清華大學
Tsinghua University

基于视听信息的音源分离与定位

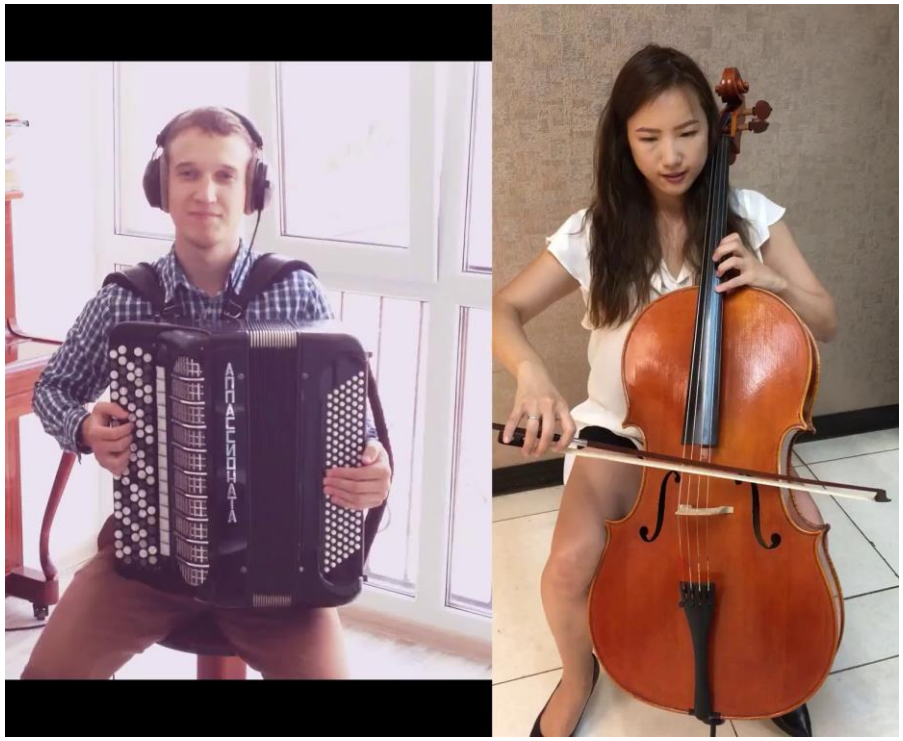
——2018-2019学年度秋季学期《视听信息系统导论》课程设计

杨浩 于诚

2018年11月2日

从独奏到合奏

- 听觉信息：幅值相加，时间并联
- 视觉信息：空间组合，时间并联



从合奏到独奏

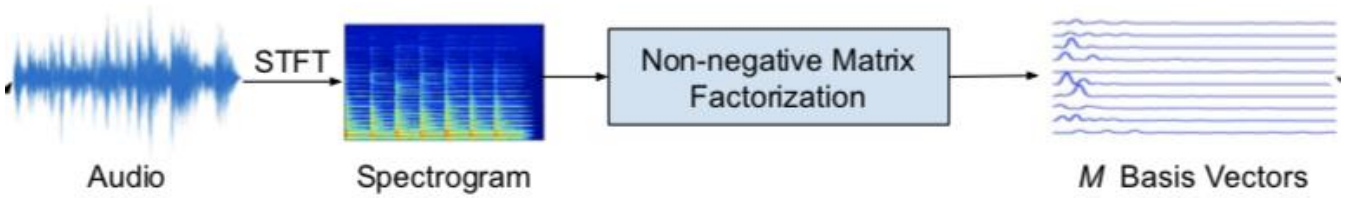
- 听觉信息：音源分离（audio source separation）
- 视觉信息：音源定位（sound localization）

基于视听信息的音源分离与定位

- **课题描述**
- 给定若干乐器合奏视频，包括自然场景下的乐器合奏视频，以及通过乐器独奏视频拼合而成的视频作为测试；同时提供若干乐器独奏与乐器合奏视频作为训练。
- 同学需使用机器学习方法，利用提供的训练视频，设计出视觉信息和听觉信息的单独或联合特征表达，在视频和音频层面获取音源信息，并实现对测试集中的乐器合奏视频的音源分离与定位。

