Learning Social Relation Traits from Face Images

Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang

Department of Information Engineering, The Chinese University of Hong Kong

zz013@ie.cuhk.edu.hk, pluo@ie.cuhk.edu.hk, ccloy@ie.cuhk.edu.hk, xtang@ie.cuhk.edu.hk

Abstract

Social relation defines the association, e.g., *warm, friendliness, and dominance, between two or more people. Motivated by psychological studies, we investigate if such fine-grained and high-level relation traits can be characterised and quantified from face images in the wild. To address this challenging problem we propose a deep model that learns a rich face representation to capture gender, expression, head pose, and age-related attributes, and then performs pairwise-face reasoning for relation prediction. To learn from heterogeneous attribute sources, we formulate a new network architecture with a bridging layer to leverage the inherent correspondences among these datasets. It can also cope with missing target attribute labels. Extensive experiments show that our approach is effective for fine-grained social relation learning in images and videos.*

1. Introduction

Social relation manifests when we establish, reciprocate, or deepen relationships with one another in either physical or virtual world. Studies have shown that implicit social relations can be discovered from texts and microblogs [7]. Images and videos are becoming the mainstream medium to share information, which capture individuals with different social connections. Effectively exploiting such socially-rich sources can provide social facts other than the conventional medium like text (Fig. 1).

The aim of this study is to characterise and quantify social relation traits from computer vision point of view. Inspired by extensive psychological studies [9, 11, 13, 18], which show that face emotional expressions can serve as social predictive functions, we wish to automatically recognise fine-grained and high-level social relation traits (*e.g.*, friendliness, warm, and dominance) from face images. Such a capability promises a wide spectrum of applications. For instance, automatic social relation inference allows for relation mining from image collection in social network, personal album, and films.
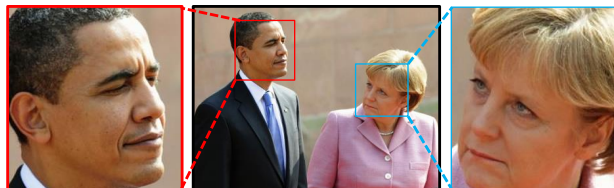

Figure 1. The image is given a caption '*German Chancellor Angela Merkel and U.S. President Barack Obama inspect a military honor guard in Baden-Baden on April 3*.' (source: www.rferl.org). Nevertheless, when we examine the face images jointly, we could observe far more rich social facts that are different from that expressed in the text.

Profiling unscripted social relation from face images is non-trivial. Among the most significant challenges are: (1) as suggested by psychological studies [9, 11, 13], relations of face images are related to high-level facial factors. Thus we need a rich face representation that captures various attributes such as expression and head pose; (2) no single dataset is presently available, which encompasses all the required facial attribute annotations to learn such a rich representation. In particular, some datasets only contain face expression labels, whilst other datasets may only contain the gender label. Moreover, these datasets are collected from different environments and exhibit different statistical distributions. How to effectively train a model on such heterogeneous data remains an open problem.

To this end, we carefully formulate a deep model to learn a face representation for social relation prediction, driven by rich facial attributes such as expression, head pose, gender, and age. We devise a new deep architecture that is capable of (1) dealing with missing attribute labels from different datasets, and (2) bridging the gap of heterogeneous datasets by weak constraints derived from the association of face part appearances. This allows the model to learn more effectively from heterogeneous datasets with different annotations and statistical distributions. Unlike existing face analyses that mostly consider single subject, our network is formulated with a Siamese-like architecture [2], it is thus capable of jointly considering pairwise faces for relation reasoning, where each face serves as the mutual

Table 1. Descriptions of social relation traits based on [17].

Relation Trait	Descriptions	Example Pair
Dominant	one leads, directs, or controls the other / dominates the conversation / gives advices to the other	teacher & student
Competitive	hard and unsmiling / contest for advancement in power, fame, or wealth	people in a debate
Trusting	sincerely look at each other / no frowning or showing doubtful expression / not-on-guard about harm from each other	partners
Warm	speak in a gentle way / look relaxed / readily to show tender feelings	mother & baby
Friendly	work or act together / express sunny face / act in a polite way / be helpful	host & guest
Attached	engaged in physical interaction / involved with each other / not being alone or separated	lovers
Demonstrative	talk freely being unreserved in speech / readily to express the thoughts instead of keep silent / act emotionally	friends in a party
Assured	express to each other a feeling of bright and positive self-concept, instead of depressed or helpless	teammates

context to the other.

The **contributions** of this study are three-fold: (1) to our knowledge, this is the first work that investigates face-driven social relation inference, of which the relation traits are defined based on psychological study [17]. We carefully investigate the detectability and quantification of such traits from a pair of face images. (2) we carefully construct a new social relation dataset labeled with pairwise relation traits supported by psychological studies [17, 18], which can facilitate future research on high-level face interpretation. (3) we formulate a new deep architecture for learning face representation driven by multiple tasks, bridging the gap from heterogeneous sources with potentially missing target attribute labels. It is also demonstrated that the model can be extended to utilize additional cues such as the faces' relative location, besides face images.

2. Related Work

Social signal processing. Understanding social relation is an important research topic in *social signal processing* [4, 29, 30, 36, 37], an important multidisciplinary problem that has attracted a surge of interest from computer vision community. Social signal processing mainly involves facial expression recognition [23] and affective behaviour analysis [28]. On the other hand, there exists a number of studies that aim to infer social relation from images and videos [5, 6, 8, 32, 39]. Many of these studies focus on the coarser level of social connection other than the one defined by Kiesler in the interpersonal circle [17]. For instance, Ding and Yilmaz [5] only discover social group without inferring relation between individuals. Fathi *et al.* [8] only detect three social interaction classes, *i.e.*, 'dialogue, monologue and discussion'. Wang *et al.* [38] define social relation by several social roles, such as 'father-child' and 'husband-wife'. Other related problems also include image communicative intents prediction [16] and social role inference [22], usually applied on news and talks shows [31], or meetings to infer dominance [15].

