Pattern Recognition Letters
journal homepage: www.elsevier.com

Deep Supervised Feature Selection for Social Relationship Recognition

Mengyin Wang[a], Xiaoyu Du[b], Xiangbo Shu[a,**], Xun Wang[c], Jinhui Tang[a]

[a]*School of Computer Science and Engineering, Nanjing University of Science and Technology, Jiangsu 210094, China*
[b]*School of Computing, National University of Singapore, Singapore 117417, Singapore*
[c]*School of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou 310018, China*

ABSTRACT

Social relationships link everyone in human society. Exploring social relationships in still images promotes researches of behaviors or characteristics among persons. Previous literature has discovered that face and body attributes can provide effective semantic information for social relationship recognition. However, they ignore that attributes contribute much differently to the recognition accuracy, and these multi-source attributes may contain redundancies and noises. This work aims to promote social relationship recognition accuracy by abstracting multi-source attribute features more efficiently. To this end, we propose a novel Deep Supervised Feature Selection (DSFS) framework to recognize social relationships in photos, which fuses the deep learning algorithm with $l_{2,1}$-norm to learn a discriminative feature subset from multi-source features by leveraging the face and body attributes. Experimental results on PIPA-relation dataset qualitatively demonstrate the effectiveness of the proposed DSFS framework.

Keywords: Social relationship recognition; Feature selection; Deep learning

1. Introduction

Social relationship is a bridge between people in society. Understanding social relationships automatically in images is beneficial to explore behaviors or characteristics among persons. It is also advantageous for many recognition tasks by providing useful supplementary information, such as human activity recognition (Lan et al., 2012) and occupation recognition (Shao et al., 2013).

For the past eight years, social relationship analysis has attracted lots of attention (Ramanathan et al., 2013; Sun et al., 2017; Goel et al., 2019; Wang et al., 2018; Zhang et al., 2015, 2018; Li et al., 2017; Liu et al., 2019; Shu et al., 2015; Zhang et al., 2019). Most literature models social relationships by auxiliary cues, which can be divided into two categories, *i.e*. contextual objects and face/body attributes. The pioneering works (Sun et al., 2017; Zhang et al., 2015) explore social information in images based on different face and body attributes, which provide comprehensive cues and make recognizing social information more accurate. According to social domain theory (Bugental, 2000), Sun *et al*. (Sun et al., 2017) collect several semantic attributes from face and body images for training their model. In their model, all face and body attributes are roughly concatenated into one feature vector. However, as our investigations shown in Fig. 5, each attribute makes different contributions to recognizing social relationships. If we remove any one attribute from all attributes, the recognition accuracy is a little lower than the baseline that concatenates all attribute features. This indicates that such concatenate feature proposed by Sun *et al*. (Sun et al., 2017) contains redundancy and noise. In addition, face and body attribute features are extracted from different models trained on their corresponding datasets. By concatenating these multi-source attribute features, a heterogeneous and high-dimensional vector is constructed. Since this concatenate feature serves as the input of computational model, it may cause some problems including over fitting, low efficiency and so on. In order to deal with such problems, we propose to utilize a feature selection model that integrates the deep learning algorithm to remove redundancy and improve the efficiency of abstracting multi-source data automatically.

**Corresponding author.
e-mail: shuxb@njust.edu.cn (Xiangbo Shu)

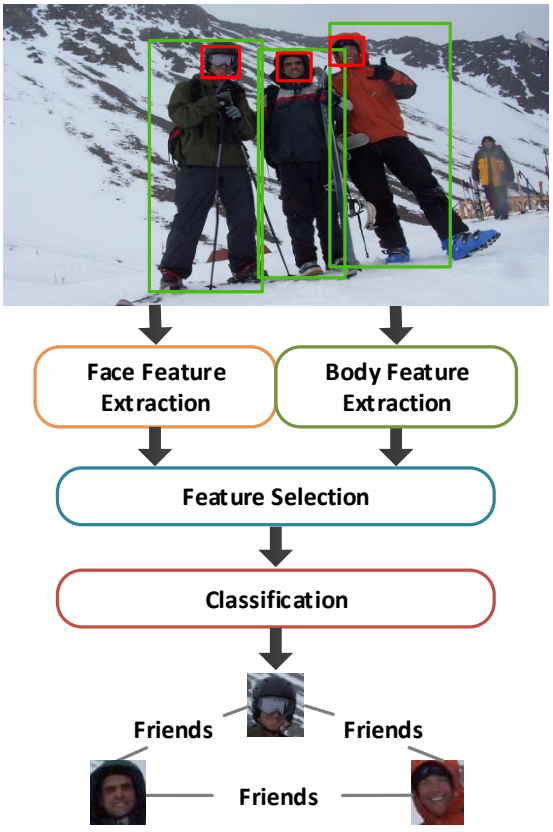


Fig. 1. Flowchart of the proposed deep supervised feature selection (DSFS) framework aims to recognize social relationships in images automatically. Face attributes representation and body attributes representation are extracted from the pre-trained face and body attributes classification models respectively. Then, the two types of representations are combined as the input of the feature selection module to select the optimal subset feature and recognize social relationships in images.

