






Character-Oriented Video Summarization With Visual and Textual Cues

Peilun Zhou , Tong Xu , *Member, IEEE*, Zhizhuo Yin, Dong Liu , *Senior Member, IEEE*, Enhong Chen , *Senior Member, IEEE*, Guangyi Lv , and Changliang Li

Abstract—With the booming of content “re-creation” in social media platforms, *character-oriented* video summary has become a crucial form of user-generated video content. However, artificial extraction could be time-consuming with high missing rate, while traditional techniques on person search may incur heavy burden of computing resources. At the same time, in social media platforms, videos are usually accompanied with rich textual information, e.g., *subtitles* or *bullet-screen comments* which provide the multi-view description of videos. Thus, there exists a potential to leverage textual information to enhance the character-oriented video summarization. To that end, in this paper, we propose a novel framework for jointly modeling visual and textual information. Specifically, we first locate characters indiscriminately through detection methods, and then identify these characters via re-identification to extract potential key-frames, in which appropriate source of textual information will be automatically selected and integrated based on the features of specific frame. Finally, key-frames will be aggregated as the character-oriented summarization. Experiments on real-world data sets validate that our solution outperforms several state-of-the-art baselines on both person search and summarization tasks, which prove the effectiveness of our solution on the character-oriented video summarization problem.

Index Terms—Character-oriented video summarization, person search, natural language processing.

I. INTRODUCTION

RECENT years have witnessed the development of online social media platforms, which leads to the boom of user-generated “re-creation” contents based on original videos. Among them, there exists a popular form called “**character-oriented summarization**”, namely the summarization of video clips in which specific character appear. Usually, fans are keen

Manuscript received June 14, 2019; revised October 18, 2019 and November 26, 2019; accepted December 6, 2019. Date of publication December 18, 2019; date of current version September 23, 2020. This work was supported in part by grants from the National Key Research and Development Program of China under Grant 2018YFB1402600, and in part by the National Natural Science Foundation of China under Grants 61703386, U1605251, and 61931014. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Shaoen Wu. (Peilun Zhou and Tong Xu contributed equally to this work.) (Corresponding author: Tong Xu; Enhong Chen.)

P. Zhou, T. Xu, Z. Yin, D. Liu, E. Chen, and G. Lv are with the Anhui Province Key Lab of Big Data Analysis and Application, School of Computer Science, University of Science and Technology of China, Hefei 230026, China (e-mail: zpl@mail.ustc.edu.cn; tongxu@ustc.edu.cn; yzz1223@mail.ustc.edu.cn; dongeliu@ustc.edu.cn; chenh@ustc.edu.cn; gylv@mail.ustc.edu.cn).

C. Li is with the Kingsoft AI Lab, Beijing 100085, China (e-mail: lichangliang@kingsoft.com).

Color versions of one or more of the figures in this article are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TMM.2019.2960594

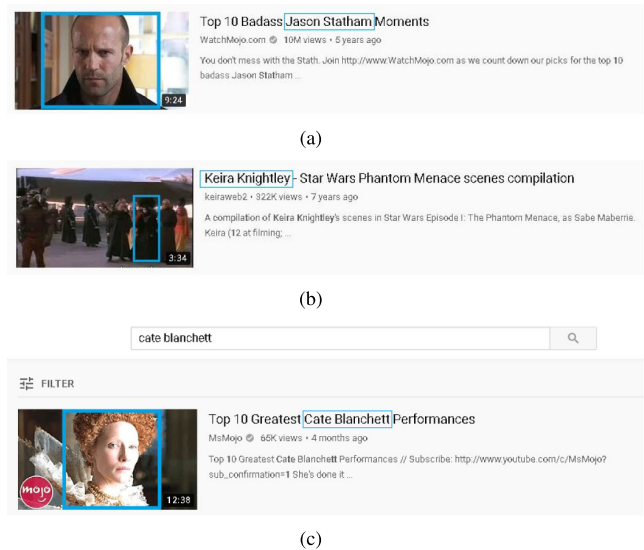


Fig. 1. Examples of character-oriented video clips.

on summarizing the clips of their favorite star from movies or TV series. For instance, as shown in Figure 1, both selective clips among multiple videos (e.g., Figure 1(a)), and scenes compilation from one certain video (e.g., Figure 1(b)), have attracted thousands or even millions of views in Youtube. Specifically, we randomly selected 20 famous characters as queries when searching in Youtube, and 1.07 compilation/clip videos appear in the top 10 results in average, even ranked as the first result as shown in Figure 1(c). This phenomenon indicates that the character-oriented summary could be attractive for massive users, while at the same time, raises the significant challenge for generating video summarization effectively and efficiently. Unfortunately, manual summarization could be time-consuming with high missing rate. Therefore, adequate techniques to automatically extract the character-oriented summarization are urgently required.

Indeed, character-oriented video summarization task is quite different from the traditional video summarization. An ordinary video summary is expected to consist of important or interesting clips of a long video [1]. However, a character-oriented video summary, as can be observed in Fig. 1, should consist of the clips in which the specific character appears. Thus, to fulfill character-oriented summarization, it is indispensable to identify the clips of a specific person from a long video, i.e.

to perform *person search* [2]–[4]. The person search task has been tackled by, for example, joint modeling of detection and re-identification [5], or the memory-guided model based on Convolutional LSTM [6]. However, these prior arts are designed for a different scenario, e.g. surveillance video analysis. Thus, they mainly focus on the person search with relatively static pose and background, or even similar clothing. In the scenario of character-oriented summarization, both background and poses of characters are always changing, as well as the different clothing in different scenarios, which extremely increase the difficulty. Obviously, current person search techniques should be further enhanced.

At the same time, we realize that in social media platforms, videos are usually accompanied with rich textual information, which may benefit the understanding of media content [7]. Especially, with the so-called “bullet-screen comments” [8], [9] (i.e., comments flying across screen like bullets), namely the time-sync feedbacks of massive users, more comprehensive or even subjective description could be achieved, which results in more explicit cue to capture the characters. For instance, when *Sheldon* (character in *The Big Bang Theory*) appears on the scene, we could always see “*Bazinga*”, the famous pet phrase in bullet-screen comments. Therefore, one can see that full advantage of textual information may benefit the character-oriented summarization with better effectiveness.

To that end, in this paper, we propose a novel framework for jointly modeling visual and textual information for character-oriented summarization. To be specific, we first locate characters indiscriminately through detection methods, then identify these characters via re-identification module to extract potential key-frames, and finally aggregate the frames as summarization. Moreover, as multi-source textual information, i.e., the *subtitles and bullet-screen comments* are utilized, we further design a selection module to automatically select and integrate the appropriate source of textual information based on the visual feature of frames, so that the function of textual information could be further refined. In general, the contribution of this paper can be summarized as follows:

- To the best of our knowledge, we are among the first ones who study the character-oriented summarization with considering textual information.
- We propose a novel framework for jointly modeling visual and textual information, in which appropriate source of text could be automatically selected.
- Experiments on real-world datasets validate that our solution outperforms several state-of-the-art baselines, and further reveal some rules of the semantic matching between characters and textual information.

II. RELATED WORK

In this section, we will summarize the prior arts in following three fields which are related to our task, namely *video summarization*, *person search* and *multimodal learning* methods.

Video Summarization. In general, video summarization task aims at producing a compact visual summary, which encapsulates the main content of given videos, e.g., the highlight

shots [10]–[12], or the clips that match a special topic [13] or description [4]. Usually, *salient detection* is utilized to measure the importance or matching degree of one certain frame, which relies on the bottom-up image cues (i.e., intensity, color, texture, etc.) [1] in the early stage. Recently, some object-driven summarization techniques like [14], [15] attempted to learn the high-level saliency to reveal some special themes with more semantic cues, or even explored the ranking of significant objects in static images based on the order of mention by artificially labeled tags [16], [17]. At the same time, some query-driven summarization techniques like [18]–[20] summarized multiple groups of video based on user queries as desired “viewpoint”. Besides, prior arts like [21] also developed the first streaming algorithm for real-time video summarization with various personalization constraints. However, these prior arts mainly focus on semantic or topic-related objects for summarization, while few of them targets at summarizing the character-oriented clips.

Person Search. Correspondingly, the person search task aims at locating a specific person in a scene given a query image [6], [22], [23], which usually can be seen as a combination of pedestrian detection and re-identification (re-ID) modules. For the *pedestrian detection* module, traditional methods usually depended on the hand-crafted features for description, which are now enhanced by the deep learning techniques. For instance, the R-CNN architectures are adapted to achieve remarkable results by applying proper adaptations [24], [25]. Recently, some methods further enhanced the performance with cascade extension [26]. Similarly, for the *re-ID* module, early works also focus on the feature designing [27], [28] and distance metric learning [29], [30]. Recently, some advanced techniques were proposed, e.g., the KPM module [31] to recover probabilistic correspondences between two images for similarity estimation, the triplet loss [32] to improve the efficiency of training, the multi-level factorization [33] to factorize the visual appearance of a person into latent discriminative factors at multiple semantic levels without manual annotation, and the jointly optimizing [34] adapted to multi-task learning via attentional network. With combining these two modules above, the person search solutions are widely used in the online instance matching (OIM) tasks in a joint learning way [5]. However, some other researches adapted these two modules separately [35], since they respectively focus on the inter-class and intra-class difference. In this paper, we follow the idea that two modules are designed separately.

