Socializing the Videos: A Multimodal Approach for Social Relation Recognition

TONG XU*, University of Science and Technology of China
PEILUN ZHOU, University of Science and Technology of China
LINKANG HU, University of Science and Technology of China
XIANGNAN HE, University of Science and Technology of China
YAO HU, Alibaba Youku Cognitive and Intelligent Lab
ENHONG CHEN†, University of Science and Technology of China

As a crucial task for video analysis, social relation recognition for characters not only provides semantically rich description of video content, but also supports intelligent applications, e.g., video retrieval and visual question answering. Unfortunately, due to the semantic gap between visual and semantic features, traditional solutions may fail to reveal the accurate relations among characters. At the same time, the development of social media platforms has now promoted the emergence of crowdsourced comments, which may enhance the recognition task with semantic and descriptive cues. To that end, in this paper, we propose a novel multimodal-based solution to deal with the character relation recognition task. Specifically, we capture the target character pairs via a search module, and then design a multi-stream architecture for jointly embedding the visual and textual information, in which feature fusion and attention mechanism are adapted for better integrating the multimodal inputs. Finally, supervised learning is applied to classify character relations. Experiments on real-world data sets validate that our solution outperforms several competitive baselines.

CCS Concepts: • **Information systems** → **Data stream mining**; *Multimedia streaming*; • **Computing methodologies** → *Visual content-based indexing and retrieval*;

Additional Key Words and Phrases: social relation recognition, multimodal learning, person search, natural language processing

ACM Reference Format:
Tong Xu, Peilun Zhou, Linkang Hu, Xiangnan He, Yao Hu, and Enhong Chen. 2020. Socializing the Videos: A Multimodal Approach for Social Relation Recognition. *ACM Trans. Multimedia Comput. Commun. Appl.* 1, 1, Article 1 (January 2020), 22 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Recent years have witnessed the rapid development of video analysis techniques. Along this line, more semantic information, rather than the simple objects or movements that are directly reflected

*Tong Xu and Peilun Zhou contributed equally to this article.
†Enhong Chen is the corresponding author.

Authors' addresses: Tong Xu, tongxu@ustc.edu.cn, University of Science and Technology of China; Peilun Zhou, zpl@mail.ustc.edu.cn, University of Science and Technology of China; Linkang Hu, hulk@mail.ustc.edu.cn, University of Science and Technology of China; Xiangnan He, xiangnanhe@gmail.com, University of Science and Technology of China; Yao Hu, yaoohu@alibaba-inc.com, Alibaba Youku Cognitive and Intelligent Lab; Enhong Chen, cheneh@ustc.edu.cn, University of Science and Technology of China.

1551-6857/2020/1-ART1 $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

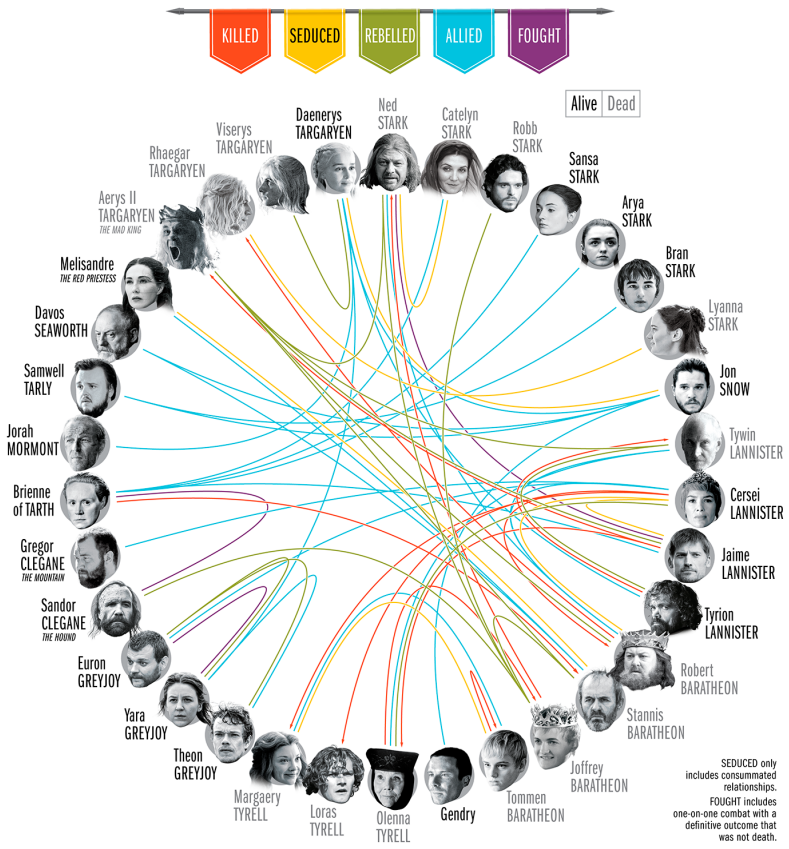

Fig. 1. Social Relation Graph for drama *Games of Thrones*, all the characters are connected in a network, in which each line indicates a specific relation between two of them (Designed by *Lon Tweeten* for Time[1]).

in visual content, is expected to be captured for more intelligent services. Among all the semantic factors, the *social relations between characters*, such as allied or enemy relations, could be one of the most crucial factor to help the audiences for better understanding the video contents. For instance, when viewing the dramas with hundreds of characters, like *Game of Thrones*, it could be a nightmare for audiences to remember hundreds of characters with complicated relations. Even worse, these relations are continuously changing as the story unfolds. In this case, a relation graph as shown in Figure 1 will definitely benefit the audiences with better experience. Moreover, with understanding the potential social relations between characters, some related intelligent applications, like fine-grained "video retrieval", "video summarization" and "visual question answering", which rely on rich semantic information, could be effectively supported. Therefore, the social relation recognition task is inevitably a crucial problem for video understanding.

To deal with this problem, large efforts have been made on relation recognition with various motivations. However, they may not fully solve this problem due to following reasons. First, most of the efforts mainly focus on the recognition of spatial relations or interactions [27, 31], e.g., "A is standing behind B" or "A is talking with B". Usually, these relations could be directly reflected via

[1]https://time.com/5560753/game-of-thrones-family-tree

visual content, but may fail to provide high-level semantic cues for describing the story due to the "semantic gap". Moreover, these approaches may heavily rely on manual intervention, e.g., manually selected images [30] or short videos [19], which focus on the target characters' co-occurrences (images or video frames in which several characters appear at the same time). However, in real-world application scenarios, most videos may contain substantial contents which are often irrelevant to the social relations. Under this condition, manually pre-processing would be an extremely heavy burden. Besides, most videos may suffer the frequently varying characters and scenes, as well as the complicated forms of relation description. As claimed by the prior study on Google AVA dataset [4], such conditions posed inevitable challenges to the current visual-based solutions. Thus, in order to address the traditional limitations, more comprehensive solutions are urgently required to provide richer semantic cues for enhancing pure visual cues, meanwhile, dependence of manual intervention will be controlled to support the real-world applications.

Indeed, we realize that the dependence of manual intervention reflects the shortage of semantic information, which could be caused by the "semantic gap" between low-level visual features and high-level semantic descriptions, e.g., social relations. Luckily, with the booming of social media platforms, videos now attracted time-sync crowdsourced comments generated by massive viewers. Usually, these comments not only provide subjective description of the story, but also contain timestamp information to map the specific frame of video [16, 17]. In this way, new cues could be extracted to reveal the social relations among characters. However, it is worth noting that this novel type of content usually contains much irrelevant or even noisy information, which may severely misguide relation recognition. Therefore, such multimodal information must be carefully filtered and embedded based on effectively semantic matching to support our task.

To that end, in this paper, we propose a novel multimodal-based framework to deal with the social relation recognition task for video characters. Our systematic framework could be directly applied to real-world scenarios via a combination of several modules as well as appropriate noise filtering operation. To be specific, first, we capture the target character pairs via a character search module. Secondly, in order to make the relation recognition not only depends on visual cues, but also on textual cues with more semantic meanings, we design a multi-stream architecture for jointly embedding the visual and textual information, in which feature fusion and self-attention mechanism are adapted for better integrating the multimodal inputs and filtering those irrelevant contents. Finally, supervised learning is aggregated to recognize social relations in an effective way. We validate our framework on real-world dataset, experimental results confirm the effectiveness of our solution. Furthermore, we visualize the different effects between visual and textual modality, which clearly shows the benefits of textual cues and further support our key idea of multimodal learning. In general, the contributions of this paper could be summarized as follows:

- We study the social relation recognition task for video characters, while considering the novel cues from crowdsourced comments from social media platforms.
- We propose an effective framework to jointly model the multimodal cues, in which noisy and irrelevant information will be filtered on different levels via attention mechanism.
- Experiments on real-world dataset have validated the effectiveness of our framework, and further reveal some interesting rules of multimodal cues.

