ISMIR

ISMIR 2019 Tutorial [T3]


# Audiovisual Music Processing

Zhiyao Duan*[1], Slim Essid*[2], Bochen Li*[1], Sanjeel Parekh*[2]

[1] Department of Electrical and Computer Engineering, University of Rochester, NY, USA
[2] Department of Images, Data and Signals, Télécom ParisTech, Paris, France

(* authors in alphabetical order)



Delft, The Netherlands

# Motivation

- Music is a multi-modal art form

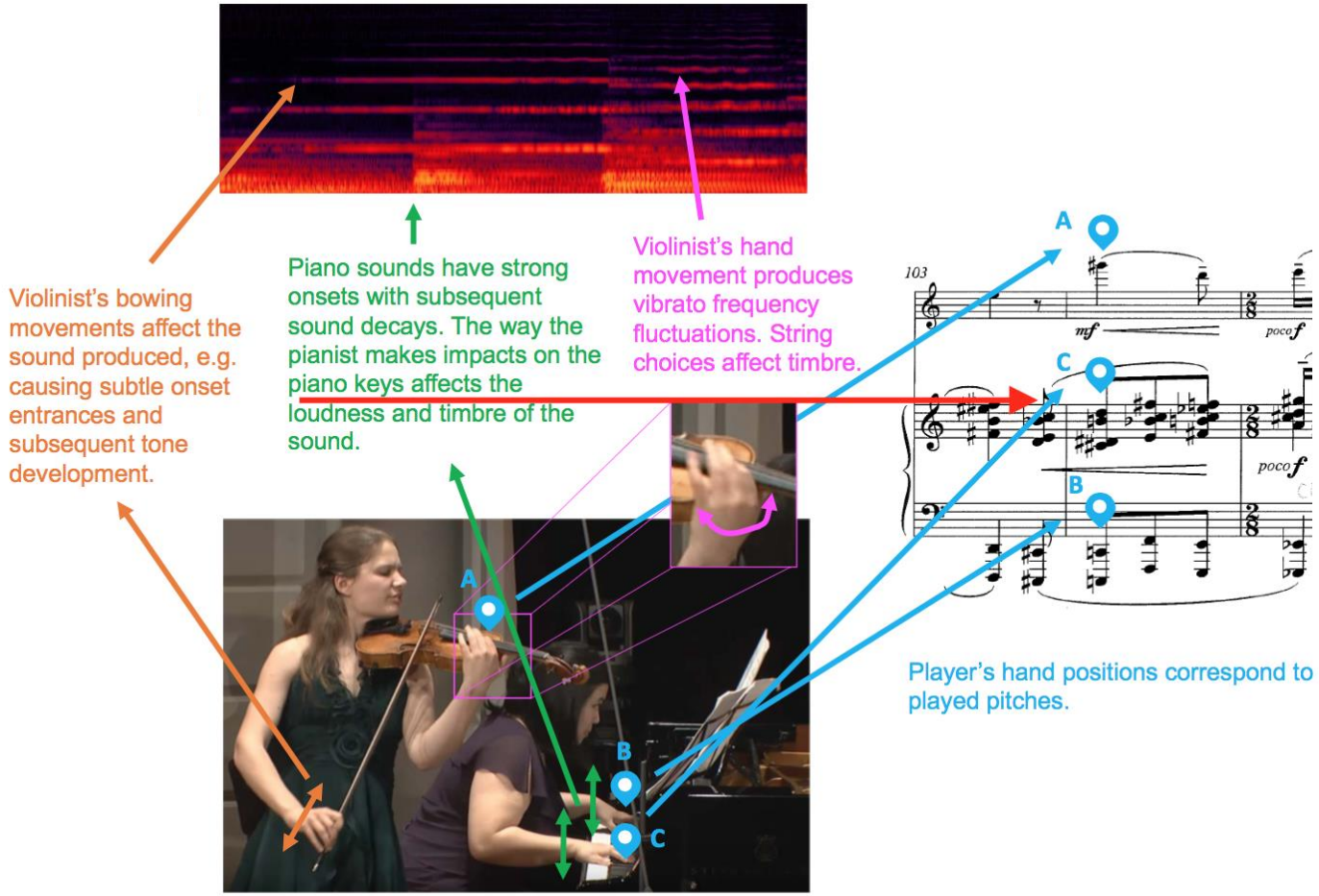
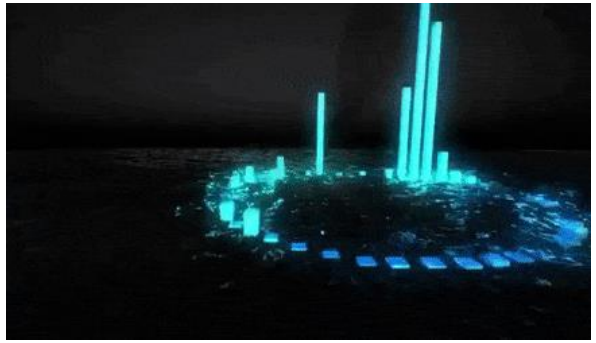



Figure from (Duan et al., 2018)

# Motivation

- Audiovisual expressions of ideas and emotions



# Motivation

- Musicians use audiovisual cues to coordinate with each other

# Motivation

- Audiences enjoy audiovisual expressions

Divje Babe Flute
(43,400 – 67,000 yrs ago)

Phonograph
(Edison, 1877)

Vinyl record
(Victor, 1931)

Compact Disc
(Phillips & Sony, 1982)

Spotify®
(2006)

MP3 (1997)

Cassette tape
(Phillips, 1962)

The audience could only enjoy live music; they would watch instead of just listen.

# Motivation

- Audiences enjoy audiovisual expressions
  - Music video streaming services become popular





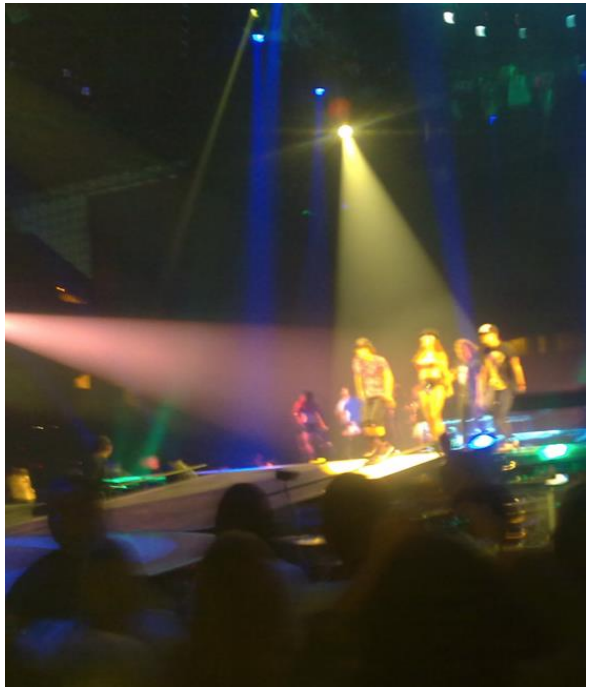

- Visual aspect is an important factor in the communication of meanings (Platz & Kopiez, 2012)
- Sight over sound in the judgement of music performance (Tsay, 2013)

# Intellectual Merit

- Broadens the scope of music signal processing research
- Connects with other areas such as image processing, computer vision and multimedia
- Creates a path towards emerging areas such as music VR/AR
- Serves as a controlled testbed for multimodal data analysis

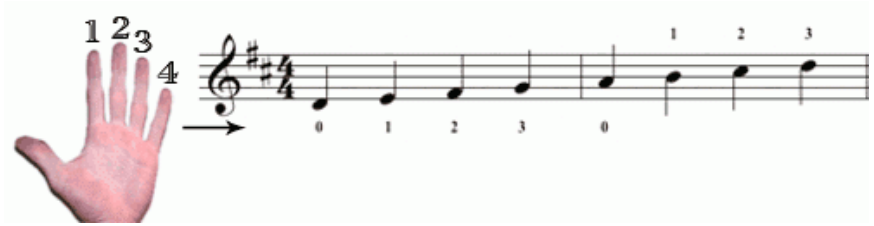
# Applications

- Concerts
  - Automatic camera/light/sound control
  - Augmented concerts with visual displays
  - Virtual concerts




# Applications

- Music Education
  - Pose analysis and feedback
  - Automatic visual demonstration
  - Automatic fingering annotation

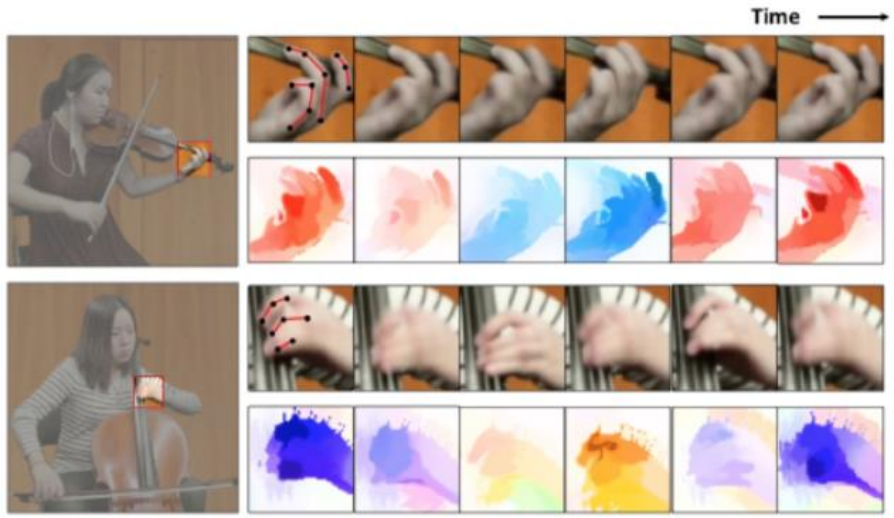


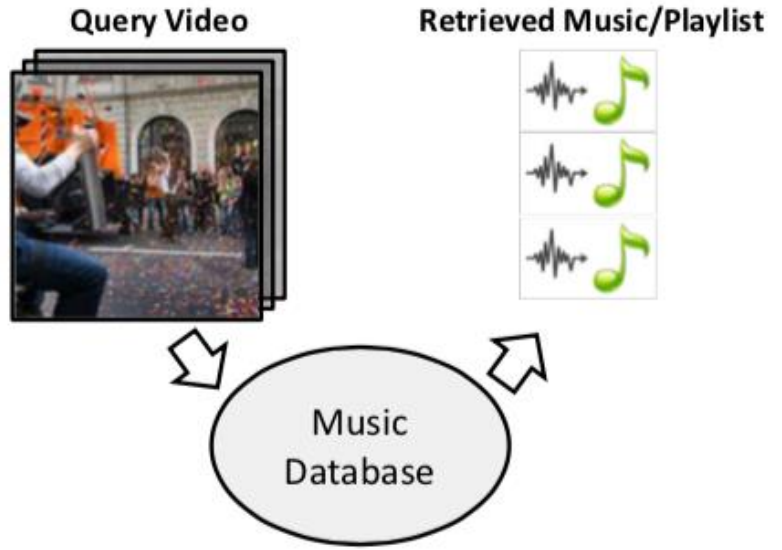

# Applications

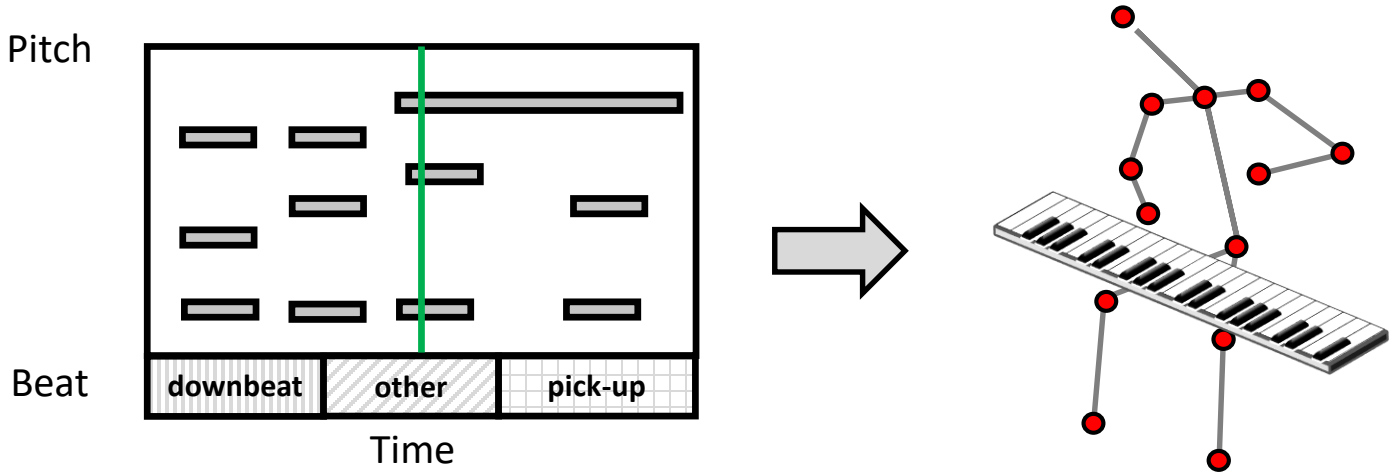
- Music Interaction
  - Human-computer collaborative music making
  - Visually informed automatic accompaniment

# Problem Categorization

- Analysis
  - Association
  - Transcription
  - Separation
- Classification & Retrieval
  - Genre classification
  - Mood recognition
  - Audiovisual matching
  - Recommendation
- Generation
  - Audio/music generation
  - Visual generation



# Tutorial Outline

- **Introduction**
- Audiovisual Music Performance Analysis
  - Overview of Analysis Tasks
  - Audiovisual Co-Factorization for Source Separation
  - Hands-on Case Study #1: Motion Informed Audio Source Separation
- Audiovisual Content Based Classification and Retrieval
  - Genre Classification
  - Emotion Analysis
  - Cross-Modal Retrieval
  - Instrument Classification
- Audiovisual Music Generation
  - Hands-on Case Study #2: Skeleton Plays the Piano
- Datasets, Tools and Other Resources
- Challenges, Opportunities and Conclusions

# Tutorial Goals

- Stimulate interests in audiovisual research in MIR
- Present a comprehensive overview of existing research
  - Research topics
  - Datasets and tools
  - Bibliography
- Provide a taste hands-on experience in two tasks
- Collect ideas and inspirations from YOU, the audience

Please interrupt us whenever you have a question or comment!

