Learning Social Relation Traits from Face Images

Abstract

社会关系定义了两个或两个以上人之间的关系，例如 warm,friendliness,dominance。在心理学研究的驱动下，我们研究了这种细粒度和高层次关系特征是否可以从人脸图像中进行表征和量化。为了解决这个具有挑战性的问题，我们提出了一个深度模型，学习丰富的人脸表征来捕捉性别、表情、头部姿势和年龄相关的属性，然后进行关系预测的推理。为了从异构的属性资源中学习，我们制定了一个新的具有桥接层的网络体系结构，以利用这些数据集之间固有的对应关系。它也可以处理丢失的目标属性标签。大量的实验表明，我们的方法对图像和视频中的这种细粒度的社会关系学习是有效的。

1.Introduction

社会关系体现在我们建立、互动或加深彼此在物质世界或虚拟世界的关系。研究表明隐含的社会关系可以从文本和微博中发现[7]。图像和视频正在成为人们共享信息的主流媒介，例如可以捕获个人各种社会关系。充分利用这种丰富的社会资源可以提供像文本等传统媒介以外的社会关系。如图1所示：

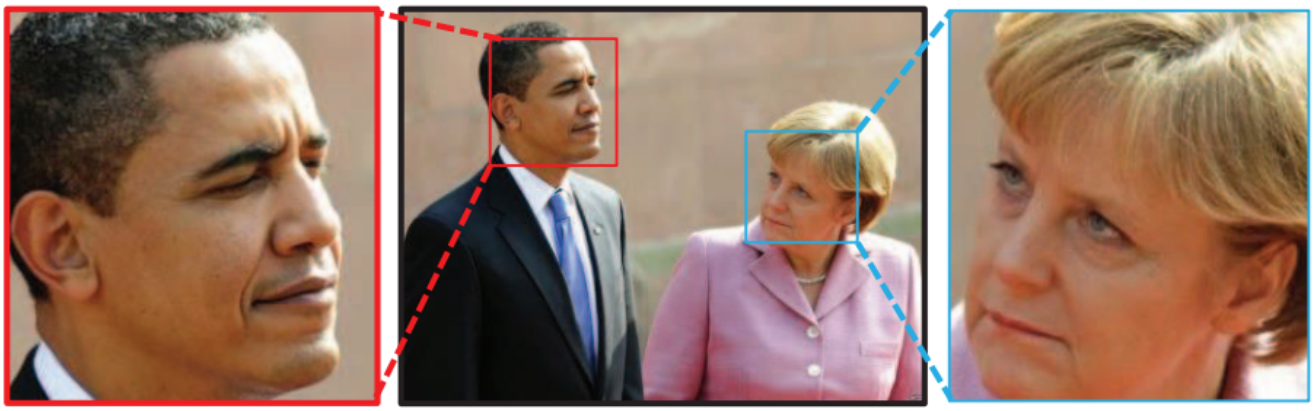


Figure 1. The image is given a caption '*German Chancellor Angela Merkel and U.S. President Barack Obama inspect a military honor guard in Baden-Baden on April 3.*' (source: www.rferl.org). Nevertheless, when we examine the face images jointly, we could observe far more rich social facts that are different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

本研究的目的是从计算机视觉的角度描述和量化社会关系。心理学研究表明面部表情可以用来进行社会关系预测[9,11,13,18]，受心理学研究的启发，我们希望能从面部图像中自动识别高层次的社会关系(如，friendliness,warm,dominance)。这有望带来广泛的应用。例如，可以从社交网络、个人相册和电影中收集图像进行关系挖掘。

从脸部图像分析社会关系是有意义的。其中有以下2个挑战：(1)正如心理学研究的提示，根据面部图像进行的关系抽取与与高层次面部特征有关。因此，我们需要一个能捕获各种属性的丰富的面部表达，例如表情和头部姿势；(2)没有一个包含所有需要的面部属性标签的数据集可用，而这些标签可以学得丰富的面部表达。特别是，一些数据集只包含面部表情标签，而其他数据集可能只包含性别标签。而且，这些数据集是从不同的环境中收集的，有不同的统计分布。如何基于这种异构数据有效地训练模型仍是一个挑战。

