IEICE TRANS. INF. & SYST., VOL.E102–D, NO.12 DECEMBER 2019

PAPER

Attentive Sequences Recurrent Network for Social Relation Recognition from Video

Jinna LV[†,††a)], ***Member***, **Bin WU**[†b)], **Yunlei ZHANG**[†c)], ***Nonmembers***, ***and*** **Yunpeng XIAO**[†††d)], ***Member***

SUMMARY Recently, social relation analysis receives an increasing amount of attention from text to image data. However, social relation analysis from video is an important problem, which is lacking in the current literature. There are still some challenges: 1) it is hard to learn a satisfactory mapping function from low-level pixels to high-level social relation space; 2) how to efficiently select the most relevant information from noisy and unsegmented video. In this paper, we present an Attentive Sequences Recurrent Network model, called ASRN, to deal with the above challenges. First, in order to explore multiple clues, we design a Multiple Feature Attention (MFA) mechanism to fuse multiple visual features (i.e. image, motion, body, and face). Through this manner, we can generate an appropriate mapping function from low-level video pixels to high-level social relation space. Second, we design a sequence recurrent network based on Global and Local Attention (GLA) mechanism. Specially, an attention mechanism is used in GLA to integrate global feature with local sequence feature to select more relevant sequences for the recognition task. Therefore, the GLA module can better deal with noisy and unsegmented video. At last, extensive experiments on the SRIV dataset demonstrate the performance of our ASRN model.

key words: *social relation recognition, video analysis, deep learning, LSTM, attention mechanism*

1. Introduction

With the development of internet and multimedia, massive video data is generated by the social media, daily surveillance, etc. A large proportion of these videos are human-centric. As an important entity type, people communicate with each other through various medias, and the social relations between them are hidden behind video information. In computer vision, social information has been exploited to improve several analytic tasks, including human trajectory prediction [1], [2], group activity recognition [3], [4], and relation network construction [5], [6]. However, social relation recognition from video has received less attention and is far from solved by the community. In this paper, we aim to address the problem of social relation recognition from video.

Fig. 1 Challenges in video data: persons are always side faces and appear in different video images. The image with red frame is an example from image dataset, meanwhile, the images with blue frames are examples from video dataset.

Most of the existing studies for social relation recognition are based on image data. These studies learn social relation traits of pairwise persons from static images, including facial, distance, and objects features [7]–[9]. However, these methods cannot be utilized in video analysis. As shown in Fig. 1, different from image data, people in video may not appear in the same image. Moreover, they usually only have the profile or the back. Therefore, new feature representation method and recognition model for the analysis of social relations in the video need to be proposed.

Understanding social relations is natural for human, however, it is still a fundamental research challenge for Artificial Intelligence (AI). Specially, it is hard to learn a mapping function from low-level video pixels to high-level social relation space due to great intra-class variance. Early methods have investigated appearance time of people for rough relation analysis [5], [10]. Later, some researches have exploited models to recognize family or kin relationship using low-level features of face images [7], [11], [12]. For example, Zhang et al. [7] employed a Siamese-like deep convolutional network to learn a mapping function from raw pixels of a pair of faces to relation traits. However, owing to the complexity of interactions and multi-angle appearance of persons in video, simple feature cannot adequately reflect the high level social relation space. Therefore, advanced methods for feature representations and map functions from raw pixels to high-level social relation space need to be proposed.

Manuscript received April 10, 2019.
Manuscript revised July 22, 2019.
Manuscript publicized September 2, 2019.

[†]The authors are with the Beijing University of Posts and Telecommunications, Beijing, 100876, China.
[††]The author is with the Beijing Information Science & Technology University, Beijing, 100192, China.
[†††]The author is with the Chongqing University of Posts and Telecommunications, Chongqing, 400065, China.
a) E-mail: lvjinna@bupt.edu.cn
b) E-mail: wubin@bupt.edu.cn
c) E-mail: yunlei0518@bupt.edu.cn
d) E-mail: xiaoyp@cqupt.edu.cn
DOI: 10.1587/transinf.2019EDP7104

Another challenge is that not all of the sequence fragments in the long video are closely related to the social relation recognition task. In recent years, some studies worked on well segmented sequence videos, or all of the available frames in video sequences [13]. Some methods simply computed summary statistics the relation traits over the whole video [14]. However, not every happening interaction during the video sequences will be relevant to recognize the social relations. Existing methods for relation analysis do not address this issue. According to psychological research [18], the perception system of the human can quickly adjust the time selection mechanism to deal with the most relevant information. However, for AI, how to locate the crucial sequences using attention mechanism to achieve more accurate recognition results is a real challenge.

In view of the challenges in social relation recognition from videos, we propose an Attentive Sequences Recurrent Network (ASRN). First, our model extracts the high-level semantic features of video frames and persons appeared in them. A Multiple Feature based on Attention module (MFA) is designed to fuse these features carefully, which weighted fuse different features (i.e. image, motion, body, and face) to get a powerful representation for social relation. Second, people always obey a large number of common sense rules and comply with social conventions when they interacting with each other [19]. For instance, when one listens to another person's speech, there will be an expression, language, or a response to the body. Therefore, we employ a recurrent sequence network to extract these features. Furthermore, in order to selectively locate prime sequences for the task of social relation recognition, we propose a recurrent network based on Global and Local Attention (GLA). In this way, the model can focus on key sequences according to the global video information for the social relation recognition.