Multimodal Learning. Besides, in order to accomplish various types of multimedia analysis tasks [36], [37], some multimodal learning methods are also adapted to visual-textual union. For instance, M-DBM [38] utilized a deep Boltzmann Machine to create fused representations across modalities, Lajugie *et al.* [39] attempted to learn a Mahalanobis distance to perform alignment of multivariate time series, and Lv *et al.* [40] designed a video understanding framework to assign temporal labels on highlighted video shots via textual summarization. Also, Structured VSE [41] proposed a contrastive learning approach for the effective learning of fine-grained alignment from image-caption pairs, and DSM [42] inferred the latent emotional state through multimodal network on sentiment recognition task. These above

TABLE I
MATHEMATICAL NOTATIONS

Notation	Description
\mathbf{C}	The target character for video summarization
\mathbf{V}	A video composed by frames
\mathbf{D}_v	The set of textual document for video \mathbf{V}
\mathbf{Q}_c	The input query for target character \mathbf{C}
f_t	The visual frame with timestamp t
d_t	The textual document with timestamp t
$\{RoI\}_t$	The set of "regions of interest" (RoIs) with timestamp t
T_t	The time window around timestamp t as $[t - m, t + n]$
D_t	The set of textual document within time window T_t

approaches inspire us to summarize videos by combining features across modalities.

Different from the prior arts, we target at summarizing the videos which focus on the specific characters, i.e., the different objective of traditional summarization task, and further jointly model the visual and textual information to resolve the difficulty of person search in dynamic scenario and status.

III. TECHNICAL FRAMEWORK FOR CHARACTER-ORIENTED SUMMARIZATION

In this section, we will formally define our problem with preliminaries, and then introduce our framework in details, including the design of modules step by step.

A. Preliminary With Problem Definition

As mentioned above, we target at solving the character-oriented summarization task, i.e., summarizing the video clips in which specific character appear. Therefore, we have $\mathbf{V} = \{f_t\}$ to present a *video*, also a streaming collection of frame f_t with related timestamp t . Along this line, each frame may contain several *regions of interest* (RoIs) as $\{RoI\}_t$, in which each *RoI* indicates a *bounding box* that contain a specific character. Definitely, $\{RoI\}_t = \emptyset$ indicates that no character appear in the frame f_t .

At the same time, as we jointly model the textual information to enhance the summarization, correspondingly, we have the time-sync documents like subtitles or bullet-screen comments, which are presented as $\mathbf{D}_v = \{d_t\}$ to indicate the set of textual documents. Specifically, considering that for the frame f_t , semantically related text may exist within the adjacent periods around time t . Thus, intuitively, we define the *time window* as $T_t = [t - m, t + n]$ to capture all the related textual documents for the frame f_t . The appropriate length of time window will be discussed in experiments.

Finally, for the character-oriented summarization task, given the target character \mathbf{C} , some frames will be captured since their *RoI* are labeled as containing \mathbf{C} . Along this line, we collect all the labeled RoI_t , with corresponding set of textual document within time window T_t as D_t , to form the query of character \mathbf{C} as $\mathbf{Q}_c = \{ \langle RoI_t, D_t \rangle \}_c$ for the video summarization task.

Specifically, all the related mathematical notations are summarized in Table I, and the character-oriented summarization problem for video \mathbf{V} and character \mathbf{C} could be formally defined as follows:

Definition 1: Problem Definition. Given the video \mathbf{V} with textual information \mathbf{D}_v , and the pre-labeled query \mathbf{Q}_c which contains the target character \mathbf{C} , we aim to generate a summarization video \mathbf{S} , which is composed by all the frames in \mathbf{V} that contain the target character \mathbf{C} .

B. Overview of Technical Framework

To deal with the problem above, in this paper, we propose a framework which contains three modules as illustrated in Figure 2, i.e., *detection*, *re-identification* and *aggregation*, whose functions are briefly introduced as follows:

- 1) First, we have **Detection** module to indiscriminately locate characters in \mathbf{V} to produce RoIs, which could be treated as pre-processing part.
- 2) As RoIs are produced, we have **re-identification** module to identify whether one RoI contains the target character \mathbf{C} based on the query \mathbf{Q}_c , and then the potential key-frames with character \mathbf{C} will be collected.
- 3) Finally, we have **Aggregation** module to summarize all the collected key-frames as \mathbf{C} -oriented summarization of \mathbf{V} , based on time interval and density.

The technical details will be introduced in the following subsections. What should be noted is that the textual information will be integrated in the re-identification module with automatic source selection mechanism. Solution for textual processing will be explained in Section IV.

C. Detection Module

As the pre-processing step, the **Detection** module targets at capturing all the potential RoIs with any character. To ensure the high recall rate, we try to adapt Faster R-CNN detector [43] due to its strong capability of detecting varying sized objects in unconstrained scenes; we also utilize the state-of-the-art Cascade R-CNN [26] detector to address the overfitting and inference-time mismatch problem for performance enhancement. It is worth noting that the detectors could be replaced flexibly. To be specific, all the characters are located indiscriminately without distinction during this stage.

At the same time, to further enhance the performance, we simplify the Detection module as a binary classifier, i.e., whether one *RoI* contains a character (no matter who) or *NOT*. Along this line, quality of classifier will be ensured with sufficient training data of character-oriented frames.

D. Re-Identification (Re-ID) Module

Given the RoIs which are extracted in *Detection* module, in order to identify target character \mathbf{C} , we formulate a textual-combined **re-identification** (re-ID) module for robust one-to-one matching based on both visual and semantic cues, which is illustrated in Figure 3. Specifically, we adapt the Multi-scale Deep Kronecker-Product Matching method (KPM) [31] as the

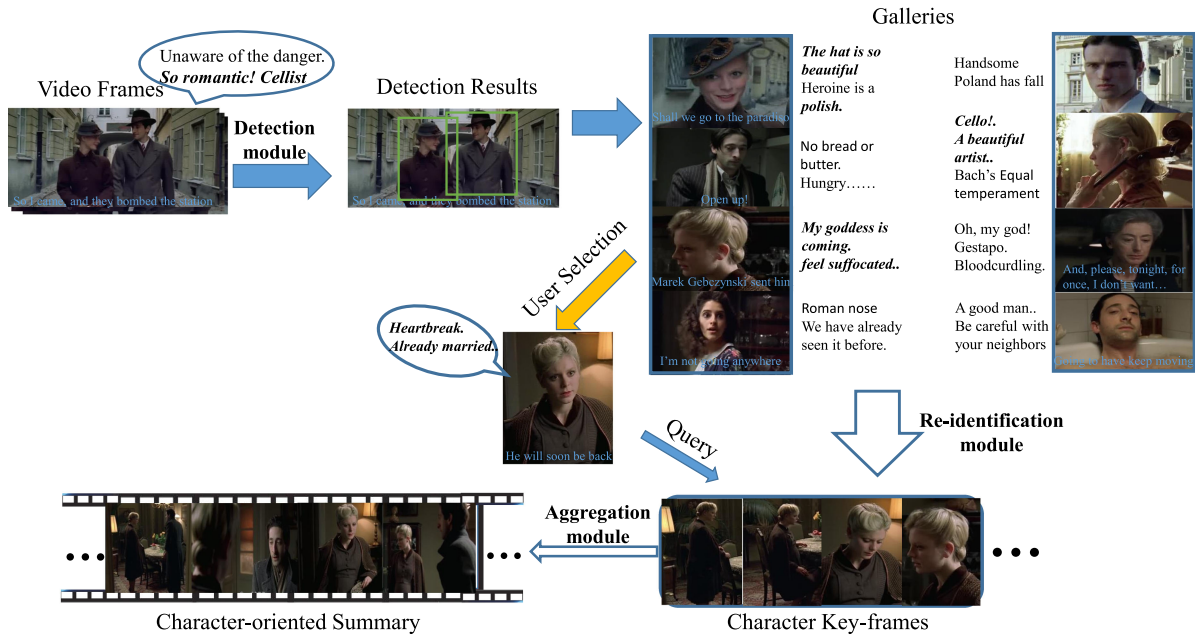


Fig. 2. Pipeline of our character-oriented video summarization system.

