Computer Science and Application 计算机科学与应用, 2022, 12(7), 1747-1755
Published Online July 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.127176

Hans汉斯

基于多模态特征记忆库的视频语音检索模型

李劼博，陈俊洪，林大润，杨振国，刘文印

广东工业大学计算机学院，广东 广州

收稿日期：2022年6月10日；录用日期：2022年7月5日；发布日期：2022年7月14日

摘　要

由于不同模态间的数据表示方式不一致，跨模态检索是多媒体领域中的一大难题。本文设计了一种基于多模态特征记忆库的视频语音检索模型，该模型主要分为三个模块，分别为特征提取模块，多模态特征映射融合模块和特征记忆库模块。在特征提取模块中，我们分别使用I3D和Bi-LSTM来提取视频中的操作动作特征和语音中的特征信息。在特征映射融合模块中，先将两种模态特征对齐到同一空间中，再进行融合。在第三个模块中，本文创新性地引入了两个对应视频和语音的特征记忆库，根据特定条件在训练和测试过程中不断更新。在经过我们拓展过的MPII Cooking 2数据集进行实验，结果表明我们的模型能够实现更好的视频语音检索效果。

关键词

跨模态检索，视频语音检索，I3D网络，Bi-LSTM，特征记忆库

Video Speech Retrieval Model Based on Multimodal Feature Memory

Jiebo Li, Junhong Chen, Darun Lin, Zhenguo Yang, Wenyin Liu

School of Computer Science and Technology, Guangdong University of Technology, Guangzhou Guangdong

Received: Jun. 10th, 2022; accepted: Jul. 5th, 2022; published: Jul. 14th, 2022

Abstract

Cross-modal retrieval is a major challenge in multimedia field due to the inconsistent data representation among different modalities. In this paper, we design a video speech retrieval model based on multimodal feature memory library, which is divided into three main modules, namely, feature extraction module, multimodal feature mapping fusion module and feature memory library module. In the feature extraction module, we use I3D and Bi-LSTM to extract the operational action features in video and feature information in speech, respectively. In the feature mapping

文章引用: 李劼博, 陈俊洪, 林大润, 杨振国, 刘文印. 基于多模态特征记忆库的视频语音检索模型[J]. 计算机科学与应用, 2022, 12(7): 1747-1755. DOI: 10.12677/csa.2022.127176

fusion module, the two modal features are first aligned into the same space and then fused. In the third module, two feature memories corresponding to video and speech are innovatively introduced in this paper, which are continuously updated during training and testing according to specific conditions. Experiments are conducted on our extended MPII Cooking 2 dataset, and the results show that our model can achieve better video-speech retrieval results.

Keywords

Cross-Modal Retrieval, Video and Speech Retrieval, I3D Network, Bi-LSTM (Bi-Directional Long Short-Term Memory), Feature Memory Libraries

1. 引言

随着互联网的迅猛发展，人们可以轻松地、随时随地与他人进行交流沟通，他们上传工作生活等方面的视频照片到互联网上进行分析，然而随之而来的便是文字类、图片类、视频类和音频类等多媒体数据的爆炸式激增。面对天量的媒体数据，跨模态检索技术能够在它们之间搭建出一座桥梁。合理利用现有的机器学习技术和深度学习算法，探索多模态内容的对应关系以实现机器对不同模态信息的理解，最终可以达到通过文本检索音频、图像和视频，以及音频检索图像和视频的目的。跨模态检索在实际的商业、教育和政务环境中，可以得到广泛的应用。比如，文案广告推送、潮流软件的推荐，能理解人类语音信息的人机交互系统，符合实际民生的服务推荐等。

