计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191000132

基于深度学习的特种车辆跨模态检索方法

邵阳雪[1,2]　孟　伟[1,2]　孔德珍[2,3]　韩林轩[2,3]　刘　扬[1,2,3]
1 河南大学河南省空间信息处理工程实验室　河南 开封 475004
2 河南大学计算机与信息工程学院　河南 开封 475004
3 河南大学河南省大数据分析与处理重点实验室　河南 开封 475004
(624535306@qq.com)

摘　要　保证正在执行任务的特种车辆的道路优先通行权,是合理配置城市交通资源、实施和保证应急救援的前提。特种车辆的跨模态识别是实现智慧交通的重要核心技术,尤其是在智能车联网尚未成熟、未来长期存在无人驾驶和有人驾驶混合交通的环境中,实现无人车对正在执行任务的特种车辆进行合理避让显得尤为重要。针对无人驾驶对特种车辆识别的需求,文中构建了跨模态检索与识别网络(Cross-Modal Retrieval and Recognition Net,CMR^{2}Net),提出了一种基于深度学习的特种车辆跨模态检索和识别方法。CMR^{2}Net由两个卷积子网络和一个特征融合网络组成,卷积子网络分别用于提取特种车的图像与音频特征,在高层语义空间中利用相似性度量的方法进行特征匹配,以达到跨模态检索和识别的目的。在特种车跨模态数据集上进行的跨模态识别实验表明,所提方法对跨模态检索和识别任务具有较高的识别率,甚至在缺失一种模态的场景下也可准确识别出特种车辆。本研究对于提升"城市大脑"的性能具有重要的理论指导意义,对设计、实现和改善未来智慧交通具有较高的工程应用价值。

关键词:跨模态检索;卷积神经网络;相似性度量;深度学习;小样本
中图法分类号　TP391

Cross-modal Retrieval Method for Special Vehicles Based on Deep Learning

SHAO Yang-xue[1,2],MENG Wei[1,2],KONG Deng-zhen[2,3],HAN Lin-xuan[2,3] and LIU Yang[1,2,3]
1 Henan Engineering Laboratory of Spatial Information Processing,Henan University,Kaifeng,Henan 475004,China
2 School of Computer and Information Engineering,Henan University,Kaifeng,Henan 475004,China
3 Henan Key Laboratory of Big Data Analysis and Processing,Henan University,Kaifeng,Henan 475004,China

Abstract　To ensure the right of way of special vehicles is the premise of rational allocation of urban traffic resources,implementation and guarantee of emergency rescue. The cross-modal identification of special vehicles is an important core technology in realization of intelligent transportation,especially in the environment where the Internet of Vehicles is not yet mature and there will be the long-term unmanned and manned mixed traffic in the future. To make way for the special vehicles reasonable that are performing the mission is particularly important. Aiming at the demand of driverless vehicle for special vehicle identification,this paper constructs a cross-modal retrieval and recognition net(CMR2Net),and proposes a method of cross-modal recognition and retrieval of special vehicles based on deep learning. CMR2Net consists of two convolution sub-networks and one feature fusion network. The convolution sub-networks are used to extract the features of the image and audio of the special vehicle,then the similarity measurement method is used in the high-level semantic space to perform feature matching to achieve cross-modal retrieval and recognition. Cross-modal identification experiments performed on special vehicle cross-modal dataset show that this method performs a high recognition rate for cross-modal retrieval and recognition tasks. Furthermore,it can be accurately identified special vehicles even one modal absence. This research has major theoretical guiding significance for improving the performance of "urban brain",and also can be used in the engineering for designing,realizing and improving the smart transportation in the future.