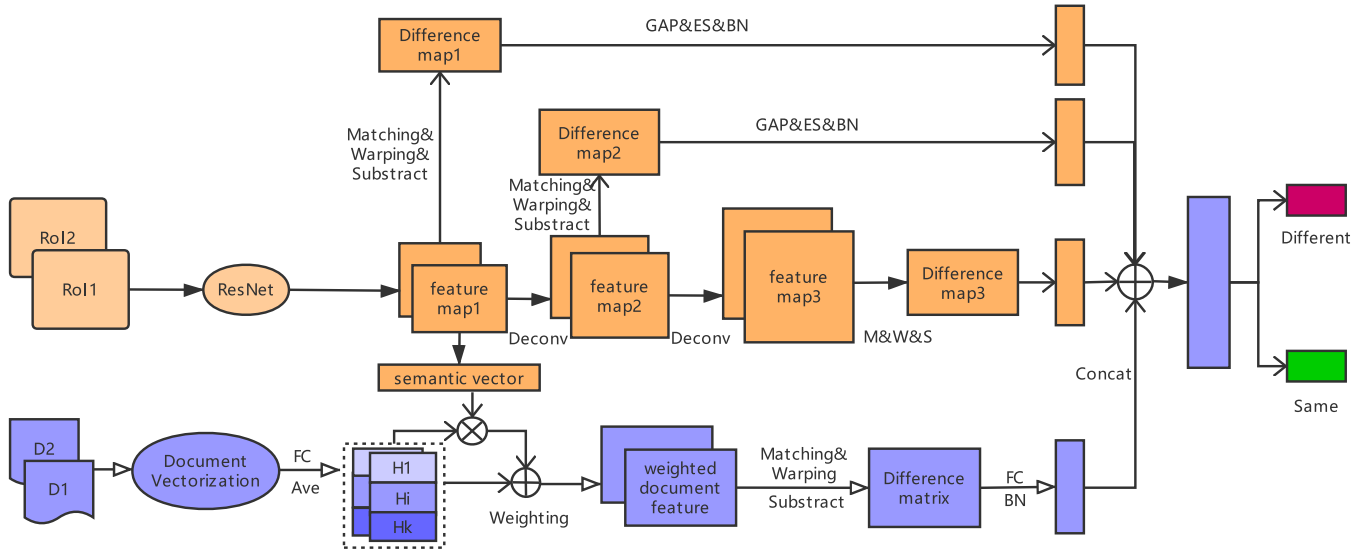


Fig. 3. Backbone for generating multi-scale feature maps with Kronecker-Product Matching, combining with semantic embedding in the branch, in which “GAP” denotes global average pooling, “Ave” denotes average operation, “ES” denotes elementwise square, and “Deconv” denotes deconvolution. The components of Document Vectorization includes: (1) Character-level LSTM (C-LSTM), (2) Neural Topic Model (NTM), and (3) Skip-gram model. Here NTM is illustrated in section IV-A, while the typical C-LSTM and Skip-gram model are detailed in section V-B.

backbone of our model to extract pair-wise visual features, which achieved a remarkable performance on pedestrian re-ID task. Along this line, the textual information is joined as extended embeddings, which will be explained in section IV.

Basically, KPM [31] adopted a deep CNN to generate multi-scale feature maps, and then match the feature pairs based on KPM module to produce feature difference maps for similarity estimation. What should be noted is that, to simplify the model, we replace the feature extractor in KPM, namely ResNet-50 with ResNet-34 [44]. Moreover, we add feature difference maps in size of 64×32 (scale-4) to capture more detailed and intuitive

features to estimate the visual difference vectors, while the rest parts of model remain unchanged. Finally, the distinction for target character C will be executed based on the joint feature vector of adapted KPM model and textual embedding.

E. Aggregation Module

After the distinction in *re-ID* module, we now obtain the potential key-frames which contain the target character C . If following the strict definition of *character-oriented summarization*, we should simply splice all the key-frames as the

summarization video. However, considering the visual effect that viewers may prefer to fluent videos with continuous story, we decide to “tolerate” some frames without target character which connect two extracted clips.

To that end, we first combine all the adjacent key-frames to form several “clips”. Then, if the interval between two clips is shorter than a pre-defined threshold, two clips as well as the interval part between them will be all merged as a new clip. Finally, to ensure the quality of clips, we investigate the “density” of clip defined as $\rho_s = \frac{|s_f|}{|s|}$, in which s_f indicates the amount of frames which contain the target character C , while $|s|$ indicates the total amount of frames in the clip. Correspondingly, if ρ_s is lower than a pre-defined threshold, the clip will be abandoned. Finally, we combine all the reserved clips as the character-oriented summarization video. Details of threshold setting are shown in section V-F.

IV. SOLUTION OF TEXT PROCESSING

As mentioned above, multi-source textual information is utilized to enhance the distinction in re-ID module. In this section, we will introduce the detailed solution of textual embedding, as well as the mechanism for automatic selection mechanism of appropriate textual source.

A. Document Vectorization

As an initial step, we first attempt to vectorize each *document*, i.e., a piece of subtitle or bullet-screen comment for following semantic embedding. Intuitively, considering the strong logic and normality of *subtitles*, they could be easily vectorized by Skip-gram model with negative sampling [45]. However, for the *bullet-screen comments* which are generated by massive users, since this novel type of comments could be short text with informal expression or even slangs, we attempt to vectorize them by following two methods:

- **Character-level LSTM**, which is utilized to deal with the informal expression in bullet-screen comments. Here we use a 3-layer character-level LSTM [46] to model the sequential characteristics in bullet-screen comments.
- **Neural Topic Model**, which is adopted to extract hidden themes from bullet-screen comments. Here we use the VAE-based [47] Neural Topic Model as extractor due to its strong capability to map documents to posterior distributions. As shown in Figure 4, we initialize each document as vector $d^v \in R^W$, where W is the vocabulary size, each element of d^v indicates the frequency of one character in the document. Followed by element-wise function $f : d^v \mapsto [0, 1]^W$ for normalization as follows:

$$g(d^v) = \log_{10}(1 + d^v) \quad (1)$$

$$f(d^v) = \frac{g(d^v)}{\|g(d^v)\|_2} \quad (2)$$

Where $g(\cdot)$ indicates an element-wise logarithmic function. After that, the normalized document vector is passed through

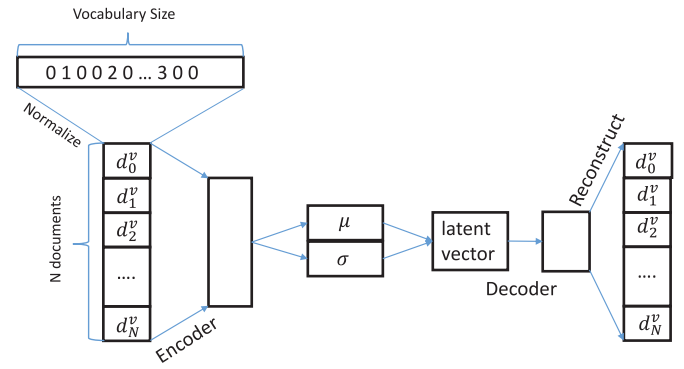


Fig. 4. Illustration of VAE-based Neural Topic Model.

an encoder to obtain hidden topics distributions (μ, σ) , in which vector z is sampled to reconstruct d^v through a decoder.

B. Attention-Based Semantic Embedding

Though documents have been initially vectorized, their correlation with visual information is still unknown. In order to enhance the association between visual and textual cues to achieve the intention of joint learning, we propose a semantic embedding approach via attention mechanism. Intuitively, two phenomena may inspire us to embed documents effectively:

- 1) Textual information may hold “temporal correlation”, i.e., nearby text should be semantically similar.
- 2) Text in different time windows may express variant significance according to the visual context.

Along this line, given the first point, we divide time window T_t into k periods with equal length. Then, documents in the same period are merged based on the average of semantic vectors, an unified representation is obtained via a fully-connected layer (FC) and denoted as $H_i \in R^{1 \times r}$.

At the same time, given the second point, each document vector H_i will be measured for its significance score α_i via attention mechanism as follow:

$$\alpha_i = \frac{\exp(H_i^T Vis)}{\sum_j \exp(H_j^T Vis)}, \quad (3)$$

in which Vis indicates the vector for visual information obtained by adapted KPMM [31] method through global average pooling and fully-connected layer. Specifically, top-level feature map is chosen to describe the context since it could express more abstract and semantic information in CNN. Then, H_i of each period will be weighted as follows:

$$\tilde{H}_i = (1 + \alpha_i)H_i \quad (4)$$

After that, with borrowing the idea of KPMM [31], we produce the semantic difference matrix Δ_H for semantic matrix pair $(\tilde{H}_x, \tilde{H}_y)$ as follows:

$$\Delta_H = \tilde{H}_x - (\tilde{H}_x \tilde{H}_y^T) \tilde{H}_y \quad (5)$$

Finally, we compress Δ_H into semantic difference vector, and then splice it with visual difference vector to support the distinction in *re-ID* module.

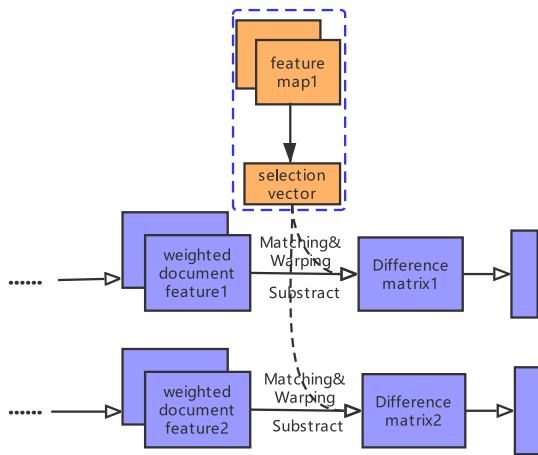


Fig. 5. Illustration of Source Selection Mechanism.

C. Source Selection Mechanism

As mentioned above, multi-source textual information is utilized in our framework, i.e., subtitles and bullet-screen comments. Generally, their characteristics could be extremely different. For instance, subtitles could be direct description of character status and behaviors in the first-person perspective, with relatively formal expression. On the contrary, bullet-screen comments are always subjective comments in the third-person perspective, with informal expression or even slangs generated by massive users. Thus, it is necessary to select the appropriate source of textual information based on the visual feature of certain frame, so that the selected textual information could better reflect the identity of character.