Our work differs significantly from the aforementioned studies. Firstly, most affective analysis approaches are based on single person therefore cannot be directly employed for interpersonal relation inference. In addition, these studies mostly focus on recognizing prototypical expressions (happy, angry, sad, disgust, surprise, fear). Social relation is far more complex involving many factors such as age and gender. Thus, we need to consider more attributes jointly in our problem. Secondly, in comparison to the existing social relation studies [5, 8], our work aims to recognize fine-grained and high-level social relation traits [17]. Thirdly, many of the social relation studies did not use face images directly for relation inference, but visual concepts [6] discovered by detectors or people spatial proximity in 2D or 3D space [3]. All these information sources are valuable for learning human interactions but social relation is fundamentally limited by the input sources.

Human interaction and group behavior analysis. Existing group behavior studies [14, 19] mainly recognize action-oriented behaviors such as hugging, handshaking or walking, but not social relations. Often, group spatial configuration and actions are exploited for the recognition. Our study differs in that we aim to recognize abstract relation traits from faces.

Deep learning. Deep learning has achieved remarkable success in many tasks of face analysis, *e.g.* face parsing [25], face landmark detection [42], face attribute prediction [24, 26], and face recognition [33, 43]. However, deep learning has not yet been adopted for face-driven social relation mining that requires joint reasoning from multiple subjects. In this work, we propose a deep model to cope with complex facial attributes from heterogeneous datasets, and joint learning from face pair.

3. Social Relation Prediction from Face Images

3.1. Definitions of Social Relation Traits

We define the social relation traits based on the interpersonal circle proposed by Kiesler [17], where human relations are divided into 16 segments as shown in Fig. 2. Each segment has its opposite side in the circle, such as "friendly and hostile". Therefore, the 16 segments can be considered as eight binary relations, whose descriptions and examples are given in Table 1. More detailed descriptions are provided in the *supplementary material*. We also provide positive and negative visual samples for each relation in Fig. 2, showing that they are visually perceptible. For instance, "friendly" and "competitive"
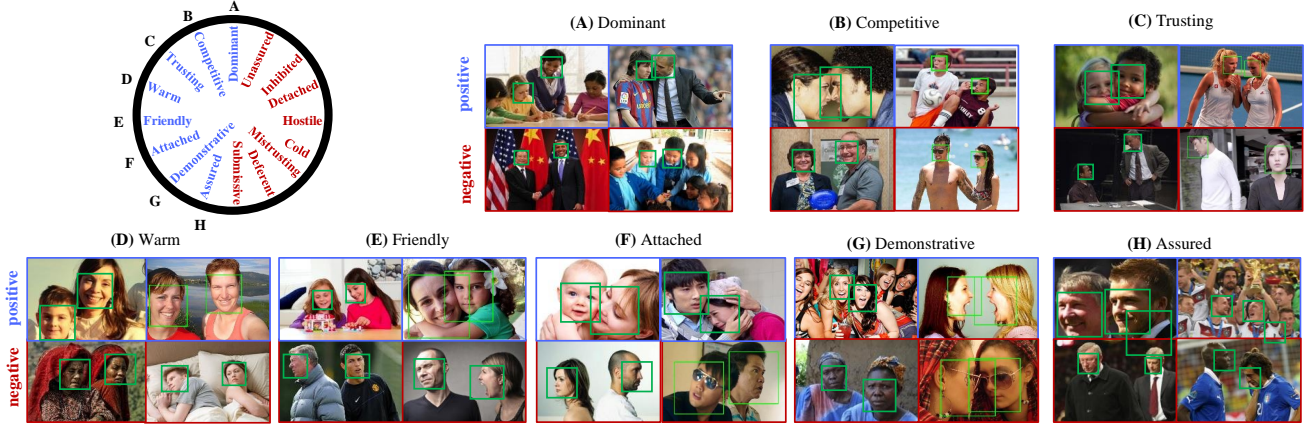
Figure 2. The 1982 Interpersonal Circle (upper left) is proposed by Donald J. Kiesle, and commonly used in psychological studies [17]. The 16 segments in the circle can be grouped into 8 relation traits. The traits are non-exclusive therefore can co-occur in an image. In this study, we investigate the detectability and quantification of these traits from computer vision point of view. (A)-(H) illustrate positive and negative examples of the eight relation traits. More detailed definition can be found in the supplementary material.

are easily separable because of the conflicting meanings. However, some relations are close such as "friendly" and "trusting", implying that a pair of faces can have more than one social relation.

3.2. Social Relation Dataset

To investigate the detectability of social relations from a pair of face images, we build a new dataset[1], containing $8,306$ images chosen from web and movies. Each image is labelled with faces' bounding boxes and their pairwise relations. This is the first face dataset measuring social relation traits and it is challenging because of large face variations including poses, occlusions, and illuminations.

We carefully built this dataset. Five performing arts students were asked to label each relation for each face image independently. Thus, each label has five annotations. A label is accepted if more than three annotations are consistent. The inconsistent samples were presented again to the five annotators to seek consensus[2]. To facilitate the annotation task, we also provide multiple cues to the annotators. First, to help them understand the social relations, we list ten related adjectives defined by [17] for the positive and negative samples on each relation trait, respectively. Multiple example images are also provided. Second, for the image frames selected from the movies, the annotators were asked to get familiar with the stories. The subtitles were presented during labelling.

---

[1] http://mmlab.ie.cuhk.edu.hk/projects/socialrelation/index.html

[2] The average Fleiss' kappa of the eight relation traits' annotation is 0.62, indicating substantial inter-rater agreement.