The innovation of this article includes three points:

- In order to build a mapping function from low-level pixels to high-level social relation space, we integrate multiple features of video frames by the MFA module. Through this approach, we can get a powerful visual representation for each time step of video sequences.
- We propose a recurrent network based on Global and Local Attention (GLA), which solves the problem of recognizing social relations from unsegmented videos. It automatically localizes the task-relevant parts to improve the performance for social relation recognition.
- Our ASRN model outperforms previous state-of-the-art algorithms on SRIV datasets. Extensive experiments demonstrate that the ASRN model achieves noticeable gains by appropriately integrating multi-feature of frames and global-local feature of video sequences into attention mechanism.

The rest of this paper is organized as follows. The related work is briefly presented from three aspects in Sect. 2. In Sect. 3, the proposed ASRN model is described in detail. The experimental results are provided and discussed in Sect. 4. Finally, Sect. 5 concludes the paper.

2. Related Work

Social relation recognition still has remained a challenging problem. The existing approaches generally fall into two categories: image-based [7], [8], [13], [21] and video-based [5], [22]–[24] methods.

For social relation recognition, some methods have investigated several low level featuresto recognize the social relations of pairwise persons from images, such as color, facial distance, and gradient histogram [13], [20]. For example, Tannisil et al. [13] explored the facial and spatial features of people for their interaction recognition. Later, several advanced recognition methods have been proposed in light of the two streams Convolutional Neural Networks (CNNs). For example, Sun et al. [8] proposed an end to end deep network based on multiple domain features. The main advantage of the social cues is that they think, to a large extent, faces features by integrating with the relative position of them is a effective approach [7], [8], [21].

Social relation recognition from video is still a field in its childhood. Some studies have analyzed the interactive behavior of the characters in the video to recognize and predicted human behavior. For example, Tran et al. [5] focused on the appearance of each character during movie play and analyzed the characters' relationships. Some studies have demonstrated the advantage of the integration of multiple modalities (vocal, text, and visual expression) in sentiment analysis [22]–[24]. Vicol et al. [25] released a video dataset for social situation analysis based on graph. However, these methods are only focus on rough relations, which are rarely based on social psychology and sociology to examine the specific social interaction of people. In addition, they are based on short or segmented videos. If we can make use of the complementarity of them, it will help to analyze the social relationships of people in complex videos.

2.1 Recurrent Neural Networks

Recurrent Neural Network (RNN) is a general term for time recursive neural networks and structural recursive neural networks. It has a wide range of applications in Natural Language Processing (NLP) [15], [17] and video description [26], [27].

Most early methods have been achieved great success and wide applications in the domain of NLP, such as language modeling and generating text [28], machine translation [29], and speech recognition. These studies have investigated the sequence feature of words in one sentence to predict the next word. Later, several advanced RNN models have been proposed [16]. Specially, Gated Recurrent Unit (GRU) and LSTM are among the most popular architectures due to their effective solutions to the vanishing gradient problem and power in modeling the dynamics and dependencies in sequential data [2], [15]. For example, Alahi et al. [2] proposed an LSTM model to learn general human movement and predict their future trajectories in video.

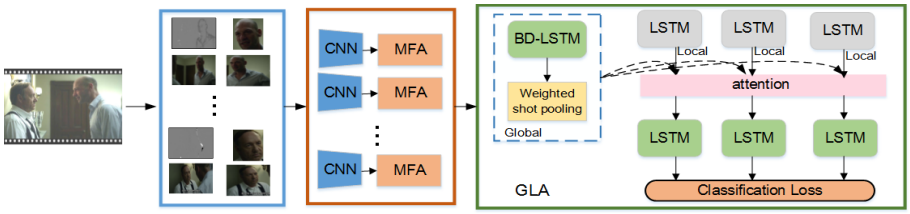


Fig. 2 Illustration of our proposed ASRN social relation recognition framework.

Most of these RNN-based models with temporal structure were designed for well-segmented sequences which have been edited to remove noisy or irrelevant information. In addition, there is no recurrent network for social relation recognition, to our best knowledge. Therefore, new sequence models based on noisy sequences expected in real-world applications for social relation recognition need to be proposed.

2.2 Attention Models

Human always concern about visual information of interest when they looking at something. Some researches have found that visual attention is attracted by the most informative regions [30], [31]. In deep learning domain, attention mechanism has been applied to video description [32], image and action classifications [34], [35], and entity disambiguation in texts [36] to learn more critical parts of the data.

On the one hand, some attention models based on CNNs have been proposed in various applications. These methods achieved more excellent performance than no-attention models. For example, Yu et al. [33] introduced Gaze Encoding Attention Network (GEAN), which can leverage gaze tracking information to provide the spatial and temporal attention for video captioning. Zhu et al. [35] proposed a spatial regularization net that using attention mechanism to learn the more related regions for different labels. On the other hand, attention mechanism was also used in the sequence learning model. For example, Pei et al. [22] proposed different attention GRU model, which can learn attention score of the sequential data. However, in their models, computation of attention weights is based on each sequence data without taking into the global video feature consideration. This limitation is mainly due to LSTM's restriction in perceiving the global contextual information, which is often crucial for the global analysis for the social relation. Our attention model adds global video features when calculating the weights of sequences, so that we can more precisely locate relevant segments.

