# 结合有监督联合一致性自编码器的跨音视频说话人标注

柳 欣[*①②] 李鹤洋[①] 钟必能[①②] 杜吉祥[①②]
[①](华侨大学计算机科学与技术学院 厦门 361021)
[②](计算机视觉与模式识别厦门市重点实验室 厦门 361021)

**摘 要**：跨模态说话人标注旨在利用说话人的不同生物特征进行相互匹配和互标注，可广泛应用于各种人机交互场合。针对人脸和语音两种不同模态生物特征之间存在明显的“语义鸿沟”问题，该文提出一种结合有监督联合一致性自编码器的跨音视频说话人标注方法。首先分别利用卷积神经网络和深度信念网络分别对人脸图像和语音数据进行判别性特征提取，接着在联合自编码器模型的基础上，提出一种新的有监督跨模态神经网络模型，同时嵌入 softmax 回归模型以保证模态间和模态内样本的相似性，进而扩展为 3 种有监督一致性自编码器神经网络模型来挖掘音视频异构特征之间的潜在关系，从而有效实现人脸和语音的跨模态相互标注。实验结果表明，该文提出的网络模型能够有效的对说话人进行跨模态标注，效果显著，取得了对姿态变化和样本多样性的鲁棒性。

**关键词**：跨模态说话人标注；有监督联合自编码器；softmax 回归模型；有监督神经网络模型
**中图分类号**： TP391.4 **文献标识码**： A **文章编号**：1009-5896(2018)07-1635-08
**DOI**: 10.11999/JEIT171011

# Efficient Audio-visual Cross-modal Speaker Tagging via Supervised Joint Correspondence Auto-encoder

LIU Xin[①②] LI Heyang[①] ZHONG Bineng[①②] DU Jixiang[①②]
[①](*Institute of Computer Science and Technology*, *Huaqiao University*, *Xiamen* 361021, *China*)
[②](*Xiamen Key Laboratory of Computer Vision and Pattern Recognition*, *Xiamen* 361021, *China*)

**Abstract**: Cross-modal speaker tagging aims to learn the latent relationship between different biometrics for mutual annotation, which can potentially be utilized in various human-computer interactions. In order to solve the “semantic gap” between the face and audio modalities, this paper presents an efficient supervised joint correspondence auto-encoder to link the face and audio counterpart, where by the speaker can be crosswise tagged. First, Convolutional Neural Network (CNN) and Deep Belief Network (DBN) are used to extract the discriminative features of the face and the audio samples respectively. Then, a supervised neural network model associated with softmax regression is embedded into a joint auto-encoder model, which can discriminatively preserving the inter-modal and intra-modal similarities. Accordingly, three different kinds of supervised joint correspondence auto-encoder models are presented to correlate the semantic relationships between the face and the audio counterparts, and the speaker can be crosswise annotated efficiently. The experimental results show that the proposed supervised joint auto-encoder is able to perform cross-modal speaker tagging with outstanding performance, and demonstrate the robustness to facial posture variations and sample diversities.

**Key words**: Cross-modal speaker tagging; Supervised joint correspondence auto-encoder; Softmax regression; Supervised neural network model

## 1 引言

随着信息技术的不断发展，分布式视频监控以及互联网多媒体应用系统中涌现出海量的音视频信息资源。其中，语音和视觉信息是人们相互交流的重要载体，也是人机交互过程中最为直接和灵活的方式。近年来，人们已逐渐认识到语音特征与视觉特征之间关联的重要性，并进行了多方面的研究，如视觉辅助的语音识别、音视频联合的说话人识别、虚拟说话人合成及动画等[1]。

基于人脸和语音的说话人身份鉴别技术以其特有的普遍性、稳定性和防伪性为人们提供了一种更为安全、方便和高效的个人身份鉴别手段，已逐渐成为替代钥匙、证件和智能卡等传统身份识别手段

收稿日期：2017-10-30；改回日期：2018-04-10；网络出版：2018-05-11
*通信作者：柳欣 xliu@hqu.edu.cn
基金项目：国家自然科学基金(61673185, 61572205, 61673186)，福建省自然科学基金(2017J01112)，华侨大学中青年创新人才培育项目(ZQN-309)
Foundation Items: The National Natural Science Foundation of China (61673185, 61572205, 61673186), The Natural Science Foundation of Fujian Province (2017J01112), The Promotion Program for Young and Middle-aged Teacher in Science and Technology Research of Huaqiao University (ZQN-309)

的更好选择[2]。在生物特征识别领域，属于同一个体的人脸和语音信息都能够有效鉴别个体身份，具有潜在的语义关联特性。调查发现，现有的音视频说话人标注方法一般通过不同的融合策略进行联合学习，并未考虑不同模态之间的关联性。然而在一些实际应用场合中，如视频会议和电视节目，通过说话人的语音信息快速定位和识别标注说话人人脸具有一定的实用性。在日常生活中，基于音视频跨模态间的相互检索方式，符合人类对身份识别的基本认知。例如，人们通过电话听到好友的语音信息后，脑海中会浮现出其人脸信息，这种基于说话人不同模态的生物特征信息进行相互匹配识别的方式称为跨模态说话人标注。近年来，虽然国内外出现了基于图像和文本的跨模态多媒体检索研究[3,4]，但基于人脸和语音的音视频跨模态说话人标注研究鲜有报道。因此，有效的音视频跨模态说话人标注能够促进身份鉴别技术创新实践的发展，具有重要的现实意义，有着广阔的应用前景。