3.3. Baseline Method

To predict social relations from face images, we first introduce a strong baseline method by using a Siamese-like deep convolutional network (DCN), which learns an end-to-end mapping from raw pixels of a pair of face images to relation traits. DCN is effective for learning shared representations as demonstrated in [34]. As shown in Fig.3(a), given an image of social relation, we detect a pair of face images, denoted as $\mathbf{I}^r$ and $\mathbf{I}^l$, from which we extract high-level features $\mathbf{x}^r$ and $\mathbf{x}^l$ using two DCNs respectively, $\forall \mathbf{x}^r, \mathbf{x}^l \in \mathbb{R}^{2048 \times 1}$. These two DCNs have identical network structures, where $\mathbf{K}^r$ and $\mathbf{K}^l$ denote the network parameters, which are tied to increase generalization ability. A weight matrix, $\mathbf{W} \in \mathbb{R}^{4096 \times 256}$, projects the concatenated feature vectors to a space of shared representation $\mathbf{x}_t$, which is utilised to predict a set of relation traits, $\mathbf{g} = \{g_i\}_{i=1}^8$, $\forall g_i \in \{0, 1\}$. Each relation is modeled as a single binary classification task, parameterized by a weight vector, $\mathbf{w}_{g_i} \in \mathbb{R}^{256 \times 1}$.

To improve the baseline method, we incorporate some spatial cues to train the deep network as shown in Fig.3(a), which includes 1) two faces' positions $\{x^l, y^l, w^l, h^l, x^r, y^r, w^r, h^r\}$, representing the x-,y-coordinates of the upper-left corner, width, and height of the bounding boxes; w^l and w^r are normalized by the image width. Similar for h^l and h^r; 2) the relative faces' positions: $\frac{x^l - x^r}{w^l}$, $\frac{y^l - y^r}{h^l}$, and 3) the ratio between the faces' scales: $\frac{w^l}{w^r}$. The above spatial cues are concatenated as a vector, $\mathbf{x}_s$, and combined with the shared representation $\mathbf{x}_t$ for learning relation traits.

As the above description, each binary variable g_i can be predicted by linear regression,

$$g_i = \mathbf{w}_{g_i}^{\mathsf{T}}[\mathbf{x}_s; \mathbf{x}_t] + \epsilon, \tag{1}$$
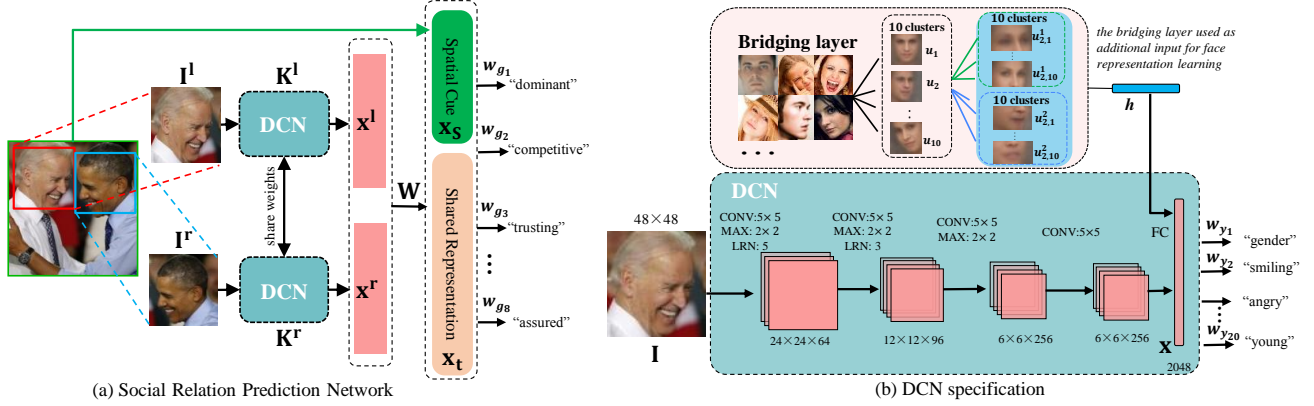
Figure 3. (a) Overview of the network for interpersonal relation learning. (b) The new deep architecture we propose to learn a rich face representation driven by sematic attributes. This network is used as the initialization for the DCN in (a) for relation learning. The operation of "CONV", "MAX", "LRN" and "FC" denote convolution, max-pooling, local response normalization and fully-connected, respectively. The numbers following the operations are the parameters for kernel size.

where ϵ is an additive error random variable, which is distributed following a standard logistic distribution, $\epsilon \sim Logistic(0,1)$. $[\cdot ; \cdot]$ indicates the column-wise concatenation of two vectors. Therefore, the probability of g_i given $\mathbf{x}_t$ and $\mathbf{x}_s$ can be written as a sigmoid function, $p(g_i = 1|\mathbf{x}_t, \mathbf{x}_s) = 1/(1 + \exp\{-\mathbf{w}_{g_i}^\mathsf{T}[\mathbf{x}_s; \mathbf{x}_t]\})$, indicating that $p(g_i|\mathbf{x}_t, \mathbf{x}_s)$ is a Bernoulli distribution, $p(g_i|\mathbf{x}_t, \mathbf{x}_s) = p(g_i = 1|\mathbf{x}_t, \mathbf{x}_s)^{g_i}\left(1 - p(g_i = 1|\mathbf{x}_t, \mathbf{x}_s)\right)^{1-g_i}$.

In addition, the probabilities of $\mathbf{w}_{g_i}$, $\mathbf{W}$, $\mathbf{K}^l$, and $\mathbf{K}^r$ can be modeled by the standard normal distributions. For example, suppose $\mathbf{K}$ contains K filters, then $p(\mathbf{K}) = \prod_{j=1}^{K} p(\mathbf{k}_j) = \prod_{j=1}^{K} \mathcal{N}(\mathbf{0}, \mathcal{I})$, where $\mathbf{0}$ and $\mathcal{I}$ are an all-zero vector and an identity matrix respectively, implying that the K filters are independent. Similarly, we have $p(\mathbf{w}_{g_i}) = \mathcal{N}(\mathbf{0}, \mathcal{I})$. Furthermore, $\mathbf{W}$ can be initialized by a standard matrix normal distribution [12], i.e. $p(\mathbf{W}) \propto \exp\{-\frac{1}{2}\mathrm{tr}(\mathbf{W}\mathbf{W}^\mathsf{T})\}$, where $\mathrm{tr}(\cdot)$ indicates the trace of a matrix.