为此，我们提出了一个面向社会关系预测的深度模型，这个模型利用丰富的面部属性学习面部表达，例如表情、头部姿势、性别和年龄。我们提出的这个深度架构能够解决以下2个问题：(1)处理来自不同数据集的缺失属性标签；(2)bridging the gap of heterogeneous datasets by weak constraints derived from the association of face part appearances. 这使得模型能够从异构数据中进行更有效地学习。与现有的大多数只考虑单一对象的人脸分析不同，我们的网络使用 Siamese-like architecture [2],它同时考虑了要进行关系推理的两个人的面部。

这项研究有3个贡献：(1)这是第一个研究面部驱动的社会关系推理的工作，其中的关系是基于心理学的研究而定义的；(2)我们构建了一个新的社会关系数据集，标签是两个人的关系；(3)我们提出了一个深度架构，用于学习人脸表示。同时证明了该模型可以利用面部图像之外的其它线索进行扩展，例如人脸的相对位置。

2.Related Work

2.1 Social signal processing

理解社会关系是社会信号处理(social signal processing)[4,29, 30, 36, 37]领域的重要研究课题，这是一个吸引了计算机视觉领域兴趣的多学科问题。社会信号处理主要涉及面部表情识别[23]和情感行为分析[28]。另外，还有一些从图像和视频中推断社会关系的研究[5, 6, 8, 32, 39]。这些研究中的许多研究关注比较粗糙的社会关系，而不关注Kiesler 在[17]中定义的人际关系。例如，Ding and Yilmaz [5]只关注社会群体(social group)而不推断个体之间的关系。Fathi et al. [8]只预测3种社交类别：对话、独白和讨论。Wang et al. [38]通过几种社会角色来定义社会关系，例如父亲、夫妻。其他相关问题还包括图像交际意图预测[16]和社会角色推断[22],通常应用于新闻和访谈节目[31]，或者应用于会议上来推断支配者。

我们的工作与上述工作有显著不同。(1)大部分情感分析方法都是基于单个人的，因此不能直接应用到人际关系预测。另外，这些研究主要集中在识别原型表情(happy,angry,sad,disgust,surprise,fear)。社会关系要复杂得多，它涉及年龄、性别等诸多因素。因此，我们要解决的问题需要同时考虑更多的属性。(2)与现有的有关社会关系的研究相比[5,8]，我们的工作旨在识别细粒度和高层次的社会关系[17]。(3)许多社会关系的研究并未使用人脸图像进行关系推理，而是使用visual concepts [6]或者使用人们在二维或三维空间中的距离[3]。

2.2 Human interaction and group behavior analysis

现有的group behavior研究[14,19]主要关注面向动作的行为识别而不是社会关系，例如hugging、handshaking、walking。我们的研究不同于他们，我们的目标是利用人脸进行关系的识别。

2.3 Deep learning

深度学习在面部分析的许多任务中取得了显著的成功，例如人脸解析(face parsing[25])、人脸标志检测(face landmark detection[42])、人脸属性预测(face attribute prediction[24,26])和人脸识别(face recognition[33,43])。然而，面部驱动的社会关系挖掘还未使用深度学习，同时它也需要多个学科的联合推理。本论文提出了一个深度模型处理来自异构数据集的复杂面部属性，并利用两个人的面部属性进行联合学习。