得益于深度学习的蓬勃发展，近些年的学者们在图文检索类的跨模态检索任务上已经获得不小的突破，例如通过环境声音去理解、识别场景[1] [2] [3] [4]，通过语音分析人类情感关系[5] [6]等，随着计算机硬件的提升，视频语音等跨模态分析成为了研究热点，例如 Rouditchenko 等人[7]提出了一种新颖的自我监督方法 AVLnet，该方法通过引入视频中自带的音频信息学习视听语言表示，并为 YouCook2，MSR-VTT 和 LSMDC 这三个数据集在视频和语言检索任务上建立了音频到视频的基准，性能超过了一些文本到视频的基准，有效地降低了对昂贵且耗时的数据注释的需求。然而由于视频数据和语音数据相互表示复杂，所以目前在跨模态检索领域内，视频语音检索受到的关注较少。针对此现状，本文提出一种基于特征记忆库的视频语音检索模型，如图 1 所示。该模型主要包括四个阶段。第一阶段，首先进行提取视频信息和音频信息的特征。第二阶段，将前面提取得到的两种不同模态特征分别送入一个全连接层神经网络中，投射到一个潜在的公共空间中并继续处理得到一个初步的分类结果。第三阶段，根据上一步的分类结果，采取特定策略更新到特征记忆库中，具体策略将在第三节详细展开介绍。特征记忆库更新完毕后，将当前样本特征与特征记忆库做矩阵乘法得到两个记忆向量。两个向量与分类结果做修正计算后得到最终结果。

本文的主要贡献如下：

1) 本文提出了一种基于特征记忆库的视频语音检索模型，它可以学习视频和语音信之间隐含的关系。

2) 本文设计了两个独立的特征记忆库，并制定了更新策略，可以提高模型的整体检索性能。

3) 本文对 MPII 烹饪活动数据集 2.0 (Max Planck Institute for Information Cooking Activities 2.0)数据集进行了必要的拓展。

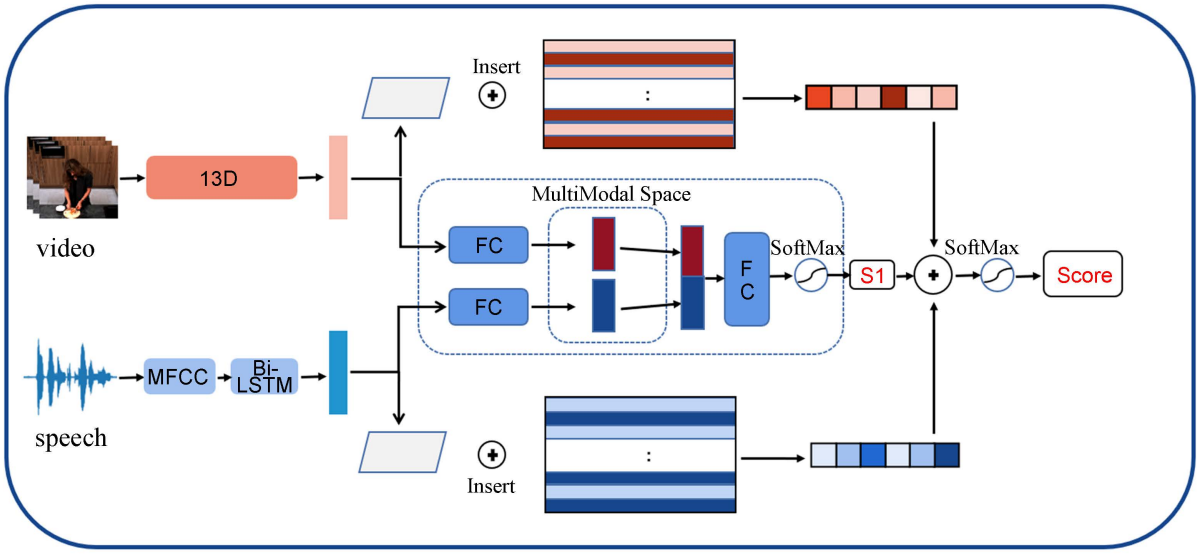

Figure 1. Feature memory network model
图 1. 特征记忆库网络模型

2. 相关研究

2.1. 视频特征提取

对于视频特征的提取，早期人们首先采用对视频进行降噪处理，然后对每一帧图像进行特征提取，最后通过 Average Pooling 和 Fisher Vector 等特征编码方法得到视频特征[8]。然而实验过程中，人们发现这种方法在长视频中表现不佳。随着深度学习的兴起，视频特征提取有了更优秀的方法出现。例如 Venugopalan 等人[9]使用长短期记忆(Long Short-Term Memory, LSTM)来采集视频中的时空信息。Tran 等人在[10]最早将卷积神经网络(Convolutional Neural Network, CNN)应用于视频数据集上，进而提出专门可用于视频的 3D CNN 模型。