# Tutorial Outline

- Introduction
- **Audiovisual Music Performance Analysis**
  - Overview of Analysis Tasks
  - Audiovisual Co-Factorization for Source Separation
  - Hands-on Case Study #1: Motion Informed Audio Source Separation
- Audiovisual Content Based Classification and Retrieval
  - Genre Classification
  - Emotion Analysis
  - Cross-Modal Retrieval
  - Instrument Classification
- Audiovisual Music Generation
  - Hands-on Case Study #2: Skeleton Plays the Piano
- Datasets, Tools and Other Resources
- Challenges, Opportunities and Conclusions

# Task Categorization

Table adapted from (Duan et al., 2019)

| Visual | Is Critical | | Is Significant | | | | |
|---|---|---|---|---|---|---|---|
| **Tasks** | **Fingering** | **Association** | **Play/Nonplay** | **Onset** | **Vibrato** | **Transcription** | **Separation** |
| Percussion | N/A | - | PP | - | N/A | PT | - |
| Piano | PF | - | - | - | N/A | - | - |
| Guitar | GF | - | - | - | - | GT | GS |
| Strings | SF | SA | SP | SO | SV | ST | SS |
| Wind | - | WA | WP | WO | - | - | WS |
| Singing | N/A | - | - | - | - | - | - |

**PF:** (Gorodnichy & Yogeswaran, 2006)
(Oka & Hashimoto, 2013)
**GF:** (Burns & Wanderley, 2006)
(Kerdvibulvech & Saito, 2007)
(Scarr & Green, 2010)
(Paleari et al., 2008)
**SF:** (Zhang & Wang, 2009)

**SA:** (Li et al., 2017a)
(Li et al., 2017b)
(Li et al., 2019)
**WA:** (Li et al., 2019)

**PP:** (Bazzica et al., 2016)
**SP:** (Bazzica et al., 2016)
(Dinesh et al., 2017)
**WP:** (Bazzica et al., 2016)

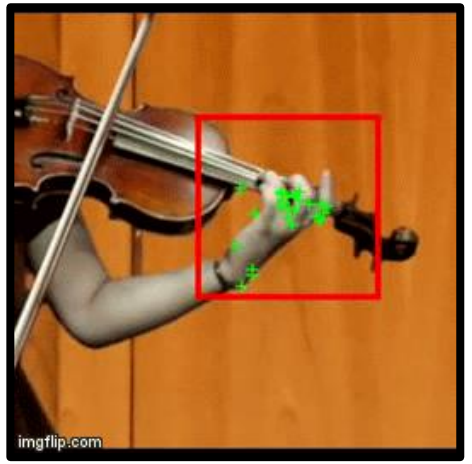
**SO:** (Li et al., 2017b)
**WO:** (Bazzica et al., 2017)
**SV:** (Li et al, 2017c)

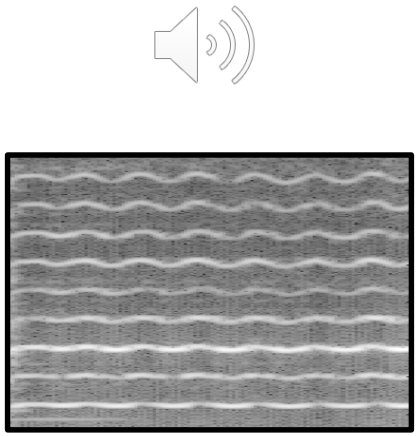
**PT:** (McGuinness et al., 2007)
**GT:** (Paleari et al., 2008)
**ST:** (Zhang & Wang, 2009)
(Dinesh, et al., 2017)
**GS/WS:** (Zhao et al., 2018)
(Zhao et al., 2019)
**SS:** (Parek et al., 2017)
(Zhao et al., 2018)
(Zhao et al., 2019)

# Audiovisual Correspondence

- Static
  - Fixed image ←→ Audio frame
  - E.g., Posture of a flutist ←→ Play/Nonplay activity
  - E.g., Piano fingering ←→ Music transcription

- Dynamic, instrument specific
  - Dynamic movement ←→ Audio feature fluctuation
  - E.g., Guitarist's strumming hand ←→ Rhythmic pattern
  - E.g., Violinist rolling left hand ←→ Vibrato

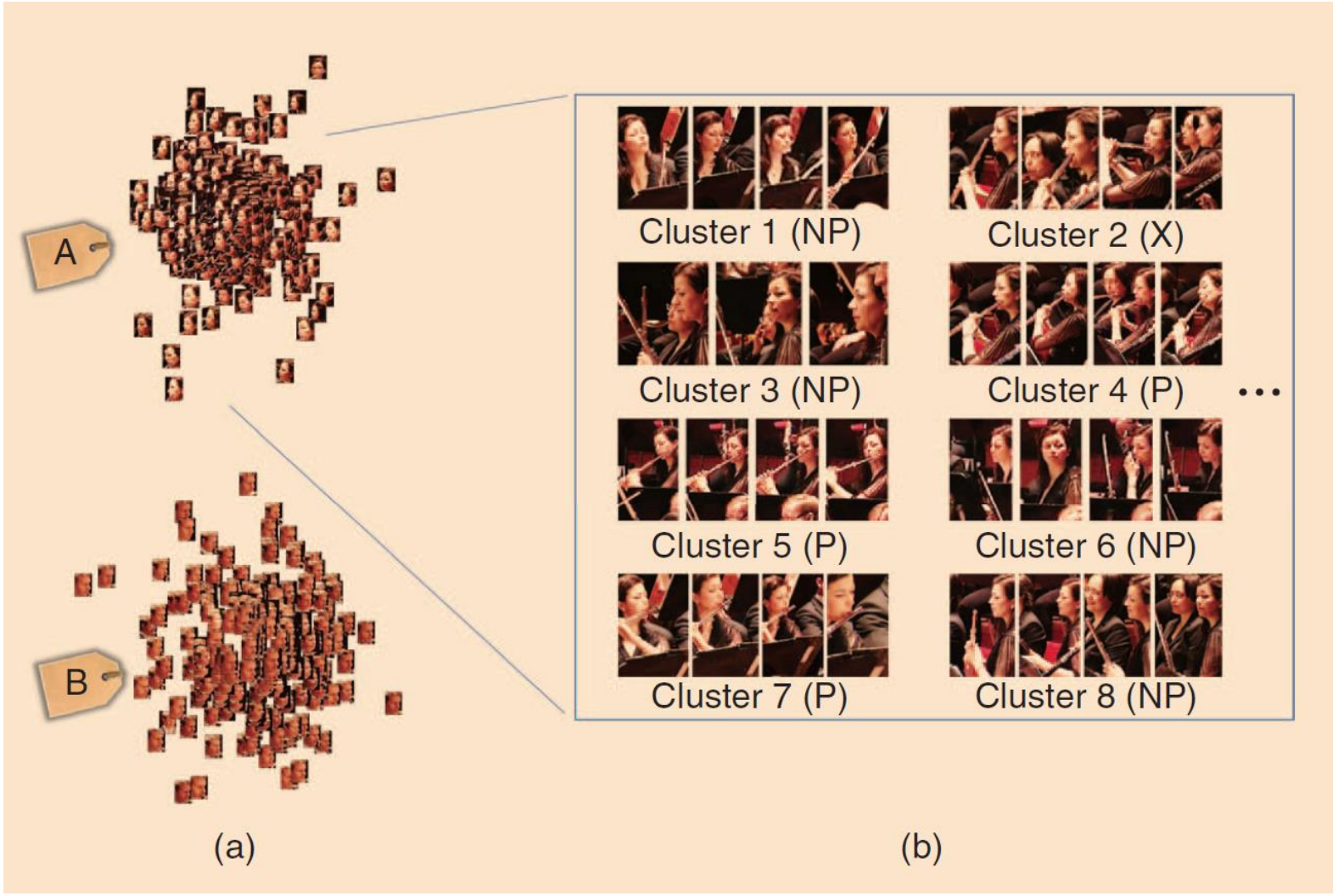


- Dynamic, general
  - Co-factorization of audio/visual fluctuations
  - Learning audiovisual embeddings

# Static Audiovisual Correspondence

- Play/Nonplay Activity Detection
  - Hierarchical clustering
    - (a) cluster musicians
    - (b) cluster P/NP activities

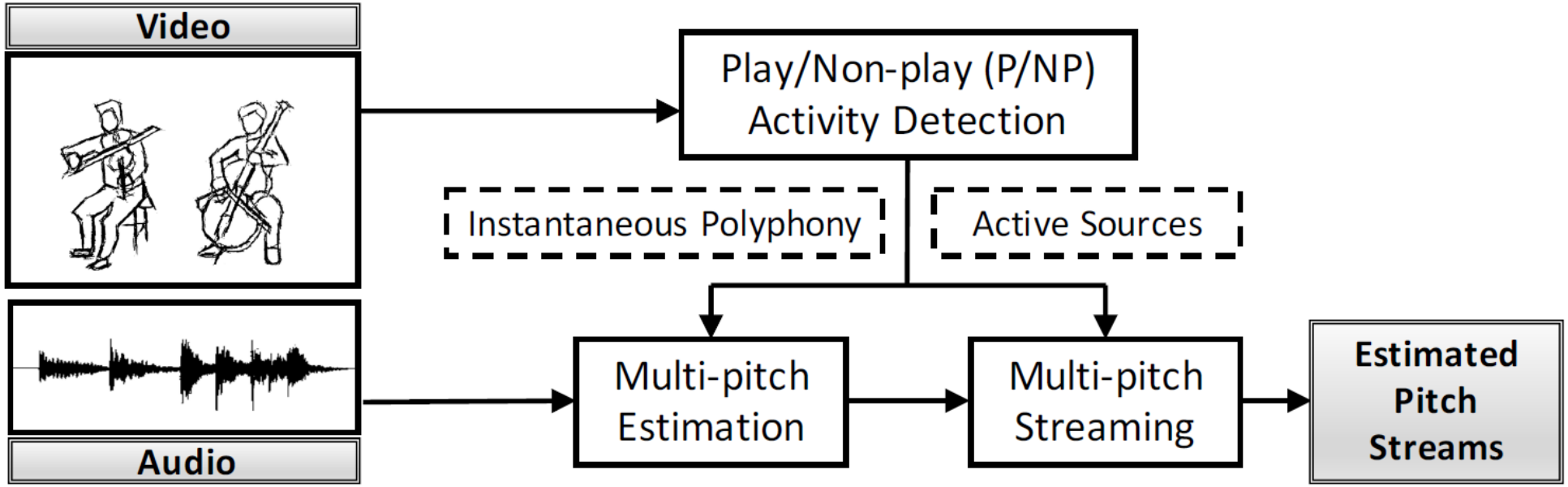



(Bazzica et al., 2016)

# Static Audiovisual Correspondence

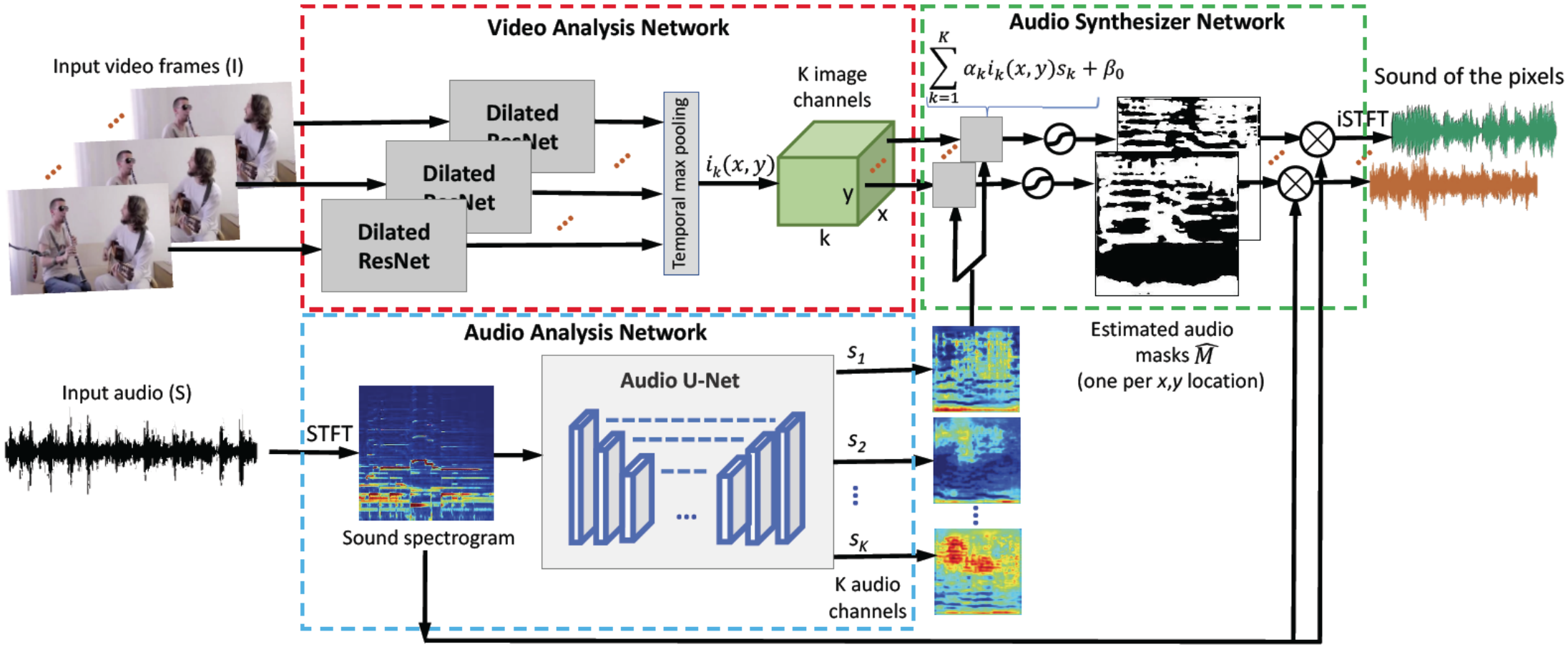
- P/NP information improves multi-pitch analysis and music transcription