To that end, we design a **source selection** mechanism (SSM) to automatically measure the significance of each source according to the visual context, which is summarized in Figure 5. To be specific, we first describe the visual context with an united feature map $U_v \in \mathbb{R}^{2 \times h \times w \times c}$ spliced by top-level feature maps (scale-1) of image pair, in which h, w, c indicates the width, height and channels of top-level feature map, respectively. Then, a GAP followed by a 2×2 pooling, as well as a fully-connected layer are adopted to learn the source selection vector $S_2 \in \mathbb{R}^2$ as $S_2 = \sigma(A_U)$, in which $\sigma(\cdot)$ indicates an element-wise sigmoid function and A_U presents the pre-activation output of source selector.

Then, we take advantage of more descriptive documents according to the selection of context. As shown in Figure 5, semantic difference matrices for both bullet-screen comments and subtitles, denoted as Δ_{H1} and Δ_{H2} obtained by Equation (5), are re-weighted through a tensor-matrix multiplication with S_2 as follow:

$$(\Delta'_{H1}, \Delta'_{H2}) = (\Delta_{H1}, \Delta_{H2}) \times S_2 \quad (6)$$

Finally, we compress semantic difference matrices from different sources into vectors, and then splice it with visual difference vector obtained by the backbone of re-ID model in Section III-D for the final distinction. In the experiments, we will discuss the function of automatic source selection with intuitive case study in section VI.

V. EXPERIMENTS

In this section, we will conduct extensive experiments on a real-world data set to validate our novel framework.

A. Dataset and Data Pre-processing

Data Preparation. In order to evaluate our framework, we extract two datasets from Bilibili,¹ one of the largest video sharing platforms in China which focuses on animation, movies, etc. Specifically, for the animation dataset, we have 148 videos from 32 animations, each lasts around 24 minutes long and contains about 3,000 bullet-screen comments (due to the amount limitation of Bilibili). Along this line, 693 characters with 10,841 bounding boxes from 8430 frames are collected for validations, in which 48 characters with 1443 bounding boxes are treated as test samples. Correspondingly, for the movie dataset, we have 19 movies, 96 characters and 2179 bounding boxes, in which 20 characters and 452 bounding boxes are treated as test samples.

Data Pre-processing. In order to ensure the effectiveness of model training and evaluation, we select those main characters with rich documents (especially on movies which contain sparser textual information), and then manually label these bounding boxes. In the labeling process, we try to ensure that each character has diverse poses, and the related frames are evenly distributed throughout the given videos.

At the same time, considering that bullet-screen comments may suffer severely informal expression, we utilize some regular rules provided by Bilibili word-filtering system. Based on the filter, those meaningless comments, e.g., time/date marker or messy code will be deleted to ensure the quality of textual information. Besides, comments with less than 5 characters in animation data set, as well as comments with less than 3 characters in movie data set are removed based on the parameter sensitivity, as mentioned in section V-E.

Besides, we realize that the timestamp recorded for bullet-screen comments are indeed the time when these comments were sent. Considering that after watching the semantical-related frame, users may require some more time to *type* these comments. Thus, obviously recorded timestamp will be late compared than the corresponding visual content. To deal with this problem, we regulate the timestamp by estimating the period for typing based on the length of comments (approximately 30 characters/minute).

B. Experimental Settings

Implementation Details. First, we will introduce the details of implementation in *Detection*, *re-ID* and *Text processing*.

- **Detection Module.** Our detectors are directly utilized for detection task in *movie* dataset. For *animation* dataset, due to the different visual style, we retrain the two detectors with 3,492 frames containing 5648 bounding boxes, additional 569 frames with 941 bounding boxes are treated for validation. The Faster R-CNN detector is initialized with the VGG-16 model, which is validated as reaching 88.76%

¹[Online]. Available: <https://www.bilibili.com/>

in AP with intersection over union (IoU) score greater or equal to 0.5, along this line, we select the RoIs with confidence score greater than 0.85 to achieve the metrics as 91.5% in Recall and 90.3% in Precision. By contrast, the Cascade R-CNN initialized with the ResNet-101 backbone brings a significant improvement, which reached 95.37% in AP. Also, we select the RoIs with confidence score greater than 0.85 to achieve the metrics as 93.44% in Recall and 92.22% in Precision.

- **re-ID Module.** All the RoIs with character labels as input are resized to 256×128 . Then, the positive-to-negative pair ratio for each batch is set to 1:4, and the temperature parameter is set as 1.0, data expansion and hard-mining strategy are adopted to further enhance the performance. Besides, cross-entropy loss and L2-regularization loss are used. Finally, we adopt Moment optimizer with momentum of 0.9 for model training, the initial learning rate is set as 0.005 and dropped exponentially.
- **Text Processing.** For the vectorization of subtitles, skip-gram model with negative sampling [45] is utilized to generate the 300-dimensional vectors, which is pre-trained by documents from Baidu Encyclopedia. At the same time, considering the informal expression of bullet-screen comments, for document vectorization, on the one hand, we set 3 layers for character-level LSTM [46] to generate 256-dimensional comment vectors; on the other hand, we utilize VAE-based neural topic model [47] with 50 hidden topics. Both LSTM [46] and NTM [47] have been pre-trained on 2 million bullet-screen documents and 1,505 characters with frequency over 1,000 were chosen to build up the vocabulary. Details for parameter sensitivity are mentioned in section V-F.
- **Aggregation Module.** Based on experiments (as shown in section V-F), for animation dataset, clip interval is set as 5 seconds and ρ_s is set as 0.45; for movie dataset, clip interval is set as 7 seconds and ρ_s is set as 0.40.

Baseline Methods. Then, to validate the performance, considering that our framework are composed by several modules, we select different methods for each module and combine them flexibly. For the detection module we have (F) Faster R-CNN and (C) Cascade R-CNN detector. For re-ID module, we select several state-of-the-art methods as follows:

- **Kronecker-Product Matching Model (KPMM)** [31], which has been adopted as the backbone in our multimodal approach for person re-identification.
- **Multi-level Factorisation Net (MLFN)** [33], which factorized the visual appearance of a person into latent discriminative factors at multiple semantic levels without manual annotation. In detail, 5 blocks are stacked in MLFN. Within each building block, 4 factor modules are aggregated, and batch normalization is adopted for effective training. The final feature dimension d is set as 256, the initial learning rate is set as 0.001.
- **Mancs** [34], which utilized a multi-task attentional network for jointly optimizing 3 different learning tasks. In detail, the dimension of final re-ID feature for each instance is set to 300, for *PK Sampling strategy*, P is set as 10 and

K is set as 5, λ_{rank} , λ_{cls} and λ_{att} is set as 1, 1 and 0.1, the soft-margin m is set as 0.5, we adopt Moment optimizer with an initial learning rate of 5×10^{-4} to minimize the three losses.

- **Textual-based Methods.** Generally, all the techniques mentioned in section IV-A could be treated as baselines, namely the vectorization based on LSTM [46] or NTM [47] for bullet-screen comments, as well as the vectorization of subtitles by Skip-gram model [45]. Specifically, we utilize the well-trained Random Forest (RF) classifier for the final distinction.
- **Deep-Semantic Model (DSM)** [42], which inferred the latent emotional state through multimodal network on sentiment recognition task. We transfer this multimodal-based approach to our task, to be specific, the dimension of visual feature is set as 256, the text information is projected into a LSTM layer with 256 units. Cross-entropy loss is adopted with learning rate setting as 10^{-3} .

Therefore, we combine these different methods in detection and re-ID module flexibly and construct several approaches for evaluation on the overall performance, namely the **F-KPMM**, **F-MLFN**, **F-Mancs**, **F-DSM**, **C-KPMM**, **C-MLFN**, **C-Mancs** and **C-DSM**. Notably, the **Textual-based** methods could be directly utilized on person search task. Also, we select several state-of-the-art person search methods for comparison:

- **Joint Convolutional Neural Network (JCNN)** [5], which adopted an Online Instance Matching (OIM) loss function for the effective training. In detail, ImageNet-pretrained ResNet-50 are exploited for parameters initialization, in which the temperature scalar is set as 0.1, the learning rate is set as 0.001 and dropped to 0.0001 after 40 K iterations.
- **Cross-Level Semantic Alignment (CLSA)** [35], which addressed the under-studied multi-scale matching problem in person search by proposing a deep learning approach capable of learning more discriminative identity feature representations in a unified end-to-end model. To be specific, the top 3 blocks are used to construct the semantic pyramid, the λ_{ce} and λ_{cls_a} is set to 1 and 1, the initial learning rate is set to 0.001.

C. Overall Performance

First of all, we plan to validate the overall performance of our framework. As mentioned before, the character-oriented summarization task aims at capturing all the frames which contain the target character. Thus, intuitively, we conduct experiments on the whole person search part, i.e., the detection and re-identification modules for objective evaluation. Along this line, we select *Recall*, *Precision* and *F1-value* as evaluation metrics, as well as the *top-1* and *top-5* cumulative matching characteristics accuracies.