2.2. 语音特征提取

语音特征提取的传统方法主要是梅尔频率倒谱系数(Mel Frequency Cepstrum Cofficients, MFCC) [11]。在 2017 年，谷歌公司(Google)发表了一篇语音识别分类论文，创新性将卷积神经网络应用到音频领域，并将网络命名为 VGGish [11]，该模型对牛津大学所发布的 VGG 卷积网络[12]做了一定修改，将其最后一层的所有单元使用批量归一化(Batch Normalization, BN)来替代了原有的局部响应归一化(Local Response Normalization, LRN)，随后在 YouTube-100M 这个大型的数据集上进行训练和测试。最终结果显示，该类 VGG 模型的 VGGish 提取的特征被用于音频分类任务上，相比基于传统特征的工作，有更为出色的表现。此后，涌现出不少的后续研究，均选用 VGGish 作为提取音频特征的模型并获得优异的表现。

2.3. 跨模态检索

跨模态检索是用一种模态数据作为查询条件来检索具有对应或相似语义的另外一种模态数据。早期，已有众多学者进行了研究。例如，Hotelling [13]提出了典型相关性分析方法(Canonical Correlation Analysis, CCA)，Li 等人[14]提出了跨模态因子分析方法(Cross-modal Factor Analysis, CFA)。在 2010 年，Liu 等人[15]便对过往跨模态检索的应用工作做了一个全面的分析和梳理。同年，Rasiwasia 等人[16]的工作指出，传统的跨模态研究方法在建立不同模态数据的深层隐含关系上存在瓶颈。近年来跨模态检索领域也迎来

了一股新生的活力。文献[17]给跨模态检索领域的建模策略做了两种分类，分别是直接建模和间接建模。Wang 等人[18]则在 2016 年对跨模态检索领域做了总结，发表了他们在模型输出采用实数值表示或二值化表示时对训练方式影响的工作。Hu 等人[19]设计了一种可以同时兼顾全局特征和局部特征的图文检索方法。Kamper 等人[20]提出了一种能够实现音频和文本之间检索的框架。Mithun 等人[21]提出了一种基于弱监督的视觉文本联合嵌入的网络模型，在图文双向检索上展示出优异效果。然而，以上的研究及相关工作，都很少有将语音信息和视频信息联合起来使用，并实现视频语音之间检索的工作。

3. 本文方法

本文提出的方法采取了一主二辅的三分支思路，设计了一个特征记忆库网络(Feature Memory Network, FMN)，其结构如图 1 所示。它由特征提取模块、特征映射融合模块和特征记忆库模块组成。下文将逐步展开介绍各个模块细节。

3.1. 视频处理

视频处理模块旨在从视频中获取到全局动作特征。对于机器而言，动作是一组时空序列。因此，要捕获动作特征必然要同时考虑到空间和时间的信息。所以，本文在此采用 I3D 双流网络(Two-Stream Inflated 3D ConvNets) [22]。该网络模型借鉴了一个极其成熟的 2D 图像分类网络——Inception-V1 框架，通过增加了 2D 卷积核和池化核的时间维度，将该网络中的 2D 卷积网络扩张成为了 3D 卷积网络。在模型的训练期间，包含 RGB 图像流和光流的原始视频帧会被重新调整成 256 × 256 的尺寸，然后再被随机剪裁成 224 × 224 的尺寸，对 RGB 图像流和光流均采用了 Softmax 损失函数。本文在应用该网络模型时，首先在 Kinetics 数据集上对 I3D 进行了预训练。然后，再在本次实验所用到的 MPII 烹饪活动数据集 2.0 上做了一定的细微调整。最后，将其作为从视频中提取特征的模块。

3.2. 音频处理

当下的深度学习模型不能直接处理原始格式的音频数据，如 MP3、WAV 和 ACC 等。于是，会先对音频做特征抽取，一个常用且有效的算法是 MFCC。然而，如果直接将 MFCC 特征送入神经网络处理，会使模型忽略掉一段完整语音中的上下文信息的关联。因此，本文采用双流长短期记忆(Bi-Directional Long Short-Term Memory, Bi-LSTM)来进一步处理前面得到 MFCC 音频特征，捕获上下文信息。Bi-LSTM 包括了一个用来计算序列前向隐含状态的 LSTM 网络和一个计算反向隐含状态的 LSTM 网络，两个方向上均得到维数为 U 的隐含状态。处理之后，得到的 1 × 2048 维特征向量，已经有效地融入前向特征和反向特征，以增强音频信息。