More specifically, we firstly perform attribute-level feature representation on any person pair which helps judge social relationships by supplying auxiliary information. For example, there are three persons in Fig. 1, they are smiling, front-facing, standing shoulder to shoulder, clothing in the same style, these attributes can give a hand to predict they are friends. Then, we propose to select an optimal feature subset from multi-source attribute features. For the task of social relationship recognition based on multi-source attributes, feature selection firstly removes noise and redundant information from the high-dimensional feature including multiple attributes. Then, the selected feature subset learns better understanding structure to represent social relationships. And, it makes computation more efficient by reducing the dimension of input features. As shown in Fig. 2, according to the contribution of each attribute feature for social relationship recognition, useful information of face and body attributes is extracted by the feature selection algorithm. Finally, we recognize the exact social relationship by using the optimal feature subset.

Accordingly, we develop a **D**eep **S**upervised **F**eature **S**election (**DSFS**) framework for social relationship recognition, as shown in Fig. 1. Our **DSFS** framework mainly contains three modules: feature extraction module, feature selection module, and classifier module. First, we employ several attribute features to boost representations of the input person pairs more discriminative. In the feature extraction module, we exploit 12 categories of the face and body attributes, including face/body age, face/body gender, face/body location & scale, face appearance, face pose, face emotion, body clothing, body proximity and body activity. The attribute features are extracted by using corresponding classification models, which act as the representation of input person pair. Second, we fuse the deep learning algorithm with $l_{2,1}$-norm to obtain an optimal feature subset from the concatenated feature automatically. In this module, we design two policies for feature selection, *i.e.* group feature selection and dimensional feature selection. Group feature selection selects the optimal group features at a coarse-grained level, which helps to discover the degree of relevance between face/body attributes and social relationship recognition. Dimensional feature selection is proposed to select a more refined feature subset and guarantee that redundancies are discarded at the dimensional level. Finally, we exploit the Softmax classifier to recognize the social relationship of each person pair.

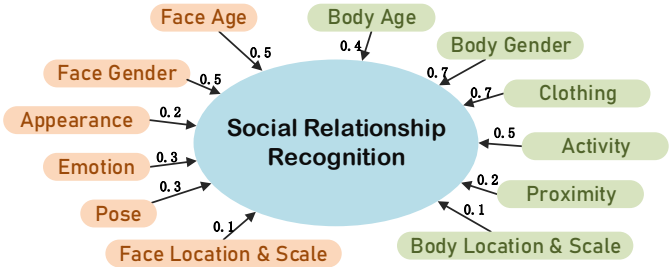


Fig. 2. Diagram of feature selection for social relationship recognition.

Our contributions in this work mainly include three-fold. First, we design a novel deep supervised feature selection framework to understand social relationships in images automatically. Second, we exploit two feature selection policies *i.e.* group feature selection and dimensional feature selection, to abstract optimal feature subset at attributive and dimensional level respectively. The selected optimal feature subset owns a better understanding structure, which is beneficial to investigate the contribution of different attributes. Third, experiments conducted on PIPA-relation dataset demonstrate our DSFS framework well boost the accuracy of social relationship recognition.

2. Related Work

In recent years, researchers pay attention to exploring social information in images or videos, such as kinship verification (Lu et al., 2014; Robinson et al., 2017), social roles (Ramanathan et al., 2013; Shu et al., 2015), social relation traits (Zhang et al., 2015, 2018; Yan and Song, 2019) and social relationships (Sun et al., 2017; Goel et al., 2019; Wang et al., 2018; Li et al., 2017; Liu et al., 2019; Zhang et al., 2019). Ramanathan *et al.* (Ramanathan et al., 2013) make the first attempt to describe social roles played by people in an event. They present the interactions along with roles by a conditional

random field, then infer weights of their model and role assignment in videos simultaneously. Zhang *et al*. (Zhang et al., 2015, 2018) propose the deep model by capturing multiple face attributes to classify social relation traits (*e.g*. warm, friendliness, and dominance) between two or more people. Besides, they formulate a bridging layer to leverage the inherent correspondences among heterogeneous attribute sources. Meanwhile, many researchers focus on analyzing kinship which is a subset of social relationships. Lu *et al*. (Lu et al., 2014) propose a new neighborhood repulsed metric learning (NRML) method for kinship verification task. Robinson *et al*. (Robinson et al., 2017) collect the largest kinship dataset (*i.e*. FIW dataset) and design several baseline frameworks on kinship verification task and family recognition task.

In terms of social relationship recognition task, most existing methods can be considered as two categories: contextual objects and multiple face or body attributes. Contextual objects have been adopted as basic cues in the existing social relationship recognition literature. Li *et al*. (Li et al., 2017) propose a dual-glance model for recognizing social relation in still images. The first glance performs coarse prediction based on appearance and geometrical information of the individual pair. The second glance integrates the attention mechanism into contextual objects generated by a region proposal network to refine the coarse prediction. Wang *et al*. (Wang et al., 2018) develop an end-to-end trainable graph reasoning model, which employs a Gated Graph Neural Network (GGNN) to propagate node message through the graph, and introduce a novel graph attention mechanism to reason key contextual objects around people in images.