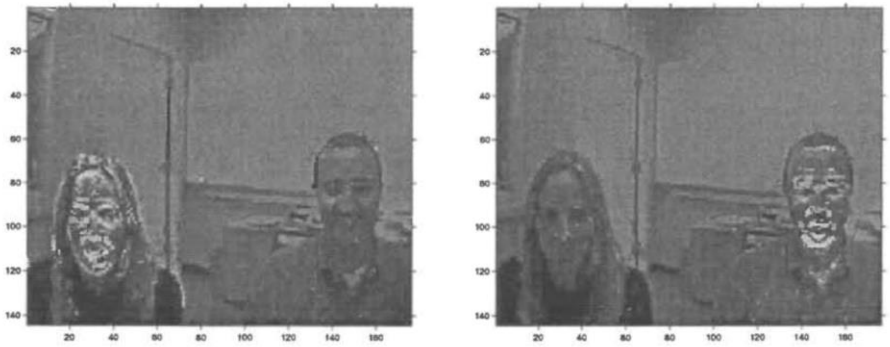
相关工作参考

- **基于无监督聚类的音源分离（纯音频输入）**
- 基本思路：音频特征提取（分解）+无监督聚类
- 常用方法：鲁棒主成分分析（RPCA）、独立成分分析（ICA）、非负矩阵分解（NMF)等
- 音源分离框架举例：STFT+NMF+MFCC+kMeans聚类
- 主要缺点：完全无监督，没有利用视觉信息



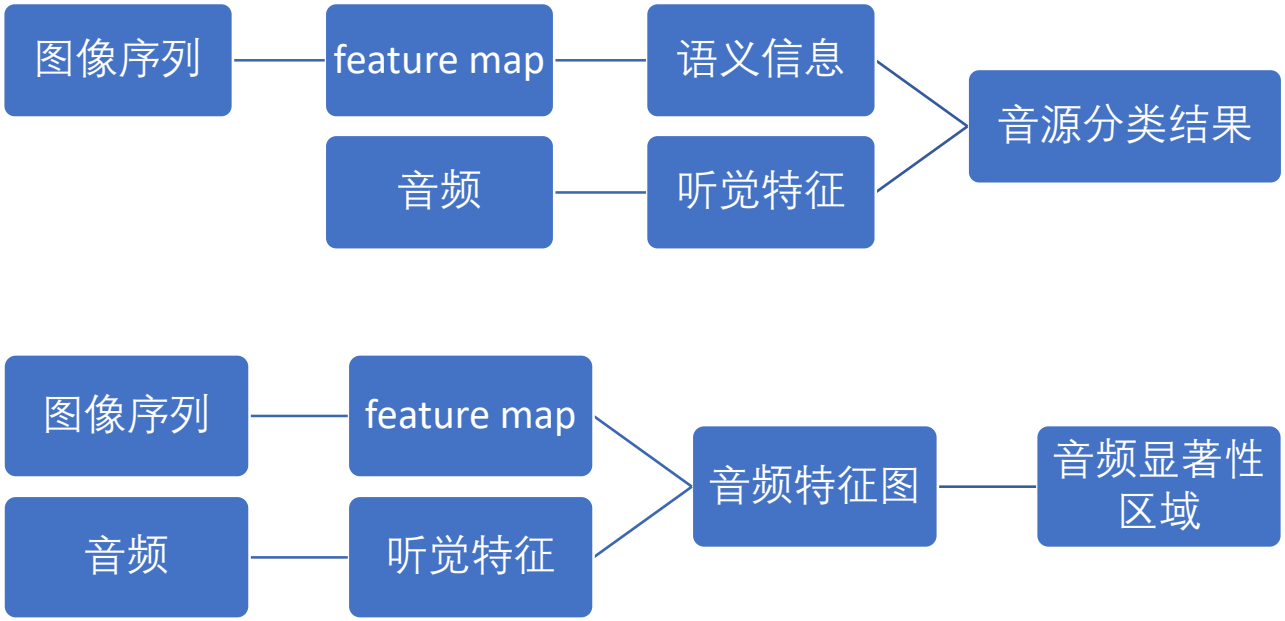
相关工作参考

- **基于音源运动信息的音源定位**
- 基本思路：将视频的图像序列进行空间分解，计算区域性像素序列与音频的某种“相关”
- 常用方法：像素点时间序列概率建模+互信息，典型相关分析（canonical correlation analysis）等
- 主要缺点：运动假设存在局限性，而且没有利用视频图像的语义信息