Combining the above probabilistic definitions, the deep network is trained by maximising a posterior probability,

$$\arg\max_{\Omega} p(\{\mathbf{w}_{g_i}\}_{i=1}^{8}, \mathbf{W}, \mathbf{K}^l, \mathbf{K}^r|\mathbf{g}, \mathbf{x}_t, \mathbf{x}_s, \mathbf{I}^r, \mathbf{I}^l) \propto$$
$$\left(\prod_{i=1}^{8} p(g_i|\mathbf{x}_t, \mathbf{x}_s)p(\mathbf{w}_{g_i})\right)\left(\prod_{j=1}^{K} p(\mathbf{k}_j^l)p(\mathbf{k}_j^r)\right)p(\mathbf{W}),$$
$$\text{s.t.} \quad \mathbf{K}^r = \mathbf{K}^l \tag{2}$$

where $\Omega = \{\{\mathbf{w}_{g_i}\}_{i=1}^{8}, \mathbf{W}, \mathbf{K}^l, \mathbf{K}^r\}$ and the constraint means the filters are tied. Note that $\mathbf{x}_t$ and $\mathbf{x}_s$ represent the hidden features and the spatial cues extracted from the left and right face images, respectively. Thus, the variable g_i is independent with $\mathbf{I}^l$ and $\mathbf{I}^r$, given $\mathbf{x}_t$ and $\mathbf{x}_s$.

By taking the negative logarithm of Eqn.(2), it is equivalent to minimising the following loss function

$$\arg\min_{\Omega} \sum_{i=1}^{8}\left\{\mathbf{w}_{g_i}^\mathsf{T}\mathbf{w}_{g_i} - (1 - g_i)\ln\left(1 - p(g_i = 1|\mathbf{x}_t, \mathbf{x}_s)\right) - \right.$$
$$\left. g_i \ln p(g_i = 1|\mathbf{x}_t, \mathbf{x}_s)\right\} + \sum_{j=1}^{K}(\mathbf{k}_j^{r\mathsf{T}}\mathbf{k}_j^r + \mathbf{k}_j^{l\mathsf{T}}\mathbf{k}_j^l) + \mathrm{tr}(\mathbf{W}\mathbf{W}^\mathsf{T}),$$
$$\text{s.t.} \quad \mathbf{k}_j^r = \mathbf{k}_j^l, \; j = 1...K \tag{3}$$

where the second and the third terms correspond to the traditional cross-entropy loss, while the remaining terms indicate the weight decays [27] of the parameters. Eqn.(3) is defined over single training sample and is a highly nonlinear function because of the hidden features $\mathbf{x}_t$. It can be efficiently solved by stochastic gradient descent [21].

3.4. A Cross-Dataset Approach

As investigated by the psychological studies [9, 11, 13], the social relations of face images are strongly related to some hidden high-level factors, such as emotion. Learning these semantic concepts implicitly from raw image pixels imposes great challenge. To explicitly learn these factors, an ideal solution is to introduce two additional loss functions on top of $\mathbf{x}^l$ and $\mathbf{x}^r$ respectively, representing that not only the concatenation of $\mathbf{x}^l$ and $\mathbf{x}^r$ learns the relation traits, but each of them also learns the high-level factors of its corresponding face image. However, this solution is impractical, because labelling both social relations and emotions of face images is too expensive.

To overcome this limitation, we extend the baseline model by pre-training the DCN with face attributes, which are borrowed from existing face databases. These attributes capture the high-level factors, guiding the predictions of relation traits. The advantages are three folds: 1) face

attributes, such as age, gender, and expressions, are highly correlated with the high-level factors of social relations, as supported by the psychological studies [9, 11, 13, 18]; 2) leveraging the existing face databases not only improves generalized capacity but also make data preparation much easier; and 3) the face representation induced by semantic attributes can bridge the gap between the high-level relation traits and low-level image pixels.

In particular, we make use of data from three public datasets, including AFLW [20], CelebFaces [33], and Kaggle [10]. Different datasets have been labelled with different sets of face attributes. A summary is given in Table 2, where the attributes are partitioned into four groups.

It is clear that the training datasets are from multiple heterogenous sources and they have been labelled with different sets of attributes. For instance, AFLW only contains gender and poses, while Kaggle only has expressions. In addition, these datasets exhibit different statistical distributions, causing issues during pre-training. It can be shown that if we perform joint training directly, each attribute is trained by the labelled data alone, instead of benefitting from the existence of the unlabelled data. Consider a simple example of three datasets, denoted as A, B, and C, where A and B are labelled with attribute y^1 and y^2 respectively, while dataset C is labelled with y^1, y^2 and y^3. Moreover, $\mathbf{x}_A$ indicates a training sample from dataset A. Given three training samples $\mathbf{x}_A$, $\mathbf{x}_B$, and $\mathbf{x}_C$, attribute classification is to maximise the joint probability $p(y_A^1, y_A^2, y_A^3, y_B^1, y_B^2, y_B^3, y_C^1, y_C^2, y_C^3 | \mathbf{x}_A, \mathbf{x}_B, \mathbf{x}_C)$. Since the samples are independent and A and B only contain attributes y^1 and y^2 respectively, the joint probability can be factorized as $p(y_A^1, y_A^2, y_A^3 | \mathbf{x}_A) \cdot p(y_B^1, y_B^2, y_B^3 | \mathbf{x}_B) \cdot p(y_C^1, y_C^2, y_C^3 | \mathbf{x}_C) = p(y_A^1 | \mathbf{x}_A) \cdot p(y_B^2 | \mathbf{x}_B) \cdot p(y_C^1, y_C^2, y_C^3 | \mathbf{x}_C)$. For example, we have $\sum_{y_A^2, y_A^3} p(y_A^1, y_A^2, y_A^3 | \mathbf{x}_A) = p(y_A^1 | \mathbf{x}_A)$. As the attributes are also independent, the joint probability can be further written as $p(y_A^1, y_C^1 | \mathbf{x}_A, \mathbf{x}_C) p(y_B^2, y_C^2 | \mathbf{x}_B, \mathbf{x}_C) p(y_C^3 | \mathbf{x}_C)$, indicating that each attribute classifier is trained by the labelled data alone. For instance, the classifier of the first attribute is trained by data from A and C.