(Dinesh et al., 2017)



# Static Audiovisual Correspondence

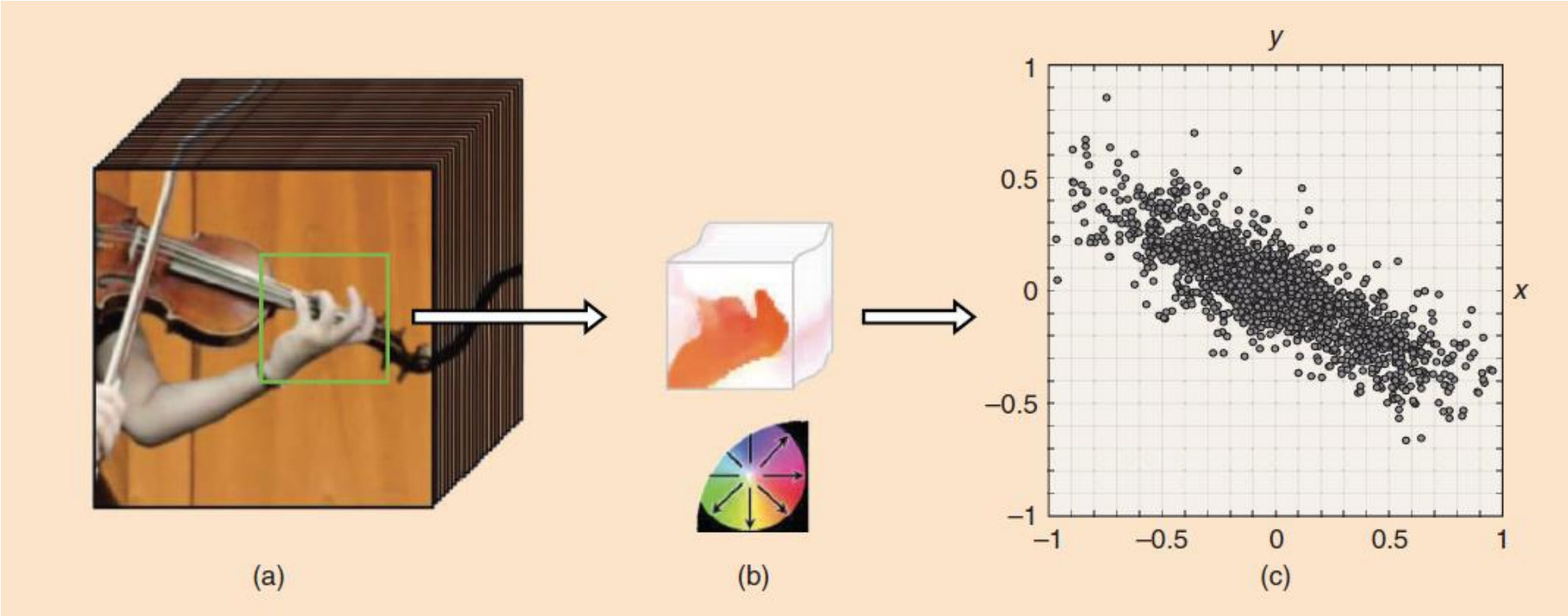
- Audiovisual Music Separation
  - The Sound of Pixels (Zhao et al., 2018)



# Dynamic Audiovisual Correspondence

- Vibrato Analysis (for strings)
  - Correlate left-hand rolling motion with pitch fluctuation

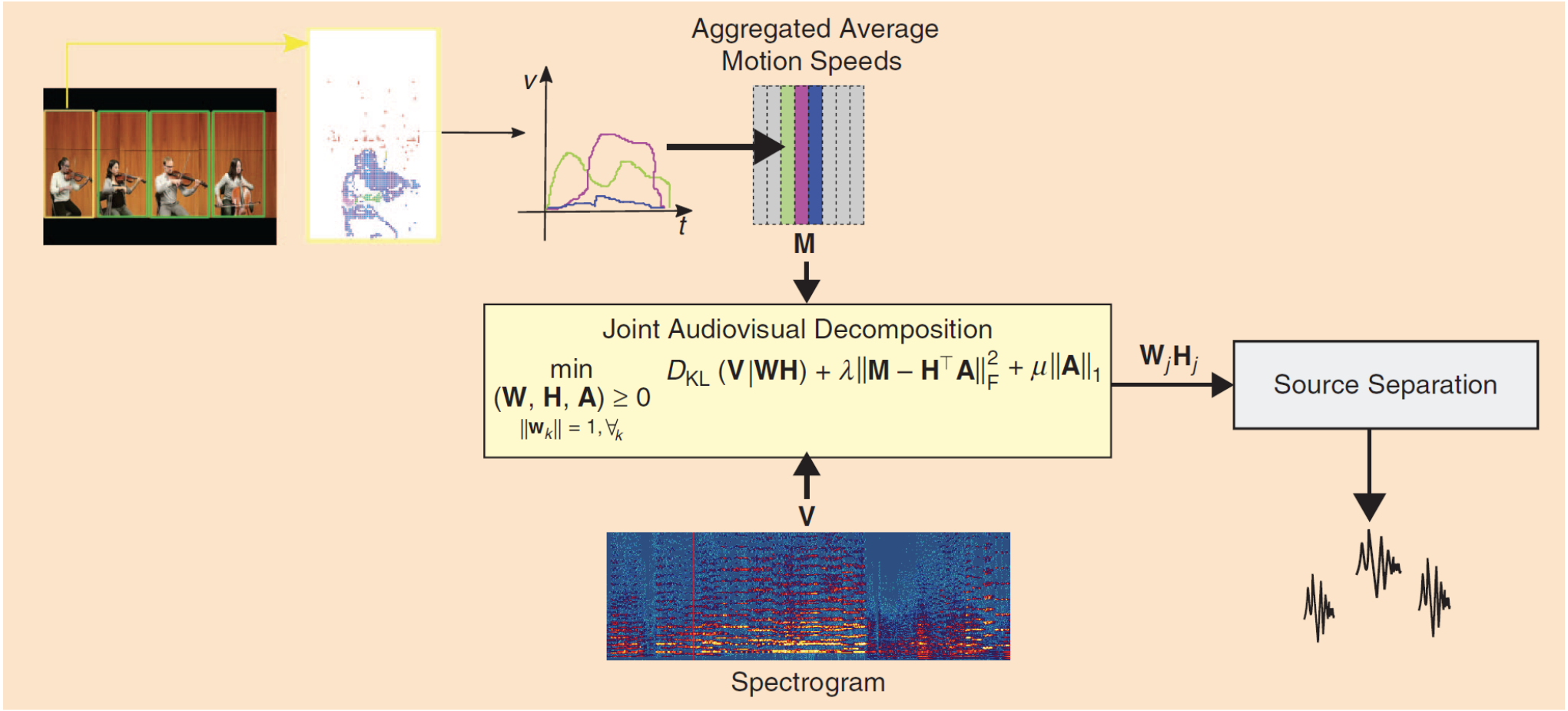
(Li et al., 2017)



# Dynamic Audiovisual Correspondence

- Co-factorization of motion data and spectrogram

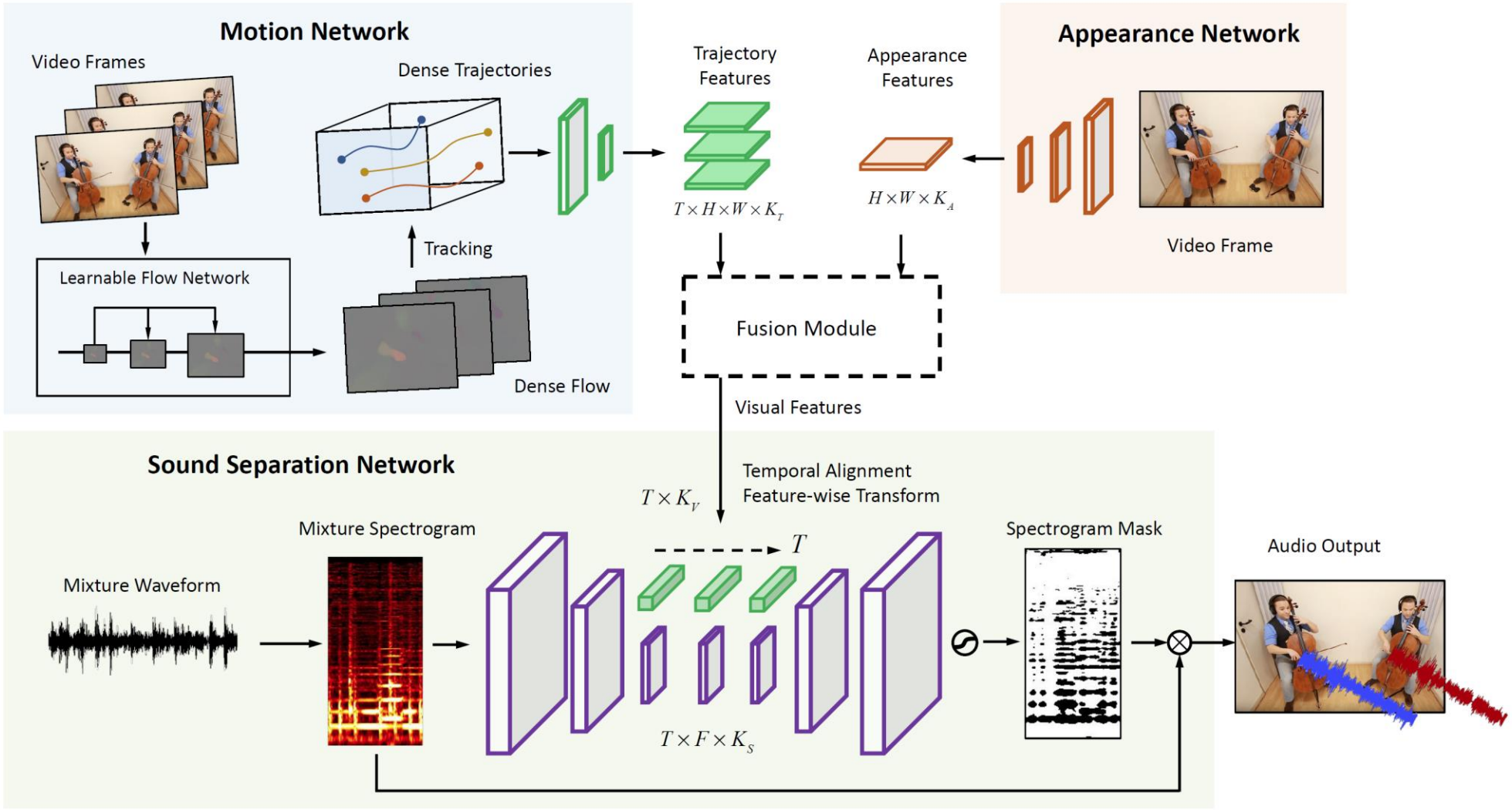
(Parekh et al., 2017)