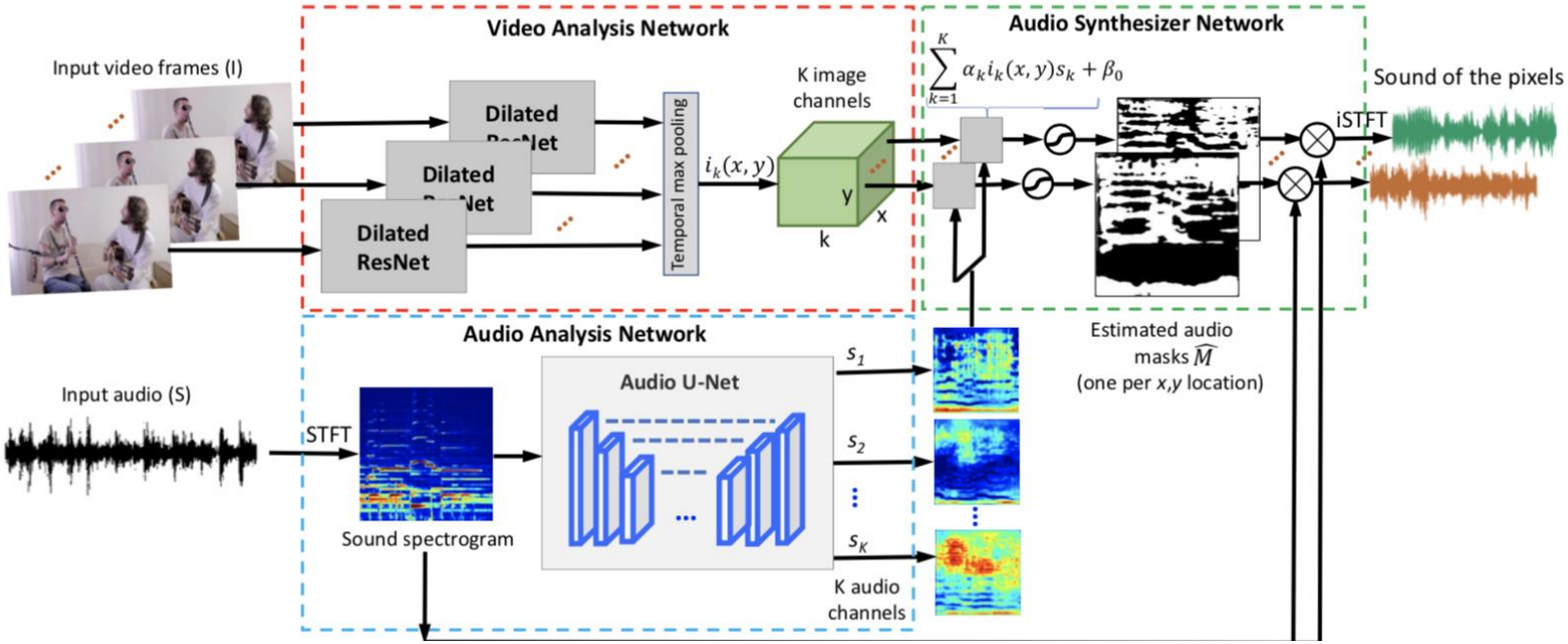
基于视听特征表达的参考框架

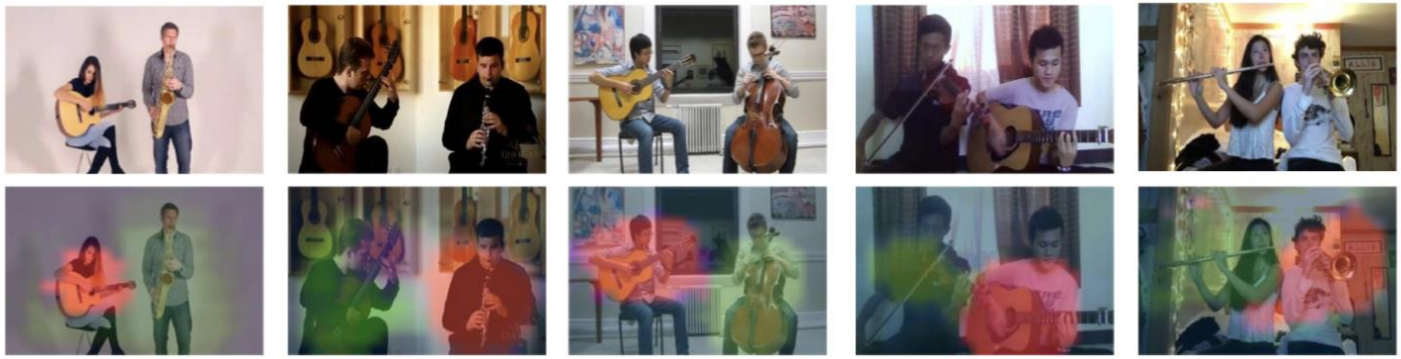
- 基本思路：利用图像语义信息指导音源分离，利用图像特征提取的空间信息计算显著性区域实现定位



清華大學 Tsinghua University

基于视听特征表达的参考框架





简化版本的参考框架

- **简化内容:**
- 1）数据集仅包括ImageNet中的乐器类别，可以直接用类别信息替代图像的语义信息；
- 2）将测试数据统一为左右部分的两种乐器的合奏视频，音频定位只需标注左右即可
- **简化后框架:**
- 1）图像特征提取得到乐器类别和显著性区域；
- 2）音频特征提取并分解；
- 3）用图像类别指导音频分类，得到两类乐器的音频，并与1）得到的显著性区域进行匹配实现音源定位

基于视听信息的音源分离与定位

- **训练数据**
- 本课题提供8类乐器的若干段独奏以及上述乐器产生的7种二重奏组合的若干段合奏：其中每组数据包括：
  - images：以每秒2.4帧提取的.jpg图像，位于images/label1/label2/*（视频名）文件夹内；
  - *.mp4：时长不定，帧率为24帧/秒，音频采样率为44100Hz，声道数为1的有声视频，位于videos/label1/label2文件夹内；
  - *.wav：音频采样率为44100Hz，声道数为1的音频，位于audios/label1/label2文件夹内；
  - 注：label1为solo/duet，label2为视频对应的乐器种类
- 全部训练数据大小约为60G，同学可从清华大学云盘下载。下载地址和使用管理规定等说明请参见网络学堂通知。

基于视听信息的音源分离与定位

- **测试数据**