Keywords　Cross-modal retrieval,Convolutional neural networks,Similarity measurement,Deep learning,Small sample

到稿日期:2019-10-21　返修日期:2020-04-24　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:河南省重点研发与推广专项(192102210096,182102310724)
This work was supported by the Key Research and Promotion Projects of Henan Province(192102210096,182102310724).
通信作者:刘扬(ly.sci.art@gmail.com)

1 引言

无人车在行驶过程中除了需要对基本的障碍物，如行人、车辆等进行识别和避让外，在遇到执行工作任务的救护车、救火车、工程抢险车等特种车辆时，根据道路通行优先权相关法律规定，也必须做出避让。因此正确识别出正在工作的特种车辆是无人驾驶车环境感知模块中不可缺少的功能。但无人车系统在识别特种车的过程中，普遍存在传感器采集数据时模态缺失，以及由于环境干扰和数据质量低而无法正常识别的问题。

为解决上述问题，研究者们在多模态辅助检索与识别的方向做了大量的探索，如 Lin 等[1]提出一种图像结合激光雷达的多模态车道场景目标检测方法(主要识别目标为车辆)，减少了单模态检测带来的错检和漏检情况。He 等[2]提出多任务分段紧凑特征的车辆检索方法，解决了白天光照过强、夜间视野模糊造成的图像质量低，不利于车辆检索的问题，其提出了跨模态检索技术方案，使用文本辅助检索。同样地，Li 等[3]提出将学习到的文本信息知识迁移到视觉信息中，生成跨模态视觉特征，有效地提高了图像检索的准确度。然而，上述研究在数据量较小的情况下无法达到很好的识别度。特种车辆的数据在日常道路场景中出现的概率较低，样本不易采集，属于小样本数据。模型拟合特种车数据时容易出现过拟合，导致识别精度低。为此，Jiang 等[4]将可见光与红外光图像融合，充分利用各光谱中蕴含的信息，在数据样本较少的情况下提高场景识别算法的识别率。但是，目前多源图像太少，能用于训练特种车识别的多源图像更是少之又少。

针对以上问题，本文结合深度学习在特征学习上的优势，选用比多源图像更易获取的特种车辆的图像与音频作为训练数据，提出了一种基于深度学习的特种车辆跨模态检索方法，并针对该检索任务进行了下述工作：

(1)利用特种车的音频和图像信息，实现跨模态识别。使用音频和图像相互辅助检索，解决检索时模态缺失和数据质量低的问题。

(2)由于特种车音频特征较为单一，即使经过添加噪声、调整音量等数据增强方法，数据量还是属于小样本范畴。因此本文使用图像-音频交叉样本配对的方式组建了特种车多模态数据集，增加了训练样本的多样性。

(3)改进了基于 VGG16 模型提出的 L^3-Net 模型[5]，提高了特征提取的效率，并在一定程度上避免了过拟合。

(4)采用改进模型，从已标注的图像-音频数据对中，提取特种车辆的图像与音频各模态的底层特征。使用余弦距离对抽取特征进行相似度度量，进一步分析语义关系来实现特种车辆的跨模态识别，解决了不同模态之间的语义计算问题。

2 相关工作

多模态数据之间有着底层特征异构、高层语义相关的特点。常用的单模态方法在异构的多模态数据之间无法产生关联，不能挖掘出不同模态数据之间的隐含关系。因此多模态任务最重要的目标是学习同语义多模态数据之间的关系。目前流行的方法可分为实值特征学习和二值特征学习，具体可细分为子空间学习、深度学习、哈希变换和主题模型 4 种类型。

子空间学习方法是利用不同模态样本对的关联信息学习得到线性组合函数，并利用该函数将不同模态样本对的特征投影到相同的潜在子空间，然后在该子空间中度量不同模态样本的相似性。子空间学习的经典算法是 Rasiwasia 等[6]提出的典型相关分析(Canonical Correlation Analysis，CCA)算法。除此之外，Liang 等[7]提出了群组不变跨模态子空间学习方法，Sharma 等[8]将无监督 CCA 推广到广义多视角判别分析。子空间学习方法是构建两个模态特征向量之间的最大相关性线性组合，但无法抽象出两个模态的高阶相关性。

