# 基于音频和图像序列的物体撞击匹配

郝千越，胡开哲，刘雨玚

## 1 原理与方法

### 1.1 总体设计

本设计中给出若干多通道音频样本和图像序列样本，最终目标为根据音频样本和图像序列样本的特征将对应音频与图像序列匹配。

一个直观的想法是直接进行端到端的匹配，即将一段音频和一段图像序列输入某种网络而直接得到其是否能匹配（或匹配概率）。但由于音频与图像序列形状、数值分布差异较大，难以设计出适合的网络模型；同时，训练过程中配对样本（正样本）容易得到，而失配样本（负样本）的生成并非平凡的随意组合即可获得；另外，端到端模型可解释性差，难以进行针对性的优化，并且对于大量图像、音频进行训练的计算量也为端到端匹配的实现增加了困难。

结合上述分析，本设计不使用直接的端到端匹配，而是首先使用深度学习方法与传统方法结合分别对音频、图像序列进行特征提取，再使用二部图匹配算法进行音图匹配。观察数据集可知，音频与图像的特征主要包含两个方面，其一是类别特征，其二是运动特征。我们使用深度学习方法获取类别特征向量，使用传统方法获取运动特征向量，由向量内积等方法得到音频与图像序列之间的距离，从而实现二部图匹配。

### 1.2 音频特征提取

音频特征提取包括两方面，即分类特征与运动特征，其中分类特征即任务一的音频分类，同时两特征均用于后续的音图匹配。

分类问题时深度学习应用相当成熟的领域，因此考虑使用深度学习模型实现音频分类。由于时域波形数据形状窄而长，且存在大段的静音片段，缺乏明显的分类特征，不适合卷积神经网络的输入；而 LSTM 等循环神经网络适用于“Sequence to Sequence”的任务，难以应用于音频分类。因此考虑将时域音频变换到频域进行分类。具体变换过程包括：

- 首先将 44000Hz 采样的音频序列降采样到 11000Hz，同时由于音频序列首尾均存在一段静音片段，降采样截去首尾一定长度；
- 对各通道依次使用窗长为 510 点，窗移动步长为 128 点的 Hanning 窗对序列做短时傅里叶变换（STFT），对应 4 个通道获得$4 \times 256 \times 256$的频谱特征图（少量长度不符合$4 \times 44000$的音频序列经过 STFT 后频谱图为$4 \times 256 \times M, M \neq 256$则通过截取或填充 0 的办法将频谱特征图处理为$4 \times 256 \times 256$形状）；
- 观察到 STFT 后数值均为极小的数字，因此取 10 为底对数将其数值进行压扩，再以$\pm 5$为门限进行整流；
- 最后将其数值仿射变换为$[0,255)$之间的数值，同时近似为 uint8 类型，以便于图像数据格式一致。

经过上述变换，每段音频转换为 4 通道图像，取部分类音频别前三个通道为 RGB 可视化结果如下：

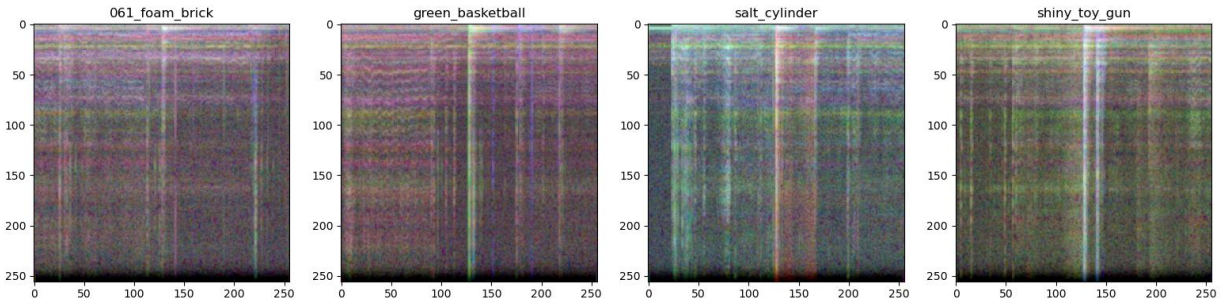


图 1 部分类音频别前三个通道可视化

可以看出，不同物体碰撞的谱线差异较大（图像颜色差异是由于不同通道发生碰撞导致的），因此可以直接使用计算机视觉中成熟的分类模型进行分类。

上述音频分类流程对于给定音频，给出长度为 10 的特征向量，分别对应该音频属于 10 个类别的概率，作为音频的分类特征。对音频进行运动特征提取时首先观察音频波形，某样本其中两个通道的波形如下：