# Dynamic Audiovisual Correspondence

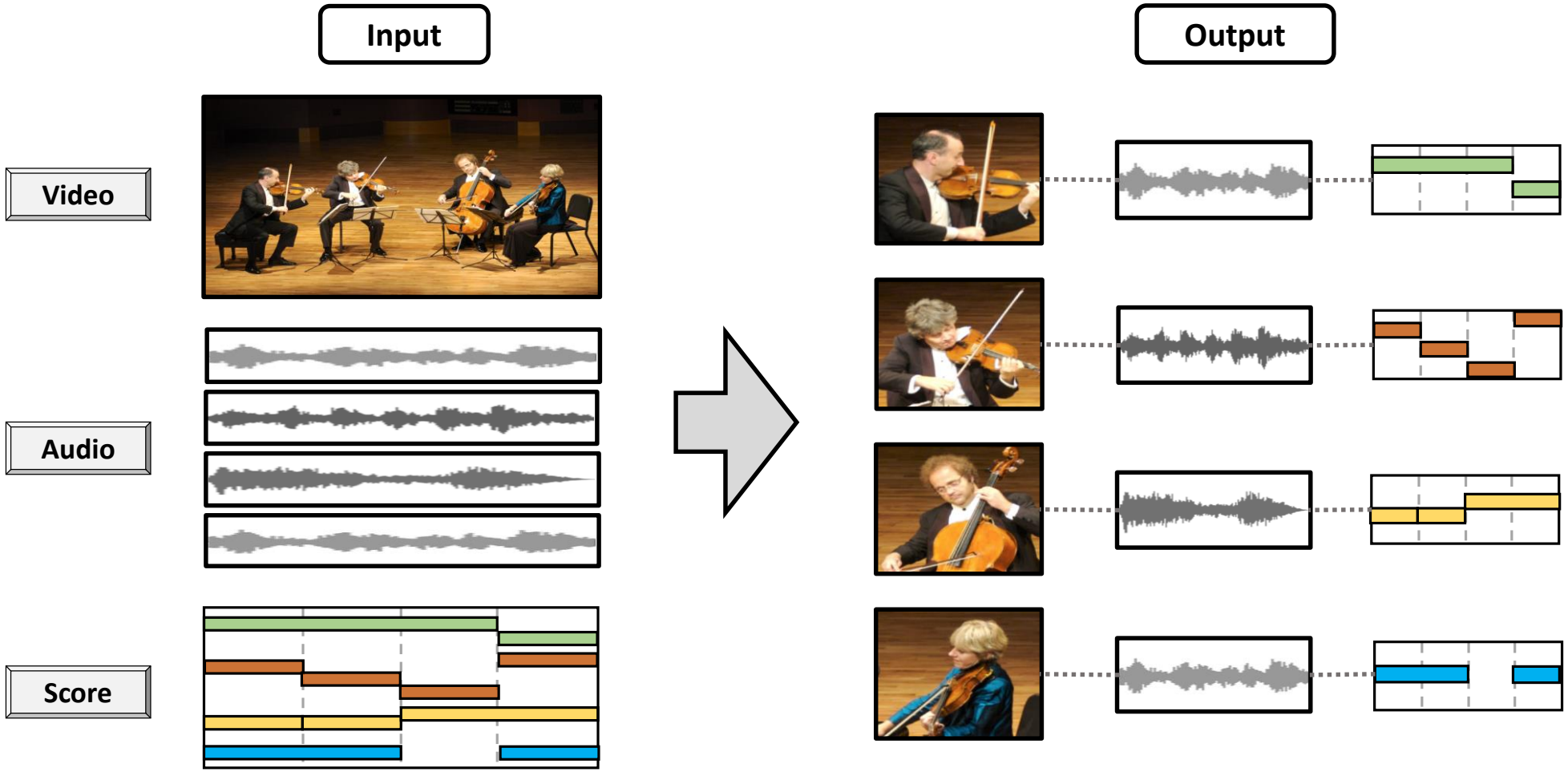
- The Sound of Motions

(Zhao et al., 2019)



# Audiovisual Source Association

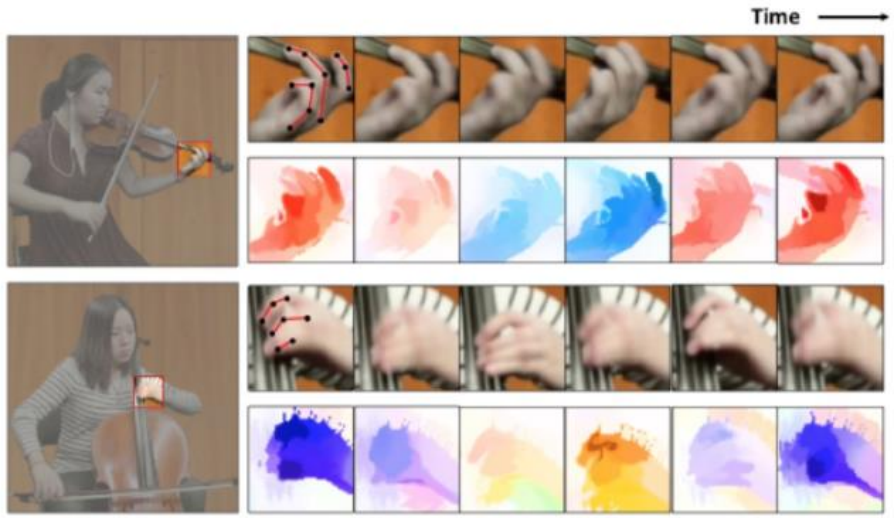
- A novel task

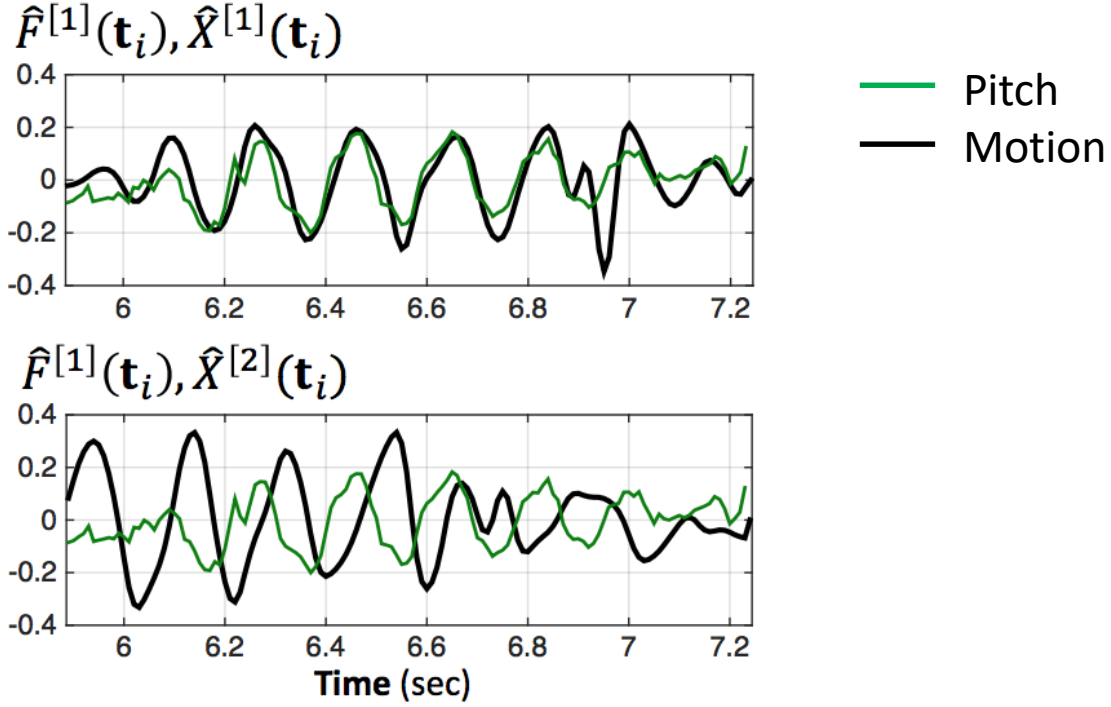


# Hand Rolling Motion ←→ Vibrato Pitch Fluctuation

- Works for strings
  - Hand tracking with OpenPose
  - Fine motion extraction with optical flow estimation
  - Pitch contour estimation with score-informed pitch estimation

(Li et al., 2017b), (Li et al., 2019)

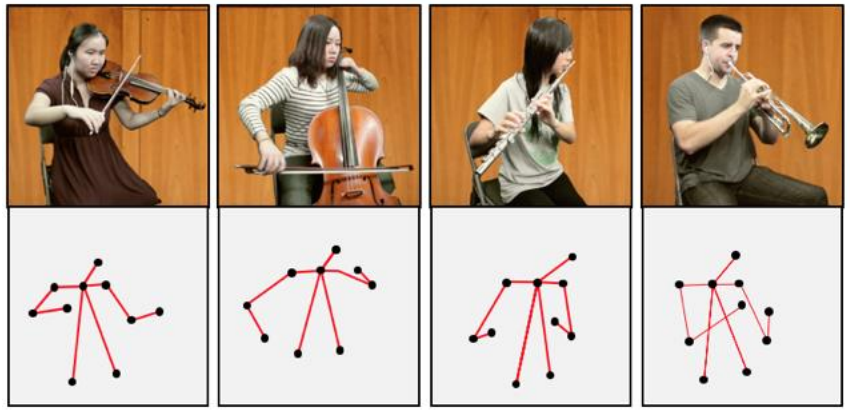


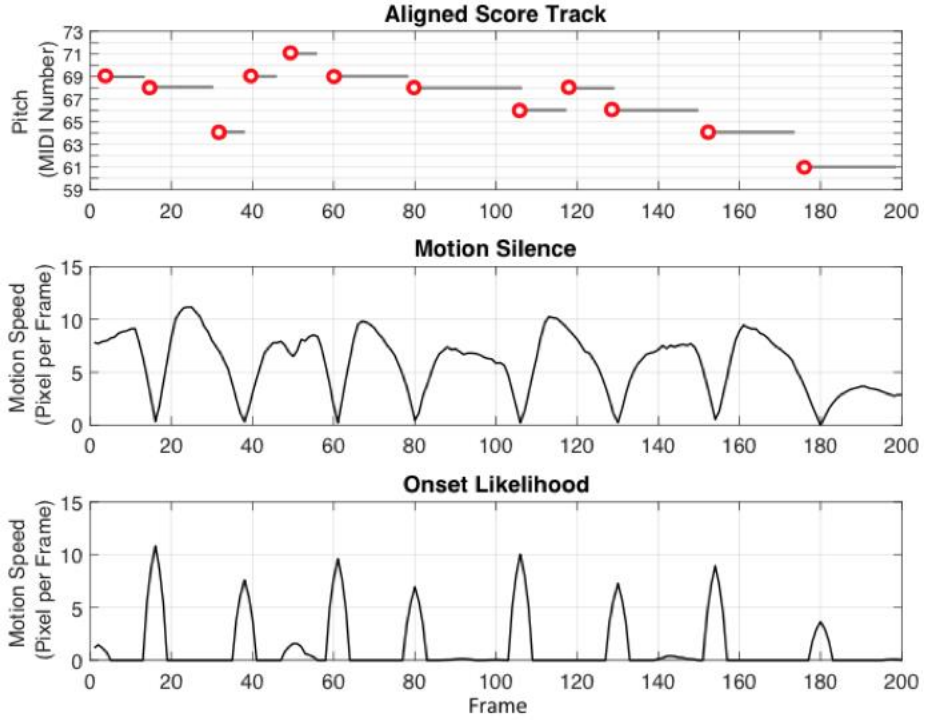



# Body Motion ←→ Note Onsets

- Works for strings
  - Body joint extraction with OpenPose
  - Detect principal motion with PCA
  - Derive onset likelihood curve from principal motion

(Li et al., 2017a), (Li et al., 2019)

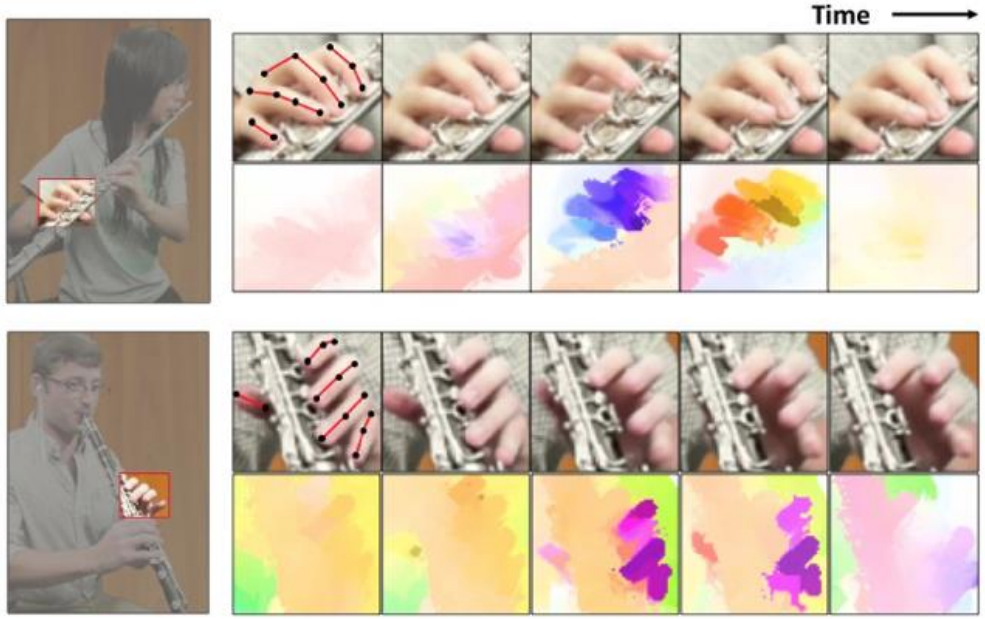


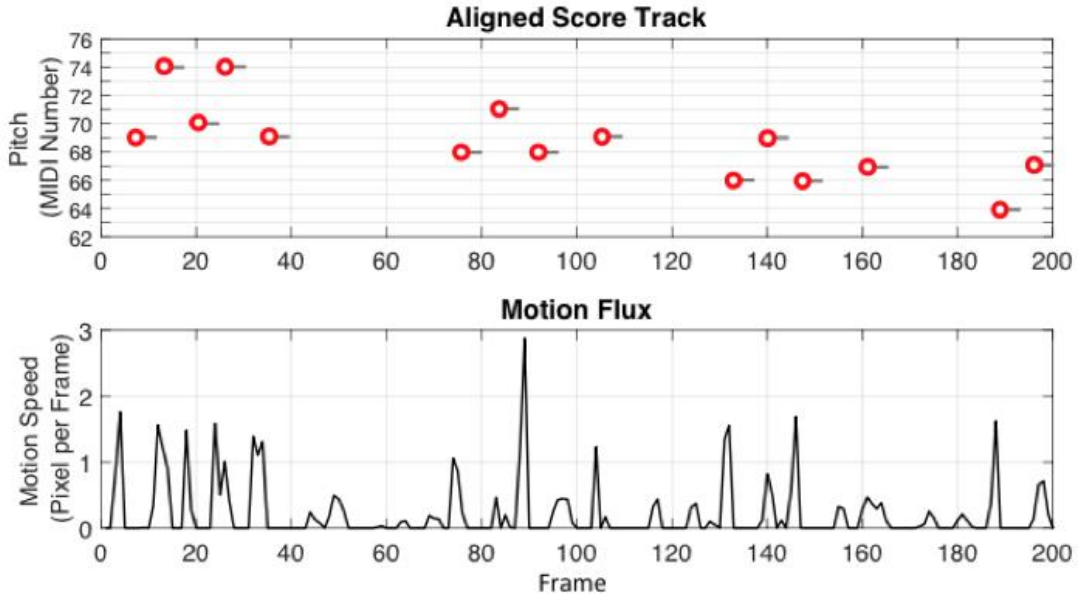


# Fingering Motion ←→ Note Onsets

- Works for woodwind/brass
  - Hand tracking with OpenPose
  - Finger motion flux extraction with optical flow estimation