3.Social Relation Prediction from Face Images

3.1 Definitions of Social Relation Traits

我们基于Kiesler [17]提出的人际关系圈定义社会关系，共分为16部分，如图2所示。

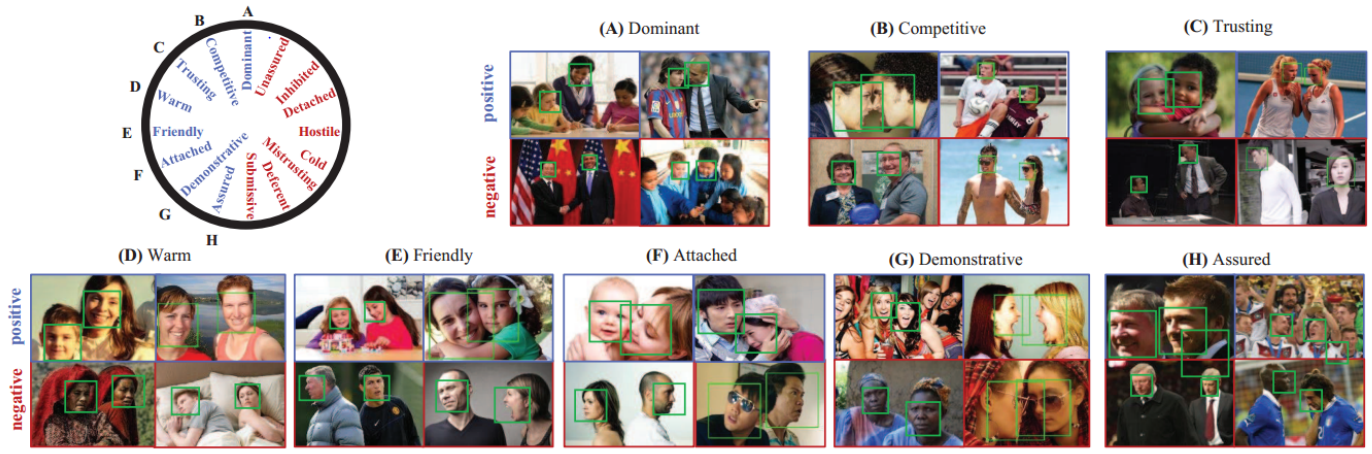


Figure 2. The 1982 Interpersonal Circle (upper left) is proposed by Donald J. Kiesle, and commonly used in psychological studies [17]. The 16 segments in the circle can be grouped into 8 relation traits. The traits are non-exclusive therefore can co-occur in an image. In this study, we investigate the 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.

如图2的环所示，每一部分都有它相反的部分，如"friendly与hostile"。因此，16部分可以认为是8个二元关系，关系的描述和示例如表1所示。补充材料中提供了更详细的说明。我们为图2中的每一类关系提供了正类和负类样本。一对人物之间可以有多种社会关系。

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

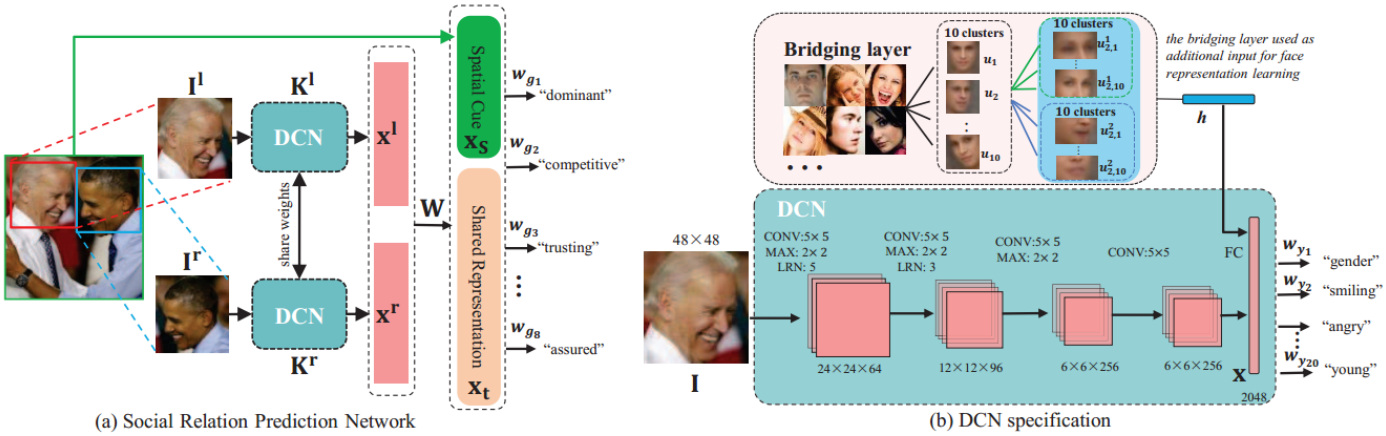
Relation Trait	Descriptions	Example Pair
Dominant	one leads, directs, or controls the other / dominates the conversation / gives advices to the other	teacher & student
Competitive	hard and unsmiling / contest for advancement in power, fame, or wealth	people in a debate
Trusting	sincerely look at each 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