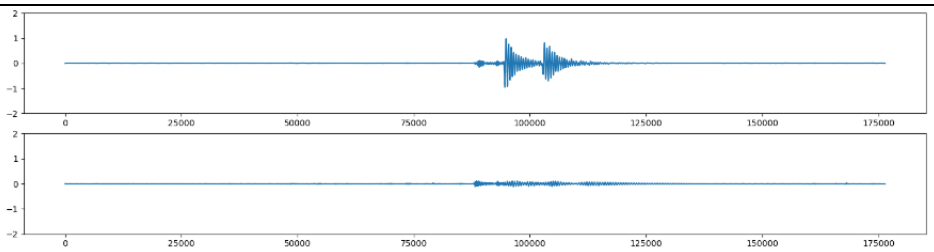

图 2 某样本两通道时域波形

观察发现，时域波形清晰度良好且基本没有噪声，有明显的碰撞特征，即被撞击一侧的麦克风记录到明显大幅度的声音信号。因此，采用平凡的传统方法进行音频的运动特征提取，即设置门限值，某个同道最大波形幅度大于该门限值则认为该通道受到撞击，由此得到 4 维运动特征向量，其每个维度均为 0 或 1，表示该通道对应的侧壁是否受到过撞击。门限值为该方法中的超参数，其具体确定方法见 2.2 中讨论。

### 1.3 图像特征提取

图像提取包括两方面，即类别特征与运动特征，两类特征均用于后续的音图匹配。

图像类别特征基于 CNN 的深度网络实现，具体地，首先使用 torchvision 库中的 transform 函数将图像重采样到(224, 224)像素，并对其进行归一化，再送入相应网络进行分类，最终将网络 10 维分类输出通过 softmax 函数得到最终各类别的置信度。实践中发现基于 AlexNet 的网络已经在手动划分(20%)的验证集上得到了较高的准确率(99%+)，因此基于简化参数、提升运算速度的考虑没有尝试如 ResNet 等更加复杂的网络结构。

对于图像运送特征的提取，虽有诸如 C3D、RC3D 等用于运动特征提取的网络结构，但其提取的特征难以与基于音频提取的运动特征进行直接匹配。考虑到算法的可解释性及训练复杂度等因素，我们并没有训练第三个网络用于输入音视频特征的匹配程度，而是基于图像掩膜数据手工提取运动信息。具体地，对于序列的掩膜数据，首先利用全部掩膜信息的叠加(即运动轨迹的掩膜)与初始位置的交并比判断其是否发生运动，对于发生运动的序列，利用 cv2 库提取其最大连通分量以去除掩膜中误识别的零星数据点，再逐帧提取物体的中心和边界位置。结合基于中心位置得到相对运动方向及基于边界位置得到的与各边接触信息综合判断是否与某一边发生碰撞，最终将结果合并为 np.array 返回.

### 1.4 音频、图像匹配

经过上述过程，记音频总数为$M$，图像序列总数为$N$，即可得到第$i$个音频、第$j$个图像序列 10 维向量的分类特征($i = \{0,1,\dots M-1\}, j = \{0,1,\dots N-1\}$)

$$a(i)_c = [a(i)_c^0, a(i)_c^1, \dots a(i)_c^9], a(i)_c^k \in [0,1], k = \{0,1,\dots 9\}$$
$$v(j)_c = [v(j)_c^0, v(j)_c^1, \dots v(j)_c^9], v(j)_c^k \in [0,1], k = \{0,1,\dots 9\}$$

及其 4 维向量的运动特征

$$a(i)_m = [a(i)_m^0, a(i)_m^1, \dots a(i)_m^3], a(i)_m^k \in \{0,1\}, k = \{1,2,3,4\}$$
$$v(j)_m = [v(j)_m^0, v(j)_m^1, \dots v(j)_m^3], v(j)_m^k \in \{0,1\}, k = \{1,2,3,4\}$$

由此定义第$i$个音频与第$j$个图像序列之间的匹配程度

$$M_{ij} = \beta a(i)_c \cdot v(j)_c + 0.25 \sum_{k=0}^{4} a(i)_m^k \, xnor \, v(j)_m^k$$

可以给类别特征相似度加不同权重，即改变$\beta$的值，实际中取$\beta = 1$。由此得到音频与图像序列的两两相似程度，考虑二部图模型一侧节点为音频，另一侧节点为图像序列，上述$\{M_{ij}\}$即为连接两侧点的边权，使用二部图匹配算法即可完成匹配任务。

对于任务二中完全匹配，即有$M = N$，设置匹配阈值$threshold \le 0$即可全部匹配；对于任务三中不完全匹配，即不保证有$M = N$，设置匹配阈值$threshold > 0$，匹配程度小于$threshold$的配对被放弃，对应音频与图像序列被标记为孤立样本。