说话人标注是通过说话人的个体生物特征或行为特征来确认一个人身份的过程[5]。在过去相当长的一段时间里，基于人脸和语音的单模态识别以及两者融合的识别吸引了国内外研究学者的普遍关注，并广泛地应用于实际生活当中。在人脸识别中，基于人脸图像的单模态生物特征识别技术取得了许多长足进步，并实现了从“浅层”特征到“深层”特征提取的突破，目前已接近甚至超过了肉眼识别的效果[6,7]。“浅层”特征提取方式一般使用子空间学习、SIFT、LBP 和 HOG 等手工提取特征的方法进行说话人人脸的特征提取[8]。近年来，深度学习通过学习一种深层非线性网络结构来实现复杂函数的逼近能力，能够分布式表征输入数据，体现了强大的特征提取能力。其中，以卷积神经网络作为“深层”方式进行人脸图像的特征提取已经取得了良好的实验效果；在声纹识别中，常用的特征表示形式有两种：一种是基于声道的线性预测倒谱参数(Linear Prediction Cepstrum Coefficient, LPCC)[9]，另一种是基于听觉特性的梅尔频率倒谱系数(Mel Frequency Cepstral Coefficients, MFCC)[10]。在文献[11]中，这些浅层的语音特征结合高斯混合模型(Gaussian Mixture Model, GMM)和隐马尔科夫模型(Hidden Markov Model, HMM)模型能够有效解决语音识别问题，取得了一定的成效。

近年来，深度学习同样应用于语音信号的特征表示学习当中。使用深度神经网络学习语音的特征空间的分布情况，从某种程度上模拟了生物的感知特性，具有较强的信息抽取能力。研究发现，单生物特征识别在实际应用中有着各自的弊端和局限性，易受环境变化的影响[12]。多生物特征融合技术利用多个可鉴别的身份信息，在一定程度上能弥补单生物特征识别的不足，从而达到降低误识率和实现高精度鉴别系统的要求。基于此，文献[13]通过训练受限玻尔兹曼机(Restricted Boltzmann Machine, RBM)将语音和视频嘴唇进行融合，实现了语音识别效果的提升。此外，文献[14]提出了说话人标注的概念，并利用卷积神经网络构建视觉和听觉的特征融合模型，实现说话人标注。相比于单模态的语音识别，多模态融合利用生物特征之间的互补关系，能够学习到更丰富充分的特征信息。然而，研究发现，现有的多模态标注方法并未考虑不同模态之间的相关性，难以直接应用到跨模态说话人标注中。

基于音视频的跨模态说话人标注是一项新颖的课题，存在的主要挑战包括以下两点：(1)传统的手工提取特征缺乏判别性自主学习，并难以直接描述高层语义信息进行有效互相标注匹配；(2)针对人脸和语音特征的异构性，目前尚缺乏有效的模型进行协同关联分析和匹配计算。针对上述挑战，本文首先利用卷积神经网络提取人脸图像特征和深度信念网络提取语音特征，并在联合一致性自编码器(Correspondence AutoEncoders, Corr-AE)[15]模型的基础上，提出有监督一致性神经网络(Supervised Correspondence Neural Network, Super-Corr-NN)模型，并加入 Softmax 回归以保证模态间和模态内样本的相似性，进而提出 3 种新的有监督联合一致性自编码器模型，以实现跨模态说话人的有效标注。

## 2 基于深度学习的音视频特征提取

人在说话时，常常包含人脸图像和语音两种异构模态信息。研究发现，人脸图像和语音信号的原始数据很难揭示这 2 个模态的身份语义信息。针对人脸图像和语音的原始数据，本节对人脸图像和语音分别进行特征提取，使得人脸图像和语音特征能够判别性地反映说话人身份的语义信息。鉴于卷积神经网络(Convolutional Neural Network, CNN)在图像处理中的广泛应用，如图 1 所示，我们利用 CNN 方法提取人脸图像的深层语义特征。对于语音而言，首先提取说话人的 MFCC 特征，再利用深度信念网络模型进一步提取说话人语音中反映身份特性的高层语义信息。

### 2.1 人脸图像特征提取

近年来，CNN 被广泛应用于众多研究领域，并在模式识别领域取得很好的成果。本节以 CNN 作为特征提取器，将人脸的原始图像作为网络的输入，相比于传统手工特征提取，避免了前期复杂的特征