2 RELATED WORK

In this section, we first summarize prior work on *social relation recognition*, then review some video analysis techniques such as *person search* and *multimodal learning*.

Social Relation Recognition. Traditionally, the prior studies [3, 32, 34] have confirmed that the co-occurrences between characters contain abundant clues reflecting their social relations. Some of

them focus on representing co-occurrence relation through a quantitative value [22, 32], however, these quantitative relations are unable to reflect the specific traits such as kinship, hostile and so on. Therefore, the others concentrate on predicting the specific social relation traits based on images or short videos clipped manually. For instance, Zhang et al. [41] proposed a deep CNN to learn a rich face representation combined with spatial cues for relation learning, and Sun et al. [30] utilized a double-stream CaffeNet to extract both head and body attributes. In [19] both visual and acoustic attributes from short videos were merged in a multi-stream network. Wu et al. [10, 21] further explore the temporal information and semantic object clues during relation recognition. Liu et al. [14] designed a novel triple graphs model to capture visual relations between persons and objects, then explored the multi-scale actions and storylines in spatial-temporal dimensions for social relation reasoning in short videos. In brief, the above studies enlighten the problem of relation recognition mainly in visual domain, in which the multimodal cues had not been considered for a full-scale understanding, they also strongly depended on manually selected short video clips, which could hardly be directly applied to real-world scenarios.

Person Search. Person search technique is usually adopted in human-centered videos, it aims at locating a specific person in a scene given a query image [36], which usually can be seen as a combination of person detection and re-identification (re-ID) modules. For the person detection module, traditional methods usually depended on the hand-crafted features for description, which are now enhanced by the deep learning techniques, e.g., the R-CNN architectures are adapted to achieve impressive performance by applying proper adaptations [39, 40]. Similarly, for the re-ID module, early works also focus on the feature designing [5, 20] and distance metric learning [13, 24]. Recently, some advanced techniques were proposed, e.g., the KPM module [28] to recover probabilistic correspondences between two images for similarity estimation, and the triplet loss [7] was exploited to improve the efficiency of training. With combining these two modules, person search technique becomes an important tool in video analysis [9], which could be directly adopted to extract the character-centered content from videos and benefit the recognition of social relations.

Multimodal Learning. Multimodal learning methods also have been widely adapted to visual-textual union in multimedia content, e.g., M-DBM [29] utilized a deep Boltzmann Machine to create fused representations across modalities, Lajugie et al. [12] attempted to learn a Mahalanobis distance to perform alignment of multivariate time series, Yinwei Wei et al. [37] presented a neural multimodal cooperative learning model to split the consistent component and the complementary component from multimodal features, and Lv et al. [16] designed a video understanding framework to assign temporal labels on highlighted video shots via textual summarization, Zhou et al. [42] utilized textual information to benefit the retrieval of video characters, while Chen et al. [2] proposed a novel multi-modal framework to achieve semantic fine-grained video search.. Also, Structured VSE [35] proposed a contrastive learning approach for the effective learning of fine-grained alignment from image-caption pairs, DSM [8] inferred the latent emotional state through the union of visual-textual cues. Valentin et al. [33] further investigated improved face descriptors based on 2D and 3D CNNs and explored several fusion methods and a novel hierarchical method for video emotion classification in the wild. Jiang et al. [11] exploited the long-range temporal dynamics in videos, and organized multimodal clues through a combination of CNN and LSTM for video classification. Long et al. [15] proposed a local feature integration framework purely based on attention clusters, and introduced a shifting operation to capture more diverse multimodal signals. Finally, Lv et al. [18] provided a novel dataset with rich semantic cues for multimodal learning tasks. These above approaches applied multimodal learning method to multimedia content and solved various classical tasks, which inspires us to analyse character relations in a multimodal way.

In summary, different from the traditional social relation recognition methods which mainly focus on the visual domain, we further utilize textual information with more abstract semantic

Table 1. Mathematical Notations.

Notation	Description
$\mathbf{V}$	A video composed by frames
$\mathbf{C_V}$	The target character set for video $\mathbf{V}$
$\mathbf{D_v}$	The set of textual document for video $\mathbf{V}$
$\mathbf{R}$	A pre-defined social relation set
I_t	The visual frame with timestamp t
d_t	The textual document with timestamp t
c_k	A specific character in $\mathbf{C_V}$
$\{r\}_{ij}$	The social relation(s) between character $\mathbf{c_i}$ and $\mathbf{c_j}$
q_i	The pre-labeled query for target character $\mathbf{c_i}$
$\{RoI\}_i$	The set of "regions of interest" (RoIs) for character $\mathbf{c_i}$
P_t	The occurrence probability of target character at timestamp t

meanings, then merge textual and visual clues to make up the shortages of each other, and finally formulate an effective solution to address this task. Besides, since our framework involves person search techniques, it is more systematic and applicable than those traditional methods and does not rely on much manual intervention, it could be directly applied to real-world scenarios with appropriate noise filtering operation.

3 TECHNICAL FRAMEWORK

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

3.1 Preliminary and Problem Definition

As mentioned above, we target at recognizing the social relation between video characters, therefore, we have $V = \{I_t\}$ to present a whole video, also a streaming collection of frame I_t with related timestamp t. At the same time, as we jointly model the textual information to enhance the task, we have the *time-sync documents* like subtitles or crowdsourced comments, which are presented as $D_V = \{d_t\}$ to present the set of textual documents.

We also have character set $C_V = \{c_k\}$, where each c_k presents a specific character in V, and a pre-defined social relation set R. For target pair $< c_i, c_j >$, several frames for each character are selected, and their character regions will be labeled to form queries q_i and q_j respectively, then relation(s) $\{r\}_{ij}$ from R will be recognized between character c_i and c_j. The social relation recognition problem for video V and character pair $< c_i, c_j >$ could be formally defined as follows:

Definition 1. ***Preliminary.*** *Given the video* $\mathbf{V}$ *with textual information* $\mathbf{D_V}$*, the pre-defined social relation set* $\mathbf{R}$*, and target character pair* $< \mathbf{c_i}, \mathbf{c_j} >$ *in form of queries* $< q_i, q_j >$*, we aim to recognize their social relation(s)* $\{\mathbf{r}\}_{\mathbf{ij}}$ *from* $\mathbf{R}$*.*

3.2 Framework Overview

To deal with the problem above, in this paper, we propose a framework which contains three modules as illustrated in Figure 2, i.e., character search module, multimodal embedding module and relation classification module, whose functions are briefly introduced as follows:

(1) First, for a target character pair $< c_i, c_j >$, we have **Character search** module to search all of their potential occurrences via detection and re-ID technique.
(2) Before **Multimodal Embedding** module, as a pre-processing step, the potential occurrences are aggregated as occurrence shots to compose the foundation for the next step. Thus, in

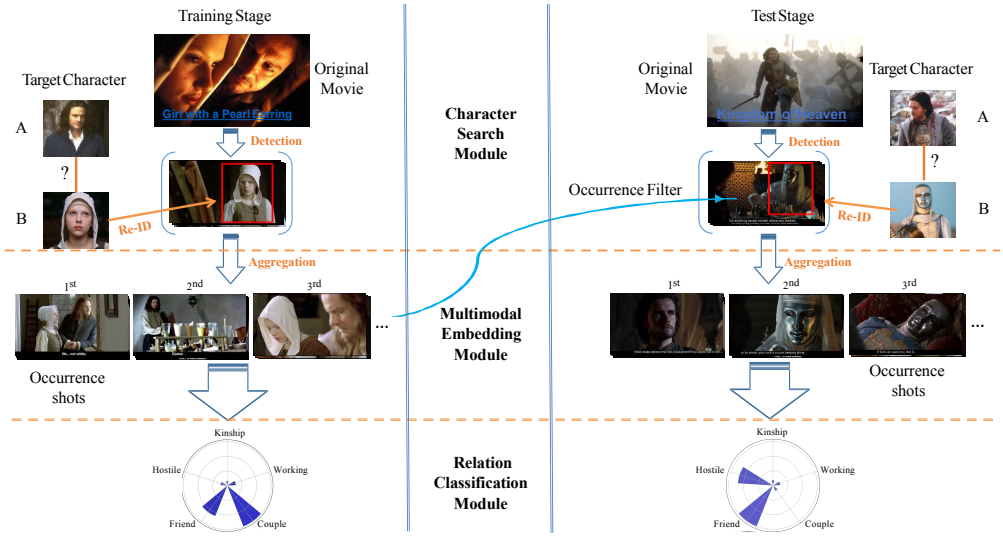

Fig. 2. Pipeline of our Social Relation Recognition Framework, which contains three important modules, i.e., character search module, multimodal embedding module and relation classification module.

the multimodal embedding module, each occurrence shot is firstly fed into a multi-stream network to extract features of each modality separately, followed by a feature fusion operation to flow them together in a coordinated way, and a self-attention mechanism designed to further integrate the features and finally obtain the shot-level multimodal representations.