Specifically, for each query, 10 frames with target characters and 40 other frames without target characters are randomly selected for validation. The results are listed in Table II and Table III for two datasets separately, in which “F-TKPMM” presents our solution enhanced by textual information. According to the results, we could observe that the improvement caused

TABLE II
OVERALL PERFORMANCE ON ANIMATION DATASET

Methods	Animation Dataset				
	Top-1(%)	Top-5(%)	R(%)	P(%)	F1(%)
LSTM [46] + RF	35.6	67.1	51.8	55.2	53.4
F-MLFN [33]	87.5	97.7	59.2	48.9	53.6
C-MLFN [33]	82.7	99.0	60.3	49.5	54.4
F-Mancs [34]	63.5	96.2	50.6	40.7	45.1
C-Mancs [34]	69.2	95.4	42.8	51.0	46.5
F-DSM [42]	75.0	93.5	58.5	58.0	58.3
C-DSM [42]	73.5	95.0	60.0	56.9	58.4
JCNN [5]	46.8	73.4	68.6	40.2	50.7
CLSA [35]	83.3	98.1	61.7	52.7	56.8
F-KPMM [31]	79.2	98.3	65.9	53.1	58.8
C-KPMM [31]	84.0	98.5	59.1	62.9	61.0
F-TKPMM	81.6	96.7	73.4	64.2	68.5
C-TKPMM	87.6	99.4	69.5	71.8	70.6

TABLE III
OVERALL PERFORMANCE ON MOVIE DATASET

Methods	Movie Dataset				
	Top-1(%)	Top-5(%)	R(%)	P(%)	F1(%)
LSTM [46] + RF	37.0	76.0	54.5	59.1	56.9
F-MLFN [33]	76.5	97.0	68.1	47.7	56.1
C-MLFN [33]	81.0	98.9	61.1	57.7	59.3
F-Mancs [34]	59.5	93.5	58.5	38.0	46.0
C-Mancs [34]	61.0	94.5	50.4	43.6	46.8
F-DSM [42]	58.9	90.3	61.0	56.1	58.4
C-DSM [42]	60.8	90.9	74.6	51.1	60.7
JCNN [5]	40.0	55.0	55.0	36.9	44.2
CLSA [35]	77.0	98.5	62.7	53.7	57.8
F-KPMM [31]	62.5	97.0	63.1	46.1	53.3
C-KPMM [31]	62.5	98.0	62.3	50.3	55.7
F-TKPMM	61.1	93.3	70.0	59.4	64.3
C-TKPMM	61.7	98.3	67.0	66.0	66.5

by textual information is significant, even more than 8.4% in F1-value. It is worth noting that although the multimodal-based method C-DSM [42] is not specially designed for re-ID, it still outperforms all the remaining visual-based methods with at least 1.4% improvement in F1-value on movie dataset. In the contrast, the classic JCNN [5] and CLSA [35] perform terrible when facing characters with changing poses in varying scenes, which proves the limitation of traditional visual-based techniques on our task. At the same time, these methods with Cascade R-CNN [26] as a detector achieve a better performance, which confirms that an accurate detector will indeed benefit the person search process.

D. Ablation Study for Re-ID Module

Then, we turn to validate the performance of separate *re-ID* module for Ablation Study. Also, we want to validate the effectiveness of SSM with comparing the various integrations of multi-source textual information. Similarly, 10 RoIs with the target characters as positive samples, and 40 RoIs with other characters as negative samples, are randomly selected to be compared with the query RoIs in each batch for validation.

TABLE IV
VALIDATION OF RE-ID MODULE ON ANIMATION DATASET

Methods	Top-1(%)	Top-5(%)	R(%)	P(%)	F1(%)
LSTM [46]+RF	37.9	74.4	57.6	59.8	58.7
NTM [47]+RF	32.2	67.6	52.4	56.9	54.6
Skip-gram [45]+RF	50.0	82.1	54.9	64.5	59.3
MLFN [33]	87.4	96.8	58.6	58.1	58.4
Mancs [34]	74.9	94.4	45.4	54.4	49.5
CLSA [35] (w/o detector)	87.5	96.6	59.3	61.2	60.2
DSM [42]	91.3	98.3	58.7	75.9	66.2
KPMM [31]	86.4	99.6	63.5	67.3	65.4
KPMM+BS - C_L	89.2	99.8	69.5	69.2	69.3
KPMM+BS - C_N	89.5	99.5	68.6	70.4	69.5
KPMM+subtitle	87.5	99.7	71.7	68.6	70.1
KPMM+SUM	90.9	99.6	69.5	69.9	69.7
KPMM+SSM	91.0	99.9	72.7	72.2	72.4

TABLE V
VALIDATION OF RE-ID MODULE ON MOVIE DATASET

Methods	Top-1(%)	Top-5(%)	R(%)	P(%)	F1(%)
LSTM [46]+RF	43.0	75.0	54.4	59.3	56.7
NTM [47]+RF	39.3	70.8	54.3	56.1	55.2
Skip-gram [45]+RF	62.0	90.0	58.3	62.9	60.5
MLFN [33]	81.6	95.8	56.5	63.1	59.6
Mancs [34]	73.2	92.2	45.8	57.8	51.1
CLSA [35] (w/o detector)	81.6	95.9	61.1	59.6	60.3
DSM [42]	91.9	98.1	59.6	62.1	60.8
KPMM [31]	72.7	98.1	55.3	58.6	56.9
KPMM+BS - C_L	78.5	98.0	64.1	61.8	62.9
KPMM+BS - C_N	76.5	98.0	62.4	60.7	61.5
KPMM+subtitle	90.5	99.5	64.3	74.6	69.1
KPMM+SUM	80.5	99.0	68.5	61.8	65.0
KPMM+SSM	88.0	99.9	63.3	77.5	69.7

The experimental results are listed in Table IV and Table V for two datasets separately, in which KPMM+BS-C (i.e., bullet-screen comments, C_L and C_N separately means textual embedding via LSTM or NTM) and KPMM+subtitle indicate the joint modeling with single textual source. At the same time, we have KPMM+SSM to present the results with source selection mechanism, and KPMM+SUM means the joint modeling with simple sum of these two textual sources as comparison.

According to the results, obviously, we could find that textual information indeed improve the performance, especially for the movie dataset, in which the visual-based technique may suffer insufficient training data and more complex scenes. Usually, characters in animations could be easily matched via some significant features, e.g., color or style of hair. However, matching between characters in movies could be much more difficult due to more fine-grained appearance, which might be the reason why KPMM performs much worse in movie dataset. Under this situation, we can find that the improvement caused by textual information is even more significant, and even some textual-based methods, especially the Skip-gram+RF approach, outperforms KPMM by 3.6% in F1-value, which further confirms our motivation. Besides, KPMM+SSM performs not only better than two solutions with single source in most cases, but also outperforms KPMM+SUM by more than 2.7% in F1-value, which verifies the function of source selection mechanism.

TABLE VI
VALIDATION OF HYPER-PARAMETERS INHERENT IN TEXTUAL-BASED APPROACHES ON ANIMATION DATASET

Methods	Top-1(%)	Top-5(%)	R(%)	P(%)	F1(%)
$NTM_{100topics}$	25.6	60.2	51.4	50.7	51.0
$NTM_{50topics}$	26.9	60.6	51.4	52.7	52.0
$LSTM_{1layer}$	30.8	62.3	52.8	53.2	53.0
$LSTM_{3layers}$	28.5	63.5	53.3	53.7	53.5

TABLE VII
VALIDATION OF HYPER-PARAMETERS INHERENT IN TEXTUAL-BASED APPROACHES ON MOVIE DATASET

Methods	Top-1(%)	Top-5(%)	R(%)	P(%)	F1(%)
$NTM_{100topics}$	25.0	72.0	51.0	50.8	50.9
$NTM_{50topics}$	21.5	62.0	51.2	51.1	51.1
$LSTM_{1layer}$	30.0	63.0	51.1	52.2	51.6
$LSTM_{3layers}$	26.1	72.0	51.3	53.6	52.4

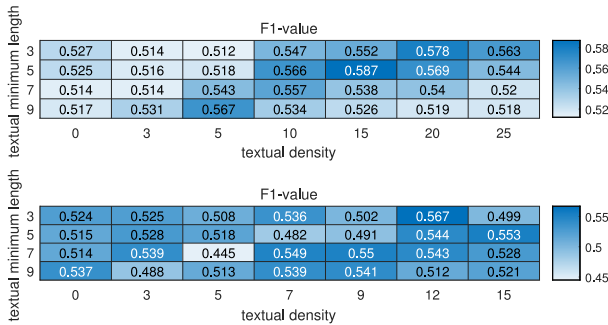


Fig. 6. Parameter sensitivity on animation (top) and movie dataset (bottom). Horizontal axis: textual density. Longitudinal axis: minimum textual length.

E. Parameter Sensitivity

Afterwards, we turn to measure the parameter sensitivity during the textual processing. Specifically, we first validate the sensitivity of *layer numbers* in LSTM and *hidden topic numbers* in NTM. The results are summarized in Table VI and Table VII, which indicate that Character-level LSTM with 3 layers might be more suitable to extract the semantic features of bullet-screen comments with better performance. Meanwhile, NTM with 50 hidden topics performs better than the one with 100 hidden topics. This phenomenon implies that the massive bullet-screen comments are distributed among a relatively small number of topics.

Secondly, for the text pre-processing step, parameter sensitivity for 2 factors, i.e., *minimum length* of each document, as well as the *textual density* which indicates the number of documents per time window, has been measured as shown in Figure 6. Specifically, we observe that the setting with textual length above 5 characters and textual density above 15 characters per window obtains the best performance in animation dataset. However, due to the sparsity of bullet-screen comments in movies, the thresholds should be set lower so that enough textual information could be extracted. In summary, we should keep a balance between the two factors to ensure the effectiveness of our solution.