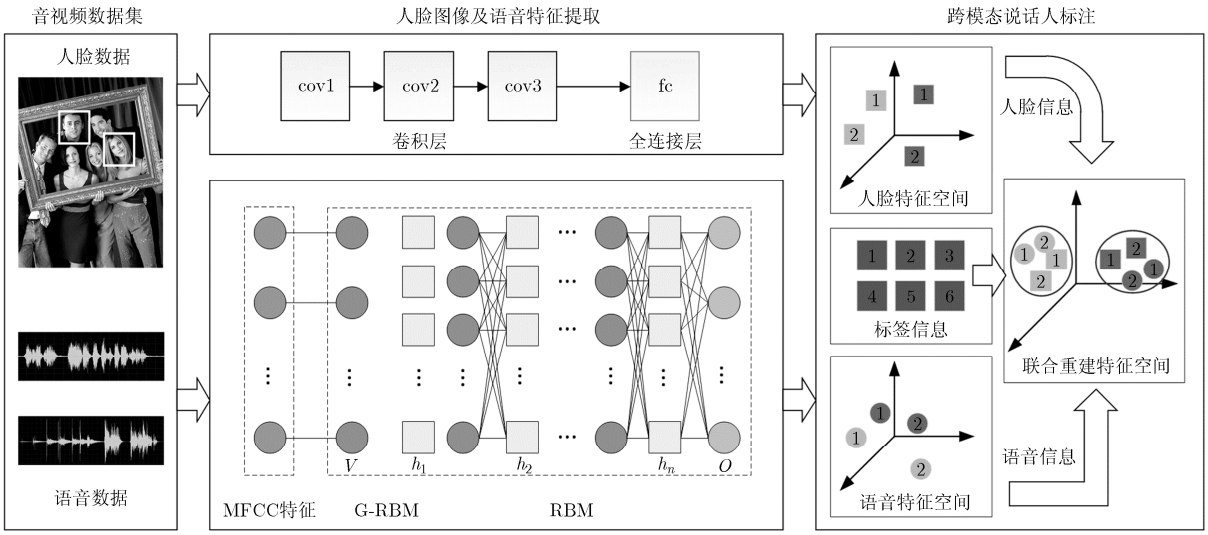



图 1 基于有监督联合一致性自编码器的跨模态说话人标注流程图

提取结构设计和提取过程中人脸信息的损失。具体地，将每一个输入的人脸图像归一化为固定尺寸大小(如 $50\times50$ )作为网络结构的输入，并使用 3 个卷积层和 1 个全连接层构成人脸特征提取器的网络结构；其中，每一个卷积层后面连接一个均值池化层，均值池化层操作能够学习前一层网络对应位置的全部感受域信息，减小信息丢失，并有效降低网络的规模，从而对人脸图像存在的旋转形变具有鲁棒性。网络模型输出仅使用一个全连接层，将全连接层的神经元作为 softmax 层的输入。通过 softmax 分类器产生一个包含 $m$ 个类别的概率分布，其 softmax 的输出函数定义为

$$p_i = \exp(o_i) \Big/ \sum_m \exp(o_m) \tag{1}$$

其中，$o_i$ 表示第 $i$ 张人脸图像在输出层的输出，$p_i$ 表示第 $i$ 张人脸图像的概率密度分布。为了能够降低光照敏感度并使卷积神经网络能够快速的收敛，一般将人脸图像的输入归一化到 $0\sim1$ 的范围内，每一个卷积层和抽样层的公式定义如式(2)所示。

$$F_l = \sigma\left(F * W^l + b^l\right),\quad l = 1,2,\cdots,n \tag{2}$$

$$dw_{(l+1)} = \mathrm{down}\left(F_l\right) \tag{3}$$

其中，$F$ 表示卷积层的输入，down 表示抽样层的采样函数。$W^l$ 和 $b^l$ 表示是第 $l$ 层的模型参数，$W$ 和 $b$ 分别表示卷积层的权重和偏置。$\sigma$ 表示神经元的激活函数 ReLU(Rectified Linear Units)，它的数学表达式为：$f(x)=\max(0,x)$ 。为了避免在训练中网络出现大量“死”的神经元，将卷积核和偏置项初始化为一个随机的很小的数值。本文使用随机梯度下降法来训练网络，并设置学习率为 0.1，冲量因子为 0.9，二者在训练过程中逐步调整。本文在此网络中使用均方损失函数，数学表达式为

$$E = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{m}\left(\boldsymbol{y}_{ij} - \boldsymbol{L}_{ij}\right)^2 \tag{4}$$

其中，$N$ 表示样本数，$m$ 表示样本类别数。$\boldsymbol{y}_{ij}$ 表示预测的概率值，$\boldsymbol{L}_{ij}$ 表示第 $i$ 个人脸样本对应于第 $j$ 类的标签值，如果第 $i$ 个人脸样本属于第 $j$ 类标签，那么设置 $\boldsymbol{L}_{ij}=1$，否则 $\boldsymbol{L}_{ij}=0$ 。

人脸图像特征提取实验框架如图 1 的第 2 部分，cov 表示卷积层，fc 表示全连接层，卷积层后紧接的抽样层未画出。各层的神经元个数信息如下：以 $2\times2$ 人脸输入图像为例，由于数据集的人脸图像的幅度较大，所以在 cov1 使用 48 个 $15\times15$ 的卷积核对原始的图像进行卷积操作；在 cov 2 层使用 256 个 $5\times5$ 卷积核对图像进行卷积操作；cov 3 层采用 1024 个 $7\times7$ 的卷积核对采样后的图像进行卷积操作；在本网络结构中，cov2 后连接局部感受域大小为 $2\times2$，步长为 2 的均值抽样层。

### 2.2 语音特征提取

深度信念网[16](Deep Belief Networks, DBN)是由单个可见层和多个随机的隐藏层组成的多层神经网络。每两个相邻的层之间可以看作是受限玻尔兹曼机(Restricted Boltzmann Machine, RBM)，并采用无监督分层的训练 RBM，学习网络参数。本文中，深度信念网络采用随机隐藏单元构成的 3 个隐藏层和随机可见单元构成的 1 个可见层构建语音特征提取器。在第 1 层采用高斯随机可见单元构成可见层 $\boldsymbol{v}$，使用随机二值隐藏单元构成 3 个隐藏层 $\boldsymbol{h}\in\{0,1\}^{\mathrm{H}}$ 。具体地，语音特征提取网络结构是由一个高斯玻尔兹曼机(Gaussian RBM, G-RBM)和 3 个普通的受限玻尔兹曼机组成。由于 RBM 是一个能量

模型，那么 G-RBM 可见层与隐藏层之间的关系可以用式(5)的能量函数表示为

$$E(\boldsymbol{v},\boldsymbol{h};\theta)=\sum_{i=1}^{D}\frac{(\boldsymbol{v}_i-\boldsymbol{d}_i)^2}{2\delta_i^2}-\sum_{i=1}^{D}\sum_{j=1}^{H}\frac{v_i}{\delta_i}\boldsymbol{W}_{ij}\boldsymbol{h}_{ij} \tag{5}$$