深度学习方法是目前多模态任务研究非常重要的方法。该方法利用深度学习具有的优秀的特征提取能力，在底层特征中提取不同模态的有效表示，之后在高层建立不同模态的语义关联。Ngiam 等[9]提出了基于深度网络的跨模态学习方法。Srivastava 等[10]提出了基于深度波尔兹曼机的跨模态学习方法。Feng 等[11]提出了 CRFMS 模型，利用深度波尔茨曼机的双向结构特点，将两个模态不同抽象层次特征关联起来，充分挖掘不同模态间的内在关联，检索精度得到进一步提升。谷歌团队 Kaiser 等[12]构建了一个包含注意力机制网络、混合专家模块(Mixture of Experts，MoE)、自编码器等处理多种类型数据模块的模型，可同时学习文本、图像和音频，并发现了即使是传统上用于解决某种形态问题的模块(例如用于自然语言处理的 attention 机制和 MOE)也能够对其他形态的问题有所帮助。Aytar 等[13]提出了多模态的深度对齐表示方法，指出了如何在高层特征共享对齐的表达。Arandjelovic 等[14]基于 VGG16 模型提出了 L^3-Net，用卷积提取特征后将特征拼接融合，即可用融合后的特征学习到音频-图像之间的共同特征。

主题模型法是借鉴自然语言领域的方法，该方法将跨模态数据的底层特征映射到一个隐性语义空间，并使用生成式模型在隐性语义空间中发掘潜在主题，构建不同模态数据的语义关联，使得该主题的可解释性较强。Hao 等[15]提出的跨模态循环对抗网络(Cross-Modal Cycle Generative Adversarial Network，CMCGAN)可实现图像-音频的互相生成。子空间学习方法、深度学习方法和主题模型都属于实值特征学习，这 3 类方法检索的准确率虽高，但速度较慢。

哈希变换属于二值特征学习，该方法使用不同模态样本对中隐含的信息，学习不同模态的哈希变换，并将不同模态特征映射到一个由哈希码组成的二值空间-汉明空间，然后在汉明空间中实现跨模态检索。哈希变换的特点是检索速度快、存储开销少，但由于其将特征映射至二值空间，造成特征信息部分丢失(如模态之间的相似度)，是一种损失准确度换来速度提升的方法。

除此之外，Liu 等在多源跨模态目标识别与视听跨媒体语义检索等方向做了大量的探索与研究[16-17]。

3 基于深度学习的特种车辆跨模态检索模型

不同模态的训练样本往往具有不均衡性，存在不对称样本和对称样本。不对称样本是指两种模态样本量不等，非一一对

应关系；对称样本则是两种模态样本量相等，且一一对应匹配。

受 L^3-Net 的启发，本文设计了一种可在小样本数据中学习、实现两种不同模态之间共同特征的 CMR^2Net。CMR^2Net 可在不对称样本中，用两个卷积子网络分别提取各模态自身的特征，并将各模态的特征映射到公共空间，进一步采用相似性度量发掘两个模态之间的关联性，同时完成基于内容的单模态检索和跨模态检索。

3.1 模型结构设计

CMR^2Net 由两个卷积子网络和特征融合模块 3 个部分组成，模型结构如图 1 所示。

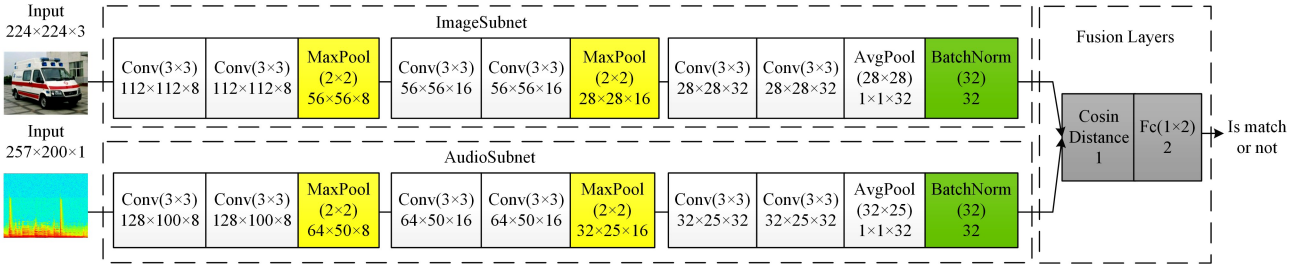

图 1 CMR^2Net 模型结构图

Fig. 1 CMR^2Net model structure diagram

CMR^2Net 在 L^3-Net 的基础上，大幅度减少了参数的数量；其特征融合层使用相似性度量的方式代替 L^3-Net 中简单的特征拼接，并将特征进行融合。具体的网络结构如下。