(3) Finally, the multimodal embedding of each occurrence shot is then fed into the **Relation Classification** module. To be specific, two fully-connected (FC) layers are utilized to output the possibility distribution on each relation trait through a softmax or sigmoid layer. The shot-level results are finally averaged as the video-level judgment, i.e., the social relation trait(s) between two target characters.

Character search module. In detail, for the first step, we attempt to retrieve all the potential occurrences for each target character. We first adopt Faster R-CNN detector [25] to indiscriminately locate each character in V and produce regions of interest (RoIs) frame by frame due to its strong capability of detecting varying sized objects in unconstrained scenes. As RoIs are obtained, 5 frames of each target character are selected, and their character regions will be labeled to form queries $q_i = \{RoI\}_i$ and q_j respectively. Then state-of-the-art KPMM identifier [28] is adopted to estimate the probability that one RoI contains a target character based on the kronecker product matching with the corresponding query. For each target character, only the RoI with the highest occurrence probability is recorded frame by frame, the probability equals zero if there is no RoI detected in one frame. Then the potential occurrences will be collected in the form of $\{< I_t, P_t >\}$ for c_i and c_j respectively, in which I_t and P_t indicate the **potential frame** and **occurrence probability**. Notably, considering that movies present relations in more complicated forms, e.g., through unilateral actions, we reserve all the potential occurrences of both two target characters no matter whether they co-occur, and their potential occurrences are going to be further processed and embedded in the next module.

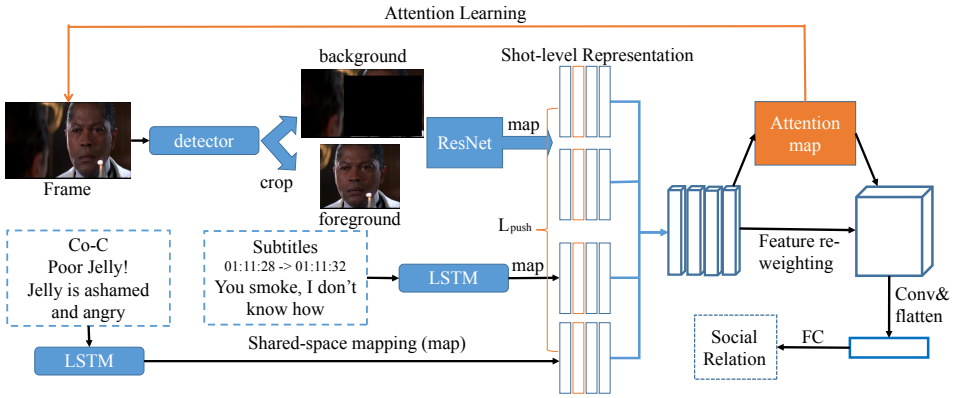

Fig. 3. Illustration of TEFM for joint modeling of visual and textual clues, which is composed by a multi-stream feature extractor, a feature fusion operation and a self-attention mechanism.

3.3 Multimodal Embedding Module

In this module, we take occurrence shots aggregated and filtered by a pre-processing step as our input, then obtain their multimodal representations through a multi-stream network.

Pre-processing. After the collection of potential occurrences for character c_i and c_j, we aggregate these frame-level clues as shot-level clues since the latter possesses better stability and integrity to increase the capacity of features' representation, followed by filtering operations to further ensure the quality. We detail this step as follows:

(1) First, in order to obtain stable representations, the occurrence probability sequence for character c_i, i.e., $\{P_t\}_i$ obtained via re-ID technique is smoothed through a Moving Average operation. Considering that consecutive video shots could provide more natural and steady clues to benefit the inference than still frames, we then aggregate these potential frames into occurrence shots based on a global threshold θ, which is set to ensure we have accurate and sufficient occurrence shots for a robust recognition. Details of hyper-parameter θ could be found in section 4.2 and section 4.5.

(2) After the aggregation of still frames, for each occurrence shot, a filter operation is adopted to keep the shot-level information sufficient and concise, to be specific, threshold (L_{min}, L_{max}) is set to filter those too short shots or segment the long ones, since the long shots may involve scene transformations, and make the feature expression become difficult. Details of these threshold settings will be discussed in section 4.2 and section 4.5.

After getting high quality occurrence shots of character c_i and c_j, we formulate a Textual-enhanced Fusion Model (TEFM) for robust multimodal embedding, as shown in Figure 3, TEFM contains several components, first, a multi-stream network is adopted to extract features of each modality separately, these features are then merged by a fusion operation in a coordinated feature space, in which a self-attention mechanism is added to further integrate these features. We also learn a weakly supervised global filter and utilize it to perform a better pre-processing in the test stage. These components are detailed one by one as follows:

Multi-stream Architecture. Basically, we adopt a multi-stream architecture to extract features from both visual and textual modalities. As shown in Figure 3, when an occurrence shot with n frames comes in, for each frame, we crop the **foreground**, i.e., the character region with the highest occurrence probability, in which sentiment-related clues are available. We also utilize the

rest part of the frame, namely the **background** which may reveal the specific occasions and benefit the relation recognition. Both of them are fed into ResNet-50 [6] to extract visual features.

In addition to the traditional visual clues, we further utilize textual information, to be specific, considering the semantically related text may exist in the neighborhood of the current frame, we use a small time window to extract these textual clues composed by **subtitles** and **crowdsourced comments** (Co-C), which are fed into two independently pre-trained LSTM to obtain textual features. Notably, those traditional methods adopted visual algorithms with high time complexity, however, even the simple utilization of textual information could make less time spent, and even better performance which will be confirmed in the experiment part.

Since these four-stream features are independent from each other and no correlation between them has been built so far, we attempt to flow these modalities together in a coordinated feature space in the next step.

Feature Fusion. In this step, as shown in Figure 3, we fuse features across multiple modalities via a shared-space mapping operation and a coordinated constraint. First, due to the fact that the visual and textual features are semantically corresponded to a certain extent (subtitles and crowdsourced comments make up the captions of current frame from both subjective and objective levels). These four-stream features are then mapped to a shared feature space R^d through a FC layer for multimodal fusion, thus, a shot-level feature map is obtained by the concatenation of frame-level features and represented as $f_i \in R^{n \times d}$, where $i \in [1, 2, 3, 4]$ indicates each one feature stream. A coordination loss function is designed to push the visual and textual features become closer as follows:

$$L_{push} = max(\alpha * ||f_V - f_T||^2, m), \tag{1}$$

where f_V and f_T indicate the average representation of visual and textual features, α is the scaling factor, considering the inevitable gap between visual and textual features, m indicates a soft-margin to avoid correcting "already correct" samples. After shot-level representations are gained, a Gaussian Blur operation is adopted to add temporal robustness for features in neighborhood as follows:

$$\widetilde{f}_{i,t} = \sum_{k=-2}^{2} f_{i,t+k} \times g(k), \tag{2}$$

where $f_{i,t}$ indicates i-stream feature of the t^{th} frame, $g(\cdot)$ is Gaussian Function with expected value 0 and variance 1.

Now we have the fused shot-level representations, however, not all modalities are equally relevant to relation recognition, and the significance of each modality is also varying temporally, therefore we filter redundant features through attention mechanism in the next step.

Self-Attention Mechanism. For a fused multi-stream feature map $F \in R^{4 \times n \times d}$, we design a self-attention mechanism to filter noisy information and select beneficial features. To be specific, the feature map F is fed into a combination of $1 \times 1 \times d$ convolution and ReLU layers to generate an attention map $a \in R^{4 \times n}$, which indicates the significance of each stream in each frame, therefore, the feature map F is re-weighted by multiplication with a for feature integrating:

$$\widetilde{F} = (\mathbf{1} + a) \times F. \tag{3}$$

As side benefit, self-attention could yield more interpretable models. We would further inspect attention distributions from self-attention mechanism and have discussions with typical cases in experiment part. The re-weighted feature map $\widetilde{F}$ is then fed into a 1D convolution layer and a flatten operation to generate the final representation, i.e., the multimodal embedding for each occurrence shot.