Thirdly, we discuss the sensitivity of *time window* length when collecting textual information. Intuitively, longer window leads

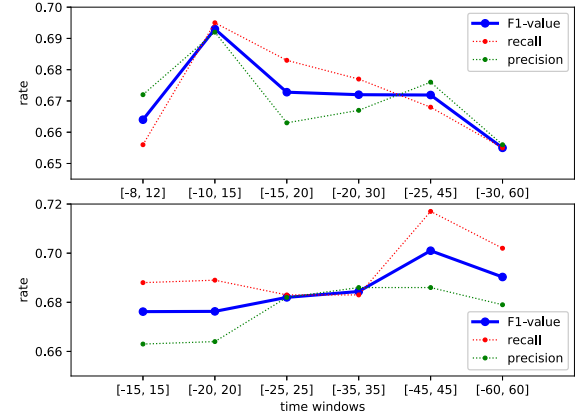


Fig. 7. Parameter sensitivity of time window for bullet-screen comments (top) and subtitles (bottom).

to rich information for better refinement, but also more noise which may disturb the results. Thus, we conduct series of experiments, which are summarized in Figure 7. Specifically, for subtitles, the windows before and after timestamp of current frame are set as equally long. But for bullet-screen comments, considering that most viewers comment after they view the frame, we set a longer window after the timestamp of current frame. The results validate our assumption that a better length should be a moderate one, since shorter window means insufficient information but less noise, and longer window indicates sufficient information but more noise. Also, for bullet-screen comments, more adjacent comments should be more semantically relevant to the frame, which may better refine the results. That's why relatively shorter window for bullet-screen comments performs better.

F. User Study on Summarization

Moreover, we would like to validate the quality of generated video summarization via user study. As mentioned in Section III-E, two pre-defined threshold, namely the *clip interval* and *clip density* may significantly impact the quality of summarization. Therefore, we first conduct several experiments to determine the proper hyper-parameters. Specifically, we adopt Silhouette Coefficient as the objective evaluation metric due to its capacity for partly reflecting the coherence of video clips, which is commonly utilized in prior arts like [48]. Moreover, in order to prevent the interference caused by too long or too short clips, which may lead to unreasonably high threshold, we directly set the Silhouette Coefficient as 0 when the amount of summarized clips for a video is out of the range [5, 50].

The experimental results are summarized in Figure 8. According to the best performance, we set the threshold of *clip interval* and *clip density* as 5 and 0.45 for animation dataset, as well as 7 and 0.4 for movie dataset, respectively.

With the pre-defined threshold, we turn to validate the quality of character-oriented video summarization via user study. To be specific, 20 viewers are required to mark 0 to 5 points for 83 animation-based videos and 17 movie-based videos, including 3

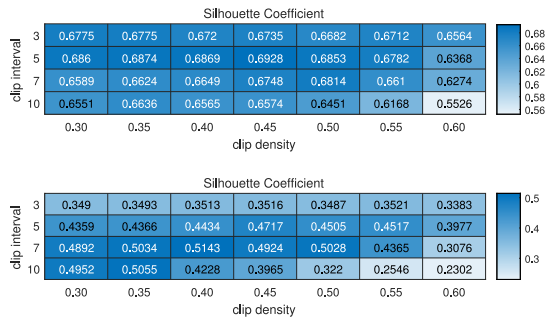


Fig. 8. Hyper-parameter experiments of Aggregation module on animation (top) and movie (bottom).

TABLE VIII
USER RATING FOR THE QUALITY OF SUMMARIZED VIDEO

Methods	Animation	Movie
JCNN [5]	2.558	1.853
CLSA [35]	3.228	1.942
C-DSM [42]	3.235	2.027
F-KPMM	3.143	2.043
F-TKPM	3.446	2.117

points for *consistency*, i.e., target characters appear in most part of summarization video without excessive gap, and 2 points for *significance*, i.e., target character is the center of this video, but not supporting actors.

The results on animation dataset and movie dataset are shown in Table VIII, in which all the methods shared the same aggregation module, as mentioned in section III-E. Obviously, our framework could summarize character-oriented videos with better user experience. Also, we found two interesting rules in these summarizations. First, the textual information could help to capture those frames where target character is insignificant. In this case, traditional techniques could probably miss the frame. Thus, textual-enhanced framework could produce more complete summarization. Second, due to the wrong distinction, traditional techniques usually produce some irrelevant clips. However, our framework could abandon those irrelevant part, which improves the quality of summarization.

VI. DISCUSSION WITH CASE STUDIES

Finally, we turn to discuss several impact factors of textual-enhanced framework, and then illustrate the benefits of textual information with a case study.

A. Impact Factors of Textual-Enhanced Framework

Impact of Attention Mechanism. First, we will discuss the impact of attention mechanism for textual document. Along this line, two pairs of cases are selected as shown in Figure 9, in which all the keywords are highlighted as bold and slanted with attention score in the brackets.

According to the results, we realize that for the first pairs which are visually difficult to judge, i.e., traditional visual-based techniques may be disturbed by the different clothing and viewpoints, in spite of this, one modality might be somewhat invariant

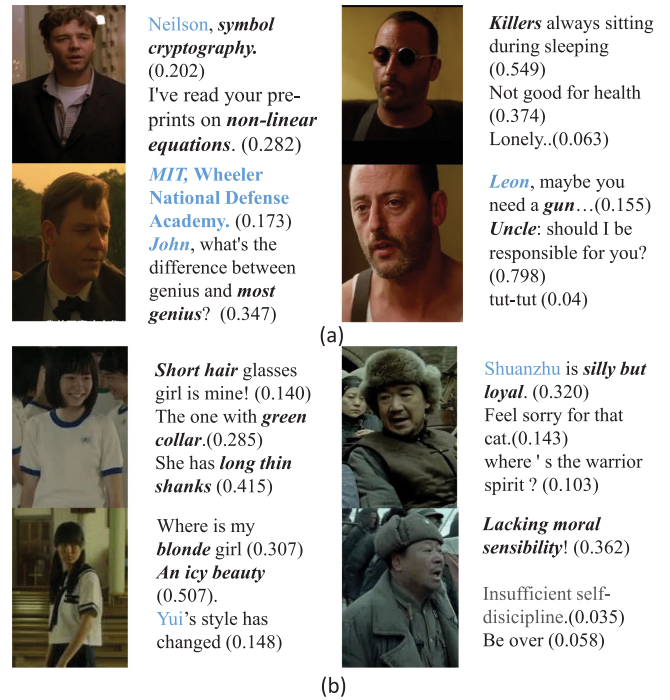


Fig. 9. Examples for the textual-enhanced distinction. (a) Frame pairs which are similar in textual description but visually difficult to judge; (b) Frame pairs which are visually confusing but different in textual descriptions.

TABLE IX
TOPIC PROPORTION OF TWO TEXTUAL SOURCES

Topic	Proportion (%)				
	Identity	Appearance	Sentiment	Plot	Others
B-S Comments	57.7	27.8	18.6	10.3	12.4
Subtitles	24.6	21.6	11.8	8.8	34.4

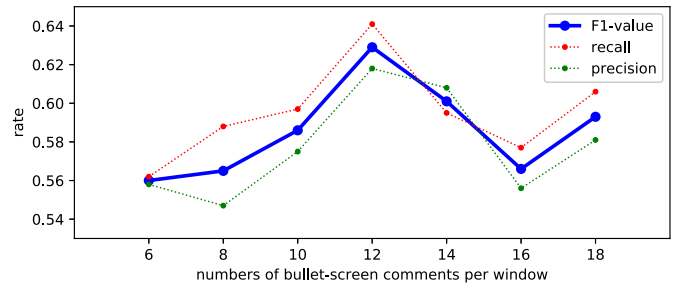


Fig. 10. Discussion on the impact of textual density for B-S comments.

to large changes in another modality, i.e., those identity-related keywords may benefit the capture of target characters. For instance, we could easily distinguish *John Nash* in *A Beautiful Mind* by the words “nonlinear equations” (indicating a mathematician) and “most genius”. Also, keywords like “killer” and “uncle” also help to identify the character, which results in higher attention score.

Similarly, for the second pairs which are visually confusing but different in textual descriptions, those keywords which highlight the difference of appearance may better improve the performance, e.g., “icy beauty” and “long shanks”. On the contrary,

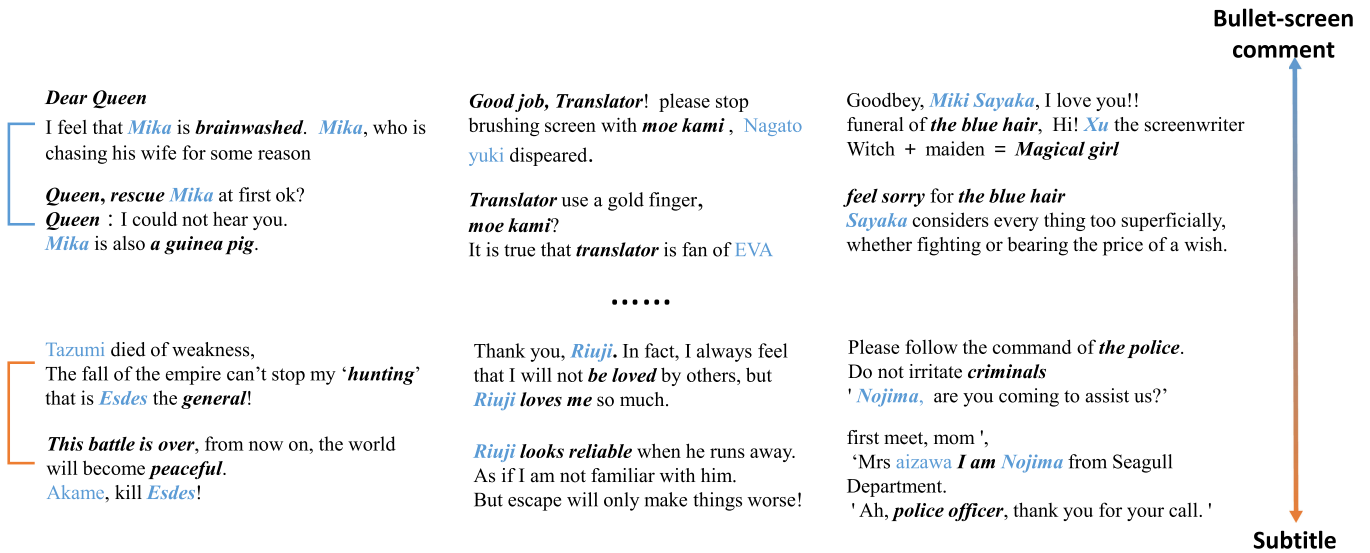


Fig. 11. Two groups of document pairs selected by SSM: bullet-screen comments (1st line) and subtitles (2nd line).

some common-used or meaningless words, may get a relatively low attention or even ignored. This phenomenon indicates that our framework could effectively capture the key information for character re-identification.