(1)图像子网络。网络的输入为 224×224×3 的 RBG 图像。卷积层为 6 层，每两层卷积和一层最大池化层[18]作为一个卷积模块，每个模块的卷积核数量相同，分别为 8，16，32。在卷积网络的最后一层使用 Lin 等[19]提出的全局平均池化(Global Average Pooling，GAP)代替 L^3-Net 中子网络的全连接层，GAP 的大小与最后一层卷积输出的特征图尺寸相同，使用特征图的平均值直接作为最后分类的依据。GAP 可有效提高特征提取的效率，同时进一步减少空间参数并压缩模型大小，使模型更加健壮，从而降低了模型过拟合的概率。每层卷积后使用线性整流函数[20](Rectified Linear Unit，ReLU)作为激活函数，并使用批标准化(Batch Normalization)对激活函数的输出进行归一化。

(2)音频子网络。网络输入为 256×200×1 的矩阵，对 1 s 音频作短时傅里叶变换并取对数，汉明窗长度为 512，每个窗口的重叠采样点数为 274。卷积网络的结构与图像子网络的结构相同，音频子网络的输出与图像子网络的输出相同，最终特征维度均为 32。

(3)特征融合模块。特征融合模块由一个相似性度量层和全连接层组成。两个模态的特征由相似性度量层联系起来，全连接层的参数由两个子网络共享。特征融合模块中相似性度量层使用余弦距离(见式(1))计算两个特征的相似度。

$$Sim=\frac{\boldsymbol{FI}\cdot\boldsymbol{FA}}{\|\boldsymbol{FI}\|\ \|\boldsymbol{FA}\|}=\frac{\sum_{i=1}^{n}\boldsymbol{FI}_i\times\boldsymbol{FA}_i}{\sqrt{\sum_{i=1}^{n}\boldsymbol{FI}_i^2}\times\sqrt{\sum_{i=1}^{n}\boldsymbol{FA}_i^2}} \tag{1}$$

其中，$\boldsymbol{FI}$ 为图像特征向量，$\boldsymbol{FA}$ 为音频特征向量，Sim 为余弦相似度。

3.2 基于 CMR^2Net 的特种车辆特征提取算法

算法 1 基于 CMR^2Net 的音频特征提取算法(CMR^2Net based Audio Feature Extraction Algorithm，CAFEA)

输入：待检索音频样本 a，CMR^2Net 权重

输出：32 维音频特征

初始化：载入 CMR^2Net 权重

步骤 1 对 a 作短时傅里叶变换，得到大小为 256×200×1 的特征矩阵 **FA**。

步骤 2 将 **FA** 作为网络的输入，进行卷积

$$\mathbf{B}(i,j)=\sum_{m=0}\sum_{n=0}\mathrm{K}(m,n)*\mathbf{A}(i-m,j-n) \tag{2}$$

其中，K(m，n)为卷积核，m=n=3；**A**(i，j)为被卷积矩阵，**A**(i，j)=**FA**；**B**(i，j)为卷积结果矩阵。

步骤 3 对 **B**(i，j)卷积结果使用 ReLU 进行非线性映射与归一化。

步骤 4 重复步骤 2－步骤 3；并对步骤 3 的输出进行最大池化，窗口大小为 2×2。

步骤 5 步骤 2－步骤 4 为一个卷积模块。重复 2 次步骤 2－步骤 4，在第 3 个卷积模块中将最大池化变为平均池化，窗口大小为 32×25，最后归一化输出。

算法 2 基于 CMR^2Net 的图像特征提取算法(CMR^2Net based Image Feature Extraction Algorithm，CIFEA)

输入：待检索图像样本 i，CMR^2Net 权重

输出：32 维图像特征

初始化：载入 CMR^2Net 权重，resize 图像为 224×224×3

步骤与 CAFEA 相同，步骤 4 中平均池化窗口大小为 28×28。

3.3 跨模态与单模态检索算法

在进行检索之前，需要先准备候选集和候选集的特征向量文件。首先候选集的选择对检索有一定的影响，需要遵循两个标准：1)准确率高，要包含经常出现的样本；2)覆盖率高，要尽量覆盖各种情况的样本，如夜间特种车辆的图像、带噪声的音频等。其次候选集的特征向量文件由[imgEmbedList，audEmbedList，label]三元组构成，其中 imgEmbedList 和 audEmbedList 是候选集通过已训练的 CMR^2Net 提取到的特种车图像与音频特征向量列表。在算法描述中 Em 为三元组，function 为余弦距离公式。跨模态与单模态检索算法(Cross-Modal and Mono-modal Retrieval Algorithm，CMMRA)的描述如算法 3 所示。