其中，$D$ 和 $H$ 分别表示可见层和隐藏层单元的个数，$\theta=\{\boldsymbol{c},\boldsymbol{d},\boldsymbol{W},\delta\}$ 表示模型参数，$\boldsymbol{c}$ 和 $\boldsymbol{d}$ 分别表示可见层和隐藏层的偏置项，$\delta$ 表示可见单元的标准差，$\boldsymbol{W}$ 表示可见层和隐藏层各节点的权重系数矩阵，RBM 可见层与隐藏层之间的关系可以用能量函数表示为

$$E(\boldsymbol{v},\boldsymbol{h};\omega)=-\sum_{i=1}^{D}\boldsymbol{d}_i\boldsymbol{v}_i-\sum_{j=1}^{H}\boldsymbol{c}_j\boldsymbol{h}_j-\sum_{i=1}^{D}\sum_{j=1}^{H}\boldsymbol{v}_i\boldsymbol{W}_{ij}\boldsymbol{h}_j \tag{6}$$

其中，$\omega=\{\boldsymbol{c},\boldsymbol{d},\boldsymbol{W}\}$ 表示模型参数，其他参数与式(5)描述中一致。

在本文中，深度信念网模型可见层的输入为 24 维 MFCC 特征。如图 1 中语音特征提取部分的框架所示，对于 DBN 的训练，由于是可以看成有多个 RBM 构成，所以采用分层的训练方法。在第 2 个隐藏层神经元的个数设置为 100，初始学习率为 0.004，共训练 500 次；其他隐藏层均设置 50 个神经元，初始学习率为 0.005，训练次数为 1000；训练过程中，在迭代开始时将冲量因子设为 0.50，在一定的迭代次数后，将其值更新为 0.99；训练时采用逐步降低学习率的方法，即经过一定迭代次数后将学习率乘以一个小的衰减因子。在迭代过程中逐步降低学习率可以加快算法的收敛速度，且所需超参数数量较少。为了能提高模型的训练效率，使用对比散度的快速学习算法[17]。模型对每一帧进行单独处理，产生的输入表示为每一个的概率分布。对于语音而言，每一帧的信息很小，不能够表示为说话人的身份信息，本实验使用投票系统，用 200 帧语音的标签作为投票系统的输入，产生每一个人的概率分布向量，将此向量作为语音的特征向量。

## 3 有监督联合一致性自编码器模型

Corr-AE 模型能够关联两种不同模态的特征语义空间。该模型使用两个自编码器来学习不同形式的特征，包含两个独立的自编码器，通过有相似性参数的编码层连接在一起。然而，这种无监督学习模型在一致性语义学习能力上有所欠缺。本文在此模型的基础上提出有监督联合一致性自编码器模型。

假设有两个不同模态的特征的数据集 $X_1=\left\{x_1^{(i)}\right\}_{i=1}^{n}$ 和 $X_2=\left\{x_2^{(i)}\right\}_{i=1}^{n}$，本文中 $X_1$ 为人脸数据集，$X_2$ 为语音数据集，$n$ 表示样本个数。两种模态有共同的标签集 $Y=\left\{y^{(i)}\right\}_{i=1}^{n}$，$y^{(i)}\in\{1,2,\cdots,c\}$，$c$ 表示类别个数。两个自编码器的隐藏层被定义为 $H_1=f\left(X_1;W_f\right)$ 和 $H_2=g\left(X_2;W_g\right)$，式中 $W_f$ 和 $W_g$ 是两个自编码器的参数，$f$ 和 $g$ 是激活函数，两种模态间的相似性定义为

$$\begin{aligned}L_H(X_1,X_2;W)&=\|H_1-H_2\|_F^2\\&=\left\|f\left(X_1;W_f\right)-g\left(X_2;W_g\right)\right\|_F^2\end{aligned} \tag{7}$$

式中，$W=\left[W_f,W_g\right]$ 是 Corr-AE 的全局参数，$\|\cdot\|_F$ 为 Frobenius 范数。两种模态的重建误差定义为

$$L_F(X_1,X_2;W)=\left\|X_1-\overline{X}_1^{(F)}\right\|_F^2 \tag{8}$$

$$L_A(X_1,X_2;W)=\left\|X_2-\overline{X}_2^{(A)}\right\|_F^2 \tag{9}$$

式中，$\overline{X}_1^{(F)}$ 表示以人脸特征为输入的自编码器输出，$\overline{X}_2^{(A)}$ 表示以语音特征为输入的语音自编码器输出，则总的输出函数可以表示为

$$L_{\text{total}}=(1-\alpha)\left(L_F+L_A\right)+\alpha L_H \tag{10}$$

式中，$\alpha\in[0,1]$ 是平衡参数。

### 3.1 Super-Corr-NN 模型

Corr-AE 模型虽然在跨媒体检索方面获得了很好的性能，其模型通过成对输入来学习模型，常常过多地关注局部信息，而损失全局信息，以至于原始数据信息并未得到充分利用。一般来说，模型中使用的信息越多，它能获得性能也越好。研究发现，有监督学习在充分利用标签信息的基础上，可以有效增强模型的判别性，从而常常获得比无监督学习方法更好的效果。研究发现，目前国内外常用的深度学习方法[18,19]常常忽略了在其模型中使用标签信息。本文以 Corr-AE 为基础，提出一种有监督联合一致性自编码器(Super-Corr-AE)模型。该方法可以通过联合自编码器保证模态间的相似性，并进一步通过有监督学习来保证模态内的相似性。如图 2 所示，本文新提出一种有监督一致性神经网络(Super-Corr-NN)模型。该模型可以利用标签信息使模型更具辨别力，每个自编码器被一个 3 层的多层感知器取代。