Similarly, face or body attributes can provide effective information for understanding social relationships in images. Sun *et al*. (Sun et al., 2017) contribute a Double-Stream Caffenet to classifying social domain or social relationship. Based on social psychology, several face and body attribute features are collected to promote the accuracy of classification. Then they concatenate these features into one vector as the input of classifier. However, such concatenated feature is high-dimensional, which may contain noise and redundancy.

In this work, we focus on multiple face or body attributes for social relationship task. Our work aims to select an optimal subset from multi-source features to remove noise and promote the accuracy of social relationship recognition.

3. Deep Supervised Feature Selection (DSFS) Framework

In this section, we introduce the **D**eep **S**upervised **F**eature **S**election (DSFS) framework in details. We first formulate the task of social relationship recognition, followed by the overview of DSFS framework. And then, we elaborate on the key designs of DSFS. Finally, we describe the optimization method of our framework. Throughout this paper, we use bold uppercase characters to denote the matrices, bold lowercase characters to denote vectors, lowercase characters are used to denote the scalars, and calligraphic uppercase letters to denote the sets.

Table 1. Attribute Details

Region	Attribute	Dimension	Extractor
Face	Age	8, 192	Double-stream CaffeNet
	Gender	8, 192	Double-stream CaffeNet
	Appearance	8, 192	Double-stream CaffeNet
	Emotion	8, 192	Double-stream CaffeNet
	Pose	8, 192	Double-stream CaffeNet
	Location & Scale	14	Spatial Information
Body	Age	8, 192	Double-stream CaffeNet
	Gender	8, 192	Double-stream CaffeNet
	Clothing	8, 192	Double-stream CaffeNet
	Activity	2, 048	CNN-CRF
	Proximity	3, 136	Multi-task RNN
	Location & Scale	14	Spatial Information

3.1. Framework Overview

Let $C = \{c_j\}_{j=1}^{M}$ denote M relationship categories. Given an input image I within N persons $\mathcal{P} = \{p_i\}_{i=1}^{N}$, the target of social relationship recognition is to predict a matrix $\mathbf{R} = C^{N\times N}$, where each element r_{ij} represents the social relationship between the persons p_i and p_j.

DSFS leverages the pre-processed data for prediction. The persons appeared in the input photo are segmented and input to DSFS by pairs. Through DSFS, the possible relationship between one pair is captured. Figure 3 shows the architecture of DSFS, including feature extraction module, feature selection module, and classification module. The feature extractor takes in the patches of the faces and bodies, and generates the correlated feature vectors based on the attributes. In order to select the optimal feature subset, DSFS leverages the feature selector based on the deep learning algorithm to model social relationships with face and body attributes. Due to different granularity levels, two feature selection policies are designed, *i.e*. group feature selection and dimensional feature selection. Throughout the Softmax layer, the classifier makes the final predictions of the relationship between two persons.

3.2. Feature Extraction

In the feature extraction module, the representations of the relationship between two persons are generated via several feature extractors. Specifically, the extractors focus on both facial and body regions, and produce the attribute features between the specific pair of persons. Following the previous work (Sun et al., 2017), we adopt 12 attributes from the human face and body. Since the attributes are different, thus their feature vectors are captured via different extractors. They are extracted by a pre-trained Double-Stream CaffeNet (Sun et al., 2017; Jia et al., 2014), a pre-trained CNN-CRF (Liu et al., 2015), or multi-task RNN (Chu et al., 2015), etc. All the attributes with their names, dimensions, and extractors are listed in Table 1. Especially, face/body location & scale attribute is made of spatial information, including location coordinates, relative distance, and relative size ratio.

Formally, one pair of persons from image I is denoted as $(p_s, p_t) \in \mathcal{P} \times \mathcal{P}$, where $p_s \neq p_t$. Their attributes are represented as a set of feature vectors $\mathcal{A} = \{\mathbf{a}\}_{i=1}^{K}$, where $\mathbf{a}_i$ denotes the feature of i-th attribute and K denotes the number of attributes. According to Table 1, the elements in the set $\mathcal{A}$ have