(Li et al., 2019)

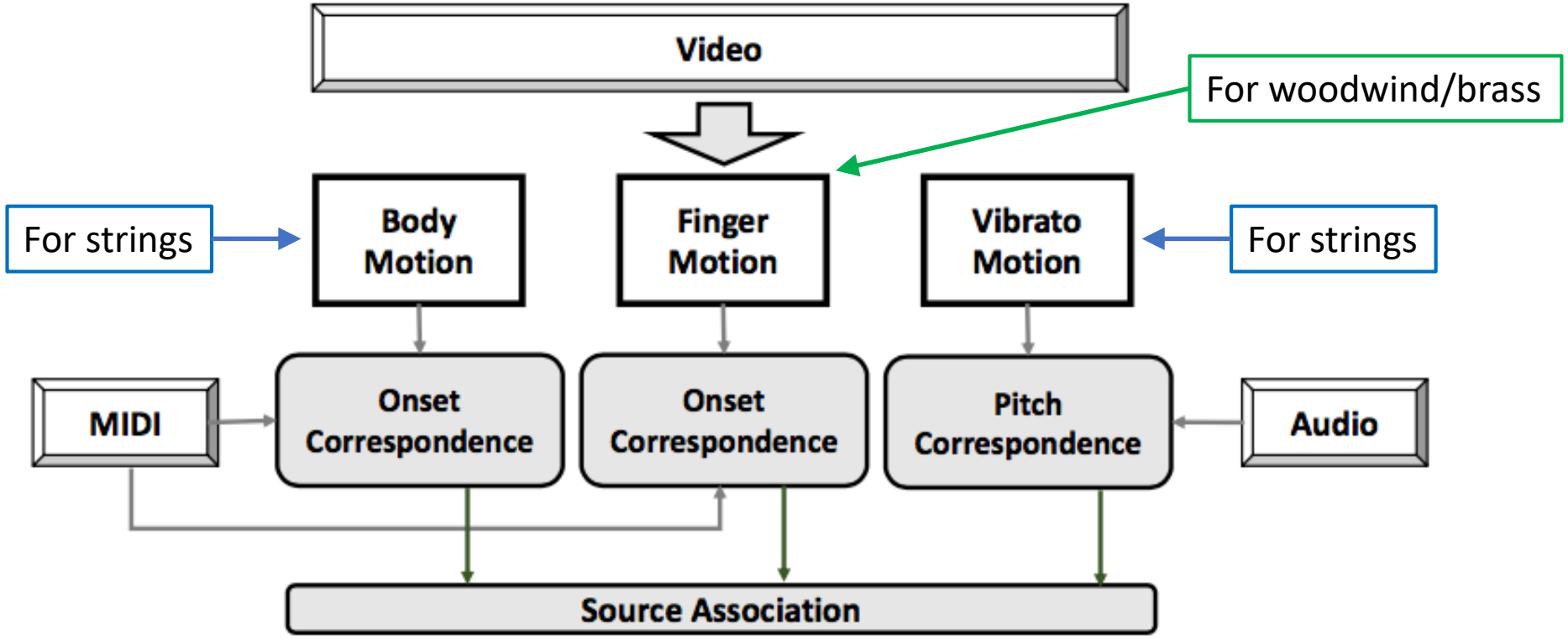


# Integrating All Components

- Linear combination with adaptive weights

(Li et al., 2019)

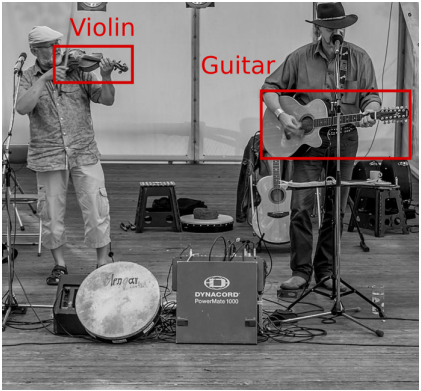


# Tasks

# Tasks

## Recognition and Detection

(Arandjelović and Zisserman, 2017; Slizovskaia et al., 2017; Parekh et al., 2019)



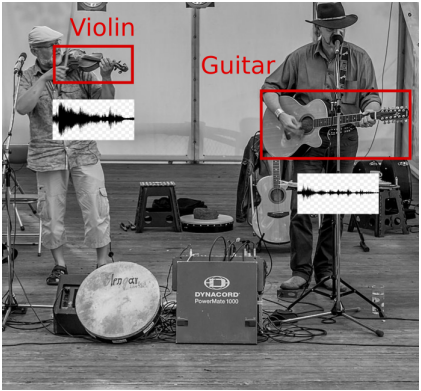
# Tasks

## Recognition and Detection

(Arandjelović and Zisserman, 2017; Slizovskaia et al., 2017; Parekh et al., 2019)

## Association and Separation

(Li et al., 2017c; Parekh et al., 2017a; Gao and Grauman, 2019b)



# Tasks

## Recognition and Detection

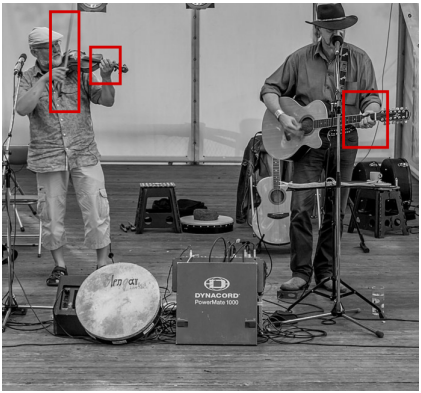
(Arandjelović and Zisserman, 2017; Slizovskaia et al., 2017; Parekh et al., 2019)

## Association and Separation

(Li et al., 2017c; Parekh et al., 2017a; Gao and Grauman, 2019b)

## Player activity analysis

- Onsets - (Bazzica et al., 2016; Li et al., 2017a)
- Fingering analysis - (Burns and Wanderley, 2006)



# Tasks

## Recognition and Detection

(Arandjelović and Zisserman, 2017; Slizovskaia et al., 2017; Parekh et al., 2019)

## Association and Separation

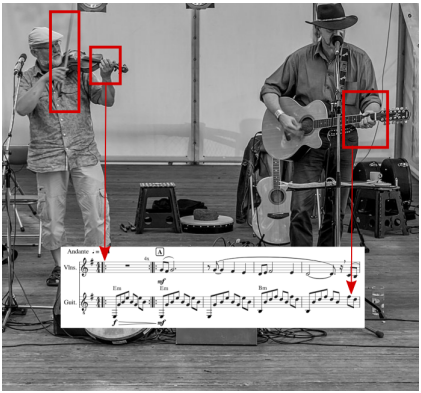
(Li et al., 2017c; Parekh et al., 2017a; Gao and Grauman, 2019b)

## Player activity analysis

- Onsets - (Bazzica et al., 2016; Li et al., 2017a)
- Fingering analysis - (Burns and Wanderley, 2006)

## Transcription

(Gillet and Richard, 2005; Paleari et al., 2008; Dinesh et al., 2017)

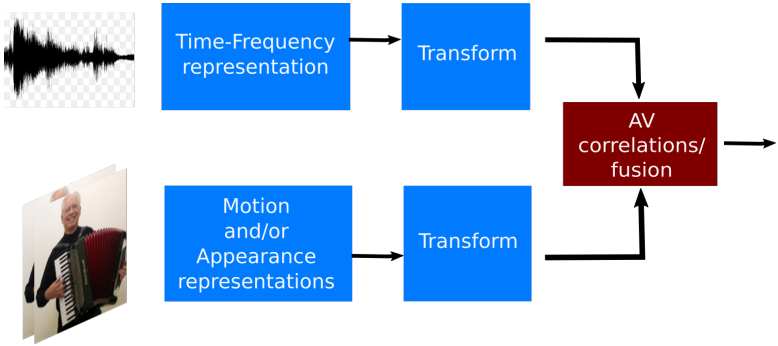


# Tasks

| Visuals | Critical | | Significant | | |
|---|---|---|---|---|---|
| | Fingering | Association | Player activity | Transcription | Separation |
|  | | | | | |
|  | | | | | |
|  | NA | | | | |
|  | | | | | |

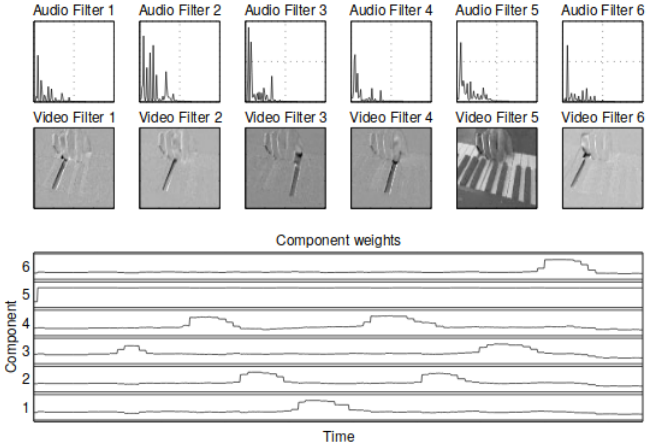
- Other tasks: Conductor gesture analysis - (Sarasúa and Guaus, 2014);
  Dance scene analysis - (Shiratori et al., 2004; Essid et al., 2012; Dremeau and Essid, 2013)

# General Framework



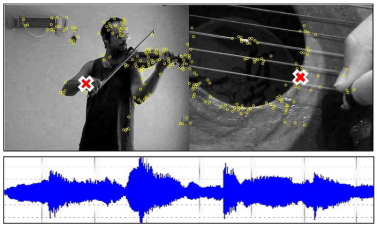

- Audio / visually informed tasks
- Simultaneous learning for other downstream tasks such as classification, detection, transcription etc.

# Illustrative examples

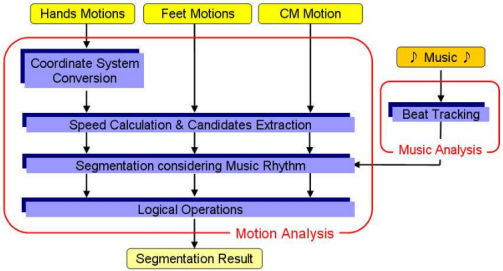



Audio Visual independent component analysis (Smaragdis and Casey, 2003)

# Illustrative examples



(Barzelay and Schechner, 2007)

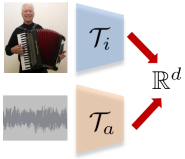



(Shiratori et al., 2004)

Audio-visual onsets co-incidence

# Modeling cross-modal dependence

- Joint subspace learning
  - Mutual Information, CCA, ICA (Fisher et al., 2001; Kidron et al., 2005)
  - Deep learning based variants (Ngiam et al., 2011; Andrew et al., 2013)
  - Misc. - cosine products, bilinear pooling (Barzelay and Schechner, 2007)

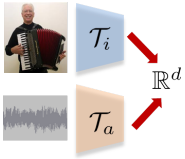



$$\mathcal{T}_i^*, \mathcal{T}_a^* = \underset{\mathcal{T}_i, \mathcal{T}_a}{\text{argmax}}\ S(\mathcal{T}_i(I)\mathcal{T}_a(A))$$

- $S$ can be correlation, covariance, mutual information
- $\mathcal{T}_i, \mathcal{T}_a$ can be modeled using deep architectures

# Modeling cross-modal dependence

- Joint subspace learning
  - Mutual Information, CCA, ICA (Fisher et al., 2001; Kidron et al., 2005)
  - Deep learning based variants (Ngiam et al., 2011; Andrew et al., 2013)
  - Misc. - cosine products, bilinear pooling (Barzelay and Schechner, 2007)




$$\mathcal{T}_i^*, \mathcal{T}_a^* = \underset{\mathcal{T}_i, \mathcal{T}_a}{\text{argmax}}\ S(\mathcal{T}_i(I)\mathcal{T}_a(A))$$