Bridging the gaps between multiple datasets. Since faces from different datasets share similar structure in local part, such as mouth and eyes, we propose a bridging layer based on the local correspondence to cope with the different dataset distributions. In particular, we establish a face descriptor h based on the mixture of aligned facial parts. As shown in Fig. 3(b), we build a three-level hierarchy to partition the facial parts' shape, where each child node groups the data of its parents into clusters, such as $u_{2,1}^1$ and $u_{2,10}^1$. In the top layer, the faces are divided into 10 clusters by K-means using the landmark locations from the SDM face alignment algorithm [41]. Each cluster captures

the topological changes due to viewpoints. Fig. 3(b) shows the mean face of each cluster. In the second layer, for each node, we perform K-means using the locations of landmarks in the upper and lower face region, and obtain 10 clusters respectively. These clusters captures the local shape of the facial parts. Then the mean HOG feature of the faces in each cluster is regarded as the corresponding template. Given a new sample, the descriptor h is obtained by concatenating its L2-distance to each template.

In this case, the descriptor h serves as a correspondence label for datasets. We use it as additional input in the fully connected layer for facial feature $\mathbf{x}$ (see Fig.3(b)). Thus the learned face representations for samples from different datasets are driven to be close if the correspondence labels are similar. It is worth noting that this bridging layer is different from the work of [1, 40], where the algorithms build some clusters from training data as an auxiliary task. Differently, the proposed method uses the aligned facial part association, which is well suited for our problem, instead of simply construct the cluster from the whole image. Moreover, since the construction of h is unsupervised, it contains noise and may harm the training if used as targets. Instead, we use the descriptor as additional input, which shows better performance than used as output (see Table. 5). The rest of the DCN structure is described in Fig.3(b), which includes four convolutional layers, three max-pooling layers, two local response normalization layers, and two fully-connected layers. The rectified linear unit [21] is adopted as the activation function.

Then the DCN objective is to predict a set of attributes $\mathbf{y} = \{y_l\}_{l=1}^{20}$, $\forall y_l \in \{0, 1\}$. Each relation is modeled as a single binary classification task, parameterized by a weight vector, $\mathbf{w}_{y_l} \in \mathbb{R}^{2048 \times 1}$. The probability of y_l can be computed by a sigmoid function. Similar to Eqn.(3), it can be formulated as minimising the cross-entropy loss.

Learning procedure. Similar to the relation prediction network, the training process can be done by back-propagation (BP) using stochastic gradient descent (SGD) [21]. The difference is that we have missing attribute labels in the training set. Specifically, we use the cross-entropy loss for attribute classification, with an estimated attribute $\widetilde{y}_l$, the back-propagation error e^l is

$$e^t = \begin{cases} 0 & \text{if } y_l \text{ is missing,} \\ y_l - \widetilde{y}_l & \text{otherwise.} \end{cases} \quad (4)$$

4. Experiments

Facial attribute datasets. To enable accurate social relation prediction, we employ three datasets to cover a wide-range of facial attributes: Annotated Facial Landmarks in the Wild (AFLW) [20] (24,386 faces), CelebFaces [33] (87,628 faces) and a facial expression dataset on Kaggle contest [10] (35,887 faces). Table 2

Table 2. Summary for the labelled attributes in the datasets: AFLW [20], CelebFaces [33] and Kaggle Expression [10].

Attributes	Gender	Pose					Expression									Age				
	gender	left profile	left	frontal	right	right profile	angry	disgust	fear	happy	sad	surprise	neutral	smiling	mouth opened	young	goatee	no beard	sideburns	5 o'clock shadow
AFLW	✓	✓	✓	✓	✓	✓														
CelebFaces	✓	✓	✓	✓	✓	✓								✓	✓	✓	✓	✓	✓	✓
Kaggle							✓	✓	✓	✓	✓	✓	✓							

Table 3. Statistics of the social relation dataset.

Relation trait	training		testing	
	#positive	#negative	#positive	#negative
dominant	418	7041	112	735
competitive	538	6921	123	724
trusting	6288	1171	609	238
warm	6224	1235	619	228
friendly	6790	669	734	113
attached	6407	1052	695	152
demonstrative	6555	904	699	148
assured	6595	864	685	162

Table 4. Balanced accuracies (%) on the movie testing subset.

Method	HOG+SVM	Baseline DCN with spatial cue	Full model with spatial cue
Accuracy	58.92%	63.76%	72.6%

and true negative. We first train the network as Sec. 3.3 (*i.e.*, Baseline DCN). After that, to examine the influences of different attribute groups, we pre-train four DCN variants using only one group of attribute (expression, age, gender, and pose). In addition, we compare the effectiveness between the full model with and without spatial cue.