3. Proposed Model

In this section, we first briefly review the ASRN model, as shown in Fig. 2. This model is introduced from two modules: multi-feature fusion module based on attention mechanism, named MFA, attention mechanism based on global and local, named GLA. First, in order to build a powerful mapping function from raw pixels to high-level social relation traits, multiple features for social relation traits are extracted, including the deep features from RGB images, optical flow images, and person's face and body images. These features are extracted by different CNNs, and then integrated by the MFA module. Second, we employ an attention mechanism GLA based on global and local features, in which the input of each time step is the result of the module MFA. The global feature of the video can be obtained by an improved LSTM network based on shot pooling. Specially, in GLA module, another LSTM based on attention mechanism is proposed to lean the temporal characteristics of the social relations based on global and local features. In this way, related parts of the video for social relations can be located, thus more accurate results of social relation recognition will be obtained by the ASRN model.

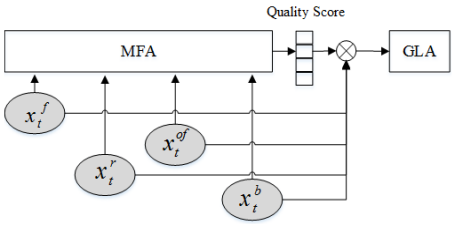


Fig. 3 Illustration of the MFA.

3.1 Multi-Feature Fusion Based on Attention Mechanism (MFA)

Humans communicate with each other using a highly complex structure of multiple signals (e.g. scene, motion, person expression). We observe that these different modalities have different contributions for the social relation recognition. Specifically, we extend the concept of attention model to measure the relevance of each observation (time step) of a sequence. In our model, a multi-feature fusion method based on attention mechanisms to solve this challenge, named MFA, as shown in Fig. 3.

Given a video, we extract RGB and optical flow frames of the video, body and face images for each person. We use the intermediate filter of the CNNs to obtain the feature representation of the frame sequence. These four types of images are fed into different CNNs, respectively. Accordingly, we can get deep features of the RGB, optical flow, person's body and face images. These feature vectors are denoted as $\mathbf{x}^r$, $\mathbf{x}^{of}$, $\mathbf{x}^b$, and $\mathbf{x}^f$, respectively. Instead of a naive av-

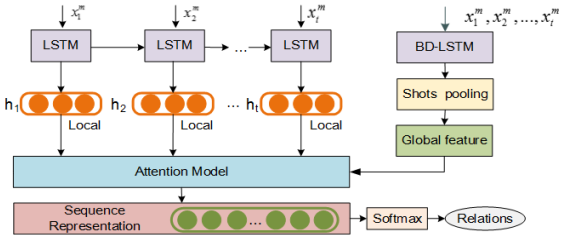


Fig. 4 Overview of the GLA.

eraging of the feature vectors, a soft attention model calculates weights c_t^i for each feature vector at time step t. In this way, important feature parts can be focused on at every time step. We employ a sequential softmax layer to obtain a set of quality scores $\{c_t^1, c_t^2, \ldots, c_t^m\}$ that quantify the relevance $\{\mathbf{x}_t^1, \mathbf{x}_t^2, \ldots, \mathbf{x}_t^m\}$, where m is the number of feature vectors. Therefore, we can obtain the fusion feature vector $\mathbf{y}_t$ for the t time step.

$$c_t^i = \frac{\exp(\mathbf{x}_t^i)}{\sum_{j=1}^m \exp(\mathbf{x}_t^j)}, \tag{1}$$

$$\mathbf{y}_t = \sum_{i=1}^m c_t^i \mathbf{x}_t^i, \tag{2}$$

Finally, we obtain the input to the LSTM $\mathbf{x}^m$ via a fully connected layer after concatenating the input with the previous outputs:

$$\mathbf{x}^m = \phi(\mathbf{W}\mathbf{Y} + b) \tag{3}$$

here, $\mathbf{W}$ is the parameter matrix to be learned, $\mathbf{Y}$ is the feature vectors $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_t, \ldots \mathbf{y}_n)$ of the video sequence, and n is the length of the time steps. We apply dropout to this multimodal layer to reduce overfitting.

3.2 Sequence Recurrent Network Based on Global and Local Attention (GLA)

In order to solve the feature learning problem in unsegmented video, a novel attention model is introduced for automatic learning to recognize social relations based on global and local features, named GLA. The overview of the GLA module is shown in Fig. 4.

The GLA module mainly solves the following two problems. On the one hand, because there are many frame sequences in one video, attention mechanism is adopted in order to highlight the task-related clips. On the other hand, we introduce not only local frame sequence feature but also global feature of video based on shot pooling in our attention model. In this way, GLA module can not only be easily trained by end to end method, but also more accurately recognize social relations.

(1) Local feature.

We use the multi-feature of relation traits in each frame as its local feature context, denoted as $(\mathbf{x}_1^m, \mathbf{x}_2^m, \ldots, \mathbf{x}_t^m, \ldots, \mathbf{x}_n^m)$. $\mathbf{x}_t^m$ is the local feature of the t-th time step, which can be computed using the formulas (1-3) in MFA module.

(2) Global feature.

The earlier content has been forgotten of long video with LSTM network, so only taking the last layer of LSTM as the global feature of the video is inaccurate. Although a video contains multiple frames, these frames are regularly distributed in each shot. In our GLA module, the global feature of video is denoted by computing shot pooling feature after video segmentation by shots.