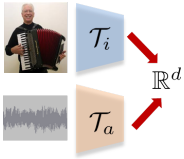
- $S$ can be correlation, covariance, mutual information
- $\mathcal{T}_i, \mathcal{T}_a$ can be modeled using deep architectures

# Modeling cross-modal dependence

- Joint subspace learning
  - Mutual Information, CCA, ICA (Fisher et al., 2001; Kidron et al., 2005)
  - Deep learning based variants (Ngiam et al., 2011; Andrew et al., 2013)
  - Misc. - cosine products, bilinear pooling (Barzelay and Schechner, 2007)

**Focus of this section**

Matrix co-factorization / cross-modal transformations

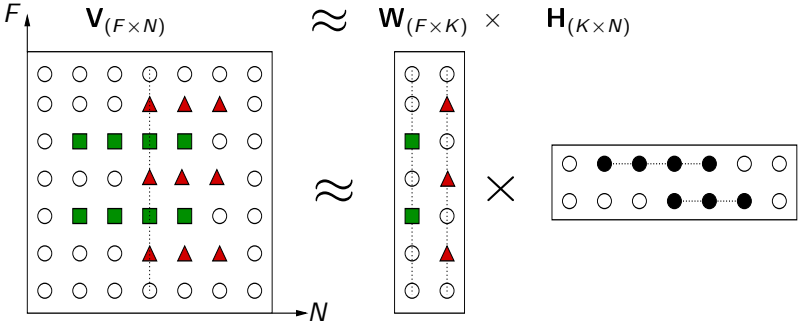



$$\mathcal{T}_i^*, \mathcal{T}_a^* = \underset{\mathcal{T}_i, \mathcal{T}_a}{\text{argmax}}\ S(\mathcal{T}_i(I)\mathcal{T}_a(A))$$

- $S$ can be correlation, covariance, mutual information
- $\mathcal{T}_i, \mathcal{T}_a$ can be modeled using deep architectures

# Explaining data by factorisation

## General formulation

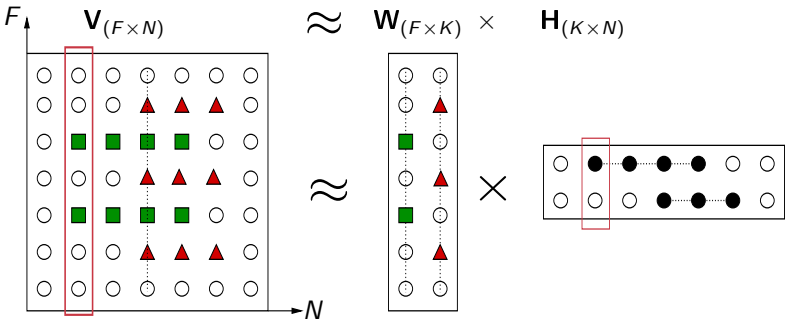



$$\mathbf{V} = [v_{fn}], \qquad \mathbf{W} = [w_{fk}] \qquad \text{and} \qquad \mathbf{H} = [h_{kn}]$$

*Based on an illustration by C. Févotte*

# Explaining data by factorisation

## General formulation

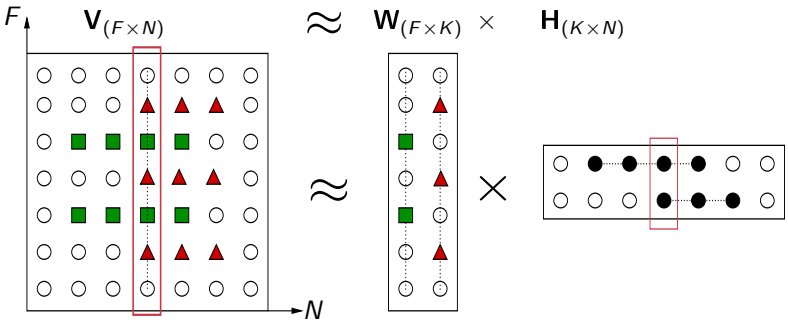



$$\mathbf{V} = [v_{fn}], \qquad \mathbf{W} = [w_{fk}] \qquad \text{and} \qquad \mathbf{H} = [h_{kn}]$$

*Based on an illustration by C. Févotte*

# Explaining data by factorisation

## General formulation

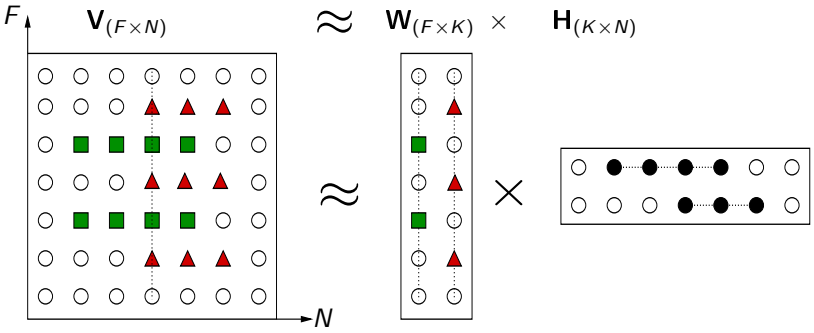



$$\mathbf{V} = [v_{fn}], \qquad \mathbf{W} = [w_{fk}] \qquad \text{and} \qquad \mathbf{H} = [h_{kn}]$$

*Based on an illustration by C. Févotte*

# Explaining nonnegative data by factorisation

A more interpretable model: **Nonnegative Matrix Factorisation (NMF)**




$$\mathbf{V} = [v_{fn}],\ v_{fn} \geq 0, \quad \mathbf{W} = [w_{fk}],\ w_{fk} \geq 0 \text{ and } \mathbf{H} = [h_{kn}],\ h_{kn} \geq 0$$

# NMF outputs

## Audio example

NMF produces **part-based** representations of the data:

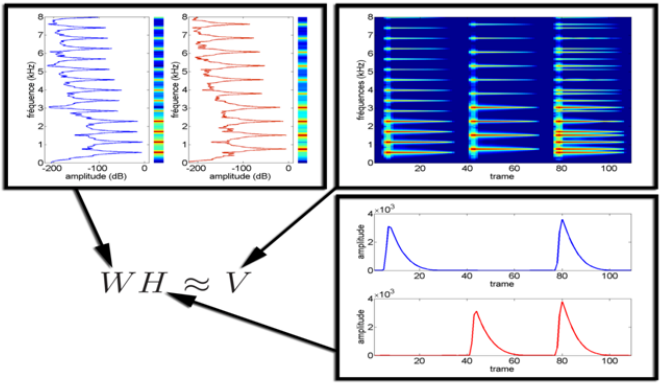
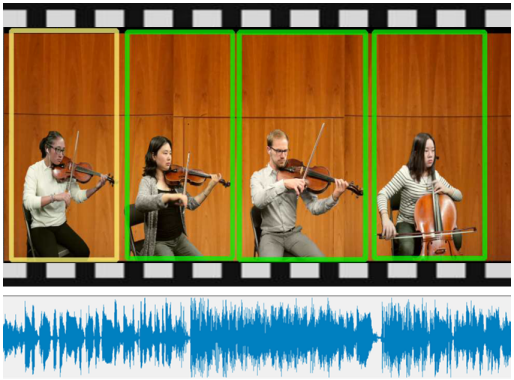



*Illustration by R. Hennequin.*

# MF for multiview data analysis

## Motivation

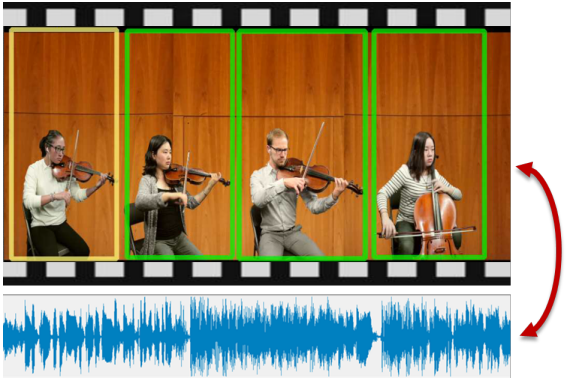
Consider **multimodal source classification/separation** in music videos



# MF for multiview data analysis

## Motivation

Consider **multimodal source classification/separation** in music videos



- The audio and visual streams are related...

# MF for multiview data analysis

## Possible approaches I

- Concatenate the feature observations:

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{bmatrix} \approx \mathbf{WH}$$

→ same cost functions used for different modalities: not always optimal.

# MF for multiview data analysis

## Possible approaches I

- Concatenate the feature observations:

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{bmatrix} \approx \mathbf{WH}$$

$\rightarrow$ same cost functions used for different modalities: not always optimal.

# MF for multiview data analysis

## Possible approaches II

- Another idea: perform **hard co-factorisation** constraining the activations to be the same:

$$\begin{cases} \mathbf{V}_1 \approx \mathbf{W}_1\mathbf{H} \\ \mathbf{V}_2 \approx \mathbf{W}_2\mathbf{H} \end{cases} \quad \text{solving:} \quad \min_{\mathbf{W}_1,\mathbf{W}_2,\mathbf{H}} D_1(\mathbf{V}_1, \mathbf{W}_1\mathbf{H}) + \beta_2 D_2(\mathbf{V}_2, \mathbf{W}_2\mathbf{H})$$

Does not account for possible local discrepancies across modalities.

→ Constrain the audio and visual data factorisations to be "**related**": temporal activations relating to these two streams of data should be **close**, not necessarily equal.

# MF for multiview data analysis

## Possible approaches II

- Another idea: perform **hard co-factorisation** constraining the activations to be the same:

$$\begin{cases} \mathbf{V}_1 \approx \mathbf{W}_1\mathbf{H} \\ \mathbf{V}_2 \approx \mathbf{W}_2\mathbf{H} \end{cases} \quad \text{solving:} \quad \min_{\mathbf{W}_1,\mathbf{W}_2,\mathbf{H}} D_1(\mathbf{V}_1, \mathbf{W}_1\mathbf{H}) + \beta_2 D_2(\mathbf{V}_2, \mathbf{W}_2\mathbf{H})$$

Does not account for possible local discrepancies across modalities.

→ Constrain the audio and visual data factorisations to be "**related**": temporal activations relating to these two streams of data should be **close**, not necessarily equal.

# MF for multiview data analysis

## Possible approaches II

- Another idea: perform **hard co-factorisation** constraining the activations to be the same:

$$\begin{cases} \mathbf{V}_1 \approx \mathbf{W}_1\mathbf{H} \\ \mathbf{V}_2 \approx \mathbf{W}_2\mathbf{H} \end{cases} \qquad \text{solving:} \quad \min_{\mathbf{W}_1,\mathbf{W}_2,\mathbf{H}} D_1(\mathbf{V}_1, \mathbf{W}_1\mathbf{H}) + \beta_2 D_2(\mathbf{V}_2, \mathbf{W}_2\mathbf{H})$$

  Does not account for possible local discrepancies across modalities.

→ Constrain the audio and visual data factorisations to be "**related**": temporal activations relating to these two streams of data should be **close**, not necessarily equal.

# Soft matrix co-factorisation

(Seichepine et al., 2014)

Consider:

$$\begin{cases} \mathbf{V}_1 \approx \mathbf{W}_1\mathbf{H}_1 \\ \mathbf{V}_2 \approx \mathbf{W}_2\mathbf{H}_2 \\ \mathbf{H}_1 \approx \mathbf{H}_2 \end{cases}$$

Solve the problem:

$$\min_{\mathbf{W}_1,\mathbf{H}_1,\mathbf{W}_2,\mathbf{H}_2} D_1(\mathbf{V}_1 \,|\, \mathbf{W}_1\mathbf{H}_1) + \beta_2 D_2(\mathbf{V}_2 \,|\, \mathbf{W}_2\mathbf{H}_2) + \beta_c P(\mathbf{H}_1, \mathbf{H}_2)$$

- $D_1$ and $D_2$ are resp. **measures of fit** for the first and second modality
- $\beta_2$ and $\beta_c$ are weighting **hyperparameters**
- $P(\mathbf{H}_1, \mathbf{H}_2)$ is a **penalization term** coupling factorizations for the first and the second modality. e.g., $P(\mathbf{H}_1, \mathbf{H}_2) = \|\mathbf{H}_1 - \mathbf{H}_2\|_p$

# Soft matrix co-factorisation

(Seichepine et al., 2014)

Consider:

$$\begin{cases} \mathbf{V}_1 \approx \mathbf{W}_1\mathbf{H}_1 \\ \mathbf{V}_2 \approx \mathbf{W}_2\mathbf{H}_2 \\ \mathbf{H}_1 \approx \mathbf{H}_2 \end{cases}$$

Solve the problem:

$$\min_{\mathbf{W}_1,\mathbf{H}_1,\mathbf{W}_2,\mathbf{H}_2} D_1(\mathbf{V}_1 \,|\, \mathbf{W}_1\mathbf{H}_1) + \beta_2 D_2(\mathbf{V}_2 \,|\, \mathbf{W}_2\mathbf{H}_2) + \beta_c P(\mathbf{H}_1, \mathbf{H}_2)$$

- $D_1$ and $D_2$ are resp. **measures of fit** for the first and second modality
- $\beta_2$ and $\beta_c$ are weighting **hyperparameters**
- $P(\mathbf{H}_1, \mathbf{H}_2)$ is a **penalization term** coupling factorizations for the first and the second modality. *e.g.*, $P(\mathbf{H}_1, \mathbf{H}_2) = \|\mathbf{H}_1 - \mathbf{H}_2\|_p$