Occurrence Filter Based on Attention Learning. In order to further filter the noisy and irrelevant information, an occurrence filter is learned in a weakly supervised way. To be specific, we observed that the intermediate results produced by Attention Mechanism, i.e., the attention scores in a, which reflects the relevance between multimodal clues and social relations, could help the retrieval of those discriminative frames and further benefit the test procedure. As shown in Figure 2, in conjunction with the character search module, the supporting information could be retrieved in a more semantic way, we handle this problem in the following three steps.

First, we preserve attention map a with its corresponded frames during the training stage, then a Support Vector Regression (SVR) model is trained to build a direct mapping from frame-level information to attention scores as shown in Figure 3, therefore, the SVR model could be treated as a global filter to efficiently predict the frame-level relevance to social relations. Finally, in order to further enhance the strength of relational expression, only those occurrence shots with top k relevance scores are selected as the foundation of social relation classification in test stage.

3.4 Optimization and Convergence Analysis

To optimize the multimodal embedding module and relation recognition module, we formulate our loss function L as follows:

$$L = L_{cro} + L_{push} + \lambda * L_{reg}. \tag{4}$$

where L_{cro} indicates the cross-entropy loss between ground-truth label and predicted label, L_{push} is the coordination loss which has been introduced during the feature fusion operation, L_{reg} is the L2-regularization loss with coefficient λ controlling the importance of this term, which is uesd to improve model's generalization ability.

Also, we would like to discuss the convergence of our solution. Specifically, our loss function consists of three parts, which could be categorized into cross-entropy constraint L_{cro} and L2-norm like functions, i.e., L_{push} and L_{reg}. Therefore, the total loss L satisfies the Lipschitz continuous condition. With this guarantee, the change in the gradient of L has an upper bound. According to the prior study [38], for a non-convex and smooth loss function, if the Moment algorithm is used for training, the convergence upper bound is the same order with SGD algorithm, i.e., $O(1/\sqrt{t})$ for the gradient's square norm, and Moment method is more uniform stable than SGD method with potentially smaller generalization error. Thus, the Moment optimizer with Heavy-Ball momentum is adopted and the convergence of our method could be ensured.

4 EXPERIMENTS

In this section, we conduct extensive experiments on a novel self-constructed dataset and a public dataset MovieGraph[2] to validate our multimodal framework. First we describe both datasets, with focusing introduction to our novel social relation dataset, after the validation of our framework, we also have discussions on the challenges of the task and summarize several future directions which should be sought.

4.1 Dataset Description

Overall Introduction. Our dataset is collected on Bilibili[3], one of the biggest social media platforms in China. It contains 70 movies with average length of 1.9 hours. We sample each movie at a certain frequency (1 frame/second) in order to obtain a streaming collection of frames.

[2]http://moviegraphs.cs.toronto.edu
[3]https://www.bilibili.com/

Table 2. Descriptions and typical examples of social relation traits in our dataset.

Relation Trait	Descriptions	Examples
Working	one leads others professionally or give advice to each other or one offers service to others	Teacher-student; Colleagues; Customer-waiter
Kinship	have blood relation or have relations of fostering and supporting	Father-child; Grandparent-grandchild
Hostile	one disapproves the other or be antagonistic to each other	Enemies; Antagonist
Friend	express sunny face or act in a polite way	Friends; Host-guest
Couple	be matched in physical attractiveness or have similar preferences	Husband-wife; Boyfriend-girlfriend

In addition, 2 types of textual information accompanied with videos are collected. One of them is the crowdsourced comments, about 3,000 comments (due to the amount limitation of Bilibili) are collected for each movie. Besides, 1213 subtitles are collected for each movie on average. This dataset will be soon released.

Annotation Description. A total of 376 main characters in these movies are selected. Two types of annotation are carried on as follows:

- For each character, 29 frames in which he or she appears are collected on average. In each frame, we manually label a bounding box for the target character with tool *TrainImageLaber* in *Matlab*. Since this dataset provides the hand-drawn ground truth bounding boxes of sundry characters, it also provides a new platform to evaluate person re-identification task.
- We also allow several experts to label the character social relations based on video content and materials from *Baidu Encyclopedia*. All social relations are classified into 5 categories referred to Liu et al. [14], however, unlike Liu's processing, we consider the sparsity of several relations (e.g., "Service" relation, "Sibling" relation) in our dataset. To be specific, in order to ensure the class balance, we aggregate "Parent-offspring" and "Sibling" relations into "Kinship" relation, and aggregate "Leader-subordinate", "Colleague" and "Service" relations into "Working" relation. The detailed descriptions are seen in table 2. Notably, it is much possible that social relations are sometimes varying and confusing due to the complicated movie plots, therefore, each candidate character pair is labeled by maximum voting to guarantee the quality, characters with no interactions between them were not labeled. In this way, each movie got 1.94 "working" relations, 1.90 "hostile" relations, 1.70 "friend" relations, 1.36 "couple" relations and 1.04 "kinship" relations on average, then a total of 556 relations were collected.

The MovieGraph dataset consists of 7637 movie clips from 51 movies annotated with graphs that represent who is in each clip, the interactions between characters, their **multiple** social relations, and various visible and inferred properties such as the reasons behind certain interactions. There are several differences between our dataset and MovieGraph dataset. First, MovieGraph dataset performed person detection and identification based on face features to extract character occurrences, followed by little human corrections to deal with the pairwise social relation recognition task. Secondly, MovieGraph dataset consists 107 relation traits, but many of them are sparse and even synonymous, therefore, we aggregate these relations into the predefined 5 classes in table 2 for a better validation. Thirdly, MovieGraph dataset makes annotations on not only the main characters

but also many small roles (32.6 characters/movie on average), and there are few interactions and behaviors, especially language behaviors (subtitles) could be found on these small roles, as a result, 17.8% occurrence shots contain less than 3 texts, this modality missing phenomenon limits the performance of textual clues.

Data Pre-processing. We divide our self-constructed dataset into a training set with 40 movies and 275 character pairs as well as a test set with 30 movies and 281 character pairs. In order to ensure the quality of training samples, only the automatically extracted co-occurrence shots in training set are utilized to fit the social relation classifier. For the MovieGraph dataset, 10 movies with 560 character pairs and 767 pairwise social relations are split as test set.

At the same time, in our dataset, considering that crowdsourced 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 are deleted to ensure the quality of textual information.

Besides, we realize that the timestamps recorded for crowdsourced comments are indeed the time when these comments were sent. Considering that after watching the semantically related frame, users may require some more time to type these comments. Thus, obviously recorded timestamp is usually later than the corresponding frame. 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).

4.2 Experimental Settings

Implementation Details. First, we introduce details of implementation in character search module, multimodal embedding module and relation classification module.

- **Character Search Module.** First, our Faster R-CNN detector is initialized with VGG-16 model, we validate it on Pascal VOC07 dataset and reach 77.05 % for MAP under intersect over union (IoU) score = 0.5. Along this line, we select the RoIs with confidence score above 0.85 to achieve 72.0 % in Recall and 88.9 % in Precision. The Faster R-CNN detector is directly utilized for character detection in our dataset. Secondly, in the re-ID part, all the character RoIs are resized to 256×128, data expansion and hard-mining strategy are adopted to further enhance the performance, we randomly sample 40 IDs with 1087 frames for validation and achieve 72.1% in Recall and 68.2% in Precision under positive-negative ratio of 1:4.
- **Multimodal Embedding Module.** First, for the pre-processing, we validate the global threshold θ from a set of numbers (0.52, 0.63, 0.70, 0.78), which indicates the average value of every character's (50-th, 62-th, 75-th, 87-th) percentile of the occurrence probabilities, we find θ setting to 0.70 performs the best, the length limitation (L_{min}, L_{max}) of each occurrence shot is set to (6, 15), details are supplied with experiments in section 4.5. The visual features in TEFM are extracted via ResNet-50 pre-trained on ImageNet dataset. All the input frames are resized to 224×224 after the character region is cropped. The dimension of shared feature space is set to 256, the (kernel size, filters) of 1D conv layer is set to (2, 10), scaling factor α is set to 0.05, the effect of soft-margin m will be discussed in experiments. Besides, considering the different characteristics between subtitles and crowdsourced comments, on our dataset, we utilize two independently pre-trained 3-layer LSTM to generate 256-dim vectors, one is pre-trained on 2 million crowdsourced documents extracted from Bilibili, the other one is pre-trained on an open-source dataset *dgk shooter* [4] which contains 34.9 million subtitles. For MovieGraph dataset, English subtitle vectors are pre-trained on Wikipedia. The SVR model

[4]https://github.com/fate233/dgk_lost_conv

with RBF kernel is trained on textual vectors which will have been confirmed to contain more relation semantics and to be more efficient.