对于每个多层感知机，给定一个数据集

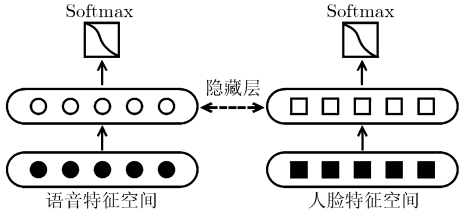



图 2 Super-Corr-NN 模型结构图

$X=\left\{x^{(i)}\right\}_{i=1}^{n}$ 和一个标签集 $L=\left\{l^{(i)}\right\}_{i=1}^{n}$，$l^{(i)}\in\{1,2,\cdots,c\}$，其损失函数如式(11)：

$$J(X,L;\Theta)=-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}I\left\{l^{(i)}=j\right\}\log\frac{e^{\theta_j^{\mathrm{T}}x^{(i)}}}{\sum_{k=1}^{c}e^{\theta_k^{\mathrm{T}}x^{(i)}}},\quad j=1,2,\cdots,c \tag{11}$$

式中，$\Theta$ 是神经网络参数，$\theta_i$ 是 $\Theta$ 集合中的第 $i$ 列，如果 $z$ 为真，$I\{z\}$ 等于 1，否则等于 0。不同于自编码器，多层感知机输出的是样本属于不同类别的可能性，可以通过标签信息保证类内相似性，从而使模型更具有判别性。$\Theta_1$ 和 $\Theta_2$ 作为两个模态的多层感知机的参数，Super-Corr-NN 的总损失函数可以表示为

$$L_S=J(X_1,Y;\Theta_1)+J(X_2,Y;\Theta_2) \tag{12}$$

本文使用多个感知机取分类样本，采用隐层间的相似性参数来进行表征学习和相关学习，其损失函数定义如式(11)所示。令 $\alpha\in[0,1]$ 为平衡参数，Super-Corr-NN 模型总损失函数定义如式(13)：

$$L_{\text{total}}=(1-\alpha)L_S+\alpha L_H \tag{13}$$

### 3.2 Super-Corr-AE 模型

Corr-AE 模型是一种无监督学习的方法，并未考虑标签信息。研究发现，使用标签信息的有监督学习方法可以有效增强模型的判别性，从而提升学习性能[20]。如图 3 所示，本文将一致性 Super-Corr-NN 模型和跨模态 Corr-AE 模型结合，得到新的有监督一致性联合自编码器模型(Super-Corr-AE)。具体地，在 Super-Corr-AE 中，Super-Corr-NN 和 Corr-AE 共享他们的输入层和隐含层，其输出层独立分离。最后，总的损失函数由 3 部分构成：重建损失部分 $L_F+L_A$，相关性损失部分 $L_H$ 和 softmax 损失部分 $L_S$。因此，Super-Corr-AE 模型总的损失函数如式(14)：

$$L_{\text{total}}=(1-\alpha)(L_F+L_A)+\alpha L_H+\beta L_S \tag{14}$$

式中，$\alpha\in[0,1]$ 和 $\beta\geq 0$ 作为平衡参数。

### 3.3 Super-Cross-AE 模型

本文进一步扩展了 Cross-AE 模型，通过结合有 Super-Corr-NN 模型，使其扩展成一个有监督跨模态自编器模型(Super-Cross-AE)，其结构如图 4 所示。Corr-AE 的输出值是通过重建相同模态的数据得到的，而 Super-Cross-AE 重建输出时用到了不同模态的输入值，其重构损失函数如式(15)和式(16)所示。

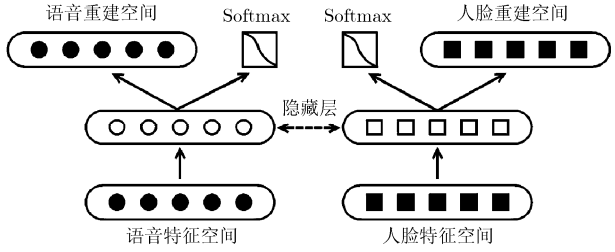



图 3　Super-Corr-AE 模型结构图

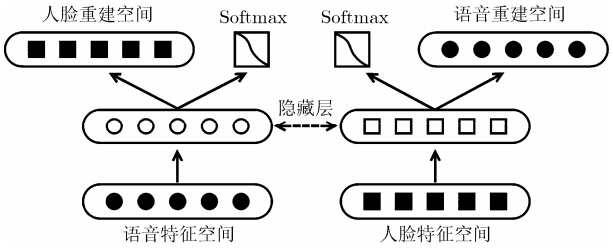



图 4　Super-Cross-AE 模型结构图

$$L_{F\text{cross}}(X_1,X_2;W)=\left\|X_2-\overline{X}_2^{(F)}\right\|_F^2 \tag{15}$$

$$L_{A\text{cross}}(X_1,X_2;W)=\left\|X_1-\overline{X}_1^{(A)}\right\|_F^2 \tag{16}$$

式中，$\overline{X}_2^{(F)}$ 表示以语音特征为输入的自编码器输出；$\overline{X}_1^{(A)}$ 表示以人脸特征为输入的自编码器输出。总损失函数定义如式(17)：

$$L_{\text{total}}=(1-\alpha)(L_{F\text{cross}}+L_{A\text{cross}})+\alpha L_H+\beta L_S \tag{17}$$

式中，$\alpha\in[0,1]$ 和 $\beta\geq 0$ 是平衡参数。

### 3.4 Super-Full-AE 模型

本文进一步扩展文献[15]中 Corr-Full-AE 模型，如图 5 所示，通过结合 Super-Corr-NN 模型，将其变为一个有监督的跨模态一致性语义全局模型(Super-Full-AE)。该模型可以看做是 Super-Corr-NN 和 Corr-Full-AE 的一个组合，使用两种模态的数据来重建每个自编码器的输出，其结构损失函数为

$$\begin{aligned}&L_{F\text{full}}(X_1,X_2;W)\\&\quad=L_F(X_1,X_2;W)+L_{F\text{cross}}(X_1,X_2;W)\\&\quad=\left\|X_1-\overline{X}_1^{(F)}\right\|_F^2+\left\|X_2-\overline{X}_2^{(F)}\right\|_F^2\end{aligned} \tag{18}$$

$$\begin{aligned}&L_{A\text{full}}(X_1,X_2;W)\\&\quad=L_A(X_1,X_2;W)+L_{A\text{cross}}(X_1,X_2;W)\\&\quad=\left\|X_2-\overline{X}_2^{(A)}\right\|_F^2+\left\|X_1-\overline{X}_1^{(A)}\right\|_F^2\end{aligned} \tag{19}$$