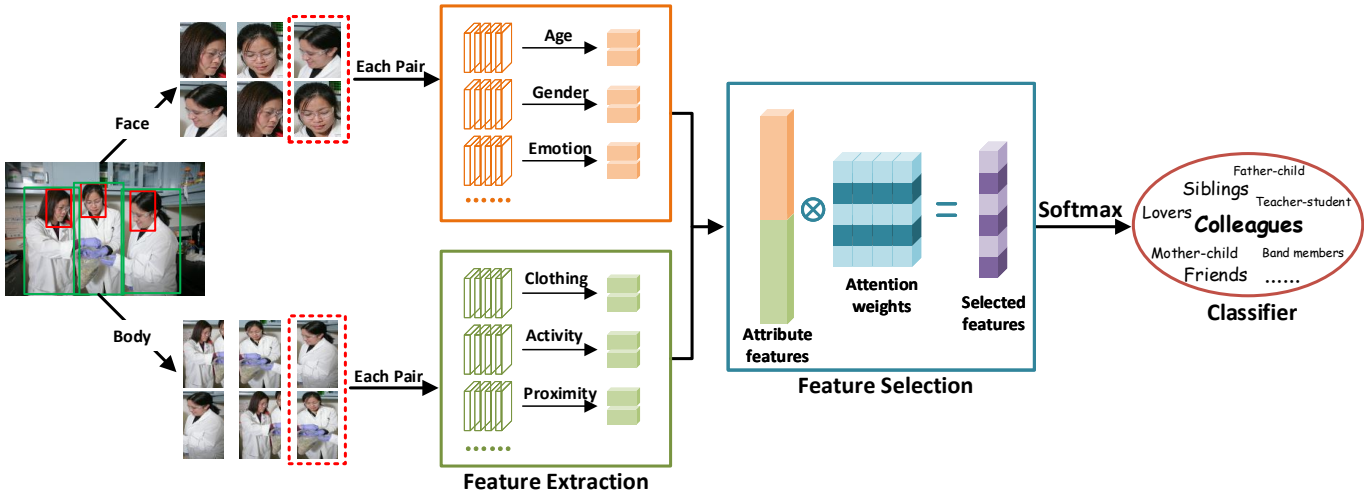


Fig. 3. Architecture of the proposed deep supervised feature selection (DSFS) framework. Our DSFS aims to recognize the exact social relationship between every person pair in the input image. The architecture of feature extraction module is shown in orange and green line box which refer to face attributes and body attributes, respectively. The architecture of feature selection module is shown in blue line box, where the input feature is made of face and body attribute features. Finally, recognizing the social relationship class between each person pair by the Softmax classifier.

different dimensions. We use $|\mathbf{a}_i|$ to denote the dimension of the feature $\mathbf{a}_i$. For example, suppose $\mathbf{a}_1$ indicate the age correlations extracted from face region, the dimension $|\mathbf{a}_1|$ is 8, 192.

3.3. Feature Selection

The pair of persons is represented by twelve attributes with 70,748 dimensions. These multi-source attributes in such high-dimensional space may contain noises and redundancies. To deal with this problem, extensive literature designs an effective model by learning a sparse matrix to improve performance in different tasks, such as automatic web-service selection (Luo et al., 2016, 2019), industrial applications (Luo et al., 2017), recommender systems (Luo et al., 2015) and so on. Thus, we propose a feature selection module that selects an optimal feature subset from the multi-source features by learning a sparse weighting matrix. In our DSFS framework, we utilize two feature selection policies: group feature selection and dimensional feature selection. Specifically, the group feature selection policy aims to select the optimal feature subset based on contributions of attributes. Dimensional feature selection learns the optimal feature subset at a fine-grained level to remove most of the redundancy.

For group feature selection, the d^s-dim selected feature is

$$\mathbf{f}^s = \sum_{i=1}^{K} \mathbf{W}_i \cdot \mathbf{a}_i, \tag{1}$$

where $\mathbf{a}_i$ denotes one attribute feature, $\mathbf{W}_i \in \mathbb{R}^{d^s \times |\mathbf{a}_i|}$ is a weight matrix $\mathbf{W}_i \in \mathbb{R}^{d^s \times |\mathbf{a}_i|}$ corresponding to this attribute, and K is the number of attributes. For dimensional feature selection, all attribute vectors are concatenated, *i.e.* $\hat{\mathbf{a}} = [\mathbf{a}_1^\top, \mathbf{a}_2^\top, \cdots, \mathbf{a}_K^\top]^\top$, the selected feature could be rewritten as,

$$\mathbf{f}^s = \hat{\mathbf{W}} \cdot \hat{\mathbf{a}}, \tag{2}$$

where $\hat{\mathbf{W}} = [\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_{|\hat{\mathbf{a}}|}]$, and $\hat{\mathbf{W}} \in \mathbb{R}^{d^s \times |\hat{\mathbf{a}}|}$.

In order to constrain the weight matrix to be sparse, we minimize $\ell_{2,1}$-norm of weights, which is detailed in section 3.5.

3.4. Classification

With the selected features, we employ the Softmax classifier (Iqbal et al., 2019) to compute the probability distribution of the input pair on social relationship categories. With the selected feature $\mathbf{f}^s$, the probability is computed by,

$$\mathbf{p} = \frac{\exp(\mathbf{f}^s)}{\sum_{j=1}^{M} \exp\left(f_j^s\right)}, \tag{3}$$

where p_j, the j-th element of vector $\mathbf{p}$, indicates the probability that the input pair belongs to the relationship c_j.