- 基础部分的测试数据为30组由独奏视频人工合成的二重奏视频，并公开其中的25组：每组数据包括：
  - images：以每秒2.4帧提取的.jpg图像，位于testimage/*（视频名）文件夹内；
  - *.mp4：时长不定，帧率为24帧/秒，音频采样率为44100Hz，声道数为1的有声视频，位于testvideo文件夹内；
  - *.wav：音频采样率为44100Hz，声道数为1的音频，位于gt_audio文件夹内；
- 提高部分的测试数据为7组真实场景下的二重奏视频，全部公开，数据格式除无独奏音频外，与基础部分相同。

基于视听信息的音源分离与定位

- 相关文件
  - 基于matlab
    - demo.m， 运行框架，生成音频以及位置信息（mat文件）
    - evaluate.m 评价函数，比对mat文件中的位置信息并读取音频计算相关指标
    - matlab文件夹， 运行框架用到的所有函数
  - 基于python
    - demo.py 运行框架，生成音频及位置信息（json文件）
    - evaluate.py 评价函数，比对json文件中位置信息并读取音频计算相关指标
    - datahelper 文件夹，存放用到的各种库/包及函数
  - 其他用到的文件会在项目文档当中详细说明

基于视听信息的音源分离与定位

- **任务要求**
- 要求同学利用训练数据集，运用提供的特征提取方法或其它改进方法，学习视频的视觉和听觉特征；
- 可以按照上述几种参考框架的步骤设计并实现算法，也可以自行设计或实现其它框架；
- 同学可以选取python或matlab框架，填充音源分解和定位函数，完成demo.py或demo.m，最后利用evaluate.py或evaluate.m评估算法的性能。如有使用深度学习框架的需要，建议同学使用PyTorch或Tensorflow训练网络模型。