First, we use the Boundary Detection LSTM (BD-LSTM) [27] to segment video into shots, which is described as the following Eqs. (4)-(6). The input to the t-th BD-LSTM is x_t^m, which is computed by the formulas (1-3) in the MFA model. Many different LSTM architectures have been proposed. In our model, we apply the following equations.

$$\begin{aligned}
\mathbf{i}_t &= \sigma(W_{iy}\mathbf{x}_t^m + W_{ih}\mathbf{h}_{t-1} + b_i) \\
\mathbf{f}_t &= \sigma(W_{fy}\mathbf{x}_t^m + W_{fh}\mathbf{h}_{t-1} + b_f) \\
\mathbf{g}_t &= \sigma(W_{gy}\mathbf{x}_t^m + W_{gh}\mathbf{h}_{t-1} + b_g) \\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t \\
\mathbf{o}_t &= \phi(W_{fy}\mathbf{x}_t^m + W_{fh}\mathbf{h}_{t-1} + b_f) \\
\mathbf{h}_t &= \mathbf{o}_t \odot \phi(\mathbf{c}_t).
\end{aligned} \tag{4}$$

We use this memory unit as the basic unit of the first LSTM network for shot segmentation. The boundary detector $s_t \in \{0, 1\}$ is obtained by the following equations:

$$\begin{aligned}
s_t &= \tau(\mathbf{v}_x^T \cdot (W_{si}\mathbf{x}_t^m + W_{sh}\mathbf{h}_{t-1} + b_s)) \\
\tau(x) &= \begin{cases} 1, if\sigma(x) > 0.5 \\ 0, otherwise, \end{cases}
\end{aligned} \tag{5}$$

where $\mathbf{v}_x^T$ is learnable row vector and W_{sh}, b_s are learned weights and biases.

Calculate information inputs to memory cell based on s_t:

$$\begin{aligned}
\mathbf{h}_{t-1} &\leftarrow \mathbf{h}_{t-1} \cdot (1 - s_t) \\
\mathbf{c}_{t-1} &\leftarrow \mathbf{c}_{t-1} \cdot (1 - s_t).
\end{aligned} \tag{6}$$

Then, due to the frames in one shot are very similar, the feature vector of one shot can be described by the average of all the features it contains using a pooling layer. According to the Eq. (5), if $s_t = 1$ denotes that the current frame is shot boundary, we take this $\mathbf{h}_t$ as the current shot feature $\mathbf{F}_t^s = \mathbf{h}_t$. So we can obtain the sequence of shot features $(\mathbf{F}_1^s, \mathbf{F}_2^s, \ldots, \mathbf{F}_t^s, \ldots, \mathbf{F}_n^s)$.

At last, the global video feature is defined as

$$\mathbf{V}_g = \sum_{s=1}^{S} \omega_s \mathbf{F}_t^s, \tag{7}$$

where ω_s is the weight of the $s - th$ shot clip.

(3) Attention model.

Obviously, for a long video, each clip sequence usually corresponds to special feature of the social relation. However,

how to find out which sequence clips better reflects the character of relations between people. For this purpose, we adopt attention mechanism to integrate the local feature with global feature.

Specifically, given a hidden vector $\mathbf{h}_t$ of a shot feature from a LSTM block at time step t, and the attention vector of the whole video $\mathbf{V}_g$, the re-weighted hidden vector $\mathbf{h}_t^{'}$ is defined by:

$$\begin{aligned} \mathbf{A}_t &= \tanh(w_v \mathbf{V}_g + w_h \mathbf{h}_t) \\ \mathbf{a}_t &= soft\max(w, \mathbf{A}_t) \\ \mathbf{h}_t^{'} &= \mathbf{h}_t \mathbf{a}_t, \end{aligned} \tag{8}$$

where w_v, w_h and w are attention parameters to be learned during training. By the formulas, $\mathbf{h}_t$ will be given more weight if it is more relevant to the attention vector $\mathbf{V}_g$. Therefore, the final video feature can be denoted as $\sum_{t=1}^{T} \mathbf{h}_t^{'}$.

As for prediction, we calculate the predicted score using sigmoid function by the following equation:

$$\hat{y} = sigmoid(w_{st}(\sum_{t=1}^{T} \mathbf{h}_t^{'}) + b_{st}). \tag{9}$$

The whole network is trained by the cross-entropy loss with the ground truth labels y,

$$\begin{aligned} F_{loss}(y, \hat{y}) = &\sum\nolimits_{i=1}^{n} y_i * \log(\hat{y}_i) \\ &+ (1 - y_i) * \log(1 - \hat{y}_i), \end{aligned} \tag{10}$$

where $\hat{y}$ is the probability of predicted class.

4. Experiments

4.1 Dataset and Methods in Comparison

(1) Dataset.

The dataset used in this paper is collected from movies and TV dramas, named SRIV [14]. The dataset is available at *https://github.com/happyheart866/SRIV*. SRIV is the first video dataset for social relation recognition from videos, to our best knowledge. It contains 3,124 videos with multi-label, about 25 hours, which is collected from 69 TV dramas and movies. The dataset contains Sub-Relation and Obj-Relation classes including 16 subclasses, which is shown in Table 1.

(2) Comparison methods.

In the experiments, we compare the ASRN model with a few state-of-the-art baselines by conducting extensive experiments on the SRIV dataset:

C3D: A network structure based on 3D convolution was proposed, which has excellent performance in video feature extraction [37].

TSN: TSN [38] is a typical two-stream CNN network which has achieved the state-of-the-art performance on many video classification datasets.

Table 1 The statistics of the number for each class on SRIV.

Sub-Relation			
Dominant	Competitive	Trusting	Warm
770	840	1614	1482
Friendly	Attached	Inhibited	Assured
2221	600	594	810
Obj-Relation			
Supervisor	Peer	Service	Parent
627	469	238	321
Mating	Sibling	Friendly	Hostile
600	141	1073	434

Multi-stream: Multiple features representing social relations were used to improve the recognition performance [14].

LSTM: The basic LSTM model [39], which is a popular technique for sequence modeling with various improved.