- **Relation Classification Module.** During the model training, importance weight λ in loss L is set to 5×10^{-3}, the momentum coefficient is set to 0.9, the learning rate is initialized as 10^{-4} and drops exponentially. Besides, we analyze the time cost of our framework in one GPU (Tesla P100). To be specific, character search module conducts person detection at a speed of 13 fps, then for each detected RoI, it takes 0.22 seconds on each target character's re-ID, after the occurrence shots are obtained, the multimodal embedding module and classification module cost 0.52 seconds for each shot with 15 frames to predict its social relation trait(s). In general, the total time cost mainly depends on the length of videos, the number of target characters and their occurrence frequencies.

Baseline Methods. Then, to validate the performance, we observe that the prior arts mainly focused on **(A1)** visual or **(A2)** textual modality (feature) and proposed an **(B1)** image-based or **(B2)** video-based algorithm. Therefore, we combine these aspects and consider several kinds of methods as baselines, all methods share the same character search module and pre-processing operation for a fair comparison. These baselines are detailed as follows:

- **A1+B1.** (1) **Double-stream CaffeNet (DSC) [30]** which used two CaffeNet with sharing parameters to extract 4096-dim body feature after *fc7* layer, we fine tune it to extract 512-dim face feature for validation in MovieGraph dataset, followed by a linear SVM for relation classification, dropout rate is set to 0.5, with learning rate setting to 5×10^{-4};
  (2) **Siamese DCN (S-DCN) [41]** further utilized 2 types of features: (i) handcrafted face position and ratio between face scales (ii) face features extracted from 2 CNN with sharing parameters, we recompile it without "cross-dataset" operation. Dropout rate is set to 0.5, cross-entropy loss is adopted with learning rate setting to 3×10^{-4}.
- **A1+B2.** (1) **MSFM [19]**: in this model, visual, optical flow and acoustic information from short videos are utilized and merged. First, 3 BN-Inception network are adopted for the extraction of visual, optical flow and acoustic features respectively, the visual and optical flow map is resized to 200×200, the scale of the frequency spectrum is set to 129×259, the learning rate of the three networks are set to $(0.01, 0.01, 0.05)$, the batch size is set to $(30, 30, 50)$ respectively, the 1000-dimension output of each BN-Inception network is then concatenated and fed into a Logistic Regression model for classification.
  (2) **Multi-scale Spatial-Temporal Reasoning (MSTR) [14]** which used a graph convolutional network (GCN) to capture visual relations, we utilize at most 40 bounding boxes of persons in our dataset or utilize at most 40 face regions in MovieGraph dataset to construct the intra-person Graph and inter-person Graph. The learning rate starts from 0.001 and multiply 0.1 by every 30 epochs.
  (3) **TEFM-V**: TEFM with only visual branch is also selected for comparison, in MovieGraph dataset, face features are treated as the foreground information.
- **A1+A2+B1. TEFM-F**: In order to validate model's capability on frame-level input, we fine tune TEFM to deal with frame-level embedding, named the TEFM-F method.
- **A2+B2.** (1) **LDA+RF**: we extract textual feature through Lda [1] model and use Random Forest (RF) classifier for classification. The number of requested latent topics is set to 15, the maximum number of iterations is 5000, stopwords from *nltk* are adopted; The number of trees in the RF classifier is set to 32.
  (2) **TEFM-T**: TEFM with only textual branch is also selected for comparison, named the TEFM-T method, in MovieGraph dataset, only subtitles are available as textual clues.

Table 3. Validation on 5-category classification (left) and relation polarity classification (right).

Methods	5-category			Binary			Params(M)
	R(%)	P(%)	F1(%)	R(%)	P(%)	F1(%)	
Lda [1]+RF	20.0	25.9	17.3	60.6	56.0	58.1	–
DSC [30]	18.4	24.0	16.8	64.5	56.4	59.0	47.5
S-DCN [41]	18.7	23.0	17.1	66.4	65.6	65.9	1.4
MSFM [19]	32.2	30.2	26.4	68.1	72.3	69.5	48.1
DSM [8]	31.1	32.7	26.3	67.7	68.7	68.2	22.7
CATF-LSTM [23]	29.6	29.1	28.3	64.5	66.4	65.6	26.7
MSTR [14]	33.6	**38.7**	31.9	72.1	63.9	66.2	2.1
TEFM-F	29.0	30.6	24.7	62.4	65.8	64.0	25.6
TEFM-T	31.4	30.5	27.8	**75.2**	64.9	67.8	0.2
TEFM-V	37.2	27.3	24.7	67.4	63.1	64.5	25.3
TEFM-CO	42.2	34.0	30.8	71.2	**74.5**	72.3	25.5
TEFM	**47.7**	35.8	**32.5**	72.2	74.4	**72.7**	25.5

- **A1+A2+B2.** Two multimodal methods based on visual and textual modalities are selected, (1) **deep semantic model (DSM) [8]** which inferred the latent emotional state from images through multimodal network, we fine tune it to receive videos as input for comparison on our task. To be specific, the shot-level representation is gained from the average of all the frame-level features. The input image resized to 224×224 is fed into the Inception network and outputs a vector of size 256, the text is projected into a LSTM layer with 1024 units. Cross-entropy loss is adopted with learning rate setting to 10^{-4}.
  (2) **contextual attention-based fusion LSTM (CATF-LSTM) [23]** which utilized attention network for multimodal fusion in the sentiment analysis of utterance. In order to fine tune this model for comparison on our task, we extract each utterance according to the subtitle's time interval, and model all the utterances in one occurrence shot through CATF-LSTM for classification. The extraction for acoustic features is unchanged. 2-layer CNN, followed by a FC layer of size 256 are used for textual feature extraction; Inception-v3 is adopted to extracted visual features; the hidden size of LSTM is set to 256, cross-entropy loss and L2 loss is adopted with learning rate setting to 5×10^{-4}. It is worth noting that all textual features have already been pre-trained on *crowdsourced documents from Bilibili*, *dgk shooter dataset* or *Wikipedia* before multimodal fusion for a fair comparison.

4.3 Evaluation of Relationship Recognition

To validate the overall performance of our framework, for the experiments in our self-constructed dataset, we conduct experiments on 5-category classification according to the relation traits in table 2. Meanwhile, since the **Hostile** relation is nearly opposite to the others, we carry on experiments on binary ("Hostile" or not) classification to recognize the relation "**polarity**" and strengthen our findings. Weighted-averaging Recall, Precision and F1-value are selected as evaluation metrics. For the experiments on MovieGraph dataset which should be treated as a multilabel classification task, we adopt F1-micro value and rank-based evaluation metric One-error for evaluation.

The results are listed in table 3 and table 4. There are two obvious findings: (1) the video-based approaches generally perform better than image-based ones, this phenomenon further confirms [27] and [26] 's opinion: As compared to still images, the videos provide a more natural and integral set of information to generate robust representation for relation recognition. (2) textual clues with more

Table 4. Validation on 5-category classification on MovieGraph.

Methods	One-error(%)	F1-micro(%)
Lda [1]+RF	67.8	30.5
DSC [30]	65.2	34.9
S-DCN [41]	64.9	35.1
MSFM [19]	62.9	43.0
DSM [8]	52.3	47.9
CATF-LSTM [23]	66.8	36.4
MSTR [14]	50.4	48.4
TEFM-F	64.3	35.6
TEFM-S	62.9	40.1
TEFM-V	52.4	47.5
TEFM	**47.8**	**51.1**

Table 5. Ablation study of different components.

Methods	5-category			Binary		
	R(%)	P(%)	F1(%)	R(%)	P(%)	F1(%)
Baseline	24.0	28.1	24.5	68.2	64.1	66.1
Baseline+FF	31.3	31.0	25.9	68.1	74.8	69.0
Baseline+AM	29.3	31.6	26.4	69.9	66.8	68.0
Baseline+FF+AM	**35.7**	**30.6**	**29.7**	**70.9**	**76.2**	**71.7**

Table 6. Ablation study of of different components on MovieGraph.