3.5. Optimization

To train our DSFS framework, in each iteration, we sample B pairs of persons denoted as $\mathcal{T}$, and minimize the objective loss function L formulated as,

$$L = -\frac{1}{B} \sum_{t \in \mathcal{T}} \log\left(p_{g_t}^t\right) + \alpha \cdot \mathrm{reg}(\mathbf{W}), \tag{4}$$

where g_t is the real relationship for the pair t, W is the weight matrix which could be either $\mathbf{W}_i$ or $\hat{\mathbf{W}}$. The first item in Eq. (4) is the cross-entropy loss to train the Softmax classifier, and the second item reg$(\cdot)$ indicates the regularization over the weight matrices. Then the stochastic gradient descent algorithm is adopted to learn the optimal weights.

In order to capture sparse weight matrix $\mathbf{W}$, the second term of L stands for $\ell_{2,1}$-norm, and α balances its influence. $\ell_{2,1}$-norm is a well-known norm function for feature selection (Li and Tang, 2015; Li et al., 2013) that it forces our weight matrix sparse. To better explore the effectiveness of feature selection, we propose two types of feature selection policies, namely group feature selection and dimensional feature selection.

Group feature selection selects the feature subset at the attributive level. According to Equation 1, we apply $\ell_{2,1}$-norm

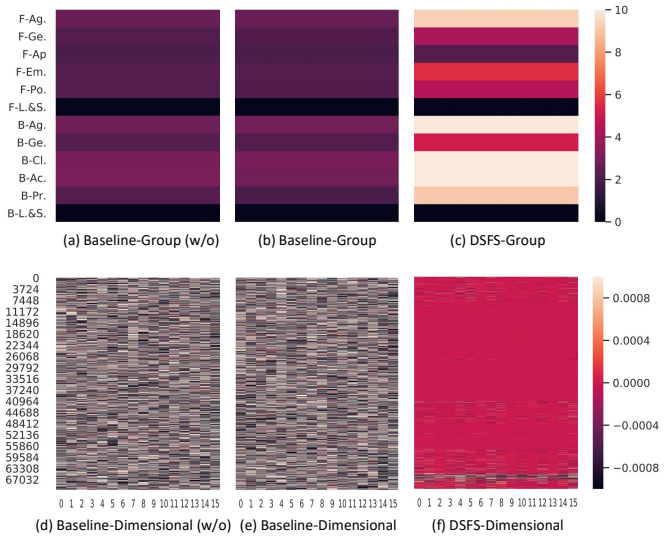


Fig. 4. Visualized weights of our proposed DSFS framework and baseline framework. (a), (b) and (c) describe weights in group level. The rows of (a), (b) and (c) are labeled by abbreviations of 12 face and body attributes, where "F" and "B" are short for face and body, the detail attributes are named by their first two letters. For example, face age is called for "F-Ag.". (d), (e) and (f) describe weights in dimensional level. The horizontal axis represents 16 social relationships, while the vertical axis represents dimensions of the input feature.

over the weight matrices about attributes. That is,

$$\mathrm{reg}^{Group}(\{\mathbf{W}_i\}_{i=1}^{K}) = \sum_{i=1}^{K} \|\mathbf{W}_i\|_F, \tag{5}$$

where $\|\mathbf{W}_i\|_F$ is the Frobenius norm of $\mathbf{W}_i$. The discussion and results are illustrated in section 4.3.

Dimensional feature selection abstracts the feature subset at the dimensional level, which makes most of the redundancy eliminated. According to Equation 2, we apply $\ell_{2,1}$-norm over the weight matrix $\hat{\mathbf{W}}$.

$$\mathrm{reg}^{Dimensional}(\hat{\mathbf{W}}) = \sum_{c=1}^{|\hat{\mathbf{a}}|} \sqrt{\sum_{j=1}^{d^s} \hat{W}_{rc}^2} = \sum_{c=1}^{|\hat{\mathbf{a}}|} \|\hat{\mathbf{W}}_c\|_2, \tag{6}$$

where $\hat{W}_{rc}$ denotes the (r, c)-th element of $\hat{\mathbf{W}}$, and $\hat{\mathbf{W}}_c$ is the c-th column vector of $\hat{\mathbf{W}}$. To facilitate feature selection at the dimensional level, the weight matrix is constrained to be sparse in columns.

4. Experiments

In this section, we conduct experiments to evaluate the effectiveness of our deep supervised feature selection (DSFS) framework on PIPA-relation dataset. Besides, we analyze the contribution of attributes and selection of parameter α.

4.1. Experimental Settings

4.1.1. Dataset

In this work, we focus on social relationship recognition based on multi-source attributes and conduct experiments on PIPA-relation dataset (Sun et al., 2017)[1] to demonstrate the superiority of our DSFS framework. PIPA-relation dataset divide social life into 5 domains build on the Bugental's domain-based theory (Bugental, 2000), and obtain 16 social relations based on these domains, including father-child, mother-child, grandpa-grandchild, grandma-grandchild, friends, siblings, classmates, lovers/spouses, presenter-audience, teacher-student, trainer-trainee, leader-subordinate, band members, dance team members, sport team members and colleagues. We target at recognizing 16 social relations in this work. Following the training configuration in (Sun et al., 2017), we utilize 13,729 person pairs for training, 709 person pairs for validating, and 5,106 person pairs for testing.