算法 3 跨模态与单模态检索算法

Input：Retrieval sample：x，Feature file：Em，trained model：CMR^2Net

Output：Result. sort(：topk) // sort the Similarity

1. initrailize Retrieval result aggregate：Result[]，topk=10

```
2. y = CMR²Net(x) //y is x's feature
3. for i,j in Length(Em):  //i is imgEmbedList index, j is audEmbedList index
4.    if image→audio:
          distance = function(y, audEmbedList[j]) // Cross-modal Retrieval
5.    if audio→image:
          distance = function(y, imgEmbedList[i])
6.    if image→image:
          distance = function(y, imgEmbedList[i])// Mono-modal Retrieval
7.    if audio→audio:
          distance = function(y, audEmbedList[j])
8. distance join Result[]
```

该算法可通过同一流程完成跨模态检索与单模态检索。在检索过程中，所有待检索模态特征与目标模态特征进行相似度计算，并将相似度进行排序即可得到检索结果，算法中 topk=1 即为识别结果。

4 实验与结果分析

4.1 数据集与数据预处理

为检验算法和模型性能，我们收集和整理了特种车多模态数据集(Special Vehicles Multimode Dataset，SVMD)，如图 2 所示。数据集与数据预处理方法已在开放科学计划 OSID 中公开。

图 2 特种车多模态数据集

Fig. 2 Special vehicles multimode dataset

4.2 训练与参数设置

CMR^2Net 的训练任务为输入一对图像-音频样本对，判断它们是否来自同一对象，即图像与音频是否配对，这是一个二分类任务。训练集通过图像-音频配对组成 4000 个正样本、4000 个负样本。模型使用 Pytorch 实现，训练时数据输入为批量输入且随机打乱，batchsize 为 8。模型使用式(5)的铰链损失[21](Hinge loss)作为损失函数，它常用在二分类任务中。其中 t 为目标值，y 为预测值。Hinge loss 使用随机梯度下降[22](Stochastic Gradient Descent，SGD)算法优化，学习率为 0.01，设置权重衰减值为 10^{-7} 以防止过拟合。

$$L(y)=\max(0,1-t\cdot y) \tag{3}$$

作为对比，CMR^2Net 使用 Softmax+CrossEntropy loss[23] 优化网络，其中 Softmax 层的输出为[0,1]两个值，0 表述两个模态信息匹配，1 表示两个模态信息不匹配。

4.3 结果与分析

实验使用 3.2 节描述的检索算法进行跨模态和单模态检索，以 L^3-Net 在特种车多模态数据集上的检索能力为基准，对比了 CMR^2Net 在 CrossEntropy loss 和 Hinge loss 损失函数下的表现，实验结果如表 1 所列。

表 1 特种车多模态数据集中不同模型的性能对比

Table 1 Performance comparison of different models in SVMD

(单位：%)

Retrieval Model	L^3-Net Top10 *mAP*	CMR^2Net+Cross Entropy Top10 *mAP*	CMR^2Net+Hinge loss Top10 *mAP*
Img→Aud	1.97	**63.95**	53.23
Aud→Img	15.51	57.79	**60.58**
Img→Img	**86.64**	72.83	74.63
Aud→Aud	94.15	**96.32**	93.15

在 3 个模型中，CMR^2Net+Hingeloss 模型的跨模态检索准确率最高，性能指标评价采用平均准确率(Mean Average Precision，MAP)，*mAP* 值越大表示模型检索能力越强。