Fig. 4 shows the accuracies of the different variants. All variants of our deep model outperform the baseline HOG+SVM. We observe that the cross dataset pre-training is beneficial, since pre-training with any of the attribute groups improves the overall performance. In particular, pre-training with expression attributes outperforms other groups of attributes (improving from 64.0% to 70.6%). This is not surprising since social relation is largely manifested from expression. The pose attributes come next in terms of influence to relation prediction. The result is also expected since when people are in a close or friendly relation, they tend to look at the same direction or face each other. Finally, the spatial cue is shown to be useful for relation prediction. However, we also observe that not every trait is improved by the spatial cue and some are degraded. This is because currently we simply use the face scale and location directly, of which the distribution is inconsistent in images from different sources. As for the relation traits, "dominant" is the most difficult trait to predict as it needs to be determined by more complicated factors, such as the social role and environmental context. The trait of "assured" is also difficult since it is visually subtle compared to other traits such as "competitive" and "friendly". In addition, we conduct analysis on the movie testing subset. Table 4 shows the average accuracy on the eight relation traits of the two baseline algorithms and the proposed method. The results correspond to that of the whole testing set. This supports the reliability of the proposed dataset.

Some qualitative results are presented in Fig. 5. Positive relation traits, such as "trusting", "warm", "friendly" are inferred between the US President *Barack Obama* and his family members. Interestingly, "dominant" trait is predicted between him and his daughter (Fig. 5(a)). The upper image

summarises the data. All the attributes are binary and labelled manually. To evaluate the performance of the cross dataset approach, we randomly select 2,000 testing faces from AFLW and CelebFaces, respectively. For the Kaggle dataset, we follow the protocol of the expression contest by using the 7,178 testing faces.

Social relation dataset. We build the social relation dataset as described in Sec. 3.2. Table 3 presents the statistics of this dataset. Specially, to reduce the potential effect from annotators' subjectivity, we select a subset (522 cases) from the testing images and build an additional testing set. The images in this subset are all from movies. As the annotators know the movies' story, they can give objective annotation assisted by the subtitle.

4.1. Social Relation Trait Prediction

Baseline algorithm. In addition to the strong baseline method in Sec. 3.3, we train an additional baseline classifier by extracting the HOG features from the given face images. The features from the two faces are then concatenated and we use a linear support vector machine (SVM) to train a binary classifier for each relation trait. For simplicity, we call this method "HOG+SVM", and the baseline method in Sec. 3.3 "Baseline DCN".

Performance evaluation. We divide the relation dataset into training and testing partitions of 7,459 and 847 images, respectively. The face pairs in these two partitions are mutually exclusive. To account for the imbalance positive and negative samples, a balanced accuracy is adopted:

$$accuracy = 0.5(n_p/N_p + n_n/N_n), \qquad (5)$$

where N_p and N_n are the numbers of positive and negative samples, whilst n_p and n_n are the numbers of true positive
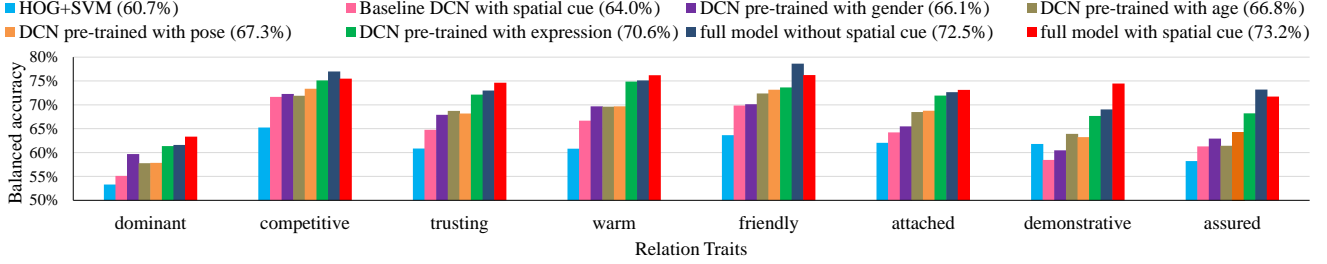
Figure 4. Relation traits prediction performance. The number in the legend indicates the average accuracy of the according method across all the relation traits.
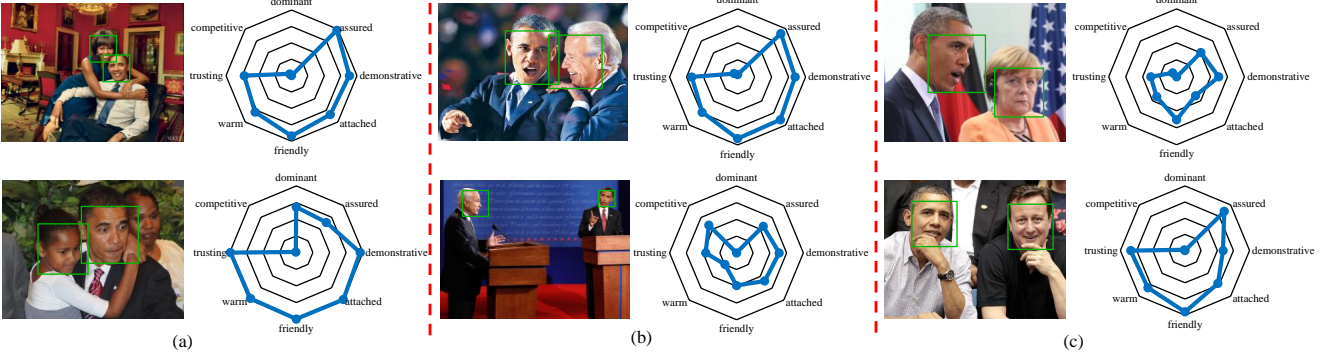

Figure 5. The relation traits predicted by our full model with spatial cue. The polar graph beside each image indicates the tendency for each trait to be positive.

in Fig. 5(b) was taken in his election celebration party with the US Vice President *Joe Biden*. We can see the relation is quite different from that of the lower image, in which Obama was in the presidential election debate. Fig. 5(c) includes the images for *Angela Merkel*, Chancellor of Germany and *David Cameron*, Prime Minister of UK. The upper image is usually used in the news articles on US spying scandal, showing low probability on the "trusting" trait. More positive and negative results on different relation traits are shown in Fig. 6 (a). In addition, we show some false positives in Fig. 6 (b), which are mainly caused by faces with large occlusions.