# Solving the problem

- We have devised a **block-coordinate majorisation-minimisation algorithms** updating $\mathbf{H}_1$, $\mathbf{H}_2$, $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{S}$ sequentially
- Update rules have thus been determined for:
  - **Kullback-Liebler** and **Itakura-Saito** cost functions;
  - $\ell_2$ and $\ell_1$-**coupling penalties**;

  see (Seichepine et al., 2014) for more details
- Scripts are available online:
  http://www.telecom-paristech.fr/~essid

# Successful applications

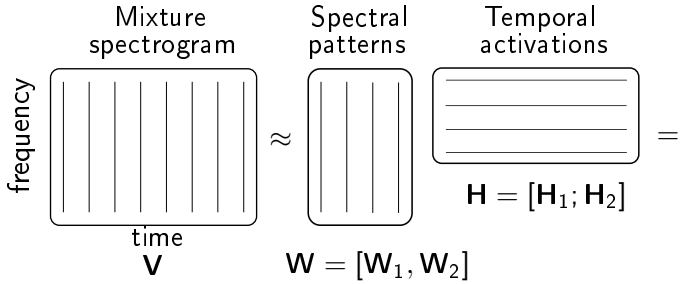
The soft co-factorisation scheme has proven effective for various tasks:

- **Multimodal speaker diarisation** ("Who Spoke When?") on videos (Seichepine et al., 2014)
- **Multi-channel speech source separation** in stereo recordings (Seichepine et al., 2014)
- **Multimodal speech separation** with lip surface data (Sedighin et al., 2017)
- **Multimodal audio source separation** in music videos (Parekh et al., 2017b,a)

- Introduction
- Audiovisual Music Performance Analysis
  - Overview of Analysis Tasks
  - Audiovisual Co-factorization for Source Separation
  - Hands-on Case Study #1: Motion Informed Audio Source Separation
- Audiovisual Content Based Classification and Retrieval
- Audiovisual Music Generation
- Datasets, Tools and Other Resources
- Challenges, Opportunities and Conclusions

# NMF-based source separation

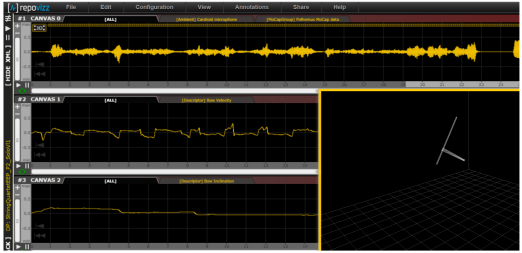
## Background



# Multiview source separation: Task specification

Motion informed source separation

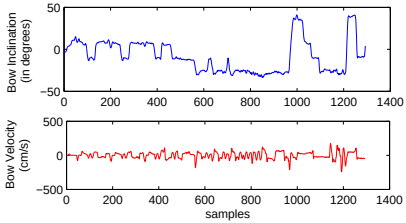
- Exploit information about the causes of vibration of each source, obtained from **motion capture**




EEP Dataset Example Video

# Motion representation

- Motion capture data from sensors placed on the instrument
- We use bow inclination and velocity

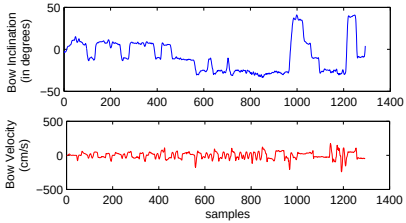



Bow inclination and velocity data for violin

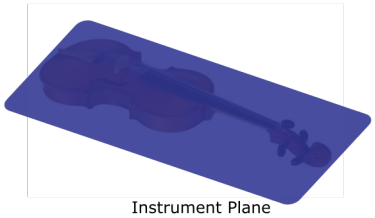
Bow velocity: computed along bow direction

# Motion representation

- Motion capture data from sensors placed on the instrument
- We use bow inclination and velocity



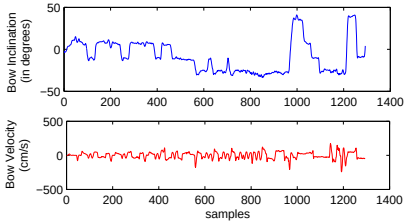

Bow inclination and velocity data for violin



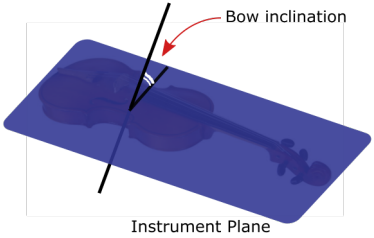

Bow velocity: computed along bow direction

# Motion representation

- Motion capture data from sensors placed on the instrument
- We use bow inclination and velocity

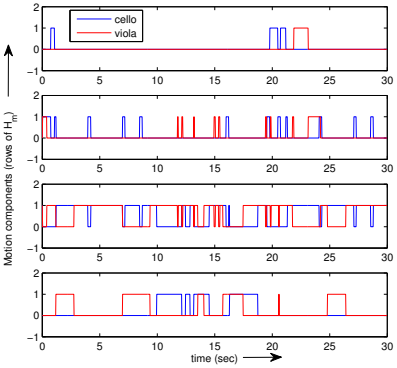



Bow inclination and velocity data for violin

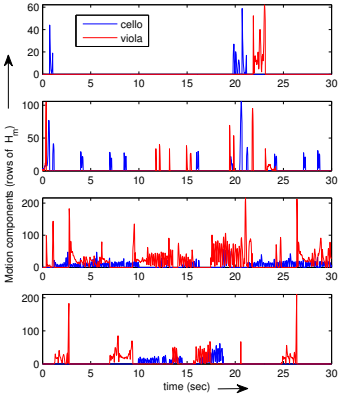



Bow velocity: computed along bow direction

# Motion representation




Quantized bow inclination




Quantized components multiplied with bow velocity: $\mathbf{H}_m \in \mathbb{R}_+^{K_m \times N}$

# Model specification

Variant of soft non-negative matrix co-factorization (sNMcF) (Seichepine et al., 2014):

$$\underset{\mathbf{W},\mathbf{H},\mathbf{S}}{\text{minimize}} \Bigg[ \underbrace{D_{KL}(\mathbf{V}|\mathbf{WH})}_{\text{spectogram factorization}} + \underbrace{\alpha\|\Lambda_a\mathbf{H} - \mathbf{S}\mathbf{H}_m\|_1}_{\text{Coupling}} + \underbrace{\beta \sum_{k=1}^{K}\sum_{n=2}^{N}(h_{kn} - h_{k(n-1)})^2}_{\text{temporal smoothing}} \Bigg]$$

$$\text{subject to } \mathbf{W} \geq 0, \mathbf{H} \geq 0.$$

Notation: $\mathbf{W} = (w_{fk})_{f,k} \in \mathbb{R}_+^{F\times K}$ , $\mathbf{H} = (h_{kn})_{k,n} \in \mathbb{R}_+^{K\times N}$, $\mathbf{S}$ - scaling diagonal matrix, $\Lambda_a$ is a diagonal matrix with $k^{th}$ diagonal coefficient $\lambda_{a,k} = \sum_f w_{fk}$, $\alpha$ and $\beta$ are hyperparameters

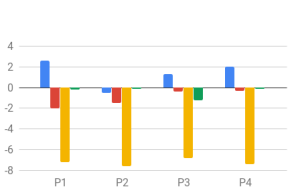
Cost function minimized using a majorization-minimization (MM) algorithm

# Dataset

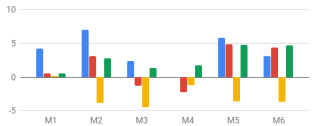
- Ensemble Expressive Performance (EEP) dataset (Marchini et al., 2014) - multimodal recordings of string instrument performances for 4 excerpts of Beethoven's Concerto


- We consider different duos and trios involving Violin, Viola and Cello

# Results and Discussion

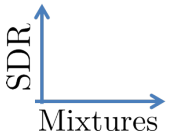



(a) 3-source mixtures



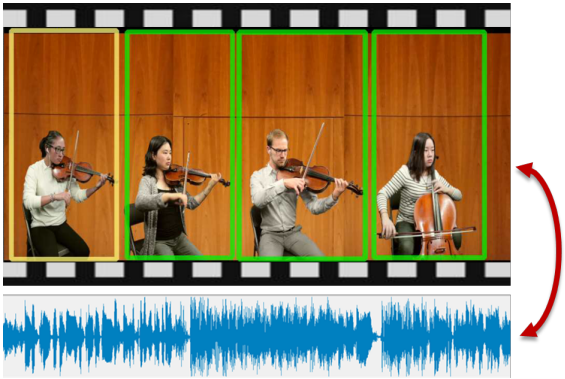

(b) 2-source mixtures


(c) same-source mixtures




Same instrument mixtures. Motion is crucial for disambiguation

# Source separation in music videos



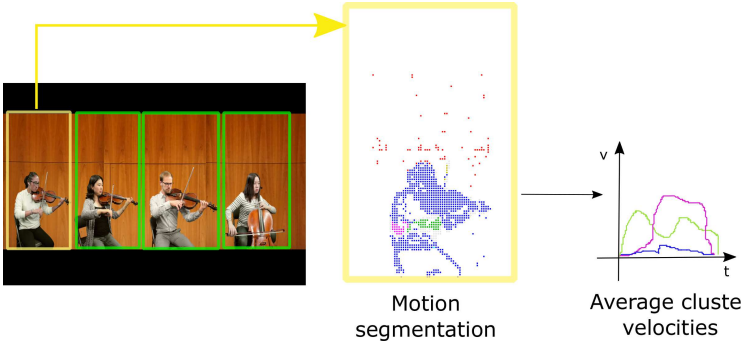
- **Difficulty:** motion representation no longer straightforward

# Cross-modal regression

(Parekh et al., 2017a)

▶ **Solution:**

1. Extract **motion clusters** (Keuper et al., 2015) in each performer's bounding box and represent them by average cluster velocities



# Cross-modal regression

(Parekh et al., 2017a)

▶ **Solution:**

1. Extract **motion clusters** (Keuper et al., 2015) in each performer's bounding box and represent them by average cluster velocities
2. **Jointly factorise** audio spectrogram and **regress** the audio activations against cluster velocities:

$$\underset{\substack{(\mathbf{W},\mathbf{H},\mathbf{A})\geqslant 0\\ \|\mathbf{w}_k\|=1,\ \forall k}}{\text{minimize}} \quad D_{KL}(\mathbf{V}|\mathbf{WH}) + \lambda\|\mathbf{M}-\mathbf{H}^\top\mathbf{A}\|_F^2 + \mu\|\mathbf{A}\|_1$$

- $\mathbf{M} \in \mathbb{R}_+^{N\times C}$: velocity signals, stacked column-wise; $C$: total number of motion clusters
- $\mathbf{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_C]$: nonnegative linear combination coefficient matrix

# Cross-modal regression

(Parekh et al., 2017a)

- **Solution:**

1. Extract **motion clusters** (Keuper et al., 2015) in each performer's bounding box and represent them by average cluster velocities
2. **Jointly factorise** audio spectrogram and **regress** the audio activations against cluster velocities:

- **Results:** Challenging case of two violins!

# Summary

- **Co-factorisation** schemes are well suited to **multiview, esp. multimodal data** analysis tasks
  - they allow for flexibly binding together related factors across different views
- **Stable** (MM) algorithms for different choices of divergences, and coupling/smoothing penalties
- Models have proven effective for **diverse multimodal applications** (music, video and physiological data)

▶ Great potential for other applications in **multiview music analysis**

# Tutorial Outline

- Introduction
- Audiovisual Music Performance Analysis
  - Overview of Analysis Tasks
  - Audiovisual Co-Factorization for Source Separation
  - Hands-on Case Study #1: Motion Informed Audio Source Separation
- **Audiovisual Content Based Classification and Retrieval**
  - Genre Classification
  - Emotion Analysis
  - Cross-Modal Retrieval
  - Instrument Classification
- Audiovisual Music Generation
  - Hands-on Case Study #2: Skeleton Plays the Piano
- Datasets, Tools and Other Resources
- Challenges, Opportunities and Conclusions

# Audiovisual Music Genre Classification

- Motivation - Visual modality conveys genre information

**Music Videos**


Jazz


Country


Rock


Pop


Classical


Heavy Metal

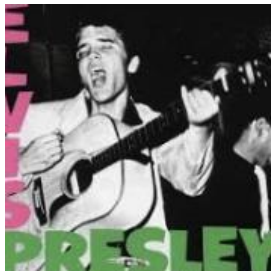
**Album Covers**

Country

New Age

Jazz


Heavy Metal



Rock

Pop

# Existing Work

- Genre tagging and artist similarity calculation from album covers and promotional photos (Libeks and Turnbull, 2011)
  - Color (RGB, HSV and LAB histograms), texture (Gabor and Haar features)
  - Distance calculation and nearest neighbor tag vector averaging