3.3. 特征映射与融合

在本小节，将介绍视频和语音特征的下一步处理细节。

首先，视频和语音特征都送入一个全连接层去映射到一个公共空间，采用三元组排序损失函数(Triplet Ranking Loss)去做参数调整，公式如(1)所示：

$$L_{triplet} = \max\left[0, S(a,p) + S(a,n) + margin\right] \tag{1}$$

其中，$S(\cdot,\cdot)$ 代表距离，在本文使用余弦相似度函数衡量特征差异，公式如(2)所示：

$$Similarity(X,Y) = Cos(\theta) = \frac{\vec{x}\cdot\vec{y}}{\|x\|\cdot\|y\|} = \frac{\sum_{i=1}^{n} X_i \times Y_i}{\sqrt{\sum_{i=1}^{n}(X_i)^2} \times \sqrt{\sum_{i=1}^{n}(Y_i)^2}} \tag{2}$$

然后，将经过映射后的视频特征与语音特征横向融合，再送入 Softmax 层，得到一个分类结果分值 S1。该分值起两个作用，一是在训练阶段，作为更新特征记忆库的一个前提条件，二是作为一个基础分值 S1，参与模型的后续运算。

3.4. 特征记忆库

在本小节，将介绍经过两个特征提取模块后得到的视频和语音特征，分别送入两个含有相应特征记忆库辅助支路的处理细节。

3.4.1. 特征记忆库的构成

首先，介绍特征记忆库的构成，无论是视频支路还是语音支路，其特征记忆库均由多条视频或语音特征并排构成。准确来说，特征记忆库是是一个的 2048 × X 的矩阵，X 表示特征记忆库所所能存储单一特征的最大条数，即该特征记忆库的容量。特别的是，X 的值是可以随着数据集的特点和实验数据去动态调整的，寻找到一个合适的数值后确定下来。

3.4.2. 特征记忆库的作用

特征记忆库用于储存，关键在于利用记忆向量的策略，辅助中间支路，做出更加细致有效的判断。下面进行介绍：

当一个样本特征向量被送入当前支路时，样本特征向量与记忆库中的每一条特征向量做一个余弦相似度计算以衡量当前样本与记忆库中某条特征的差异，最终得到一维向量去表示样本与特征记忆库中所有特征的相似度，称作记忆向量。在视频和语音支路均得到该记忆向量后，两两相乘，取最大值 Max，用该值表示两条辅助支路对样本特征向量的分歧。最后再用 Max 与 $S1$ 做修正计算，得到最终可用于衡量视频和语音匹配程度的得分 S。公式如(3)所示：

$$S = \frac{S1}{\mathrm{Max}_x\left(M_{x-video} \times M_{x-audio}\right)} \tag{3}$$

同时，整个模型采用的二分类交叉熵损失(Binary Cross Entropy Loss, BCE Loss)来计算，公式如(4)所示：

$$L_{binary} = -\left(y \times \log\left(\hat{y}\right) + \left(1-y\right) \times \log\left(1-\hat{y}\right)\right) \tag{4}$$

其中，表示模型预测样本为正样本的概率，y 表示实际的样本标签(Ground-Turth)。

3.4.3. 特征记忆库的建立和更新策略

特征记忆库初始化的时候，整个矩阵设定初始值为零。训练过程中，对于每一对进入模型的样本，经过特征提取后，首先进入中间支路经过映射操作并得到一个 S1，本文将通过多次实验找到一个阈值 T1，当 S1 > T1，即认为通过中间支路也能判断出来视频和语音是匹配的。此时，将未映射之前的原始视频特征和音频特征分别与视频和语音特征记忆库中的全部已有特征计算余弦相似度，取得最大值 m2，若 m2 仍小于一个设定的阈值 T2，则认为当前样本特征对于特征记忆库而言，是一种崭新的、未见过的特征，于是将当前的样本特征添加到记忆库中。反之，若 m2 大于阈值 T2，则认为该样本特征已经存在于特征记忆库中，便仅对取得最大值的记忆库特征做加权平均更新。