4.2. Further Analyses

Facial expression recognition. Given the essential role of expression attributes, we further evaluate our cross dataset approach on the challenging Kaggle facial expression dataset. Following the protocol in [10], we classify each face into one of the seven expressions, (*i.e.* angry, disgust, fear, happy, sad, surprise, and neutral). The Kaggle winning method [35] reports an accuracy of 71.2% by applying a CNN with SVM loss function. Our method achieves a better performance of 75.10%, through fusing data from multiple sources with the proposed bridging layer.

The effectiveness of bridging layer. We examine the effectiveness of the bridging layer from two perspectives. First, we show some clusters discovered by using the face

Figure 6. (a) Positive and negative prediction results on different relation traits. (b) False positives on "assured", "demonstrative" and "friendly" relation traits (from left to right).

descriptor (Sec. 3.4). It is observed that the proposed approach successfully divides samples from different datasets into coherent clusters of similar face patterns.

Table 5. Balanced accuracies (%) over different attributes with and without bridging layer (BL).

| Attributes	average	Gender	Pose					Expression									Age				
		gender	left profile	left	frontal	right	right profile	angry	disgust	fear	happy	sad	surprise	neutral	smiling	mouth opened	young	goatee	no beard	sideburns	5 o'clock shadow
HOG+SVM	72.6	81.2	86.8	71.7	88.3	**74.5**	90.1	61.2	63.7	59.2	77.8	60.2	74.8	66.3	83.2	78.9	67.1	60.8	67.8	70.3	67.2
Without BL	78.3	92.4	90.2	69.8	87.8	67.3	88.7	64.5	74.5	55.2	87.9	57.3	80.1	66.7	90.9	92.0	79.1	77.4	88.1	76.5	79.3
BL as output	81.3	92.9	91.7	70.1	90.0	70.6	90.2	69.1	77.0	**64.0**	**91.0**	**66.1**	**86.6**	73.9	91.5	**92.4**	**83.5**	74.5	**91.2**	79.5	**80.6**
BL as input	**82.4**	**93.8**	**92.2**	**73.4**	**95.4**	72.5	**90.4**	**69.8**	**79.4**	63.3	90.9	65.4	85.3	**74.9**	**92.8**	91.7	83.2	**82.1**	90.3	**81.7**	80.0
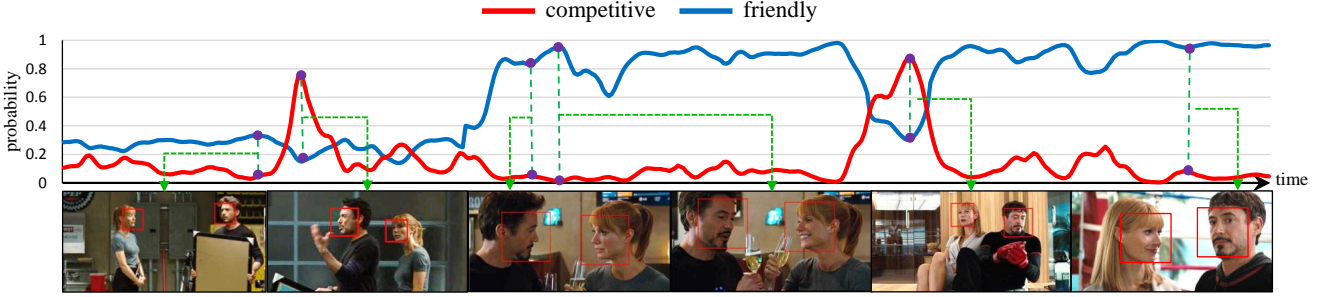
Figure 7. Prediction for relation traits of "friendly" and "competitive"for the movie *Iron Man*. The probability indicates the tendency for the trait to be positive. It shows that the algorithm can capture the friendly talking scene and the moment of confliction.

Figure 8. Test samples from different datasets are automatically grouped into coherent clusters by the face descriptor of bridging layer (Sec. 3.4). Each row corresponds to a cluster.

Second, we examine the balanced accuracy (Eqn. (5)) of attribute classification with and without the bridging layer (Table 5). It is observed that bridging layer benefits the recognition of most attributes, especially the expression attributes. The results suggest the bringing layer an effective way to combine heterogeneous datasets for visual learning by deep network. Moreover, treating bridging layer as input provides higher accuracy than as output.

4.3. Application: Character Relation Profiling

We show an example of application on using our method to profile the relations among the characters in a movie automatically. Here we choose the movie *Iron Man*. We focus on different interaction patterns, such as conversation and conflict, of the main roles "*Tony Stark*" and "*Pepper Potts*". Firstly, we apply a face detector to the movie and select the frames capturing the two roles. Then, we apply our algorithm on each frame to infer their relation traits. The predicted probabilities are averaged across 5 neighbouring frames to obtain a smooth profile. Fig. 7 shows a video segment with the traits of "friendly" and "competitive". Our method accurately captures the friendly talking scene and the moment when Tony and Pepper were in a conflict (where the "competitive" trait is assigned with a high probability while the "friendly" trait is low).

5. Conclusion

In this paper we investigate a new problem of predicting social relation traits from face images. This problem is challenging in that accurate prediction relies on recognition of complex facial attributes. We have shown that deep model with bridging layer is essential to exploit multiple datasets with potential missing attribute labels. Future work will integrate face cues with other information such as environment context and body gesture for relation prediction. We will also investigate other interesting applications such as relation mining from image collection in social network. Moreover, we can also explore modelling relations of more than two people, which can be implemented by voting or graphical model, where each node is a face and edge is relations between faces.