基于视听信息的音源分离与定位

- **接口描述（完成音源分解与定位函数）**
  - audio_decomp（音源分解函数）：
    - Input：video、audio（图像序列和音频的numpy数组或matlab矩阵）、OutputPath、file（视频名）
    - Output：返回值按照要求填写，主要是生成音频的名字以及音频数据，除此以外，需将两段分解音频分别写入file+’_seg1.wav’与file+’_seg2.wav’文件且放入对应文件夹；
  - audio_video_sim2（音源定位函数）：
    - Input：audio1、audio2、video（图像序列和音频的numpy数组或matlab矩阵，分别代表待定位的两段音频和图像序列）、file（视频名）
    - Output：音频和位置对应关系，位置使用[0,1]表示（代表audio1为视频的左半部分，audio2为视频的右半部分）或[1,0]（相反情况），注意文件名和位置的对应关系，matlab版本的同学输出文件名顺序和位置顺序对应
  - 其他：如果现有帧率图片数量无法满足算法需求，可以直接读视频，并修改帧率参数，报告中说明且提交相关code

基于视听信息的音源分离与定位

- **结果评价指标**
- 音源分解测试：
  - SDR：根据分解的音频和实际音频比对计算，指标与信噪比类似，单位为dB；
  - 音源分解测试不受分解的两段音频的顺序所影响，评测时自动将分解结果与实际音频进行比对；
- 音源定位测试：
  - 用音源分解得到的音频进行测试：结果音频与真实音频的匹配方式取决于音源分解测试的比对结果，result1=mean_acc1；
  - 用实际音频进行测试：result2=mean_acc2*60%;
  - 最终的测试结果为result=max(result1,result2).

- **设计分组**
- 本课程设计以小组为单位进行，自由分组，每小组成员不超过2人。
- 每小组应独立完成课程设计。

- **设计报告**
- 撰写设计报告，篇幅不超过4页（A4纸），应至少包含如下内容:
  - 课程设计团队成员及分工情况。小组成员评分可能因分工及工作量产生一定差异。
  - 提交文件清单。
  - 工作开展与完成情况（原理、实现、性能分析、问题不足）。

基于视听信息的音源分离与定位

- **提交清单**
- 需提交命名为“提交同学学号_提交同学姓名.rar/zip”的压缩文件。压缩文件应当至少包括如下内容:
  - report.pdf/doc/docx 课程设计报告;
  - 所有code文件，保证环境没有问题时可运行（重要）
  - 生成的音频文件（放到result_audio文件夹，并按照指定方式命名）
  - 如果使用matlab，包含位置信息的mat文件，具体内容见项目文档
  - 如果使用python，包含位置信息的json文件，具体内容见项目文档
  - 如果使用用到神经网络模型且模型较大（超过200M)，最好上传网盘提供下载链接，并说明如何使用
  - 如果有其他程序运行需要注意的地方，请另外说明，写入Readme文件当中

- **提交日期**
- 每个小组应当仅由1名小组成员在2018年12月31日24时前通过网络学堂提交课程设计文件。
- 助教将对2018年11月30日24时前通过网络学堂提交的课程设计文件进行中期评定，并反馈评定结果；助教将在截止时间后对提交的全部课程设计文件进行最终评定。
- 经历两次评定的小组，最终成绩取两次较高的评定成绩；仅经历最终评定的小组，最终成绩取最终评定的评定成绩。

基于视听信息的音源分离与定位

- **评定方法**
- 1) 课程设计报告，30%。
- 2) 设计需求完成情况（基础部分），70%。综合音源分解与定位的测试结果以及算法复杂度进行计算。
  - 音源分解与定位测试，60%。二者各占30%，具体测试方法同“测试指标”。测试数据为基础部分的30段视频。若与小组成员沟通后确认代码运行结果与公开视频所提交的结果不匹配，则分数乘以60%。
  - 视频匹配复杂度，10%。根据算法的时间开销评定。
- 3）设计需求完成情况（提高部分），10%。测试采取主观测试，取效果最好的5组进行加分。
- 如有设计文件延期提交，设计报告、程序实现中存在抄袭行为等，将根据情节程度，扣除课程设计的部分或全部分数。

清華大學
Tsinghua University

谢谢!

杨浩 于诚

2018年11月2日