4. 实验评估

4.1. 数据集

为了对所提出方法的性能进行评估，本文将使用以下介绍的数据集进行训练和测试：

MPII 烹饪活动数据集 2.0 (Max Planck Institute for Information Cooking Activities 2.0) [23]囊括了 273 个由不同的人参与录制的烹饪视频构成。数据集发布方已经根据动作的分类把视频分割成了多个的视频小片段。本文为用于本次饰演的动作视频扩充了相应的语音数据。

在此次实验中，本文选取了 20 种不同类别的视频片段及其相对应的语音片段，一共有 50,000 条数据，每种类别的视频和语音都包含不同的主体、动作和受体。对于每个类别，视频和语音都分别包含了 50 条数据，即每个类别的视频或语音会有 2500 条的组合数据。如表 1 所示：

Table 1. The category information involved in the MPII Cooking 2.0 dataset
表 1. MPII 烹饪活动数据集 2.0 涉及到的类别信息

动作	主体		受体
倒	苹果	油	餐叉
切	碗	洋葱	煎锅
撒	萝卜	橙子	餐刀
挤压	黄瓜	盘子	香料
搅拌	煎锅	罐子	榨汁机
取出	切菜板	菠萝	抹刀
洗	餐刀	盐	盘子
磨碎	柠檬	抹刀	罐子
剥	青柠	香料	
	牛奶	罐头	

4.2. 评估指标

评估指标：本文使用跨模态检索领域的标准评估指标来验证所提框架的学习与评估性能，主要是使用了召回率(Recall)这一主要指标。

召回率是指检索系统返回的查询样本相关音视频数据与数据集所有匹配条目之比。召回率的计算公式如(5)所示：

$$R(recall)=\frac{TP}{TP+FN} \tag{5}$$

其中，*TP* (True Positive)，即预测为真，实际为真，具体来说就是检索返回的与查询样本匹配的正样本数量，*FN* (False Negative)，即预测为假，实际为真，具体而言就是数据集中没有返回的与查询样本匹配的正样本数量。

对于音视频之间的检索，常用的一个评价标准是 R@k，即为 Recall@k。R@k 计算在前 k 个检索的语音或视频中找到至少一个正确结果的测试视频或测试语音的百分比。对于视频检索语音，计算前 k 个检索的视频的百分比，即测量在前 k 个结果中检索到正确语音条目的查询比例，反之亦然。其中，“R@1”、“R@10”、“R@25”分别表示前 1、10、25 个结果的召回率。

实施细节：对于网络的训练，本文选择 Adam 作为网络的优化器并将其学习率设置为，训练的迭代周期 epoch 设置为 500，批处理 batch 的大小则设置为 64。本文的所有实验都是使用 NVIDIA GeForce RTX 3080 GPU 进行训练和测试。

4.3. 视频语音检索性能评估

对于视频语音检索性能的评估，我们将其分为如下两个任务：

1) 语音检索视频：根据输入的视频片段，检索出对应的语音片段；

2) 视频检索语音：根据输入的语音片段，检索出对应的视频片段；

以上两个任务，我们都采用 k 处的召回率(R@k)来衡量所提算法的性能，具体为查询结果最近的 k 点处检索到正确结果的比例。在下列对比试验中，我们给出了各种音视频检索算法之间的性能对比，其中 CCA [24]，KCCA [25]，DCCA [26]是传统的音视频检索算法，而 MuSimNet [27]，MusiCNN [28]，VM-Net [29]则是近几年较为成熟的深度学习模型。

检索音视频的性能评估

为了测试我们的模型对音视频的检索性能，本文将随机选择 20 种类别中 70%的数据用于训练，并在训练过程中寻找合适的特征记忆库维度，在本次实验中，特征记忆库的维度为 50。同时，将剩余 30%的数据作为测试集参与测试。

从表 2 中我们可以看出，基于传统方法的音视频检索准确率均不如基于深度学习的方法，并且当 k = 25 时，深度学习模型比传统模型的综合检索的召回率平均提升了近一倍。同时，我们模型在语音检索视频任务中的 R@1，R@10，R@25 也比目前最好的深度学习模型分别提高了 40.42%，29.15%和 19.93%，相应地，视频检索语音任务中则分别提高了 40.16%，30.66%，14.15%。