At the same time, we would like to discuss how different topics may benefit our framework. Specifically, we select the top 100 bullet-screen comments and subtitles based on their attention score, and then count the proportion of different topics, which are shown in Table IX. Generally, we make similar conclusions like identity-related contents benefit the most for both two data sets. Besides, we find that subtitles with pet phrase usually play an important role, e.g., Sheldon with “Bazinga” as mentioned in Introduction part.

Impact of Textual Density. Then, we would like to discuss the impact of textual density in jointly modeling of textual and visual cues. Specifically, as the density of subtitles is usually fixed, we only regulate the density of bullet-screen comments, i.e., we divide all the time windows based on their amount of bullet-screen comments, and then count their average performance. The result is summarized in Figure 10, under the time window as $[-10, 15]$ which has been proven as the best setting in the parameter sensitivity part.

In this case, we get similar conclusion with the parameter sensitivity that a moderate value of density could result in better performance. This phenomenon teaches us two lessons. For the popular characters with rich text, we should overcome the diversity to carefully select those semantically correlated documents, so that our solution will not be disturbed by irrelevant content. On the contrary, for those supporting characters with less textual description, we should attempt to enrich more textual information, e.g., longer time window to collect more documents for ensuring the performance.

Impact of Textual Source. Along this line, we also discuss the impact of textual source, i.e., when the subtitles or bullet-screen comments could better improve the performance. Thus, we select the cases from both sources with top 3 scores via source selection mechanism, as shown in in Figure 11.

According to the results, we realize that different sources may focus on different type/topic of content. On the one hand, as mentioned above, bullet-screen comments usually hold subjective feedback of characters. Thus, if target character is *popular* with some distinguishing feature, or even nickname, usually bullet-screen comments could help more. For instance, in the cases from the first line, we can find some features like “*blue hair*”, or “*moe kami*” as the nickname of one character, which clearly reveal the character. On the other hand, subtitles usually provide objective description of characters, thus, if the character has some particular identity, e.g., “*police*” and “*general*”, we could link them with semantically correlated words like “*criminal*” and “*war*” to match characters, as shown in the cases from the second line.

B. Illustration for Benefits of Textual Information

Finally, we would like to illustrate how our framework, especially the textual information, could improve the character-oriented video summarization task. To be specific, in Figure 12, we select five short clips extracted from the movie *A Man Called Ove* as an example, in which the leading character *young Ove* is selected as target character, who is labeled in the orange boxes in Figure 12. Along this line, Clip 1, 3 and 5 should be treated as positive samples, while Clip 2 and 4 should be negative samples.

Furthermore, for each clip, we draw the curves of *saliency score* for each frame, i.e., the similarity score estimated by re-ID module based on target character. To be specific, curves in the top line illustrate the results based on re-ID module with the enhancement of textual information, and curves in the bottom lines are re-ID module without textual information. Obviously, frames in positive samples should have higher saliency scores, and vice versa. According to the results, we can find that for positive samples, namely Clip 1, 3 and 5, re-ID module with textual information gets higher saliency scores than the module without textual information. Also, in negative samples like Clip 2

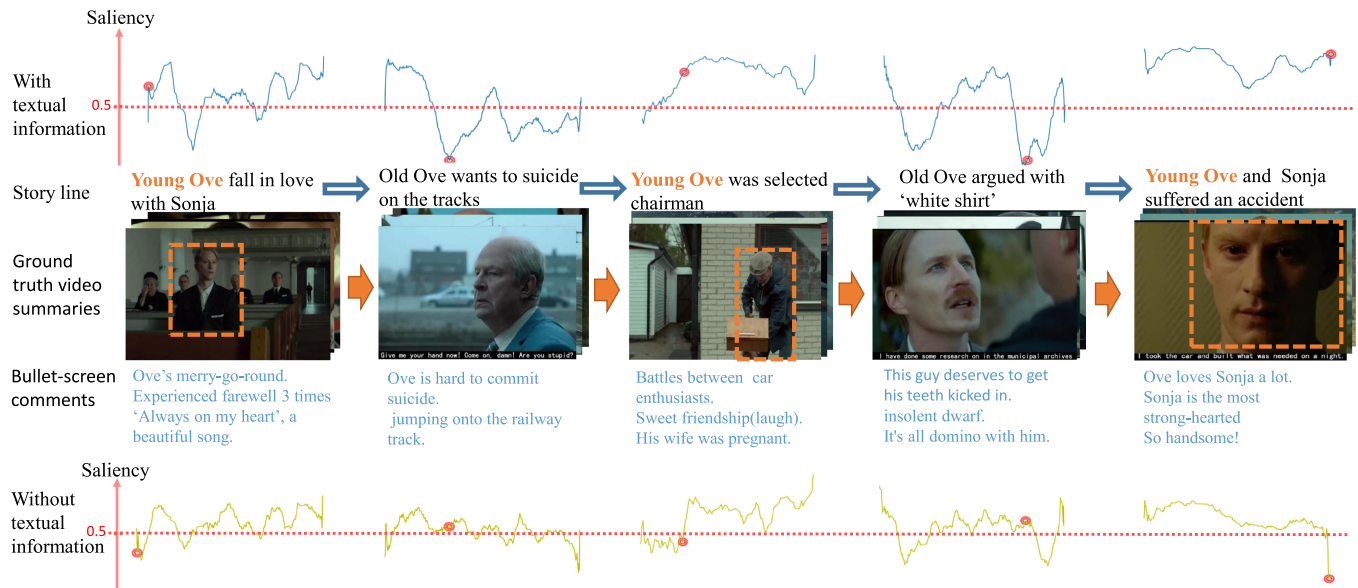


Fig. 12. Character-oriented video summarization examples on movie *A Man Called Ove*, with integrating textual information (top) or not (bottom).

and 4, re-ID module with textual information usually gets lower saliency scores and lower error rate. This phenomenon indicates that textual information could better distinguish the target character, with more significant distinction between positive and negative samples.

For deeply understanding the differences, we pick up several frames with opposite opinions by two methods (with/without textual information), which are labeled as *red points* in curves in Figure 12. According to the frames, we could find that visual-based approach will be easily misled by different clothing (e.g., the 1st frame) or a covered face (e.g., the 3rd frame), which lead to wrong judgments. However, under this situation, story-related comments may help. For instance, textual information in Clip 3 and 5 is talking about *love*, *marriage* and *friendship* about *young Ove*, our target character. On the contrary, in Clip 2 and 4, the textual description is about *suicide*, or *argument* with “white shirt” which could be reflected by those critical comments. In these cases, stories are significantly different between positive and negative samples, which definitely help to distinguish our target character.

In summary, textual information may help to describe the target characters more completely, even provide some more clues which could hardly be revealed by visual features. Obviously, these clues could enlarge the gap between positive and negative samples, which results in an easier distinction. This phenomenon also inspires us to fully utilize the potential of time-sync textual information for better understanding the videos, which may support some more applications, e.g., retrieval and recommendation of semantic-sensitive video clips.

VII. CONCLUSION

In this paper, we proposed a novel framework to jointly model the visual and textual cues for character-oriented summarization. To be specific, we first located characters indiscriminately through detection methods, and then identified the target

characters via re-identification module to extract potential key-frames, in which appropriate source of textual information will be automatically selected and integrated. Finally, key-frames were aggregated to form the summarization video of target characters. Experiments on real-world data sets validated that our solution outperformed several state-of-the-art baselines on both search and summarization tasks, which proved the effectiveness of our framework on the character-oriented video summarization problem.