Methods	One-error(%)	F1-micro(%)
Baseline	53.3	47.6
Baseline+FF	49.5	48.9
Baseline+AM	50.1	50.0
Baseline+FF+AM	**47.8**	**51.1**

direct semantic meanings indeed benefit the relation recognition, more than 6% in F1-value on our dataset and 3.6% in F1-micro on MovieGraph dataset are gained. We find that the combination of subtitles and crowdsourced comments are effective and efficient, as a result, even TEFM-T method performs better than many other baselines on our dataset with rather small size of parameters. TEFM-S ("S" indicates subtitles) method performs relatively poor on MovieGraph dataset due to the modality missing problem mentioned in section 4.1. Generally, the whole TEFM performs better than two solutions with single modality, which admits the motivation of multimodal embedding.

Besides, the prior arts [3, 32, 34] mainly utilized the co-occurrences between characters as the evidence of social relation recognition. However, movies sometimes present relations in more complicated forms, e.g., through unilateral actions or bystander's statements. Therefore, in order to inspect the benefit of the "non-co-occurrence" shots, we compare TEFM with "TEFM-CO" approach, which was validated only based on the co-occurrence shots. The result on our dataset shows an improvement of 1.7% in F1-value on 5-category task, which confirms that the relaxation of co-occurrence constraint in test stage is sensible.

4.4 Ablation Study

After the validation of overall performance, we first verify the effectiveness of Feature Fusion and Attention Mechanism, which are treated as the significant components in TEFM.

The experimental results are listed in Table 5 and Table 6, in which **FF** and **AM** represent the **Feature Fusion** operation and **Attention Mechanism** respectively. We treat TEFM without FF and AM operations as baseline model in this subsection. According to the results, we could find that both of the two components indeed improve the performance. Although either of them brings limited growth compared with the baseline, the cooperation between them makes a bigger improvement. This phenomenon indicates that the fusion operation across modalities makes attention mechanism more effectively integrate the features, which verifies the function of our design.

We also explore the different effects between visual and textual features, especially the differentials between crowdsourced comments (Cro-C) and subtitles. Comparison experiments on our dataset shown in Figure 4 reveal the differentials on each relation class. Firstly, we observe that the textual information outperforms a lot on "working" relation which could be hardly inferred from the visual content, on the contrary, relations with more emotional expressions (e.g., "friend" and "hostile" relations) are easier to be recognized via visual features. Secondly, overall the crowdsourced comments outperform the subtitles, especially on the "couple" relation which is most talked about by viewers, rich clues could be found in their comments. We also find the "kinship" relation is easier inferred from subtitles with more notable clues (e.g., person deixis). These phenomenon explains the different characteristics of different modalities.

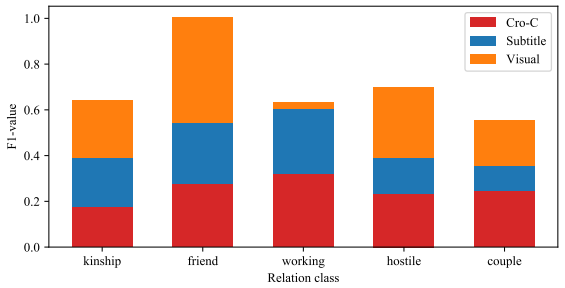

Fig. 4. Different effects between crowdsourced comments, subtitles and visual features.

4.5 Parameter Sensitivity

In this part, we test the effects of several important hyper-parameters. First, as we have claimed in section 3, the global threshold θ is set to ensure both quality and adequacy when aggregating potential frames, therefore, it plays an important role in our whole system. Intuitively, a high threshold means we only keep a few occurrence shots with high re-ID scores, which makes the model become relatively fragile since the accurate occurrences of characters does not necessarily lead to the easy recognition of social relations between them (although it is easier in many cases), meanwhile, a low threshold may bring in some opposite effects. Thus, we conduct experiments on hyper-parameter θ on our dataset, the results are shown in table 7, we find that the better threshold should be a slightly lower value, which considers not only the quality but also the adequacy of occurrence shots for a robust recognition.

Then we validate the influence of length limitation (L_{min}, L_{max}) in the pre-processing in **Multimodal Embedding** module. The experiments on our dataset are carried out on 5-category

Table 7. Validation of global threshold θ on 5-category classification.

θ	R(%)	P(%)	F1(%)
0.52	33.9	29.5	24.6
0.63	31.5	30.2	28.2
0.70	**47.7**	**35.8**	**32.5**
0.78	34.6	34.5	30.4

classification task as shown in Figure 5. We observed that hyper-parameter L_{min} is much more important than L_{max} since it has a major impact on the performance, it reminds us that the occurrence shots with more than 6 frames may have relatively sufficient information for relation recognition.

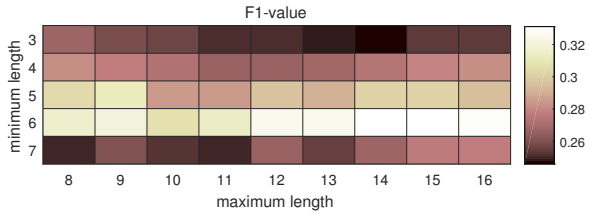

Fig. 5. Validation of occurrence shots' length limitation.

Afterwards, in the training stage, we use a coordination loss function to push the visual and textual features become closer, therefore, we test soft-margin m in this loss function. Intuitively, too small margin leads to overfitting between visual and textual features, however, large margin also makes the learning inefficient. Thus, several experiments are carried out and summarized in table 8 and table 9, where ∞ represents that loss function L_{push} is not considered. The results validate our assumption that a better margin should be a moderate one, since we should consider not only the affinity but also the gap between visual and textual features. It is also worth noting that, in the experiments on MovieGraph dataset, the performance is much poor due to the textual modality missing phenomenon, since the coordinate nature between textual and visual features may hardly be learnt in such conditions. This phenomenon also reminds us how to adjust the coordinated learning strategy to combat the modality missing challenge, and give full play to the potentials of textual clues in social relation recognition task.

Table 8. Validation of hyper-parameter soft-margin m.

m	5-category			Binary		
	R(%)	P(%)	F1(%)	R(%)	P(%)	F1(%)
∞	30.1	28.1	26.5	67.7	69.0	68.0
1	31.0	28.2	26.8	68.4	70.3	69.2
0.2	**35.7**	**30.6**	**29.7**	**70.9**	**76.2**	**71.7**
0.04	26.7	29.5	24.3	66.2	66.7	66.2

Finally, in the test stage, we validate the effect of threshold k during the filtering operation by SVR, which is designed based on the self-attention mechanism and adopted to benefit the test

Table 9. Validation of hyper-parameter soft-margin m on MovieGraph.

m	One-error(%)	F1-micro(%)
∞	**46.4**	50.9
1	47.8	**51.1**
0.2	54.1	48.8
0.04	58.4	43.1

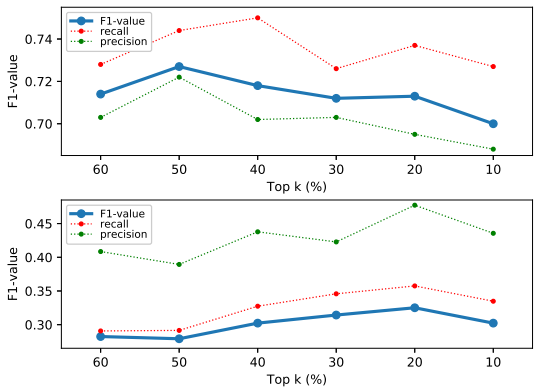

Fig. 6. Validation of SVR's filtering threshold on binary (top) and 5-category (bottom) classification.

procedure. Since threshold k indicates the strength of relational expression, low threshold leads to rich information for better refinement, but also more noise which has nothing to do with social relations and may disturb the results. Thus, we conduct experiments summarized in Figure 6. We observe that since the recognition of relation polarity is much easier, the binary task gains more robust performance with a better precision-recall balance. However, considering the relatively low robustness on 5-category task, more false alarms introduced by the decreasing of threshold would cause more perturbations and worse effect, therefore, a high threshold is needed. In detail, occurrence shots with top 20% relevance scores make the best performance and high precision. In the contrast, on binary task, a better refinement is feasible and occurrence shots with top 50% relevance scores gain the best effect.

4.6 Case Study

In this subsection, we first conduct several intuitive cases to discuss the benefits of textual information. Then we statistically analyze the effect of attention mechanism.