从表中数据中可看出，L^3-Net 对于各个模态特征提取效果较为优秀，但在跨模态检索任务中的检索效果并不好，因其特征融合层含有全连接层，参数较多，在小样本数据的训练下出现过拟合，无法有效地学习到高层语义信息。经过改进后的 CMR^2Net 无论是使用 CrossEntropy loss 还是 Hinge loss 作为损失函数，在跨模态检索任务和单模态检索任务中均能达到较好的检索精度，有效避免了模型过拟合。同时实验结果还验证了小样本的音频可通过与样本量较大的图像共享损失函数，实现知识的迁移学习，在 Aud→Aud 和跨模态检索任务中均有较好的表现。

表 2 列出了 3 种模型跨模态检索各类特种车的 *mAP* 和 CMR^2Net+Hingeloss 与 CMR^2Net+CrossEntropy 总体数据的方差值，因 L^3-Net 的 *mAP* 较低而不予比较。图 3 为 Img→Aud 与 Aud→Img 两种跨模态检索任务的 *mAP* 统计情况。可观察到，CMR^2Net 在各种特种车上检索准确度均高于 L^3-Net。从方差值(Variance)可分析出，CMR^2Net+Hinge loss 在各种特种车的两种跨模态相互检索任务中表现较为稳定。综合以上分析和数据，CMR^2Net+Hingeloss 表现最佳。

表 2 不同模型跨模态检索各类特种车的 mAP@Top10

Table 2 mAP@Top10 of different models cross-modal retrieval each class special vehicles

(单位：%)

Methods / Classes	L^3-Net		CMR^2Net+CrossEntropy		CMR^2Net+Hingeloss	
	Img→Aud	Aud→Img	Img→Aud	Aud→Img	Img→Aud	Aud→Img
Ambulance	0.00	0.00	20.30	56.82	23.77	20.49
Engineering	3.33	11.28	97.66	**2.48**	98.67	85.62
Fire	0.00	56.22	42.42	96.97	41.78	89.32
Police	0.00	8.37	77.10	46.98	66.58	62.79
Variance	—	—	—	0.1031	—	**0.0763**

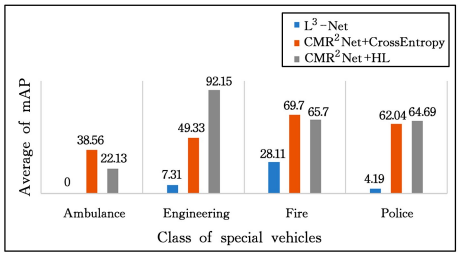

图 3 不同跨模态检索任务的 *mAP*

Fig. 3 Differente cross-modal retrieval search *mAP*

结束语 本文提出的基于深度学习的特种车辆跨模态检索方法，利用深度神经网络无需特征工程和端到端训练的优势，构建了由两个深度卷积神经子网络和一个融合模块组成的跨模态检索模型 CMR² Net。实验结果表明，CMR² Net 模型在特种车多模态数据集上可训练达到较高的跨模态检索精度。

本文进一步验证了同一对象的不同模态数据可代替原目标模态数据以实现检索，很好地解决了检索时模态缺失和数据质量低的问题。同时 CMR² Net 模型比现有的方法更适合数据不易获得的小样本数据与各个模态样本数量不均匀的场景。本文重点研究跨模态检索的精度提升问题，模型检索的时效性还有待进一步深入研究。

参 考 文 献

[1] LIN Z H. Multimodal Deep Learning Object Detecting and Application[D]. Chengdu: University of Electronic Science and Technology of China. 2018.

[2] HE X, TANG Y P, CHEN P. Fast hash vehicle retrieval method based on multitasking [J]. Journal of Image and Graphics, 2018, 23(12): 1801-1812.

[3] LI X Y, NIE X S, CUI C R, et al. Image Retrieval Algorithm Based on Transfer Learning [J]. Computer Science, 2019, 46(1): 73-77.

[4] JIANG Z T, QIN J Q, HU S. Multi-spectral Scene Recognition Method Basedon Multi-way Convolution Neural Network[J]. Computer Science, 2019, 46(9): 265-270.

[5] ARANDJELOVI R, ZISSERMAN, et al. Look, Listen and Learn [J/OL]. https://ui. adsabs. harvard. edu/abs/2017arXiv170508168A.