3.2 Social Relation Dataset

数据集标签有头部groundtruth和关系标签。

3.3 Baseline Method

为了从脸部图像预测社会关系，我们首先引入baseline方法，baseline方法使用DCN网络，学习从一对人脸图像到社会关系的端到端映射。如[34]所述，DCN对学习共享表示是很有效的。如图3(a)所示，给定一张图片，我们首先检测两个人物的脸部，表示为I^l和I^r；然后分别用DCN网络从I^l和I^r图像中抽取高层次特征X^l和X^r，$\forall X^l, X^r \in R^{2048\times 1}$。这两个DCN网络有相同的网络结构，其中$\mathbf{K}^l$和$\mathbf{K}^r$表示网络参数；之后，将$X^l$和$X^r$连接成4096维的特征向量，通过权重矩阵$\mathbf{W} \in R^{4096\times 256}$映射到共享表示空间$X_t$；最后$X_t$被用来预测关系类别$g = \{g_i\}_{i=1}^{8}$，$\forall g_i \in \{0, 1\}$。每一类关系都被建模成一个二分类任务，每个二分类任务的权重向量参数为$\mathbf{w}_{g_i} \in R^{256\times 1}$。



(a) Social Relation Prediction Network (b) DCN specification

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.

为了优化baseline方法，我们考虑加入了一些空间特征，如图3(a)所示。共包括3种类型的空间特征：(1)两人脸部位置$\{x^l, y^l, w^l, h^l, x^r, y^r, w^r, h^r\}$，分别表示两人bounding box的左上角$x, y$坐标以及宽度和高度。$w^i$，$w^r$用图片宽度进行了规范化，$h^l$，$h^r$用图片高度进行了规范化。(2)两人面部相对位置：$\frac{x^l - x^r}{w^l}, \frac{y^l - y^r}{h^l}$。(3)面部宽度之间的比例$\frac{w^l}{w^r}$。上述空间特征连接形成特征向量$X_s$。然后将$X_s$与$X_t$连接，一起用于关系预测。

综上所述，每个二分类变量g_i可通过如下线性回归进行预测：

$$g_i = \mathbf{w}_{g_i}^T [X_s; X_t] + \epsilon \tag{1}$$

其中，ϵ是加性误差随机变量，遵循标准逻辑分布，$\epsilon \sim Logistic(0, 1)$。因此，给定$X_s, X_t$的条件下，$g_i$可以表示为sigmoid函数，$p(g_i = 1|X_t, X_s) = \frac{1}{1 + exp(-\mathbf{w}_{g_i}^T [X_s; X_t])}$，又$p(g_i|X_t, X_s)$服从伯努利分布，又可表示为：$p(g_i|X_t, X_s) = p(g_i = 1|X_t, X_s)^{g_i}(1 - p(g_i = 1|X_t, X_s))^{1-g_i}$。