ASRN (no GLA): In our ASRN model, without attention module, we use LSTM network replace the GLA module with the multi-feature as input.

ASRN (no MFA): In our ASRN model, we only employ image feature without the multi-feature attention (MFA) module.

ASRN: The ASRN model, which fusion of multi-feature (i.e. image, motion, body, and face) and can be easily trained by end to end.

4.2 Feature Extraction and Parameter Settings

When training a sequence relation classification model, as the first step, we need to extract feature vectors that serve as input to the sequence model. CNNs have shown their powerful representation learning abilities in various image classification tasks. In our model, we first extract the video frames, including RGB and optical flow frames. Then, different CNNs are used to extract multiple kinds of video features. For RGB feature, we use Resnet101 (trained on the ImageNet dataset) to extract frame features, with results in 2048 features. For motion feature, the TSN network (trained on the SRIV dataset) is used. We employ a fixed length feature vector every 5 pair optical flow frames, which encodes motion features computed around the middle of the window. In our experiment, we use Faster-CNN to detect the bodies and faces of persons appeared in the video frames. In addition, we choose the body or face that occupying the larger proportion of the current image to extract deep features. For body feature, we use VGG19 (trained on the Market101 person re-id dataset) to extract body feature of person appearing in each frame. For face feature, we use Deepid (trained on the YouTube face dataset) to extract deep face feature of person, with results in 260 dimensions feature.

Our experiment environment is comprised of a Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz processors running at 2.10GHz with two TITAN X (Pascal) GPUs, and the system is Ubuntu 16.04. Our model is trained in an end to end fashion with Stochastic Gradient Descent (SGD) as the optimizer. In our experiment, the batch size of temporal networks is set to 32, the basic learning rate is set to 0.01 and

Table 2 Performance of different methods on Sub-Relation classes.

Methods	F_1_mi	F_1_ma	Acc	Sub_{acc}
C3D [37]	0.3958	0.3018	0.5568	0.1451
TSN [38]	0.6034	0.4894	0.5412	0.3045
Multi-Stream [14]	0.7019	0.6383	0.6136	0.5291
LSTM [39]	0.4714	0.4193	0.6547	0.3792
ASRN(no GLA)	0.6754	0.5613	0.5862	0.4045
ASRN(no MFA)	0.7258	0.6351	0.6528	0.4627
ASRN	**0.7353**	**0.6812**	**0.6722**	**0.5392**

Table 3 Performance of different methods on Obj-Relation classes.

Methods	F_1_mi	F_1_ma	Acc	Sub_{acc}
C3D [37]	0.4383	0.3886	0.0557	0.0347
TSN [38]	0.7142	0.6142	0.7089	0.3482
Multi-Stream [14]	0.8119	0.6683	0.7436	0.5213
LSTM [39]	0.6780	0.5776	0.6667	0.2797
ASRN(no GLA)	0.7358	0.6188	0.7286	0.4124
ASRN(no MFA)	0.7858	0.6351	0.7412	0.4641
ASRN	**0.8141**	**0.6766**	**0.7692**	**0.5259**

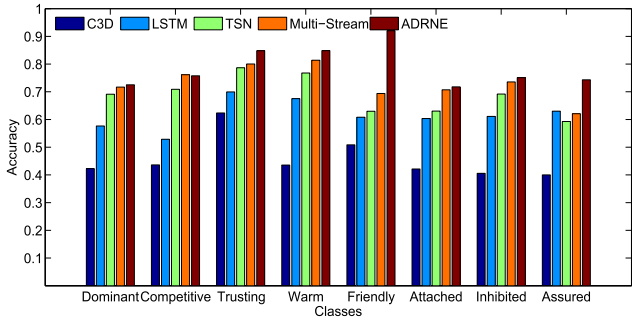
decreased to its 0.01 every one epoch. The max epoch is 200. In order to reduce overfitting, we apply dropout with probability 0.7 before the final fully connected layer. Owing to not every frame has facial and body features, we choose a sliding window with three frames and select the frame that all four features existing to form the video sequence.

4.3 Result and Discussion

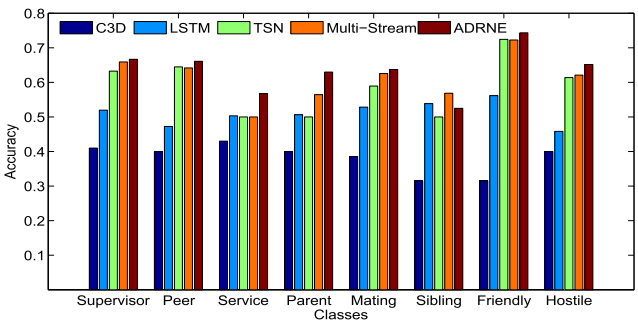
In this section, the performance of our method is evaluated from three viewpoints. First, we evaluate the performance of our method by comparing it with other baseline methods. Next, we compare the results of our method on different classes. Finally, we verify the impact of multi-feature on recognition results by comparing single feature. According to the above three viewpoints, the superiority of our method can be verified. For performance evaluation, we use the F_1_mi, F_1_ma, Acc, and Sub_{acc} [14].

Table 2 and 3 show the comparison results of our model with evaluation. Our ASRN achieves the best performance. This is because MFA and GLA modules can provide more hints information for social relation analysis. Thus powerful representation with multi-feature and attention for key sequences of the video can be obtained. The performance of the C3D method is pretty poor, which indicates that the whole feature of the video cannot carefully represent the detail characteristics of social relations between persons. In addition, compared with the best counterpart (i.e. Multi-stream) which takes three kinds of the whole video features without learning the sequence feature based on shot, our method has 3.44% improvement on accuracy. Specially, ASRN (no MFA) is out performed by ASRN. This indicates that the fusion of multiple features is crucial in understanding social relations. At last, we notice that the ASRN (no GLA) baseline under-performs compared to ASRN. This supports the importance of the GLA attention module.