Impact of textual information. Some qualitative results are presented in Figure 7, we select 3 frames for each occurrence shot, in each frame, the most significant character is detected in red box, the attention scores are marked temporally beside each type of clues ("S:" indicate subtitles while "C:" indicates crowdsourced comments), which quantifies how they are semantically related to relation recognition.

These five cases correspond to "working", "kinship", "couple", "friend" and "hostile" relation respectively. We observed that the outstanding textual clues (highlighted in orange color) benefit the classification with relatively high attention scores. For instance, the first case is about the teacher-student relation between *Mathieu* and *Pépinot* in movie *Les Choristes*, though both working

Fig. 7. Several cases predicted by TEFM with multimodal information. The polar graph beside each case indicates the tendency for each trait to be positive.

and kinship relation seem to be possible, we made true judgment according to some keywords such as "little fellow". At the same time, in the second and third cases, some identity-related keywords (e.g., "Mom", "daughter") and sentiment-related ones ("beautiful", "Handsome") also suggest the relations at different aspects. Moreover, the visual content, especially the background is also beneficial by revealing specific occasions (e.g., classroom, Dojo). These phenomena indicates that our attention mechanism could effectively capture the key information, especially the key textual clues for social relation recognition.

Especially, in the last two cases we could still inference the relations through unilateral actions. For example, the last case is about *Hyoma Utsuki* and *Ryunosuke* in movie *The Sword of Doom*. Even though they did not co-occur, their duel is talked about by *Hyoma Utsuki* and others. Therefore, the relation between them could be inferred from some keywords (e.g., "*winning*", "*risk*"), which intuitively confirmed the benefits of "non-co-occurrence" shots.

Impact of attention mechanism. After the intuitive presentation, we would like to discuss how different topics may benefit the relation discrimination statistically. Therefore, we selected not

only the top 100 but also the last 100 crowdsourced comments (Cro-C) and subtitles based on their attention scores, and then counted the proportion of different topics listed in table 10. We observe that more identity-related and sentiment-related text are captured, which confirms the intuition that these clues are beneficial to relation judgment. In the contrary, more common-used or meaningless words (e.g., "*Okay*", "*tut-tut*", "*one day*") are usually ignored.

Table 10. Topic proportion of textual information.

Topic	Subtitles		Cro-C	
	Top 100	Last 100	Top 100	Last 100
identity (%)	**21**	13	**28**	25
sentiment (%)	**16**	7	**25**	17
noise (%)	5	**13**	14	**23**

4.7 Discussion

From the experimental results, we could observe that explicit recognition of social relations from movies is a very challenging task. We further reveal the challenges and difficulties through several typical failure cases as well as analysis of confusion matrix in this part.

We first show some failure cases to reveal how challenging this task is, and to indicate which direction should be sought. Figure 8 points out three typical challenges. First, many occurrence shots could not reflect characters' relations, for example, in the first case, the young man is the nephew of the old man (so they should be labeled as "kinship" relation), however, they act like friends and there is no clear evidence to distinguish. Secondly, some cases rather than provide no evidence, they even mislead the judgment. This phenomenon becomes more frequent due to the plot twists and conflicts in movies, as shown in the third case, even friends could have a quarrel, the unpleasant topics lead judgment becomes "hostile" instead of "friend". Thirdly, the expressions of relations could be pregnant with meaning. As shown in the second case, the officer and the girl have almost no communications with each other, therefore, even ordinary people could be confused without the overall understanding, especially the deep insight of movies, which is particularly needed in this condition.

These three typical challenges mainly indicate two directions to the better recognition of social relations. First, the story line is urgently required to handle the overall understanding, there are conflicts and plot twists, which may be treated as inevitable noise in the current experiments, however, if these somewhat confusing occurrences could be effectively organized with other occurrences according to the timeline and then be aggregated as a story pattern (instead of a simple average operation), the true relation could be mined more accurately based on the overall understanding. Secondly, since different modalities prefer to describe the same event at different aspects and even make up for each other, more modalities are required to achieve a full-scale modeling of movies, for example, in the second case (in movie *Le Silence de la Mer*), the background music could benefit a lot on the overall understanding, since the officer *Werner* and the girl *Jeanne* have few interactions and words with each other, the frequently occurred peaceful music of *Bach* plays an important role, it dispels the breath of war, indicates the emotional tone and confides trust and love between the target characters.

Two confusion matrices shown in Figure 9 further point out the strengths of the current system. In brief, it is difficult to distinguish similar or even somewhat overlap relations. For example, "friend", "couple" and "kinship" relations may be ambiguous, in this condition, more detailed and more abstract visual clues like genders, ages and human activities may be important for relation

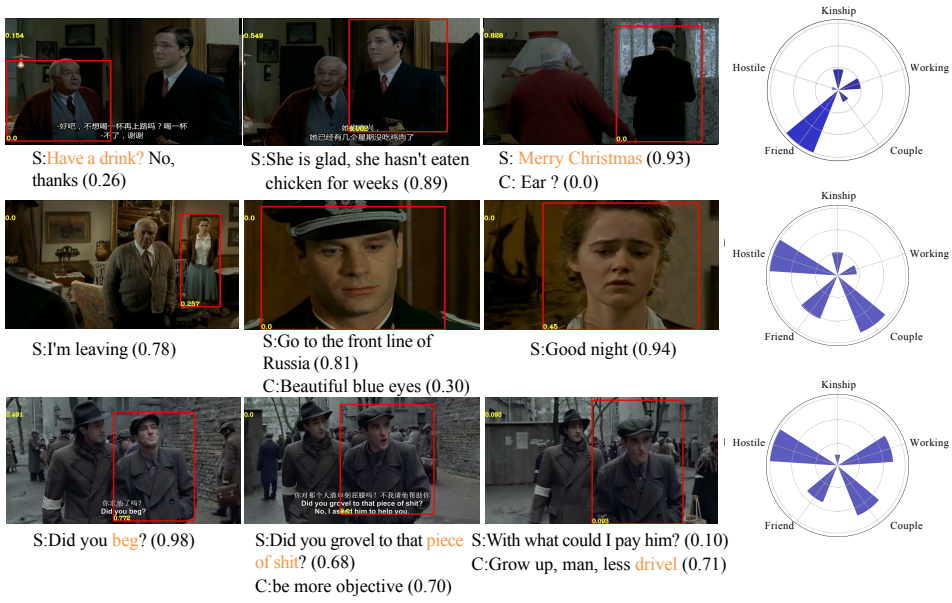

Fig. 8. Failure cases predicted by TEFM with multimodal information. The polar graph beside each case indicates the tendency for each trait to be positive.

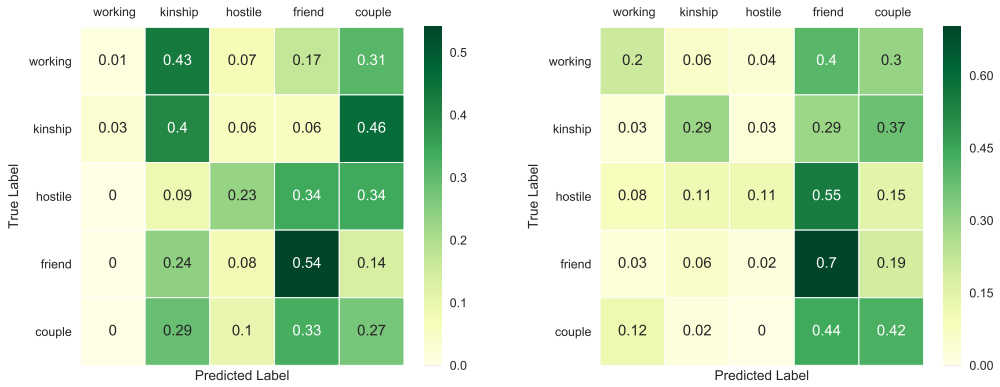

Fig. 9. The normalized fusion matrix of TEFM model With (right) or without (left) textual feature.

reasoning. Currently, we simply adopt a standard ResNet-50 to learn general visual cues. In future, more semantic visual concepts should be learned to benefit social relation recognition in videos.

Besides, the comparison between these two matrices also support the benefit brought by textual information. To be specific, be consistent with the argument mentioned in section 4.4, text promotes the performance of F1-value on three relation classes, especially when distinguishing "working" relation. However, the utilization of text does not bring an improvement on "hostile" relation due to a violent imbalance between precision and recall, this phenomenon teaches us the multimodal fusion should be further studied and be adapted to the different characteristics of visual and textual clues, and bring into full play the advantages of each modality.