Table 2. Performance comparison of different audio and video retrieval algorithms
表 2. 不同音视频检索算法的性能对比

	AudiotoVideo			VideotoAudio		
	R@1	R@10	R@25	R@1	R@10	R@25
CCA	0.92	6.83	9.59	0.87	6.34	9.78
KCCA	1.59	7.98	10.82	1.12	7.44	10.31
DCCA	1.83	8.35	15.65	1.55	7.87	12.78
MuSimNet	1.93	8.21	16.24	1.85	8.35	16.80
MusiCNN	2.13	8.73	19.78	2.05	8.41	19.52
VM-Net	2.40	9.64	21.37	2.54	9.36	21.35
Ours	**3.37**	**12.45**	**25.63**	**3.56**	**12.23**	**24.37**

5. 结论

本文提出了一种深度学习模型，它基于视频和语音的特征，实现视频和语音之间的互相高效检索。该模型首先通过 I3D 卷积神经网络和 Bi-LSTM，分别提取视频和语音中的特征信息。然后将两种模态特征映射到同一个隐含的公共子空间中，再进行融合。随后，本文充分考虑了人类具有长期记忆的特点，创新性地该特点引入到模型的设计当中，构建了两个特征记忆库，目的是为该模型提供更加细致准确的样本识别和学习新类样本的能力。经过本文的性能评估实验，可以表明本文方法相较于传统和目前流行的视频语音检索方法，能够达到一个更加优异的效果。

参考文献

[1] Yang, H. and Meinel, C. (2014) Content Based Lecture Video Retrieval Using Speech and Video Text Information.

IEEE Transactions on Learning Technologies, **7**, 142-154. https://doi.org/10.1109/TLT.2014.2307305

[2] Owens, A., Wu, J.J., McDermott, J.H., Freeman, W.T. and Torralba, A. (2016) Ambient Sound Provides Supervision for Visual Learning. *Proceedings of the European Conference on Computer Vision* (*ECCV*), Amsterdam, 11-14 October 2016, 801-816. https://doi.org/10.1007/978-3-319-46448-0_48

[3] Gaver, W.W. (1993) What in the World Do We Hear? An Ecological Approach to Auditory Event Perception. *Ecological Psychology*, **5**, 1-29. https://doi.org/10.1207/s15326969eco0501_1

[4] McDermott, J.H. and Simoncelli, E.P. (2011) Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis. *Neuron*, **71**, 926-940. https://doi.org/10.1016/j.neuron.2011.06.032

[5] Darwin, C. and Prodger, P. (1998) The Expression of the Emotions in Man and Animals. Oxford University Press, Oxford.

[6] Tian, Y.-I., Kanade, T. and Cohn, J.F. (2001) Recognizing Action Units for Facial Expression Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**, 97-115. https://doi.org/10.1109/34.908962

[7] Rouditchenko, A., Boggust, A., Harwath, D., Joshi, D., Thomas, S., Audhkhasi, K., Feris, R., Kingsbury, B., Picheny, M., Torralba, A. and Glass, J. (2020) AVLnet: Learning Audio-Visual Language Representations from Instructional Videos. *INTERSPEECH* 2021, Brno, 30 August-3 September 2021, 1584-1588. https://doi.org/10.21437/Interspeech.2021-1312

[8] 董建锋. 跨模态检索中的相关度计算研究[D]: [博士学位论文]. 杭州: 浙江大学, 2018.

[9] Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T. and Saenko, K. (2015) Sequence to Sequence-Video to Text. *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, 11-18 December 2015, 4534-4542. https://doi.org/10.1109/ICCV.2015.515

[10] Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M. (2015) Learning Spatiotemporal Features with 3d Convolutional Networks. *IEEE International Conference on Computer Vision*, Santiago, 7-13 December 2015, 4489-4497. https://doi.org/10.1109/ICCV.2015.510

[11] Hershey, S., Chaudhuri, S., Ellis, D.P.W., *et al*. (2017) CNN Architectures for Large-Scale Audio Classification. 2017 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), New Orleans, 5-9 March 2017, 131-135. https://doi.org/10.1109/ICASSP.2017.7952132

[12] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition.