Table 2. Accuracy(%) of social relationship recognition on PIPA-relation

Methods	Accuracy(%)
Face+Softmax	51.74
Body+Softmax	58.25
Face+Body+Softmax	58.54
Face+Body+SVM (Sun et al., 2017)	57.20
DSFS-Group	60.14
DSFS-Dimensional	**61.51**

4.1.2. Implementation details

In the face and body representation modules of our framework, following the experimental setting of (Sun et al., 2017), we directly exploit neural network models trained with face and body attributes released by (Sun et al., 2017; Chu et al., 2015; Liu et al., 2015) and fix the weights of these neural network models. We train the feature selection module by stochastic gradient descent algorithm. During training, the learning rate and batch size are set as 0.001 and 1,000, respectively. We run 30 epochs for training our frameworks by costing about 3 to 6 minutes. Our implementation is based on Pytorch framework with Torch library. Our experiments are carried on Intel(R) Xeon(R) CPU with 64GB memory.

For the following attributes, *i.e.* face age, face gender, face appearance, face pose, face emotion, body age, body gender, body clothing, body proximity and body activity, we extract attribute features from fc7 layer of Siamese CaffeNets. The feature dimension of body proximity and body activity are 3,136D and 2,048D respectively. Face/body location & scale is a 7D vector. So the concatenate feature dimension $|\hat{\mathbf{a}}|$ is equal to $8,192 \times 8 + 2,048 + 3,136 + 14 \times 2 = 70,748$. The dimension of selected feature d_s is set as 16.

4.1.3. Baselines

To evaluate the efficacy of our proposed DSFS framework, we build a baseline framework for recognizing social relationships. In the baseline framework, we replace the $l_{2,1}$-norm by l_2-norm which is classical penalty term of weights, and the other configurations are the same as our DSFS framework. Taking the

[1] https://www.mpi-inf.mpg.de/social-relation

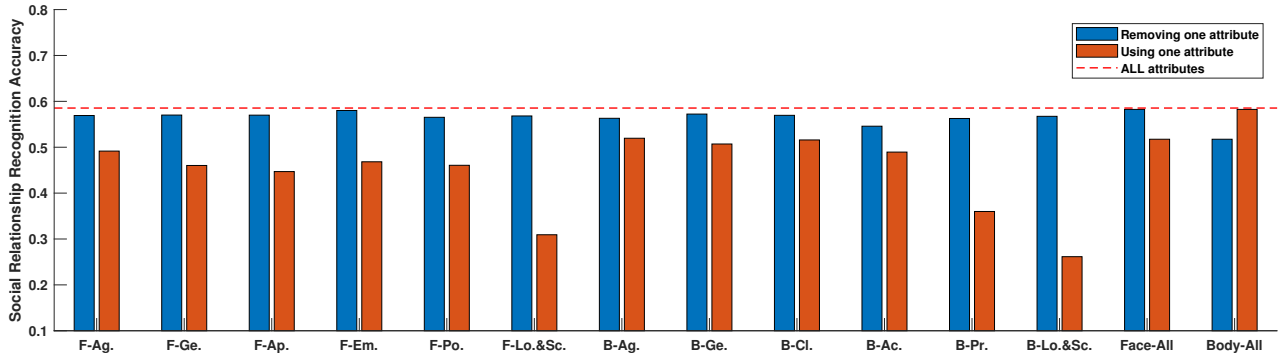


Fig. 5. Accuracy of social relationship recognition on PIPA-relation dataset with the baseline framework. As same as Fig. 4, the first 12 marks of X coordinate axis denote 12 face and body attributes. "Face-all" and "Body-all" represent accuracies are computed by baseline frameworks with all face attributes and all body attributes as input, respectively. The red dotted line describes the result computed by the baseline framework with all face and body attributes as input. Accuracy of using one attribute and that of removing one attribute from all attributes, are denoted as orange and blue, respectively.

baseline framework as the backbone, we design three baselines (shown in first to the third row of Table 2) for comparison.

In Table 2, the details of these three baselines are: **B1: Face+Softmax.** The model is trained by using face attributes only, and employs Softmax algorithm as classifier. **B2: Body+Softmax.** Like baseline B1, we utilize body attributes only. **B3: Face+Body+Softmax.** We train the model with all face and body attributes. **B4: Face+Body+SVM.** The method is proposed by (Sun et al., 2017), which exploits all face and body attributes for training and employs SVM as classifier. **DSFS-Group** and **DSFS-Dimensional** are our proposed DSFS framework by using group feature selection and dimensional feature selection policies, respectively.