另外，参数$\mathbf{w}_{g_i}, \mathbf{W}, \mathbf{K}^l, \mathbf{K}^r$都符合标准正态分布。假设$\mathbf{K}$包含$K$个filters，$p(\mathbf{K}) = \prod_{j=1}^{K} p(\mathbf{k}_j) = \prod_{j=1}^{K} N(\mathbf{0}, \mathbf{I})$，其中$\mathbf{0}, \mathbf{I}$表示全0矩阵和单位矩阵。同时，$p(\mathbf{w}_{g_i}) = N(\mathbf{0}, \mathbf{I})$。另外，$\mathbf{W}$被初始化为标准矩阵正态分布[12]，例如$p(\mathbf{W}) \propto exp(-\frac{1}{2} tr(\mathbf{W}\mathbf{W}^{\mathbf{T}}))$，其中，$tr(\cdot)$表示矩阵的迹。

基于上述概率定义，深度网络通过最大化后验概率进行训练：

$$arg\,max_{\Omega}\ p(\{\mathbf{w}_{g_i}\}_{i=1}^{8},\mathbf{W},\mathbf{K}^l,\mathbf{K}^r|g,X_t,X_s,I^r,I^l)\propto(\prod_{i=1}^{8}p(g_i|X_t,X_s)p(\mathbf{w}_{g_i}))(\prod_{j=1}^{K}p(\mathbf{k}_j^l)p(\mathbf{k}_j^r))p(\mathbf{W})$$

$$s.t.\ \mathbf{K}^r=\mathbf{K}^l \tag{2}$$

其中$\Omega=\{\{\mathbf{w}_{g_i}\}_{i=1}^{8},\mathbf{W},\mathbf{K}^l,\mathbf{K}^r\}$。

通过求等式(2)的负对数，可得到如下等价损失函数：

$$arg\,min_{\Omega}\ \Sigma_{i=1}^{8}\{\mathbf{w}_{g_i}^T\mathbf{w}_{g_i}-(1-g_i)\ln(1-p(g_i=1|X_t,X_s))-g_i\ln p(g_i=1|X_t,X_s)\}+\Sigma_{j=1}^{K}(\mathbf{k}_\mathbf{j}^{\mathbf{r}T}\mathbf{k}_\mathbf{j}^{\mathbf{r}}+\mathbf{k}_\mathbf{j}^{\mathbf{l}T}\mathbf{k}_\mathbf{j}^{\mathbf{l}})+tr(\mathbf{W}\mathbf{W}^T)$$

$$s.t.\ \mathbf{k}_\mathbf{j}^{\mathbf{r}}=\mathbf{k}_\mathbf{j}^{\mathbf{l}} \tag{3}$$

等式(3)是定义在单个训练样本上的非线性函数，可通过梯度下降法进行求解。

3.4 A Cross-Dataset Approach

根据心理学研究[9,11,13]，利用人脸图像进行社会关系抽取与与一些隐藏的高层次因素(如情绪)密切相关。从原始图像像素学习这些语义特征是很大的挑战。为了学习这些相关因素，一个理想的解决方案是分别为X^l和X^r引入额外的损失函数，使得X^l和X^r的连接不仅学习了关系特征，而且还能学到人脸图像对应的高层次特征。然而，这样的解决方案是不切实际的，因为标记人脸图像的社会关系和情绪代价太大。

为了克服这个局限性，我们通过用从现有的人脸数据库中借鉴的人脸属性对DCN进行预训练来扩展baseline模型。这些属性捕捉高层次的因素，指导关系的预测。其优点有3个：(1)心理学研究表明[9,11,13,18],年龄、性别和表情等面部属性与社会关系的高层次特征密切相关；(2)利用现有的人脸数据库不仅提高了数据容量，而且使数据准备更容易；(3)由语义属性引发的人脸表示可以弥合高层关系特征与低层次像素之间的差距。