Figure 5 shows the recognition performance of the sixteenth relations with our ASRN method. We can see that



(a) Performance of different methods on the Sub-Relation.

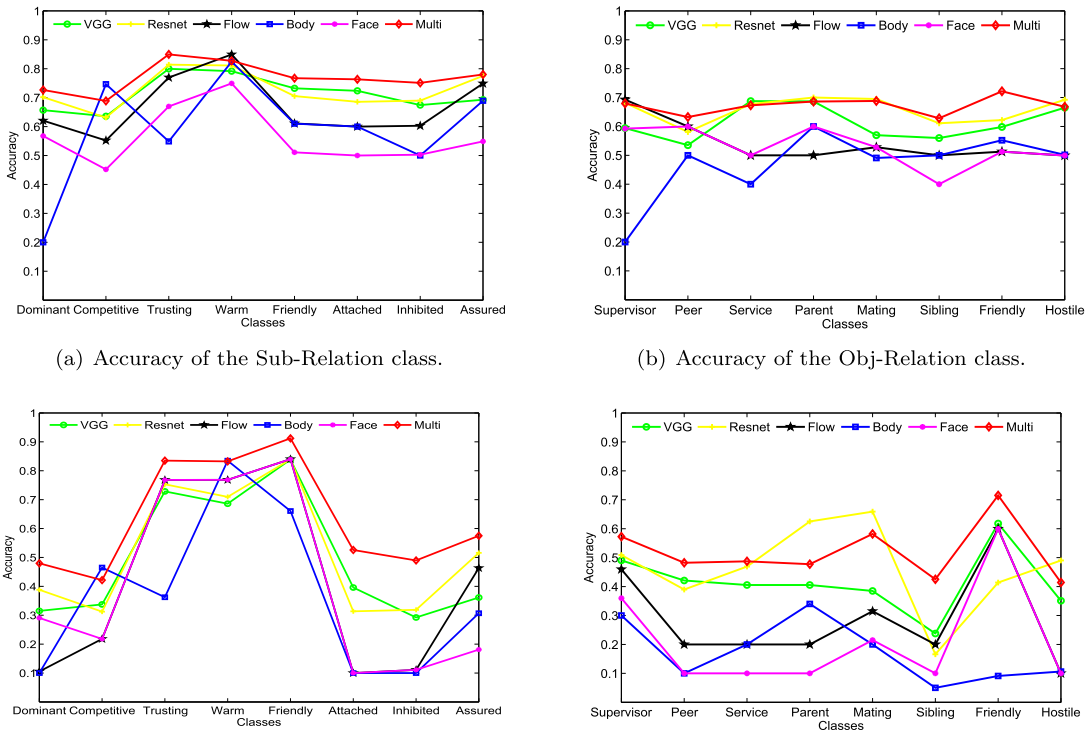


(b) Performance of different methods on the Obj-Relation.

Fig. 5 Social relation prediction performance of each classes.

the ASRN method has the best performance on almost all classes, which shows that ASRN has well generalization ability to classify relations more accurately. However, there are a few classes that are not the best performance with ASRN model. For example, the performance in term of accuracy on classes of "competitive" and "sibling" are lower than Multi-stream method. This explains that these classes do not have fine sequence features. Specially, the predictions of "friendly" class are significantly better than other classes. The reason is that "friendly" class has more training samples than others.

In order to verify our multi-feature fusion module MFA, we compare different features which are shown in Fig. 6. Form the comparison, we can find that the ASRN model using MFA can significantly improve the recognition performance on most classes. It demonstrates that the fusion of different information of video is definitely various useful for relation recognition. From the Fig. 6, we can see that the values of the classification accuracy obtained by different feature on ASRN model vary widely. As an example, the accuracy of body feature is very poor on the class "dominant", conversely very high on the class "warm". It suggests that different features may express the features of relationship between characters from different perspectives. Therefore, our ASRN model with multi-feature based on attention mechanics can better describe the character of the relations and scenes in the video, such as "inhibited", "friendly" and so on. However, some classes have not high accuracy enhancement with integration, because it is visually subtle compared to other relation classes, such as "peer". In some



(a) Accuracy of the Sub-Relation class.

(b) Accuracy of the Obj-Relation class.

(c) F1 of the Sub-Relation class.

(d) F1 of the Obj-Relation class.

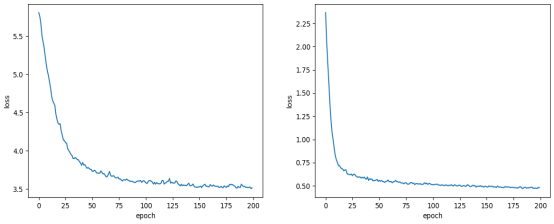
Fig. 6 The performance of accuracy and F1 curves of different features.

Fig. 7 Visualization of attention weights. The images are the keyframe of each shot. In the color bar, the light color denotes high weights while dark color denotes low weights.

classes, such as "Parent" and "Mating", the image feature can achieve the best performance on F_1. It suggests that the image feature can better recognize the "Parent" and "Mating" classes than multi-feature.