[13] Hotelling, H. (1992) Relations between Two Sets of Variates. In: Kotz, S. and Johnson, N.L., Eds., *Breakthroughs in Statistics*, Springer, New York, 162-190. https://doi.org/10.1007/978-1-4612-4380-9_14

[14] Li, D., Dimitrova, N., Li, M., *et al*. (2003) Multimedia Content Processing through Cross-Modal Association. *Proceedings of the Eleventh ACM International Conference on Multimedia*, Berkeley, 2-8 November 2003, 604-611. https://doi.org/10.1145/957013.957143

[15] Liu, J., Xu, C. and Lu, H. (2010) Cross-Media Retrieval: State-of-the-Art and Open Issues. *International Journal of Multimedia Intelligence and Security*, **1**, 33-52. https://doi.org/10.1504/IJMIS.2010.035970

[16] Rasiwasia, N., Costa Pereira, J., Coviello, E., *et al*. (2010) A New Approach to Cross-Modal Multimedia Retrieval. *Proceedings of the* 18*th ACM international conference on Multimedia*, Firenze, 25-29 October 2010, 251-260. https://doi.org/10.1145/1873951.1873987

[17] Feng, F., Wang, X. and Li, R. (2014) Cross-Modal Retrieval with Correspondence Autoencoder. *Proceedings of the* 22*nd ACM International Conference on Multimedia*, Orlando, 3-7 November 2014, 7-16. https://doi.org/10.1145/2647868.2654902

[18] Wang, K., Yin, Q., Wang, W., *et al*. (2016) A Comprehensive Survey on Cross-Modal Retrieval.

[19] Hu, R., Xu, H., Rohrbach, M., *et al*. (2016) Natural Language Object Retrieval. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 4555-4564. https://doi.org/10.1109/CVPR.2016.493

[20] Kamper, H., Shakhnarovich, G. and Livescu, K. (2018) Semantic Speech Retrieval with a Visually Grounded Model of Untranscribed Speech. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Salt Lake City, 18-22 June 2018, 2514-2517.

[21] Mithun, N.C., Panda, R., Papalexakis, E.E., *et al*. (2018) Webly Supervised Joint Embedding for Cross-Modal Image-Text Retrieval. *Proceedings of the* 26*th ACM International Conference on Multimedia*, Seoul, 22-26 October 2018, 1856-1864. https://doi.org/10.1145/3240508.3240712

[22] Carreira, J. and Zisserman, A. (2017) Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 6299-6308. https://doi.org/10.1109/CVPR.2017.502

[23] Lin, T.Y., Maire, M., Belongie, S., *et al*. (2014) Microsoft COCO: Common Objects in Context. In: *European Confe-*

rence on Computer Vision, Springer, Cham, 740-755. https://doi.org/10.1007/978-3-319-10602-1_48

[24] Thompson, B. (2000) Canonical Correlation Analysis. In: Grimm, L.G. and Yarnold, P.R., Eds., *Reading and Understanding MORE Multivariate Statistics*, American Psychological Association, Washington DC, 285-316.

[25] Hwang, S.J. and Grauman, K. (2012) Learning the Relative Importance of Objects from Tagged Images for Retrieval and Cross-Modal Search. *International Journal of Computer Vision*, **100**, 134-153. https://doi.org/10.1007/s11263-011-0494-3

[26] Andrew, G., Arora, R., Bilmes, J., *et al*. (2013) Deep Canonical Correlation Analysis. *International Conference on Machine Learning*, Atlanta, 17-19 June 2013, 1247-1255.

[27] Prétet, L., Richard, G. and Peeters, G. (2020) Learning to Rank Music Tracks Using Triplet Loss. *ICASSP* 2020 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Barcelona, 4-8 May 2020, 511-515. https://doi.org/10.1109/ICASSP40776.2020.9053135

[28] Pons, J. and Serra, X. (2019) Musicnn: Pre-Trained Convolutional Neural Networks for Music Audio Tagging.

[29] Prétet, L., Richard, G. and Peeters, G. (2021) Cross-Modal Music-Video Recommendation: A Study of Design Choices. 2021 *International Joint Conference on Neural Networks* (*IJCNN*) *IEEE*, Shenzhen, 18-22 July 2021, 1-9. https://doi.org/10.1109/IJCNN52387.2021.9533662