## 2 结果与分析

### 2.1 task1：音频分类

本设计中使用 ResNet-18 作为频谱图的分类器，训练过程中划分 10%的样本作为验证集，经过 300 个 epoch 的训练后，训练集上的准确率接近 100%，验证集上的准确率达到 97.8%，音频分类效果良好。

## 2.2 特征提取

用任务一中得到的音频分类模型作为音频分类器，分类特征提取效果良好。使用 AlexNet 作为图像分类器，划分20%的样本作为验证集，经过训练，训练集、验证集上的分类准确率均在 99%以上，可以见图像分类特征提取效果良好。

下面讨论运动特征提取的效果。理想运动特征应当具有如下特点：配对样本之间运动特征相似度高，不同样本之间差异程度大。由此定义相似度$P_s$和区分度$P_d$如下（$N_t$为训练集中样本数量）：

$$P_s = \frac{1}{4N_t}\sum_{i=0}^{n}\sum_{k=0}^{4} a(i)_m^k \; xnor \; v(i)_m^k$$

$$P_d = 1 - \frac{1}{4N_t(N_t-1)}\sum_{\substack{i=0,j=0\\ i\neq j}}^{n}\sum_{k=0}^{4} a(i)_m^k \; xnor \; v(j)_m^k$$

两者均为[0,1]间的变量，且数值越大代表运动特征提取越理想。使用前述算法分别提取音频、图像序列运动特征，搜索音频特征提取中不同的阈值，结果如下。可见相似度指标存在一个最优峰值点，而区分度指标随着阈值增大呈单调下降，综合考虑两项指标，最终使用 0.35 作为音频提取阈值。

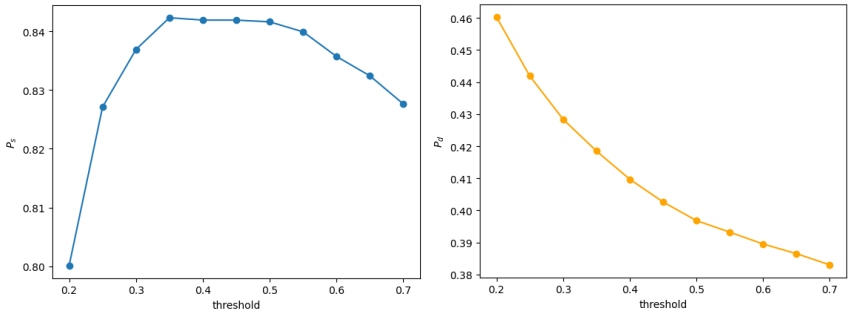


图 4 音频特征提取阈值搜索结果

下面展示某样本运动特征提取结果，左图为其音频波形图，右图为其图像序列图。从图像序列中可以看到该样本撞击了上侧壁，同时在音频对应通道也可以看到撞击留下的音频波形。图像运动特征提取算法和音频运动特征提取算法都正确输出了特征向量[1,0,0,0]，表示上侧壁发生撞击，可见运动特征提取效果良好。