4.2. Social Relationship Recognition with DSFS Framework

We firstly compare our proposed DSFS with several baselines on PIPA-relation dataset. The results are shown in Table 2. We evaluate the effect of each component of DSFS, *i.e.* feature extraction module, feature selection module, and classifier. We can see that, Body+Foftmax (**B2**) beats Face+Foftmax (**B1**), which indicates that body attributes provide more useful information for social relationship recognition than face attributes. Face+Body+Softmax (**B3**) performs better than Face+Softmax (**B1**) and Body+Softmax (**B2**). This is because more attributes can be extracted more recognition cues. For feature selection module, our DSFS-Group and DSFS-dimensional both give better accuracy than Face+Body+Softmax (**B3**). This is because $l_{2,1}$-norm makes more contributions to removing redundancy for selecting an optimal feature subset than l_2-norm. Moreover, DSFS-Dimensional framework achieves an accuracy of 61.51%, improving DSFS-Group by 1.5%. This is because that DSFS-Dimensional framework conducts a fine-level feature selection policy to abstract more discriminative information for the social relationship recognition task. For classifier, Face+Body+Softmax (**B3**) beats Face+Body+SVM (**B4**). It indicates that Softmax gives better accuracy than SVM for recognizing social relationships. It can be observed that our proposed DSFS-Group and DSFS-Dimensional frameworks outperform the best baseline Face+Body+Softmax (**B3**) by about 2%-3% in accuracy.

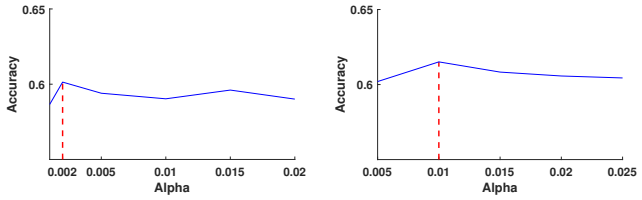
To observe the effectiveness of the feature selection module, we visualize the weights of the our proposed framework, as shown in Fig. 4. In Fig. 4 (a), (b) and (c), each row represents visualization of the Frobenius norm of weights, *i.e.* $\|\mathbf{W}_i\|_F$ (shown in Eqn. 1). In Fig. 4 (d), (e) and (f), the color of the ith row and jth corresponds to the (i, j)th-entry of optimal weights. Fig. 4 (a) and (d) represent the weights of a naive baseline framework which is similar to face+Body+Softmax (**B3**) but without any penalty term. Fig. 4 (b) and (e) describe the weights of Face+Body+Softmax (**B3**) with l_2-norm as the penalty term. Fig. 4 (c) and (f) describe the weights of DSFS-Group and DSFS-Dimensional frameworks with $l_{2,1}$-norm as the penalty term. In the proposed DSFS framework, we reduce redundancies and noises from multi-source features to obtain a feature subset by learning a sparse weighting matrix. In a word, the weight of framework is more sparser, which indicates the more redundancies are removed from the multi-source features. Obviously, the weights of our DSFS-Group and DSFS-Dimensional frameworks are sparser than those of other baselines. This demonstrates the effectiveness of reducing redundancies and noises of our DSFS-Group and DSFS-Dimensional frameworks on social relationship recognition at group and dimensional level.

4.3. Analysis on Contribution of Attributes

In our framework, twelve face and body attributes are involved to provide useful auxiliary information. To investigate the contribution of each attribute to social relationship recognition performance. First, we design two simple methods:

- *Using one attribute.* We train the baseline framework by using only one attribute.
- *Removing one attribute.* We train the baseline framework by removing one attribute from all face and body attributes.

The results of these manual two methods are shown in Fig. 5, we can observe that the methods with face age, body age, body gender, body activity achieve higher accuracy than others, and face/body location & scale works poorly. This is because age,



(a) α for DSFS-Group (b) α for DSFS-Dimensional

Fig. 6. Evaluations of DSFS-Group (a) and DSFD-Dimensional (b) frameworks over variations in parameter α of $l_{2,1}$-norm.

gender, body clothing, and body activity provide more discriminative information. For example, parents and their children have a significant age difference, colleagues always dress formally and sport team members join in the same activity. Face/body location & scale is composed of location coordinates, relative distance and relative size ratio, which may not present social relationships effectively. Overall, body attributes outperform face attributes. One possible reason is that face attributes describe more details, like pose, emotion, which may be difficultly learned by machine for social relationship recognition.

The abovementioned conclusions are consistent to the weight-visualization of DSFS-Group (presented in Fig. 4 (c)). This is because DSFS-Group framework regards each attribute as a group, and selects the optimal feature subset at the attribute-level. For our DSFS, the shade of color of weight-visualization reflects the contribution degree of attributes. For example, body attributes are lighter than face attributes in color, which indicates body attributes make more contributions than face attributes to social relationship recognition.

Besides, it can be observed that the impact of removing one attribute from all attributes on the recognition accuracy is relatively small. Further, if we remove all face attributes from all attributes (*i.e*. using body attributes only), the performance is a little lower than the baseline computed with all attributes. We conclude that different attributes may provide repeated information. This also demonstrates the concatenate feature combined by all face and body attributes has redundancy.

4.4. Analysis on Parameter α of $l_{2,1}$-norm

The parameter α in Eqn. 4 is used for weighting $l_{2,1}$-norm. Obviously, a small value may not reduce redundancy, and a high value may miss some face and body attribute cues. Hence, we conduct experiments with different values for finding the proper value. As shown in 6, when the parameter α set as 0.002 and 0.01 for DSFS-Group and DSFS-Dimensional respectively, these two methods perform the best accuracy.