我们用了三个公共数据集：AFLW[20],CelebFaces[33],Kaggle[10]。不同的数据集标注了不同的人脸属性，如表2所示：

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							√	√	√	√	√	√	√							

由上表可知，我们的训练数据集来自多个异构数据源，而且各自有不同的属性标记。例如，AFLW数据集只有Gender和Pose标记；Kaggle数据集只有Expression标记。另外，这些数据集存在不同的统计分布。可知每个属性只通过带此属性标签的数据集进行训练。假设我们有3个数据集A，B，C。其中，数据集A有属性y^1，数据集B有属性y^2，数据集C有属性y^1,y^2,y^3。$\mathbf{x}_A$表示来自数据集A的一个训练样本。给定三个训练样本$\mathbf{x}_A,\mathbf{x}_B,\mathbf{x}_C$，属性分类器是最大化联合概率$p(y_A^1,y_A^2,y_A^3,y_B^1,y_B^2,y_B^3,y_C^1,y_C^2,y_C^3|\mathbf{x}_A,\mathbf{x}_B,\mathbf{x}_C)$。由于样本是相互独立的，同时数据集A和B分别只包含y^1，y^2，因此联合概率可分解为$p(y_A^1,y_A^2,y_A^3|\mathbf{x}_A)\cdot p(y_B^1,y_B^2,y_B^3|\mathbf{x}_B)\cdot p(y_C^1,y_C^2,y_C^3|\mathbf{x}_C)=p(y_A^1|\mathbf{x}_A)\cdot p(y_B^2|\mathbf{x}_B)\cdot p(y_C^1,y_C^2,y_C^3|\mathbf{x}_C)$。由于属性也是相互独立的，因此联合概率可进一步写为：$p(y_A^1,y_C^1|\mathbf{x}_A,\mathbf{x}_C)\cdot p(y_B^2,y_C^2|\mathbf{x}_B,\mathbf{x}_C)\cdot p(y_C^3|\mathbf{x}_C)$，由概率公式可知，每一个属性分类器只由被标记了的数据进行训练。例如，第一个属性分类器仅由来自数据集A和数据集C的数据进行训练。

Bridging the gaps between multiple datasets. 由于来自不同数据集的人脸在局部部分（如嘴和眼睛）具有相似的结构，我们提出基于局部对应的桥接层(bridging layer)来处理具有不同分布的数据集。具体来说，我们基于对齐的面部部分的组合建立一个面部描述符h。如图3(b)所示，我们建立了一个三层的层次结果来划分面部各个部分的形状，其中没个子节点都将父节点的数据分组成簇，例如$\mu_{2,1}^1,\mu_{2,10}^1$。在最上层，我们使用SDM人脸对齐算法[4]获得的关键点坐标，通过K-means算法将人脸分成10簇。每个簇捕获因角度改变而引起的拓扑变化。图3(b)展示了每簇的平均面部。在第二层，针对每个节点，我们利用上下面部区域的关键点坐标执行K-means算法，并分别获得10个簇。这些簇能获得面部的局部形状。然后将每一簇中面部的平均HOG特征作为相应的template。给定一个新的样本时，通过连接该样本到其他template的L2距离得到面部描述符h。

我们将h作为全连接层的一个输入(见图3(b))。因此，如果来自不同数据集的样本标签是相似的，那么得到的面部表示也将会非常一致。值得一提的是，bridge layer与[1,40]的工作是不同的，我们的算法做了一些聚类作为辅助。而且，由于h是无监督的，所以它包含噪音，如果直接将它用作target可能会取得更坏的训练结果。相反，我们使用h作为附加输入，从而获得了更好的实验结果(如表5所示)。其余的DCN结构如图3(b)所示，包括4个卷积层、3个max-pooling层和2个全连接层。最后，rectified linear unit[21] 作为激活函数。

DCN网络的目标是预测属性值$y=\{y_l\}_{l=1}^{20},\forall y_l\in\{0,1\}$。每一类关系都是一个二分类任务，权重向量为$\mathbf{w}_{y_l}\in R^{2048\times 1}$。$y_l$概率可通过sigmoid函数来计算。如等式(3)所示，可通过最小化交叉损失进行学习。