5 CONCLUSION

In this paper, we proposed a novel multimodal-based solution to deal with the social relation recognition task. Specifically, we capture the target characters via a search module, and then designed a multi-stream architecture for jointly embedding the visual and textual information, in which feature fusion and attention mechanism were adapted for better integrating the multimodal inputs, finally, supervised learning is aggregated to recognize social relations. Experiments on real-world dataset proved the effectiveness of our framework, revealed some interesting rules of multimodal cues, and further indicated challenges as well as future directions.

ACKNOWLEDGEMENT

This research was partially supported by grants from the National Key Research and Development Program of China (Grant No. 2018YFB1402600), and the National Natural Science Foundation of China (Grant No. 61727809, 61703386, U19A2079).

REFERENCES

[1] David M Blei, Andrew Y Ng, Michael I Jordan, and John Lafferty. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

[2] Joya CHEN, Hao DU, Yufei WU, Tong XU, and Enhong CHEN. 2020. Cross-modal video moment retrieval based on visual-textual relationship alignment. *SCIENTIA SINICA Informationis* 50, 6 (2020), 862–876.

[3] Lei Ding and Alper Yilmaz. 2010. Learning Relations among Movie Characters: A Social Network Perspective. *European Conference on Computer Vision* (2010), 410–423.

[4] Chunhui Gu, Chen Sun, Sudheendra Vijayanarasimhan, Caroline Pantofaru, David A. Ross, George Toderici, Yeqing Li, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jagannath Malik. 2017. AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2017), 6047–6056.

[5] Omar Hamdoun, Fabien Moutarde, Bogdan Stanciulescu, and Bruno Steux. 2008. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. *Second ACM/IEEE International Conference on Distributed Smart Cameras* (2008), 1–6.

[6] K. He, X. Zhang, S. Ren, and J. Sun. 2015. Deep Residual Learning for Image Recognition. *ArXiv e-prints* (Dec. 2015). arXiv:cs.CV/1512.03385

[7] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In Defense of the Triplet Loss for Person Re-Identification. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* abs/1703.07737 (2017).

[8] Anthony Hu and Seth Flaxman. 2018. Multimodal Sentiment Analysis To Explore the Structure of Emotions. *KDD* abs/1805.10205 (2018).

[9] Qingqiu Huang, Wentao Liu, and Dahua Lin. 2018. Person Search in Videos with One Portrait Through Visual and Temporal Links. *CoRR* abs/1807.10510 (2018).

[10] Lv J and Wu B. 2019. Spatio-temporal attention model based on multi-view for social relation understanding. *International Conference on Multimedia Modeling* (2019), 390–401.

[11] Yu-Gang Jiang, Zuxuan Wu, Jinhui Tang, Zechao Li, Xiangyang Xue, and Shih-Fu Chang. 2017. Modeling Multimodal Clues in a Hybrid Deep Learning Framework for Video Classification. *IEEE Transactions on Multimedia* 20 (2017), 3137–3147.

[12] Rémi Lajugie, Damien Garreau, Francis R. Bach, and Sylvain Arlot. 2014. Metric Learning for Temporal Sequence Alignment. *NIPS* abs/1409.3136 (2014).

[13] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. 2015. Person re-identification by Local Maximal Occurrence representation and metric learning. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 2197–2206.

[14] Xinchen Liu, Wu Liu, Meng Zhang, Jingwen Chen, and Lianli Gao. 2018. Social Relation Recognition from Videos via Multi-scale Spatial-Temporal Reasoning. *CVPR* (2018).

[15] Xiang Long, Chuang Gan, Gerard de Melo, Jiajun Wu, Xiao Liu, and Shilei Wen. 2017. Attention Clusters: Purely Attention Based Local Feature Integration for Video Classification. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2017), 7834–7843.

[16] Guangyi Lv, Tong Xu, Enhong Chen, Qi Feng Liu, and Yi Zheng. 2016. Reading the Videos: Temporal Labeling for Crowdsourced Time-Sync Videos Based on Semantic Embedding. *AAAI*, 3000–3006.

[17] Guangyi Lv, Tong Xu, Qi Liu, Enhong Chen, Weidong He, Mingxiao An, and Zhongming Chen. 2019. Gossiping the Videos: An Embedding-Based Generative Adversarial Framework for Time-Sync Comments Generation. *PAKDD* (2019).

[18] Guangyi Lv, Kun Zhang, Le Wu, Enhong Chen, Tong Xu, Qi Liu, and Weidong He. 2019. Understanding the Users and Videos by Mining a Novel Danmu Dataset. *IEEE Transactions on Big Data* (2019).

[19] Jinna Lv, Wu Liu, Linghong Linda Zhou, Bai Jun Wu, and Huadong Ma. 2018. Multi-stream Fusion Model for Social Relation Recognition from Videos. *MMM* (2018).

[20] Bingpeng Ma, Yu Su, and Frédéric Jurie. 2012. Local Descriptors Encoded by Fisher Vectors for Person Re-identification. *ECCV Workshops*.

[21] Dai P, Lv J, and Wu B. 2019. Two-Stage Model for Social Relationship Understanding from Videos. *IEEE International Conference on Multimedia and Expo (ICME)* (2019), 1132–1137.

[22] Seung-Bo Park, Kyeong-Jin Oh, and GeunSik Jo. 2011. Social network analysis in a movie using character-net. *Multimedia Tools and Applications* 59 (2011), 601–627.

[23] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Multi-level Multiple Attentions for Contextual Multimodal Sentiment Analysis. *IEEE International Conference on Data Mining (ICDM)* (2017), 1033–1038.

[24] Bryan James Prosser, Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. 2010. Person Re-Identification by Support Vector Ranking. *BMVC*.

[25] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (2015), 1137–1149.

[26] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. 2019. Annotating Objects and Relations in User-Generated Videos. *ICMR* (2019).

[27] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. 2017. Video Visual Relation Detection. *ACM Multimedia* (2017).

[28] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. 2018. End-to-End Deep Kronecker-Product Matching for Person Re-identification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), 6886–6895.

[29] Nitish Srivastava and Ruslan Salakhutdinov. 2012. Multimodal Learning with Deep Boltzmann Machines. *The Journal of Machine Learning Research.* 15 (2012), 2949–2980.

[30] Qianru Sun, Bernt Schiele, and Mario Fritz. 2017. A Domain Based Approach to Social Relation Recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 435–444.

[31] Gokhan Tanisik, Cemil Zalluhoglu, and Nazli Ikizler-Cinbis. 2015. Facial Descriptors for Human Interaction Recognition In Still Images. *ArXiv* abs/1509.05366 (2015).

[32] Quang Dieu Tran and Jai E. Jung. 2015. CoCharNet: Extracting Social Networks using Character Co-occurrence in Movies. *J. UCS* 21 (2015), 796–815.

[33] Valentin Vielzeuf, Stephane Pateux, and Frederic Jurie. 2017. Temporal multimodal fusion for video emotion classification in the wild. *ICMI 2017*.

[34] Chung-Yi Weng, Wei-Ta Chu, and Ja-Ling Wu. 2009. RoleNet: Movie Analysis from the Perspective of Social Networks. *IEEE Transactions on Multimedia* 11 (2009), 256–271.

[35] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. 2019. Unified Visual-Semantic Embeddings: Bridging Vision and Language With Structured Meaning Representations. *CVPR*, 6609–6618.

[36] Yuanlu Xu, Bingpeng Ma, Rui Huang, and Liang Lin. 2014. Person Search in a Scene by Jointly Modeling People Commonness and Person Uniqueness. *ACM Multimedia* (2014).

[37] Wei Y, Wang X, and et al. Guan W. 2019. Neural multimodal cooperative learning toward micro-video understanding. *IEEE Transactions on Image Processing* 29 (2019).

[38] Yan Yan, Tianbao Yang, Zhe Li, Qihang Lin, and Yi Yang. 2018. A Unified Analysis of Stochastic Momentum Methods for Deep Learning. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence* (2018).

[39] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. 2017. CityPersons: A Diverse Dataset for Pedestrian Detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 4457–4465.

[40] Shanshan Zhang, Jian Xi Yang, and Bernt Schiele. 2018. Occluded Pedestrian Detection Through Guided Attention in CNNs. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), 6995–7003.

[41] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2015. Learning Social Relation Traits from Face Images. *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), 3631–3639.

[42] Peilun Zhou, Tong Xu, Zhizhuo Yin, Dong Liu, Enhong Chen, Guangyi Lv, and Changliang Li. 2019. Character-oriented Video Summarization with Visual and Textual Cues. *IEEE Transactions on Multimedia* (2019).