其中，$\overline{X}_1^{(F)}$，$\overline{X}_2^{(F)}$，$\overline{X}_1^{(A)}$，$\overline{X}_2^{(A)}$ 与前文公式中定义相同，总损失函数可以定义为

$$L_{\text{total}}=(1-\alpha)(L_{F\text{full}}+L_{A\text{full}})+\alpha L_H+\beta L_S \tag{20}$$

式中，$\alpha\in[0,1]$ 和 $\beta\geq 0$ 是平衡参数。

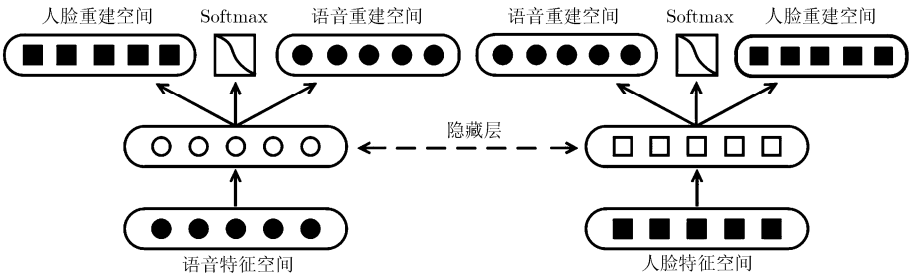



图 5 Super-Full-AE 模型结构图

## 4 实验结果

### 4.1 数据集介绍

为验证本文提出的方法，本文选用公开的老友记(Friends)和生活大爆炸(BigBang)音视频数据集进行测试[13]，两种数据集中都存在着人脸表情多样化以及不均匀光照影响。特别地，Friends 数据集选取 S01E03(第 1 季第 3 集)，S04E04, S05E05, S07E07 和 S10E15 的人脸图像和语音组成。本文选取 1200 组人脸和语音数据对作为训练集，360 数据对作为验证集，240 数据对作为测试集。类似地，BigBang 数据集选取 5000 组人脸和语音数据对作为训练集，1000 数据对作为验证集，500 组数据对作为测试集。

### 4.2 衡量标准

本文采用两种跨模态标注(检索)形式，使用人脸标注语音，或使用语音标注人脸。本文采用 mAP (mean Average Precision)来衡量标注的质量，取前个 $R$ 标注数据进行排序，其标注质量计算公式如式(21)：

$$\mathrm{mAP}=\frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{L}\sum_{r=1}^{R}P(r)\times\delta(r)\right) \tag{21}$$

其中，$M$ 为样本总量，$L$ 为标注正确总量，$P(r)$ 为前 $r$ 个准确率最高的标注数据，当标注正确时 $\delta(r)$ 为 1，反之为 0，在本文中 $(R=50)$。

### 4.3 实验效果对比

典型相关分析(Canonical Correlation Analysis, CCA)是常用的跨模态识别方法。本文参照文献[3]中跨模态匹配方法，选取关联子空间(Correlation Matching, CM)，语义空间(Semantic Matching, SM)，语义关联子空间(Semantic Correlation Matching, SCM)3 种跨模态方法进行试验对比和分析。在参数设置 $\alpha=0.01,0.10\sim0.90,0.99$ 和 $\beta=1$ 情况下，本文提出的 3 种模型，在 Friends 数据集上与不同 CCA 模型方法对比结果如表 1 所示，实验结果表明本文提出的 3 种有监督一致性联合自编码器模型与 3 种传统 CCA 模型方法相比，其不同形式的跨模态说话人标注准确率有明显提高，mAP 结果平均提升约 0.1 左右，其中 Super-Full-AE 方法标注效果最好，达到 0.992。主要原因在于 Friends 数据集人脸表情变化差异较大，传统 CCA 方法只是利用浅层方法对原始提取的特征进行简单的语义匹配和相关性分析，并没有充分挖掘高层的语义的一致性和判别性分析。本文提出的 Super-Full-AE 方法共享人脸语音两个模态的重建层约束，相比于 CM 方法，人脸和语音模态间关联分析更好；相比于 SM 的方法，更能体现高层的语义一致性特征；相比于 SCM 方法，加入 softmax 回归和标签信息使得本文模型更具判别性。

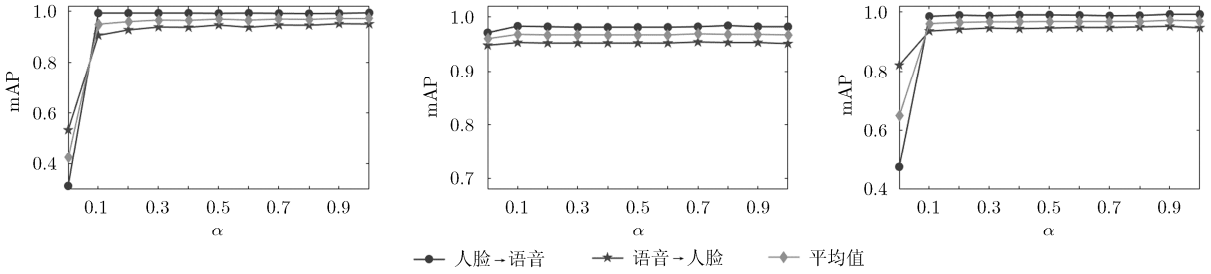
进一步，图 6 表示在选取不同参数情况下在 Friends 数据集上得到的不同实验结果。可以看出，在相同的数据集 Friends 上，$\alpha=0.01,0.10\sim0.90,0.99$，其中当参数选取 0.01 时可以近似表示忽略该部分损失的影响，随着 $\alpha$ 的增大，损失函数中跨模态重建部分比重不断增大，其一致性语义约束权重较大，实验结果越来越好，其中在 $\alpha=0.9$ 时取得相对稳定和较好的结果。