We show the attention mechanism results of our ASRN model on some examples in Fig. 7. These examples for temporal attention weights are visualized using thermodynamic diagram below the sample video frames. The color is from dark to light, the weight value is from small to large. We select five frames visualization from video frame sequence using the black arrows. Form the Fig. 7, we can see the last frame image of the first example has higher attention weight for the label "Mating". The reason may be that the intimate hug action feature can better reflect the characteristic of "Mating" relation. In the below example, the second and last images have higher weights for the relation "Sibling". This phenomenon may be because the face and body features play more important role, which can describe the visual feature of children, parents, etc. These phenomena indicate that our model can focus on key frames relating to the social relation recognition.

In our experiment environment, the training time for Sub-Relation and Obj-Relation is 2,735 and 2,345 seconds, respectively. Meanwhile, the testing time for them is 26 and 19 seconds respectively. The loss curves for ASRN model are shown in Fig. 8. Form the Fig. 8, we can see the ASRN model has good convergence. In addition, the performance of the ASRN model is analyzed when batch size and iteration parameters change, which is shown in Table 4. We can see that the ASRN model has the best performance when the batch size is set to 32 and the iteration is set to 200. At the top of the table shows the accuracy with different batch



(a) The Sub-Relation class. (b) The Obj-Relation class.

Fig. 8 The loss curves for training.

Table 4 Performance of different parameters.

Batch size	4	16	32	64
Accuracy	0.3142	0.4512	0.7692	0.5547
Iteration	50	100	200	500
Accuracy	0.4524	0.5147	0.7692	0.6836

size when epoch is set to 200. At the bottom of the Table, it shows the accuracy performance of the ASRN model with different iteration when batch size is set to 32.

5. Conclusion

In this paper, we propose the Attention-based Sequences Recurrent Network (ASRN) model for social relation recognition from video. Specially, the Multiple Feature Attention (MFA) module integrates multi-feature of video frame based on attention model. In this way, a powerful mapping function from raw pixels to high-level social relation traits can be built. More importantly, taking the advantage of attention mechanism, we develop a sequence recurrent network based on Global and Local Attention (GLA) module for selective attention on key sequences of video for social relation recognition. Compared with other attention methods based on LSTM, our GLA module can selectively focus on more important sequences at different time while keeping global video information through integrating local sequence features with global features via attention mechanism. Consequently, our ASRN model generates more relevant and coherent clip sequences which can describe the context of social relation. In the experiments, our model achieves the state-of-the-art performance for social relation recognition from video.

Acknowledgments

This research is supported by the National Key R&D Program of China (No. 2018YFC0831500), and the National Natural Science Foundation of China (NSFC) under Grant (No. 61972047).

References

[1] Y.-J. Lin and S.-K. Weng, "Trajectory Estimation of the Players and Shuttlecock for the Broadcast Badminton Videos," IEICE Transactions, vol.E101-A, no.10, pp.1730–1734, 2018.

[2] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," Proc. Int. IEEE Conf. Computer Vision and Pattern Recognition, pp.961–971, 2016.

[3] P.-S. Kim, D.-G. Lee, and S.-W. Lee, "Discriminative context learning with gated recurrent unit for group activity recognition," Pattern Recognition, vol.76, pp.149–161, 2018.

[4] T. Lan, Y. Wang, W. Yang, S.N. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," IEEE Trans. Pattern Anal. Mach. Intell, vol.34, no.8, pp.1549–1562, 2012.

[5] Q.D. Tran and J.E. Jung, "Cocharnet: Extracting social networks using character co-occurrence in movies," Journal of Universal Computerence, vol.21, no.6, pp.796–815, 2015.

[6] J. Lv, B. Wu, L. Zhou, and H. Wang, "StoryRoleNet: Social Network Construction of Role Relationship in Video," IEEE Access, vol.6, pp.25958–25969, 2018.

[7] Z. Zhang, P. Luo, C.-C. Loy, and X. Tang, "Learning social relation traits from face images," Proc. IEEE Int. Conf. on Computer Vision, pp.3631–3639, 2015.

[8] Q. Sun, B. Schiele, and M. Fritz, "A domain based approach to social relation recognition," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.435–444, 2017.

[9] M. Zhang, X. Liu, W. Liu, A. Zhou, H. Ma, and T. Mei, "Multi-Granularity Reasoning for Social Relation Recognition from Images," CoRR ABZ/1901.01067, 2019.

[10] C.-Y. Weng, W.-T. Chu, and J.-L. Wu, "RoleNet: Movie analysis from the perspective of social networks," IEEE Trans. Multimedia, vol.11, no.2, pp.256–271, 2009.

[11] S. Xia, M. Shao, J. Luo, and Y. Fu, "Understanding kin relationships in a photo," IEEE Trans. Multimedia, vol.14, no.4, pp.1046–1056, 2012.

[12] X. Tang, F. Guo, J. Shen, and T. Du, "Facial Landmark Detection by Semi-supervised Deep Learning," Neurocomputing, vol.297, pp.22–32, 2018.

[13] Z. Sun, Z.-P. Hu, R. Chiong, M. Wang, and W. He, "Combining the Kernel Collaboration Representation and Deep Subspace Learning for Facial Expression Recognition," Journal of Circuits, Systems, and Computers, vol.27, no.8, pp.1–16, 2018.

[14] J. Lv, W. Liu, L. Zhou, B. Wu, and H. Ma, "Multi-stream Fusion Model for Social Relation Recognition from Videos," Proc. IEEE Conf. Conference on Multimedia Modeling, pp.355–368, 2018.

[15] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective lstms for target-dependent sentiment classification," Coling, pp.3298–3307, 2016.