[6] RASIWASIA N, PEREIRA J C, COVIELLO E, et al. A New Approach to Cross-Modal Multimedia Retrieval [C]// International Conference on Multimedia. 2010: 521-535.

[7] JIAN L, RAN H, SUN Z, et al. Group-Invariant Cross-Modal Subspace Learning [C]// International Joint Conference on Artificial Intelligence. Seattle, WA, USA: IEEE Press, 2016: 1739-1745.

[8] SHARMA A, KUMAR A, DAUME H, et al. Generalized Multiview Analysis: A discriminative latent space [C]// IEEE Conference on Computer Vision & Pattern Recognition. 2012: 2160-2167.

[9] NGIAM J, KHOSLA A, KIM M, et al. Multimodal deep learning [C]// International Conference on Machine Learning. Washington, USA, 2011: 689-696.

[10] SRIVASTAVA, NITISH, SALAKHUTDINOV, et al. Multimodal Learning with Deep Boltzmann Machines [C]// Advances in Neural Information Processing Systems. 2012: 2222-2230.

[11] FENG Y G. CAI GY. Cross-modal Retrieval Fusing Multilayer Semantic [J]. Computer Science, 2019, 46(3): 227-233.

[12] KAISER L, GOMEZ A N, SHAZEER N, et al. One Model To Learn Them All[J/OL]. https: // ui. adsabs. harvard. edu/abs/2017arXiv170605137K.

[13] AYTAR Y, VONDRICK C, TORRALBA A. See, Hear, and Read: Deep Aligned Representations [J/OL]. https: // ui. adsabs. harvard. edu/abs/2017arXiv170600932A.

[14] ARANDJELOVIĆ, RELJA, ZISSERMAN, et al. Look, Listen and Learn[EB/OL]. https: // ui. adsabs. harvard. edu/abs/2017arXiv170508168A.

[15] HAO W, ZHANG Z, HE G. CMCGAN: A Uniform Framework for Cross-Modal Visual-Audio Mutual Generation[C]// AAAI Conference on Artificial Intelligence (AAAI). New Orleans, LA, USA: AAAI, 2018: 6886-6893.

[16] LIU Y, CAI K, LIU C, et al. CSRNCVA: a Model of Cross-media Semantic Retrieval based on Neural Computing of Visual and Auditory Sensations [J]. Neural Network World, 2018, 28(4): 305-323.

[17] LIU Y, TU C L, ZHENG F B. Research of Neural Cognitive Computing Model for Visual and Auditory Cross-media Retrieval [J]. Computer Science, 2015, 42(3): 19-25, 30.

[18] JIN K. H, Maccan M. T, Froustey E, et al. Deep Convolutional Neural Network for Inverse Problems in Imaging [J]. IEEE Transactions on Image Procession, 2016, 26(9): 4509-4522.

[19] LIN M, CHEN Q, YAN S. Network In Network [J/OL]. https://ui. adsabs. harvard. edu/abs/2013arXiv1312. 4400L.

[20] HAHNLOSER RICHARD H R, SEBASTIAN S H, JACQUES S J. Permitted and forbidden sets in symmetric threshold-li-near networks. [J]. Neural Computation, 2003, 15(3): 621-638.

[21] VAPNIK V N. Statistical Learning Theory[J]. Encyclopedia of the ences of Learning, 1998, 41(4): 3185.

[22] HAO Y, QI C. Robust virtual frontal face synthesis from a given pose usingregularized linear regression[C]// International Conference on Image Processing (ICIP). Paris: IEEE Press, 2014: 702-4706.

[23] LIU W, WEN Y, YU Z, et al. Large-margin softmax loss for convolutional neural networks[C]// International Conferencer on International Conference on Machine Learning. Vienna, Austria: ICML, 2016: 69-75.

SHAO Yang-xue, born in 1994, postgraduate, is a member of China Computer Federation. Her main research interests include cross-modal retrieval, machine learning and brain-like computing.

LIU Yang, born in 1971, Ph. D, associate professor, M. S. supervisor. is a member of China Computer Federation. His main research interests include brain-inspired computing (i. e. , multimedia neural cognitive computing, multisource cross-modal target recognition, and audio-visual cross-media semantic retrieval), and temporal-spatial information high-performance computing in remote sensing.