同样，在 Bigbang 数据集上以相同参数配置运行本文方法和 3 种文献[3]CCA 跨模态匹配方法对比试验结果如表 2 所示，并且在不同 $\alpha$ 值的设置情况下得到的实验结果图 7 所示。

如表 2 所示，实验在 Bigbang 数据集上测试结果表明，传统基于 CCA 跨模态匹配的方法获得音视

**表 1 本文方法(深度特征)与原始 CCA 方法(浅层特征)在 Friends 数据集实验对比**

| 方法 | CM | SM | SCM | Super-Corr-AE | Super-Cross-AE | Super-Full-AE |
|---|---|---|---|---|---|---|
| 人脸-语音 | 0.818 | 0.833 | 0.823 | 0.991 | 0.982 | **0.992** |
| 语音-人脸 | 0.739 | 0.752 | 0.746 | 0.951 | **0.954** | 0.952 |
| 平均值 | 0.779 | 0.793 | 0.785 | **0.972** | 0.969 | **0.972** |

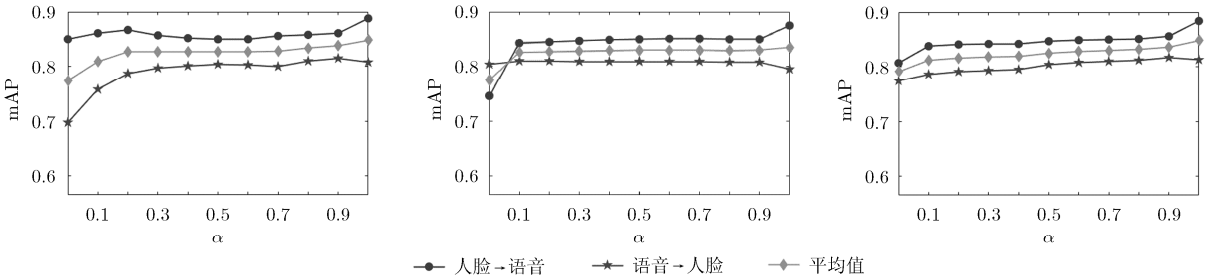



(a) Super-Corr-AE模型　(b) Super-Cross-AE模型　(c) Super-Full-AE模型

图 6　本文 3 种模型在不同参数下的实验结果对比图(Friends 数据集)

**表 2 基于深度特征的不同方法在 Bigbang 数据集上的对比结果**

| 方法 | CM | SM | SCM | Super-Corr-AE | Super-Cross-AE | Super-Full-AE |
|---|---|---|---|---|---|---|
| 人脸-语音 | 0.748 | 0.853 | 0.852 | **0.888** | 0.875 | 0.884 |
| 语音-人脸 | 0.731 | **0.860** | 0.850 | 0.808 | 0.795 | 0.813 |
| 平均值 | 0.740 | **0.857** | 0.851 | 0.848 | 0.835 | 0.848 |




(a) Super-Corr-AE模型　(b) Super-Cross-AE模型　(c) Super-Full-AE模型

图 7 本文 3 种模型在不同参数下的实验结果(Bigbang 数据集)

频互标注结果明显出现了比较大的波动，其中 CM 匹配方法在不同跨模态标注中效果相对有所欠缺，但 SM 方法和 SCM 方法都取得了相对较好的结果，其主要原因在于本文前期利用卷积神经网络和深度信念网络分别对人脸图像和语音数据进行判别性特征提取，其判别性语义表征能够较好相互兼容，从而使得跨模态标注取得了较好的效果。相比之下，本文提出的 3 种模型在不同模态形势下都取得到了相对稳定的跨模态说话人标注结果，平均 mAP 结果都超过了 80%，其中由人脸标注语音时稍好于由语音标注人脸，最终平均标注准确率与 3 种 CCA 方法持平。在参数设置上与 Friends 相同，由图 7 统计的实验数据可知，随着 $\alpha$ 增大，标注准确率越来越高，且在 $\alpha = 0.99$ 时取得相对较好效果。该结果表明本文深度学习模型能够较好的学习人脸和语音两种模态的一致性语义表达，在高层语义表达上较好地解决了两种模态之间的“语义鸿沟”问题，从而跨模态重建部分约束影响较小。从不同数据集对比试验可以看出，本文提出的人脸和语音深度特征提取方法能够判别性地揭示身份语义信息，并且有监督联合一致性自编码器模型能够挖掘不同模态之间的关联特性，从而有效实现两种模态的灵活跨越，实验结果验证了本文方法的有效性和稳定性。

## 5　结束语

本文针对说话人的人脸图像和语音数据结构及表现形式不一致的问题，首先分别利用卷积神经网络和多个受限玻尔兹曼机组成的神经网络提取人脸和语音判别性特征，在一致性自编码器学习框架下，利用有监督的神经网络并加入 Softmax 回归使模型以保证模态间和模态内样本的相似性，并扩展为 3 种有监督一致性自编码器神经网络模型来有效挖掘音视频异构特征之间的潜在关系，从而有效实现人脸和语音的跨模态相互标注。在两个公开数据集上的实验结果表明本文提出的网络模型能够对不同场景下说话人进行跨模态标注，效果显著，取得了对面部姿态变化和样本多样性的鲁棒性，其标注准确率最高达到 0.972。除了人脸和语音两种模态外，本文方法预期也同样适用于其他不同种类生物特征进行跨模态匹配。

## 参 考 文 献


[1] 陈存宝, 赵力. 嵌入自联想神经网络的高斯混合模型说话人辨认[J]. 电子与信息学报, 2010, 32(3): 528–532. doi: 10.3724/SP.J.1146.2008.00275.
CHEN Cunbao and ZHAO Li. Speaker identification based on GMM with embedded AANN[J]. *Journal of Electronics & Information Technology*, 2010, 32(3): 528–532. doi: 10.3724/SP.J.1146.2008.00275.