5. Conclusion

In this work, we propose a Deep Supervised Feature Selection (DSFS) framework that selects an optimal feature subset from the multi-source attribute features to control redundancy and take advantage of effective information for social relationship recognition task. Specifically, we design two feature selection policies based on $l_{1,2}$-norm, *i.e*. group feature selection and dimensional feature selection. We evaluate our proposed framework on PIPA-relation dataset and achieve superior performance compared with the state-of-the-art method.

References

Bugental, D.B., 2000. Acquisition of the algorithms of social life: A domain-based approach. Psychological bulletin 126, 187.

Chu, X., Ouyang, W., Yang, W., Wang, X., 2015. Multi-task recurrent neural network for immediacy prediction, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3352–3360.

Goel, A., Ma, K.T., Tan, C., 2019. An end-to-end network for generating social relationship graphs, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11186–11195.

Iqbal, M., Sameem, M.S.I., Naqvi, N., Kanwal, S., Ye, Z., 2019. A deep learning approach for face recognition based on angularly discriminative features. Pattern Recognition Letters 128, 414–419.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding, in: Proceedings of the 22nd ACM international conference on Multimedia(ACM MM), ACM. pp. 675–678.

Lan, T., Sigal, L., Mori, G., 2012. Social roles in hierarchical models for human activity recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 1354–1361.

Li, J., Wong, Y., Zhao, Q., Kankanhalli, M.S., 2017. Dual-glance model for deciphering social relationships, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2650–2659.

Li, Z., Liu, J., Yang, Y., Zhou, X., Lu, H., 2013. Clustering-guided sparse structural learning for unsupervised feature selection. IEEE Transactions on Knowledge and Data Engineering 26, 2138–2150.

Li, Z., Tang, J., 2015. Unsupervised feature selection via nonnegative spectral analysis and redundancy control. IEEE Transactions on Image Processing 24, 5343–5355.

Liu, X., Liu, W., Zhang, M., Chen, J., Gao, L., Yan, C., Mei, T., 2019. Social relation recognition from videos via multi-scale spatial-temporal reasoning, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3566–3574.

Liu, Z., Luo, P., Wang, X., Tang, X., 2015. Deep learning face attributes in the wild, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3730–3738.

Lu, J., Zhou, X., Tan, Y., Shang, Y., Zhou, J., 2014. Neighborhood repulsed metric learning for kinship verification. IEEE Transactions on Pattern Analysis and Machine Intelligence 36, 331–345.

Luo, X., Sun, J., Wang, Z., Li, S., Shang, M., 2017. Symmetric and nonnegative latent factor models for undirected, high-dimensional, and sparse networks in industrial applications. IEEE Transactions on Industrial Informatics 13, 3098–3107.

Luo, X., Zhou, M., Li, S., Xia, Y., You, Z., Zhu, Q., Leung, H., 2015. An efficient second-order approach to factorize sparse matrices in recommender systems. IEEE Transactions on Industrial Informatics 11, 946–956.

Luo, X., Zhou, M., Wang, Z., Xia, Y., Zhu, Q., 2019. An effective scheme for qos estimation via alternating direction method-based matrix factorization. IEEE Transactions on Services Computing 12, 503–518.

Luo, X., Zhou, M., Xia, Y., Zhu, Q., Ammari, A.C., Alabdulwahab, A., 2016. Generating highly accurate predictions for missing qos-data via aggregating non-negative latent factor models. IEEE Transactions on Neural Networks and Learning Systems 27, 579–592.

Ramanathan, V., Yao, B., Fei-Fei, L., 2013. Social role discovery in human events, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2475–2482.

Robinson, J.P., Shao, M., Zhao, H., Wu, Y., Gillis, T., Fu, Y., 2017. Rfiw: Large-scale kinship recognition challenge, in: Proceedings of the 25th ACM international conference on Multimedia (ACM MM), pp. 1971–1973.

Shao, M., Li, L., Fu, Y., 2013. What do you do? occupation recognition in a photo via social context, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3631–3638.

Shu, T., Xie, D., Rothrock, B., Todorovic, S., Zhu, S.C., 2015. Joint inference of groups, events and human roles in aerial videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4576–4584.

Sun, Q., Schiele, B., Fritz, M., 2017. A domain based approach to social relation recognition, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3481–3490.

Wang, Z., Chen, T., Ren, J., Yu, W., Cheng, H., Lin, L., 2018. Deep reasoning with knowledge graph for social relationship understanding , 1021–1028.

Yan, H., Song, C., 2019. Semantic three-stream network for social relation recognition. Pattern Recognition Letters 128, 78–84.

Zhang, M., Liu, X., Liu, W., Zhou, A., Ma, H., Mei, T., 2019. Multi-granularity reasoning for social relation recognition from images, in: arXiv preprint arXiv:1901.03067.

Zhang, Z., Luo, P., Loy, C.C., Tang, X., 2015. Learning social relation traits from face images, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3631–3639.

Zhang, Z., Luo, P., Loy, C.C., Tang, X., 2018. From facial expression recognition to interpersonal relation prediction. International Journal of Computer Vision 126, 550–569.