Learning procedure. 与关系预测网络类似，可利用SGD算法通过BP进行训练。区别在于训练集中有些缺失的属性标签。我们利用交叉损失进行属性分类，得到预测属性值$\widetilde{y_l}$,反向传播误差e^l为：

$$e^l = \begin{cases} 0 & \text{if } y_l \text{ is missing,} \\ y_l - \widetilde{y_l} & \text{otherwise.} \end{cases} \tag{4}$$

4.Experiments

4.1 Social Relation Trait Prediction

Baseline algorithm. 除了3.3部分介绍的baseline方法“Baseline DCN”，我们还训练了另一个baseline分类器“HOG+SVM”:我们抽取给定面部图像的HOG特征，然后将两个人的HOG特征连接，最后用线性SVM为每一类关系训练一个二分类器。

Performance evaluation. 我们将关系数据集的训练集和测试集分别分为7459张、847张。训练集和测试集互不相交。为了解决正负样本的不平衡性，一个新的准确率计算方法被采用：

$$accuracy = 0.5(n_p/N_p + n_n/N_n), \tag{5}$$

其中，N_p, N_n分别代表正负样本的个数，n_p表示true positive的个数,n_n表示true negative的个数。我们首先训练Baseline DCN网络。然后，为了检验不同attribute groups对实验的影响，我们预训练了4个DCN网络，每个DCN网络使用一组attribute(expression,age,gender,pose)进行训练。另外，我们比较了有空间特征和无空间特征的模型。

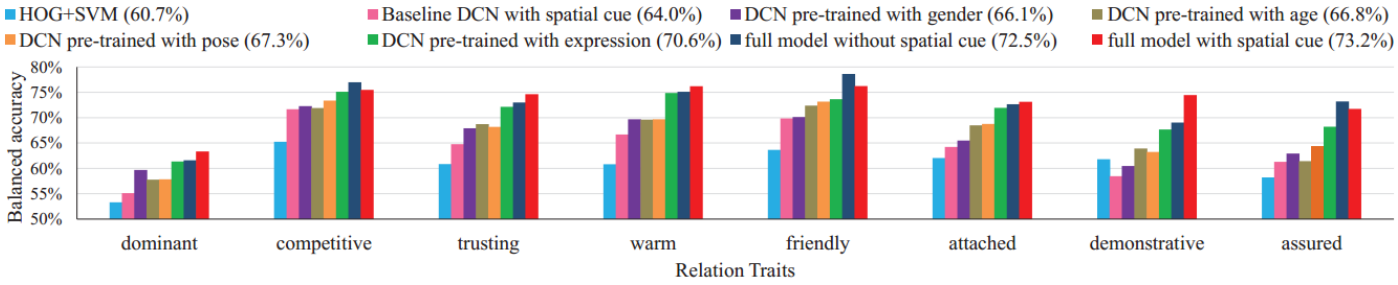


Figure 4. Relation traits prediction performance. The number in the legend indicates the average accuracy of the according method across all the relation traits.

图4展示了各种实验条件下的实验结果。我们的各个深度模型的实验结果都优于“HOG+SVM”。实验结果表明，跨数据集的预训练是有利的，因为用任意一个attribute group预训练的网络的实验结果都得到了提升。用expression attributes预训练的网络是四个预训练网络中效果最好的(从64.0%提升到了70.6%)。其次是用pose attributes预训练的网络得到了次好的结果。最后，spatial cue对社交关系预测也是有利的。表4是在电影上的实验结果。

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%

4.2. Further Analyses

此部分不再叙述。

5.Conclusion

本论文研究了从人脸图像预测社会关系的问题。未来的工作，可以融入其他特征，例如上下文环境、身体姿势等。而且，我们可以挖掘多个人的关系。