[2] 郭武, 戴礼荣, 王仁华. 采用因子分析和支持向量机的说话人确认系统[J]. 电子与信息学报, 2009, 31(2): 302–305. doi: 10.3724/SP.J.1146.2007.01289.
GUO Wu, DAI Lirong, and WANG Renhua. Speaker verification based on factor analysis and SVM[J]. *Journal of Electronics & Information Technology*, 2009, 31(2): 302–305. doi: 10.3724/SP.J.1146.2007.01289.

[3] RASIWASIA N, PEREIRA J C, COVIELLO E, *et al*. A new approach to cross-modal multimedia retrieval[C]. ACM International Conference on Multimedia, Firenze, Italy, 2010: 251–260.

[4] ZHANG Liang, MA Bingpeng, LI Guorong, *et al*. Cross-modal retrieval using multiordered discriminative structured subspace learning[J]. *IEEE Transactions on Multimedia*, 2017, 19(6): 1220–1233. doi: 10.1109/TMM.2016.2646219.

[5] ZOU Hui, DU Jixiang, ZHAI Chuanmin, *et al*. Deep learning and shared representation space learning based cross-modal multimedia retrieval[C]. International Conference on Intelligent Computing. Lanzhou, China, 2016: 322–331.

[6] SUN Yi, WANG Xiaogang, and TANG Xiaoou. Hybrid deep learning for face verification[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, 38(10): 1997–2009. doi: 10.1109/TPAMI.2015.2505293.

[7] SUN Yi, WANG Xiaogang, TANG Xiaoou, *et al*. Deep learning face representation from predicting 10,000 classes[C]. IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 1891–1898.

[8] KARAABA M F, SURINTA O, SCHOMAKER L R B, *et al*. Robust face identification with small sample sizes using bag of words and histogram of oriented gradients[C]. International Joint Conference on Computer Vision Imaging and Computer Graphics Theory and Applications, Rome, Italy, 2016: 582–589.

[9] TAIGMAN Y, YANG M, RANZATO M, *et al*. DeepFace: Closing the gap to human-level performance in face verification[C]. IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA, 2014: 1701–1708.

[10] YUAN Xiaochen, PUN Chiman, and CHEN C L. Robust Mel-Frequency cepstral coefficients feature detection and dual-tree complex wavelet transform for digital audio watermarking[J]. *Information Sciences*, 2015, 29(8): 159–179. doi: 10.1016/j.ins.2014.11.040.

[11] PATHAK M A and RAJ B. Privacy-preserving speaker verification and identification using Gaussian mixture models [J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2013, 21(2): 397–406. doi: 10.1109/TASL.2012. 2215602.

[12] HINTON G, LI Deng, DONG Yu, *et al*. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups[J]. *IEEE Signal Processing Magazine*, 2012, 29(6): 82–97. doi: 10.1109/MSP.2012.2205597.

[13] NGIAM J, KHOSLA A, KIM M, *et al*. Multimodal deep learning[C]. IEEE International Conference on Machine Learning, Bellevue, USA, 2011: 689–696.

[14] HU Yongtao, REN J S, DAI Jingwen, *et al*. Deep multimodal speaker naming[C]. ACM International Conference on Multimedia, Brisbane, Australia, 2015: 1107–1110.

[15] FENG Fangxiang, WANG Xi, LI Ruifan, *et al*. Correspondence autoencoders for cross-modal retrieval[J]. *ACM Transactions on Multimedia Computing Communications & Applications*, 2015, 12(1s): 1–22. doi: 10.1145/2808205.

[16] MOHAMED A, DAHL G E, and HINTON G. Acoustic modeling using deep belief networks[J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2012, 20(1): 14–22. doi: 10.1109/TASL.2011.2109382.

[17] WANG Kaiye, HE Ran, WANG Liang, *et al*. Joint feature selection and subspace learning for cross-modal retrieval[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, 38(10): 2010–2023. doi: 10.1109/TPAMI. 2015.2505311.

[18] CASTREJÓN L, AYTAR Y, VONDRICK C, *et al*. Learning aligned cross-modal representations from weakly aligned data[C]. IEEE International Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 2940–2949.

[19] KIM J, NAM J, and GUREVYCH I. Learning semantics with deep belief network for cross-language information retrieval[C]. International Conference on Computational Linguistics, Dublin, Ireland, 2013: 579–588.

[20] TANG Jun, WANG Ke, and SHAO Ling. Supervised matrix factorization hashing for cross-modal retrieval[J]. *IEEE Transactions on Image Processing*, 2016, 25(7): 3157–3166. doi: 10.1109/TIP.2016.2564638.



柳 欣： 男，1982 年生，博士，副教授，研究方向为生物特征识别和机器学习.

李鹤洋： 男，1994 年生，硕士生，研究方向为计算机视觉与模式识别.

钟必能： 男，1981 年生，博士，教授，研究方向为机器学习和模式识别.

杜吉祥： 男，1977 年生，博士，教授，研究方向为计算机视觉和机器学习.