[16] J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic, and H.T. Shen, "From Deterministic to Generative: Multimodal Stochastic RNNs for Video Captioning," IEEE Transactions on Neural Networks and Learning Systems, vol.30, no.10, pp.3047–3058, 2019.

[17] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," CoRR abs/1508.01991, pp.1–10, 2015.

[18] L. Oksama and J. Hyönä, "Dynamic binding of identity and location information: a serial model of multiple identity tracking," Cognitive Psychol, vol.56, no.4, pp.237–283, 2008.

[19] A. Cureton, "Solidarity and social moral rules," Ethical Theory Moral, vol.15, no.5, pp.691–706, 2012.

[20] Y. Guo, H. Dibeklioglu, and L.V.D. Maaten, "Graph-based kinship recognition," Proc. Int. Conf. Pattern Recognition, pp.4287–4292, 2014.

[21] J. Li, Y. Wong, Q. Zhao, and M.S. Kankanhalli, "Dual-glance model for deciphering social relationships," Proc. IEEE Int. Conf. Computer Vision, pp.2669–2678, 2017.

[22] W. Pei, T. Baltrusaitis, D.M.J. Tax, and L.P. Morency, "Temporal attention-gated model for robust sequence classification," Proc. IEEE Int. Conf. Computer Vision, pp.820–829, 2016.

[23] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis," Psychol Bull, vol.11, no.2, pp.256–274, 1992.

[24] H. Yu, L. Gui, M. Madaio, A. Ogan, J. Cassell, and L.-P. Morency, "Temporally selective attention model for social and affective state

recognition in multimedia content," Proc. ACM on Multimedia Conference, pp.1743–1751, 2017.
[25] P. Vicol, M. Tapaswi, L. Castrejón, and S. Fidler, "MovieGraphs: Towards Understanding Human-Centric Situations From Videos," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.8581–8590, 2018.
[26] J. Xu, T. Yao, Y. Zhang, and T. Mei, "Learning multimodal attention lstm networks for video captioning," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.537–545, 2017.
[27] L. Baraldi, C. Grana, and R. Cucchiara, "Hierarchical boundary-aware neural encoder for video captioning," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.3185–3194, 2017.
[28] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," Proc. Int. Conf. Speech Communication Association, pp.1045–1048, 2010.
[29] M. Tufano, J. Pantiuchina, C. Watson, G. Bavota, and D. Poshyvanyk, "On Learning Meaningful Code Changes via Neural Machine Translation," arXiv preprint arXiv:1901.09102, 2019.
[30] J. Xu, T. Yao, Y. Zhang, and T. Mei, "Learning multimodal attention lstm networks for video captioning," Proc. ACM on Multimedia, pp.537–545, 2017.
[31] Y. Li, Z. Miao, M. He, Y. Zhang, and H. Li, "Deep Attention Residual Hashing," IEICE Transactions, vol.E101-A, no.3, pp.654–657, 2018.
[32] Y. Bin, Y. Yang, F. Shen, N. Xie, H.T. Shen, and X. Li, "Describing Video With Attention-Based Bidirectional LSTM," IEEE Transactions on Cybernetics, vol.49, no.7, pp.2631–2641, 2018.
[33] T. Rao, X. Li, H. Zhang, and M. Xu, "Multi-level region-based Convolutional Neural Network for image emotion classification," Neurocomputing, vol.333, pp.429–439, 2019.
[34] R. Girdhar and D. Ramanan, "Attentional pooling for action recognition," Proc. Conf. Neural Information Processing Systems, pp.34–45, 2017.
[35] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.2027–2036, 2017.
[36] M.C. Phan, A. Sun, Y. Tay, J. Han, and C. Li, "Neupl: Attention-based semantic matching and pair-linking for entity disambiguation," Proc. Conf. Information and Knowledge Management, pp.1667–1676, 2017.
[37] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.4489–4497, 2015.
[38] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L.V. Gool, "Temporal segment networks: towards good practices for deep action recognition," Proc. European Conference on Computer Vision, pp.20–36, 2016.
[39] N.V. Findler, "Short note on a heuristic search strategy in long-term memory networks," Inf. Process. Lett, vol.1, no.5, pp.191–196, 1972.

Jinna Lv received BE and ME degrees from Zhengzhou University, Zhengzhou, China, in 2006 and 2009, respectively. She received the PhD degrees from Beijing University of Posts and Telecommunications in 2019. She joined the Beijing Information Science & Technology University as a lecturer in 2019. Her research interests include multimedia content analysis, social relation extraction and social network analysis.

Bin Wu received B.S. degree from Beijing University of Posts and Telecommunications in 1991, and the M.S. and PhD degrees from the ICT of Chinese Academic of Sciences in 1998 and 2002, respectively. He joined the Beijing University of Posts and Telecommunications as a lecturer in 2002, and is currently a professor there. His research interests include data mining, complex network, and cloud computing. He has published more than 100 papers in refereed journals and conferences.

Yunlei Zhang received the B.E. and M.E. degrees in Computer Science and Technology from Hebei University of Science and Technology and Liaoning University of Technology in 2005 and 2009, respectively. He is a Ph.D. candidate in Computer Science and Technology at Beijing University of Posts and Telecommunications since 2014. His research interests focus on data mining, social computing, and social network analysis.

Yunpeng Xiao received the Ph.D. degree in computer science from Beijing University of Posts and Telecommunications, Beijing, China, in 2013. He is currently a Professor with Chongqing University of Posts and Telecommunications, Chongqing, China. He is the winner of Chongqing young scientific innovation talents. His research interests include social networks and machine learning.