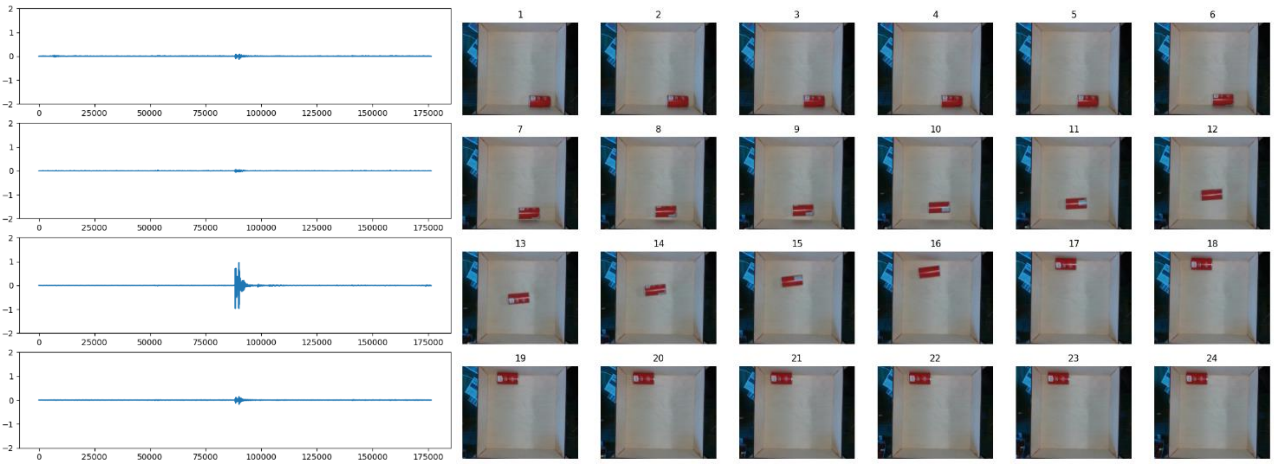


图 5 某样本音频波形与图像序列

## 2.3 task2&3：音图匹配

按照前述算法完成任务二中的完全匹配。随机从 train 中抽取 50 对样本作为验证，验证集准确率为 70%左右（由于抽取样本不同而变化，最好情况达到 92%，最差情况 54%，平均约 70%）。完成任务三时需要首先确定不完全匹配的$threshold$设置，先将任务三考虑为完全匹配，观察全部配对之间的匹配程度，其分布如下图。可以看出大部分配对样本匹配程度在 1.5 以上，结合前述匹配度度量方式，1.5 以上可以看做较好的匹配，由此确定$threshold$设置为 1.5 较为合理，完成任务三不完全匹配。

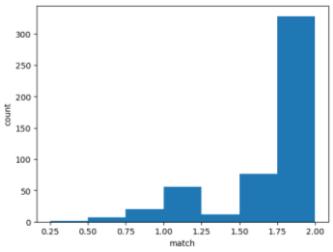


图 6 配对相似程度分布

# 3 问题与讨论

## 3.1 深度学习方法应用于音频运动特征提取

在本设计最终方案中使用了较为简单的阈值检测方法判断音频运动特征，在前期探索尝试中，我们也尝试了深度学习方法提取音频运动特征。具体操作为：首先使用前述算法获取运动特征，一次以此作为标签进行音频运动特征提取网络的训练。音频运动特征提取网络以与音频分类网络相同的 STFT 得到的频谱图为输入，输出为 4 个二分类单元，即判断四壁是否受到撞击。

经过训练，发现该方法计算量远大于最终采用的阈值检测方法，但特征提取效果较差（仅获得 75%左右的相似度 $P_s$值），分析原因如下：图像预处理中使用 STFT 变换到频域的时间窗操作损失了时域精度而获取了频域特征。然而，检测撞击事件是否发生并不需要频域特征，因此网络接收到的冗余信息过多，难以有效训练。从中可以看出，并非所有问题都可以通过端到端的学习得到解决，一些传统的方法仍有应用价值。

## 3.2 运动特征中的时序性

本设计最终使用的运动特征仅包含是否撞击，而不描述撞击四壁的顺序。在探索过程中，我们尝试了加入撞击时序的描述，但最后未采用该方案，原因分析如下。在匹配任务中，类别特征即可极大缩小匹配范围，而同类样本中，撞击四壁的情况差异较大，同时仅有少部分样本存在侧壁反弹和多次撞击的情况。因此，只通过是否撞击即可有效区分大多数样本，不必引入撞击顺序。如果引入撞击顺序，在少部分混淆情况下可能产生积极作用，但更复杂的算法引起的精确度损失反而不利于在全局样本上的表现。

## 3.3 更精细图像运动特征的提取

本设计最终使用的运动特征仅包含四个维度，即上下左右四壁是否受到撞击。在探索过程中，我们尝试了更加精细的运动特征，如将每个侧壁细分为两段或三段，获取更高维度的运动特征。这样的运动特征在图像中是相对容易获取的，但从音频中区分侧壁上受到撞击的方位涉及到多个麦克风协同等问题，加之不同物体撞击特性不同，难以准确区分。同时，更高维度的运动特征中大部分元素为 0，这增加了非配对音图特征之间的相似性，从而降低了区分度。因此，最终选择 4 维运动特征，足以完成匹配任务。

**附录 1 成员分工**

| 姓名 | 班级 | 学号 | 分工 |
|---|---|---|---|
| 郝千越 | 无 85 | 2018011153 | 音频分类、音频特征提取、联合测试、报告撰写 |
| 胡开哲 | 无 84 | 2018013326 | 图像分类、图像特征提取、报告撰写 |
| 刘雨玚 | 无 81 | 2018010651 | 音频与图像序列匹配、报告撰写 |

**附录 2 文件清单**

| 文件名 | 说明 | 文件名 | 说明 |
|---|---|---|---|
| 设计报告.pdf | 本文件 | video_process.py | 图像特征提取 |
| test.py | 指定测试接口 | model/resnet18.pth | 音频分类网络模型 |
| ResNet.py | 音频分类网络定义 | model/alexnet.pt | 图像分类网络模型 |
| alexnet.py | 图像分类网络定义 | pairing.py | 匹配算法 |
| utils.py | 图像特征提取依赖文件 | requirements.txt | 环境说明文件 |
| *********以下为模型构建所需代码，不被测试接口调用********* | | | |
| audio_class_DataSet.py | 音频文件自定义数据集 | audio_process.py | 音频预处理 |
| audio_class_train.py | 音频分类模型训练 | audio_move.py | 搜索音频运动检测阈值 |