References

[1] A. Ahmed, K. Yu, W. Xu, Y. Gong, and E. Xing. Training hierarchical feed-forward visual recognition models using transfer learning from pseudo-tasks. In *ECCV*, pages 69–82. Springer, 2008.

[2] J. Bromley, I. Guyon, Y. Lecun, E. Säckinger, and R. Shah. Signature verification using a siamese time delay neural network. In *NIPS*, 1994.

[3] Y.-Y. Chen, W. H. Hsu, and H.-Y. M. Liao. Discovering informative social subgraphs and predicting pairwise relationships from group photos. In *ACM MM*, pages 669–678, 2012.

[4] M. Cristani, R. Raghavendra, A. Del Bue, and V. Murino. Human behavior analysis in video surveillance: A social signal processing perspective. *Neurocomputing*, 100:86–97, 2013.

[5] L. Ding and A. Yilmaz. Learning relations among movie characters: A social network perspective. In *ECCV*, 2010.

[6] L. Ding and A. Yilmaz. Inferring social relations from visual concepts. In *ICCV*, pages 699–706, 2011.

[7] N. Fairclough. *Analysing discourse: Textual analysis for social research*. Psychology Press, 2003.

[8] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *CVPR*, 2012.

[9] J. M. Girard. Perceptions of interpersonal behavior are influenced by gender, facial expression intensity, and head pose. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 394–398, 2014.

[10] I. Goodfellow, D. Erhan, P.-L. Carrier, A. Courville, Mirza, et al. Challenges in representation learning: A report on three machine learning contests, 2013.

[11] J. Gottman, R. Levenson, and E. Woodin. Facial expressions during marital conflict. *Journal of Family Communication*, 1(1):37–57, 2001.

[12] A. K. Gupta and D. K. Nagar. *Matrix variate distributions*. CRC Press, 1999.

[13] U. Hess, S. Blairy, and R. E. Kleck. The influence of facial emotion displays, gender, and ethnicity on judgments of dominance and affiliation. *Journal of Nonverbal Behavior*, 24(4):265–283, 2000.

[14] M. Hoai and A. Zisserman. Talking heads: detecting humans and recognizing their interactions. In *CVPR*, 2014.

[15] H. Hung, D. Jayagopi, C. Yeo, G. Friedland, S. Ba, J.-M. Odobez, K. Ramchandran, N. Mirghafori, and D. Gatica-Perez. Using audio and video features to classify the most dominant person in a group meeting. In *ACM MM*, 2007.

[16] J. Joo, W. Li, F. Steen, and S.-C. Zhu. Visual persuasion: Inferring communicative intents of images. In *CVPR*, pages 216–223, 2014.

[17] D. J. Kiesler. The 1982 interpersonal circle: A taxonomy for complementarity in human transactions. *Psychological Review*, 90(3):185, 1983.

[18] B. Knutson. Facial expressions of emotion influence interpersonal trait inferences. *Journal of Nonverbal Behavior*, 20(3):165–182, 1996.

[19] Y. Kong, Y. Jia, and Y. Fu. Learning human interaction by interactive phrases. In *ECCV*, pages 300–313. 2012.

[20] M. Kostinger, P. Wohlhart, P. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *ICCV Workshops*, pages 2144–2151, 2011.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.

[22] T. Lan, L. Sigal, and G. Mori. Social roles in hierarchical models for human activity recognition. In *CVPR*, 2012.

[23] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted deep belief network. In *CVPR*, pages 1805–1812, 2014.

[24] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[25] P. Luo, X. Wang, and X. Tang. Hierarchical face parsing via deep learning. In *CVPR*, 2012.

[26] P. Luo, X. Wang, and X. Tang. A deep sum-product architecture for robust facial attributes analysis. In *ICCV*, 2013.

[27] J. Moody, S. Hanson, A. Krogh, and J. A. Hertz. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4:950–957, 1995.

[28] M. A. Nicolaou, V. Pavlovic, and M. Pantic. Dynamic probabilistic CCA for analysis of affective behaviour. In *ECCV*, pages 98–111, 2012.

[29] M. Pantic, R. Cowie, F. D'Errico, D. Heylen, M. Mehu, C. Pelachaud, I. Poggi, M. Schroeder, and A. Vinciarelli. Social signal processing: the research agenda. In *Visual analysis of humans*, pages 511–538. Springer, 2011.

[30] A. Pentland. Social signal processing. *IEEE Signal Processing Magazine*, 24(4):108, 2007.

[31] B. Raducanu and D. Gatica-Perez. Inferring competitive role patterns in reality TV show through nonverbal analysis. *Multimedia Tools and Applications*, 56(1):207–226, 2012.

[32] V. Ramanathan, B. Yao, and L. Fei-Fei. Social role discovery in human events. In *CVPR*, pages 2475–2482, 2013.

[33] Y. Sun, X. Wang, and X. Tang. Hybrid deep learning for face verification. In *ICCV*, pages 1489–1496, 2013.

[34] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. DeepFace: Closing the gap to human-level performance in face verification. In *CVPR*, 2014.

[35] Y. Tang. Deep learning using linear support vector machines. In *ICML Workshop on Challenges in Representation Learning*, 2013.

[36] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.

[37] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schröder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1):69–87, 2012.

[38] G. Wang, A. Gallagher, J. Luo, and D. Forsyth. Seeing people in social context: Recognizing people and social relationships. In *ECCV*, pages 169–182. 2010.

[39] C.-Y. Weng, W.-T. Chu, and J.-L. Wu. Rolenet: Movie analysis from the perspective of social networks. *IEEE Transactions on Multimedia*, 11(2):256–271, 2009.

[40] J. Weston, F. Ratle, and R. Collobert. Deep learning via semi-supervised embedding. In *ICML*, 2008.

[41] X. Xiong and F. De La Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013.

[42] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Learning deep representation for face alignment with auxiliary attributes. In *TPAMI*, 2015.

[43] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity-preserving face space. In *ICCV*, 2013.