REFERENCES

- [1] W. Hu, N. Xie, L. Li, X. Zeng, and S. J. Maybank, “A survey on visual content-based video indexing and retrieval,” *IEEE Trans. Syst., Man, Cybern., C, Appl. Rev.*, vol. 41, no. 6, pp. 797–819, Nov. 2011.
- [2] S.-I. Yu, Y. Yang, X. Li, and A. G. Hauptmann, “Long-term identity-aware multi-person tracking for surveillance video summarization,” 2016, *arXiv:1604.07468*.
- [3] T. Zhang, D. Wen, and X. Ding, “Person-based video summarization and retrieval by tracking and clustering temporal face sequences,” *Proc. SPIE*, vol. 8664, 2013, Art. no. 866400.
- [4] A. Sharghi, J. S. Laurel, and B. Gong, “Query-focused video summarization: Dataset, evaluation, and a memory network based approach,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 2127–2136.
- [5] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, “Joint detection and identification feature learning for person search,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 3376–3385.
- [6] H. Liu *et al.*, “Neural person search machines,” in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 493–501.
- [7] T. Xu *et al.*, “Learning to annotate via social interaction analytics,” *Knowl. Inf. Syst.*, vol. 41, no. 2, pp. 251–276, 2014.
- [8] G. Lv *et al.*, “Gossiping the videos: An embedding-based generative adversarial framework for time-sync comments generation,” in *Proc. Pacific-Asia Conf. Knowl. Discovery Data*, 2019, pp. 412–424.
- [9] G. Lv *et al.*, “Understanding the users and videos by mining a novel Danmu dataset,” *IEEE Trans. Big Data*, to be published.
- [10] B. Zhao, X. Li, and X. Lu, “HSA-RNN: Hierarchical structure-adaptive RNN for video summarization,” in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7405–7414.
- [11] S. Lee, J. Sung, Y. Yu, and G. Kim, “A memory network approach for story-based temporal summarization of 360 videos,” in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 1410–1419.

- [12] R. Panda and A. K. Roy-Chowdhury, "Multi-view surveillance video summarization via joint embedding and sparse optimization," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2010–2021, Sep. 2017.
- [13] B. A. Plummer, M. Brown, and S. Lazebnik, "Enhancing video summarization via vision-language embedding," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 1052–1060.
- [14] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, 2012, pp. 1346–1353.
- [15] M. Gygli, H. Grabner, and L. V. Gool, "Video summarization by learning submodular mixtures of objectives," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3090–3098.
- [16] M. Spain and P. Perona, "Some objects are more equal than others: Measuring and predicting importance," in *Proc. Eur. Conf. Comput. Vision*, 2008, pp. 523–536.
- [17] S. J. Hwang and K. Grauman, "Accounting for the relative importance of objects in image retrieval," in *Proc. Brit. Mach. Vision Conf.*, 2010, pp. 58.1–58.12.
- [18] A. Kanehira, L. V. Gool, Y. Ushiku, and T. Harada, "Viewpoint-aware video summarization," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7435–7444.
- [19] Y. Zhang, M. C. Kampffmeyer, X. Liang, M. Tan, and E. P. Xing, "Query-conditioned three-player adversarial network for video summarization," in *Proc. Brit. Mach. Vision Conf.*, 2018.
- [20] P. Varini, G. Serra, and R. Cucchiara, "Personalized egocentric video summarization of cultural tour on user preferences input," *IEEE Trans. Multimedia*, vol. 19, no. 12, pp. 2832–2845, Dec. 2017.
- [21] B. Mirzasoleiman, S. Jegelka, and A. Krause, "Streaming non-monotone submodular maximization: Personalized video summarization on the fly," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1379–1386.
- [22] Y. Xu, B. Ma, R. Huang, and L. Lin, "Person search in a scene by jointly modeling people commonness and person uniqueness," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 937–940.
- [23] L. Zheng, H. Zhang, S. Sun, M. K. Chandraker, and Q. Tian, "Person re-identification in the wild," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 3346–3355.
- [24] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 4457–4465.
- [25] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in CNNs," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 6995–7003.
- [26] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 6154–6162.
- [27] O. Hamdoun, F. Moutarde, B. Stanculescu, and B. Steux, "Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences," in *Proc. 2nd ACM/IEEE Int. Conf. Distrib. Smart Cameras*, 2008, pp. 1–6.
- [28] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *Proc. Eur. Conf. Comput. Vision Workshops*, 2012, pp. 413–422.
- [29] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 2197–2206.
- [30] B. J. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *Proc. Brit. Mach. Vision Conf.*, 2010, pp. 21.1–21.11.
- [31] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "End-to-end deep Kronecker-product matching for person re-identification," *IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, pp. 6886–6895, 2018.
- [32] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*.
- [33] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," *IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, pp. 2109–2118, 2018.
- [34] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Manacs: A multi-task attentional network with curriculum sampling for person re-identification," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 384–400.
- [35] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai, "Person search via a mask-guided two-stream CNN model," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 764–781.
- [36] P. K. Atrey, M. A. Hossain, A. El-Saddik, and M. S. Kankanhalli, "Multi-modal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, pp. 345–379, 2010.
- [37] Y. Guo, Z. Cheng, L. Nie, X.-S. Xu, and M. S. Kankanhalli, "Multi-modal preference modeling for product search," in *Proc. ACM Int. Conf. Multimedia*, 2018, pp. 1865–1873.
- [38] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," *J. Mach. Learn. Res.*, vol. 15, pp. 2949–2980, 2012.
- [39] R. Lajugie, D. Garreau, F. R. Bach, and S. Arlot, "Metric learning for temporal sequence alignment," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, vol. abs/1409.3136, 2014, pp. 1817–1825.
- [40] G. Lv, T. Xu, E. Chen, Q. F. Liu, and Y. Zheng, "Reading the videos: Temporal labeling for crowdsourced time-sync videos based on semantic embedding," in *Proc. Assoc. Advancement Artif. Intell.*, 2016, pp. 3000–3006.
- [41] H. Wu *et al.*, "Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 6609–6618.
- [42] A. Hu and S. Flaxman, "Multimodal sentiment analysis to explore the structure of emotions," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, vol. abs/1805.10205, 2018, pp. 350–358.
- [43] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
- [44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *IEEE Conf. Comput. Vision Pattern Recognit.*, pp. 770–778, 2015.
- [45] S. Li *et al.*, "Analogical reasoning on Chinese morphological and semantic relations," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 138–143.
- [46] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. INTERSPEECH*, 2014, pp. 338–342.
- [47] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.
- [48] I. Mademlis, N. Nikolaidis, and I. Pitas, "Stereoscopic video description for key-frame extraction in movie summarization," in *Proc. 23rd Eur. Signal Process. Conf.*, Aug. 2015, pp. 819–823.



Peilun Zhou received the B.S. degree from the University of Science and Technology of China (USTC), Hefei, China, in 2017. He is currently working toward the M.S. degree with USTC, Hefei, China. He is working with Key Laboratory of Big Data Analysis and Application. His major research interests include computer vision and natural language processing.



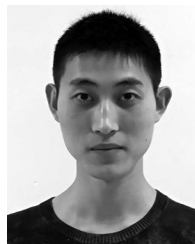
Tong Xu (M'17) received the Ph.D. degree from the University of Science and Technology of China (USTC), Hefei, China, in 2016. He is currently an Associate Researcher of the Anhui Province Key Laboratory of Big Data Analysis and Application, USTC. He has authored more than 40 journal and conference papers in the fields of social network and social media analysis, including TKDE, TMC, KDD, AAAI, ICDM, SDM, etc.



Zhizhuo Yin is currently working toward the B.S. degree with the University of Science and Technology of China, Hefei, China. He is working with Key Laboratory of Big Data Analysis and Application. His research areas are mainly in computer vision, and high performance computing.



Dong Liu (SM'19) received the B.S. and Ph.D. degrees in electrical engineering from the University of Science and Technology of China (USTC), Hefei, China, in 2004 and 2009, respectively. He joined USTC as an Associate Professor in 2012. He has authored or coauthored more than 100 papers in international journals and conferences. He has 16 granted patents, and one technical proposal adopted by AVS. His research interests include image and video coding, multimedia signal processing, and multimedia data mining. He was a Registration Co-Chair for ICME 2019, and as a Symposium Co-Chair for WCSP 2014. He was the recipient of the 2009 IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY Best Paper Award, and the Best 10% Paper Award at VCIP 2016. He and his team were winners of several technical challenges that were held in ACM MM 2018, ECCV 2018, CVPR 2018, and ICME 2016, respectively.



Guangyi Lv received the B.E. degree in computer science and technology from Sichuan University, Chengdu, China, in 2013. He is currently working toward the Ph.D. degree with the School of Computer Science and Technology, University of Science and Technology of China, Hefei, China. His major research interests include deep learning, natural language processing and recommendation systems. He has authored or coauthored several papers in refereed conference proceedings, such as AAAI, ICDM, PAKDD.



Enhong Chen (SM'07) is currently a Professor and Vice Dean of the School of Computer Science, University of Science and Technology of China, Hefei, China. His research is supported by the National Science Foundation for Distinguished Young Scholars of China. His general areas of research include data mining and machine learning, social network analysis, and recommender systems. He has authored or coauthored more than 100 papers in refereed conferences and journals, including *Nature Communications*, *IEEE/ACM Transactions*, *KDD*, *NIPS*, *IJCAI*,

AAAI, etc. He was on program committees of numerous conferences including *KDD*, *ICDM*, and *SDM*. He was the recipient of the Best Application Paper Award on *KDD-2008*, the Best Research Paper Award on *ICDM-2011*, and the Best of *SDM-2015*.



Changliang Li received the Ph.D. degree from the Institute of Automation, Chinese Academy of Science, Beijing, China, in 2015. Since 2018, he has been the Head with the Kingsoft Institute of Artificial Intelligence. He has published widely in artificial intelligence and deep learning research. His current research interests include deep learning, natural language processing, and data mining. He has authored or coauthored several papers in refereed conferences and journals, including *EMNLP*, *IJCNN*, *PAKDD*, *NLPCC*, etc.