.

## Influence of Parameters

In our experiments, we set $\lambda = 200$ and choose $w = 4$ as window size on all the datasets. Here we investigate the influence of parameters $\rho$ and $\gamma$ on the performance of the proposed algorithms. We take “LYB-EP26”, and in each of the following experiment, we vary one parameter while keeping the others fixed.

**Influence of** $\rho$. In our framework, $\rho$ controls the tradeoff between text reconstruction and topic reconstruction. To examine the effect of $\rho$, we gradually increase $\rho$ from $0$ to $1$ with a step size $0.1$. Figure 2(a) shows the curve of the F-measure versus the $\rho$ value, where we can observe that the F-measure reaches the peak and remains reasonably stable when $\rho \in [0.4, 0.7]$. In our experiments, we take $\rho = 0.5$ as the balance factor on all the datasets.

**Influence of** $\gamma$. In the proposed model of temporal summarization, $\gamma$ controls the shape of exponential decay and hence the influence from adjacent summaries. To examine the effect of $\gamma$, we increase $\gamma$ from $0.25$ to $100$ and plot the curve of the F-measure in Figure 2(b). We can observe that very small $\gamma$ values imply intense influence from adjacent segments which may cause performance decay; while when $\gamma > 20$, the temporal influence gradually diminishes. In our experiment, we choose $\gamma = 0.8$ on all the datasets.

## Case Study

As further illustration, we randomly pick a video segment in “LYB-EP26” to visualize the practical results. In Table 3 we list the top 5 selected comments and display the corresponding plots of the video in the first row. For comparison, we also include keywords extracted by LDA in the last row. From Table 3, we can observe that the selected comments provide very informative and consistent description of the video content in a time-sync manner, including the roles “Jing Rui”, “Nian Nian”, “Su”. On the other hand, keywords are not sufficient to describe the complicated video content which includes a number of characters and scenes.

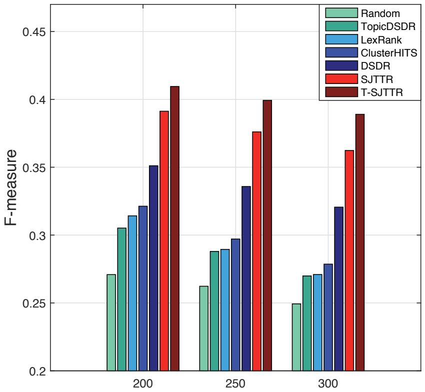
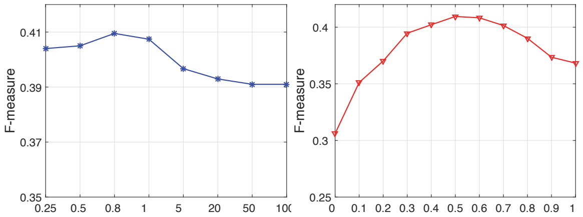


Figure 1: F-measure with different segment sizes



(a) $\rho$ (b) $\gamma$

Figure 2: Influence of parameters

## Conclusion

In this paper, we propose a novel method that generates temporal descriptions of videos by summarizing crowdsourced time-sync comments. The proposed model integrates text and topic reconstruction simultaneously to resolve the issues of informal, noisy and redundant information contained in the time-sync comments. In addition, we consider the temporal nature of the problem with dynamic summarization of comments in sequential segments. Experimental results and case study justify the idea of exploiting crowdsourced time-sync comments to describe videos.

## Acknowledgments

Research supported by the National Natural Science Foundation of China (No. 61375060 and No. 61673364).

## References

Arora, R., and Ravindran, B. 2008. Latent dirichlet allocation based multi-document summarization. In *Proceedings of the 2nd Workshop on Analytics for Noisy Unstructured Text Data*, 91–97.

Boyd, S. P., and Vandenberghe, L. 2004. *Convex optimization*. Cambridge university press.

Chen, Z.; Xu, L.; Chen, E.; Chang, B.; Wang, Z.; and Li, Y. 2015. Selecting social media responses to news: A convex framework based on data reconstruction. In *Proceedings of the 15th SIAM International Conference on Data Mining*.

Erkan, G., and Radev, D. R. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* 22:457–479.

Gong, Y., and Liu, X. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 19–25.

He, Z.; Chen, C.; Bu, J.; Wang, C.; Zhang, L.; Cai, D.; and He, X. 2012. Document summarization based on data reconstruction. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, 620–626.

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 168–177.

Hu, M., and Liu, B. 2006. Opinion extraction and summarization on the web. In *Proceedings of the 21st AAAI Conference on Artificial Intelligence*, 1621–1624.

Inouye, D., and Kalita, J. K. 2011. Comparing twitter summarization algorithms for multiple post summaries. In *SocialCom/PASSAT*, 298–306.

Lin, C.-Y., and Hovy, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of Human Language Technology Conference (HLT-NAACL 2003)*, 71–78.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 74–81.

Lv, G.; Xu, T.; Chen, E.; Liu, Q.; and Zheng, Y. 2016. Reading the videos: Temporal labeling for crowdsourced time-sync videos based on semantic embedding. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 721–730.

Sha, F.; Lin, Y.; Saul, L. K.; and Lee, D. D. 2007. Multiplicative updates for nonnegative quadratic programming. *Neural Computation* 19(8):2004–2031.

Shou, L.; Wang, Z.; Chen, K.; and Chen, G. 2013. Sumblr: continuous summarization of evolving tweet streams. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 533–542.

Siersdorfer, S.; San Pedro, J.; and Sanderson, M. 2009. Automatic video tagging using content redundancy. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 395–402.

Ulges, A.; Schulze, C.; Koch, M.; and Breuel, T. M. 2010. Learning automatic concept detectors from online video. *Computer Vision and Image Understanding* 114(4):429–438.

Wan, X., and Yang, J. 2008. Multi-document summarization using cluster-based link analysis. In *Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval*, 299–306.

Wang, D.; Li, T.; Zhu, S.; and Ding, C. 2008. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval*, 307–314.

Wu, Z., and Ito, E. 2014. Correlation analysis between user's emotional comments and popularity measures. In *IIAI 3rd International Conference on Advanced Applied Informatics (IIAIAAI)*, 280–283.

Wu, B.; Zhong, E.; Tan, B.; Horner, A.; and Yang, Q. 2014. Crowdsourced time-sync video tagging using temporal and personalized topic modeling. In *Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 721–730.

Xian, Y.; Li, J.; Zhang, C.; and Liao, Z. 2015. Video highlight shot extraction with time-sync comment. In *Proceedings of the 7th International Workshop on HotPOST*, 31–36.

Yan, R.; Kong, L.; Huang, C.; Wan, X.; Li, X.; and Zhang, Y. 2011. Timeline generation through evolutionary trans-temporal summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 433–443.

Yu, K.; Zhu, S.; Xu, W.; and Gong, Y. 2008. Non-greedy active learning for text categorization using convex ansductive experimental design. In *Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval*, 635–642.

Zhang, Z.; Li, H.; and Huang, L. 2013. Topicdsdr: combining topic decomposition and data reconstruction for summarization. In *Proceedings of the 14th International Conference on Web-Age Information Management*. 338–350.