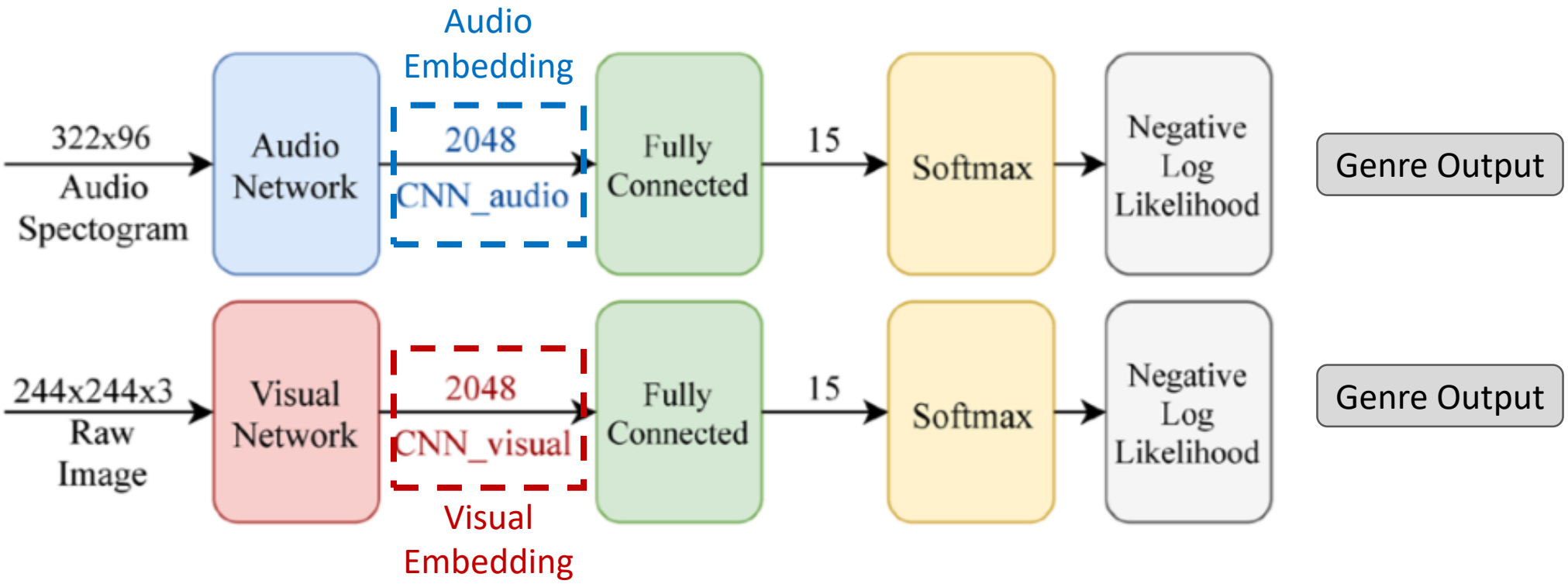
- Genre classification from music videos (Schindler and Rauber, 2015)
  - Audio feature: psychoacoustic music descriptors, MFCCs
  - Visual (color) features: color statistics, emotion values, colorfulness, Wang emotional factors, Itten's contrast, color names, lightness fluctuation patterns
  - Classifiers: SVM, KNN, Random Forest, Naïve Bayes

- Genre classification from audio, album covers, and text reviews (Oramas et al., 2018)
  - Deep neural networks for audio, visual and text representation learning, fusion and multi-label classification

# Multimodal Representation Learning and Fusion

(Oramas et al., 2018)

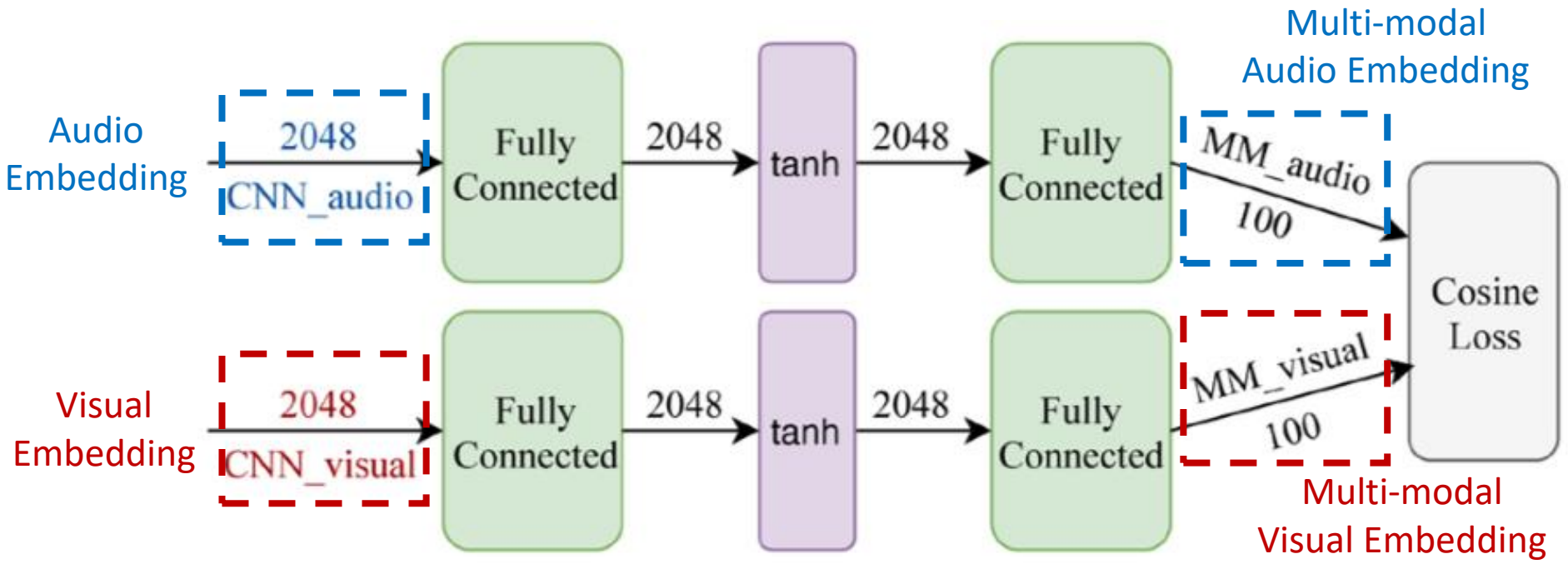
- Training Phase 1: learning single modal features through genre classification



# Multimodal Representation Learning and Fusion

(Oramas et al., 2018)

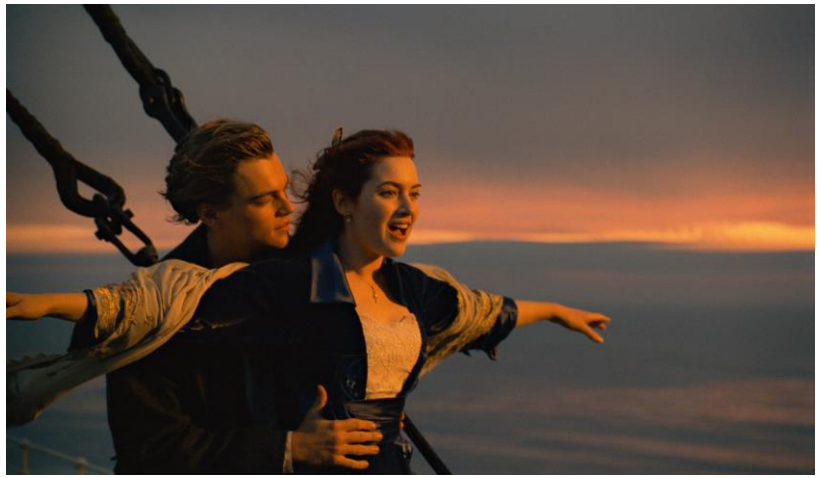
- Training Phase 2: learning multi-modal features through distance minimization



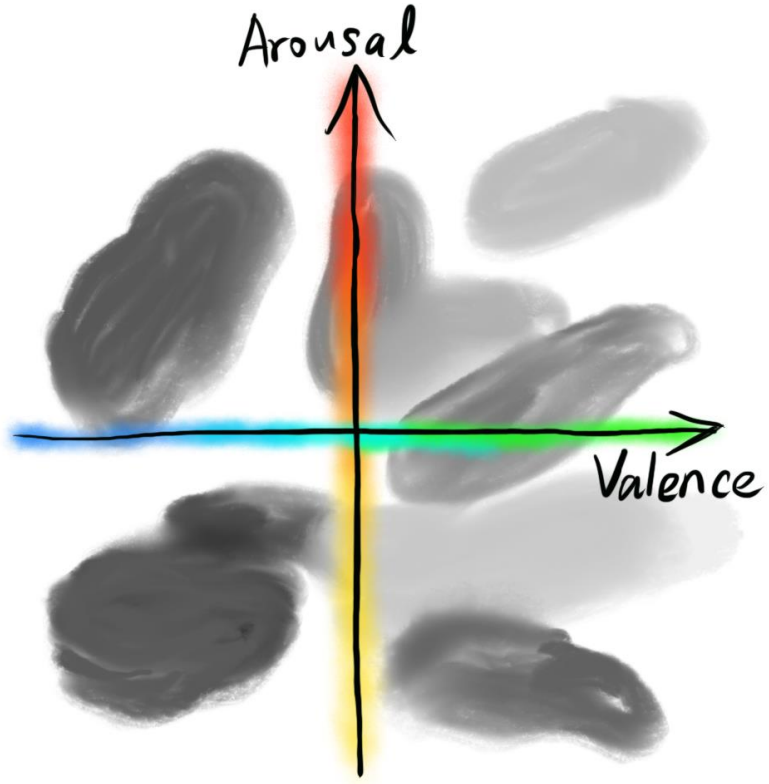

- Training Phase 3: training a feedforward network for multi-label genre classification based on a concatenation of the pre-learned multi-modal features

# Audiovisual Music Emotion Analysis

- Motivation
  - Visual information influences emotional experience of music listening
    - (Coutinho and Scherer, 2017), (Boltz et al., 2009)
  - Music is used to heighten the emotional impact of movie scenes
    - (Parke et al., 2007)
- Emotionally Relevant Factors
  - Audio: tempo, rhythm, pitch, mode, dissonance, articulation, instrument, etc.
  - Video: color, saturation, brightness, motion, action, event, instrument, etc.



# Emotion Representations


- Categorical
  - More categories: ambiguous and subjective (Juslin and Laukka 2004)
  - Fewer categories: not enough
- Dimensional
  - Typically two dimensions
  - Difficult to represent some sentiments/emotions, e.g., nostalgia, awkward, envy, disgust, etc.

# Single-Modality Emotion Analysis Work

- Using **categorical** emotion representations

| Approach | | # Emotion Categories | Data Type |
|---|---|---|---|
| Audio | (Lu et al., 2006) | 4 | Classical music |
| | MIREX 2010 (Hu and Downie, 2007) | 5 | Pop music |
| | (Wu and Jeng, 2006) | 8 | Pop music |
| | (Skowronek et al., 2007) | 12 | 12 music genres |
| | (Leman et al., 2005) | 15 | Pop music |
| | MIREX 2016 | 18 | 7 music genres (80% Pop) |
| Visual | (Kahou et al., 2016), (Kaya et al., 2017) | 7 | Video of facial expressions |
| | (Zhao et al., 2014) | 8 | Unconstrained photos |
| | (Xu et al., 2016) | 8 (3 variations on each) | Unconstrained videos |
| | (Cowen and Keltner, 2017) | 27 | Unconstrained videos |

# Single-Modality Emotion Analysis Work

- Using **dimensional** emotion representations

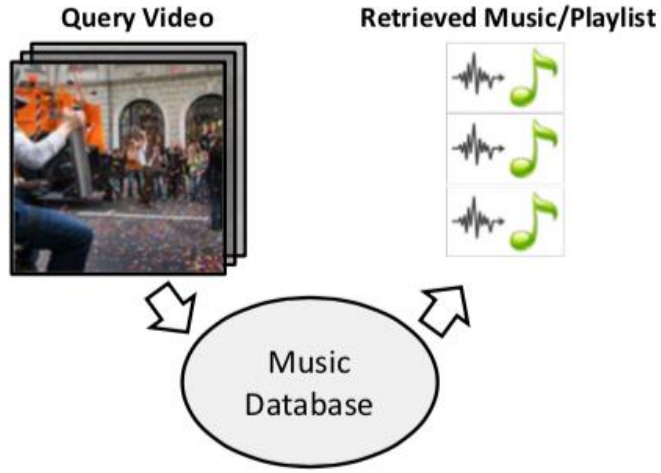
| Approach | | Dimensions | Data Type |
|---|---|---|---|
| Audio | (Wang et al., 2015) | Arousal-Valence | Music |
| | (Soleymani et al., 2013) | Arousal-Valence | Music (8 genres) |
| | (Yang et al., 2008) | Arousal-Valence | Pop Music |
| Visual | (Baveye et al., 2015) | Arousal-Valence | Movie scenes |
| | (Hanjalic and Xu, 2005) | Arousal-Valence-Dominance | Unconstrained videos |
| | (Koelstra et al., 2011) | Arousal-Valence-Dominance | Music videos |

# Cross-Modality Emotion Analysis Work

- Predicting emotions that can induce in people from music videos (Yazdani et al., 2011)
  - Audio features: zero-crossing rate, energy, silence ratio, pitch, MFCC, LPC, spectral centroid
  - Video features: lighting, key frame and shot boundary, color, motion vectors
  - Concatenation, normalization, PCA and kNN classification

- Affective visualization and retrieval (Zhang et al., 2010)
  - Arousal features: audio - *zero cross rate, tempo, beat strength*; video - *motion intensity, shot switch rate*
  - Valence features: audio - *rhythm regularity,* and *pitch;* video - *lighting*, *saturation*, *color energy*
  - Personalized affective analysis using Support Vector Regression

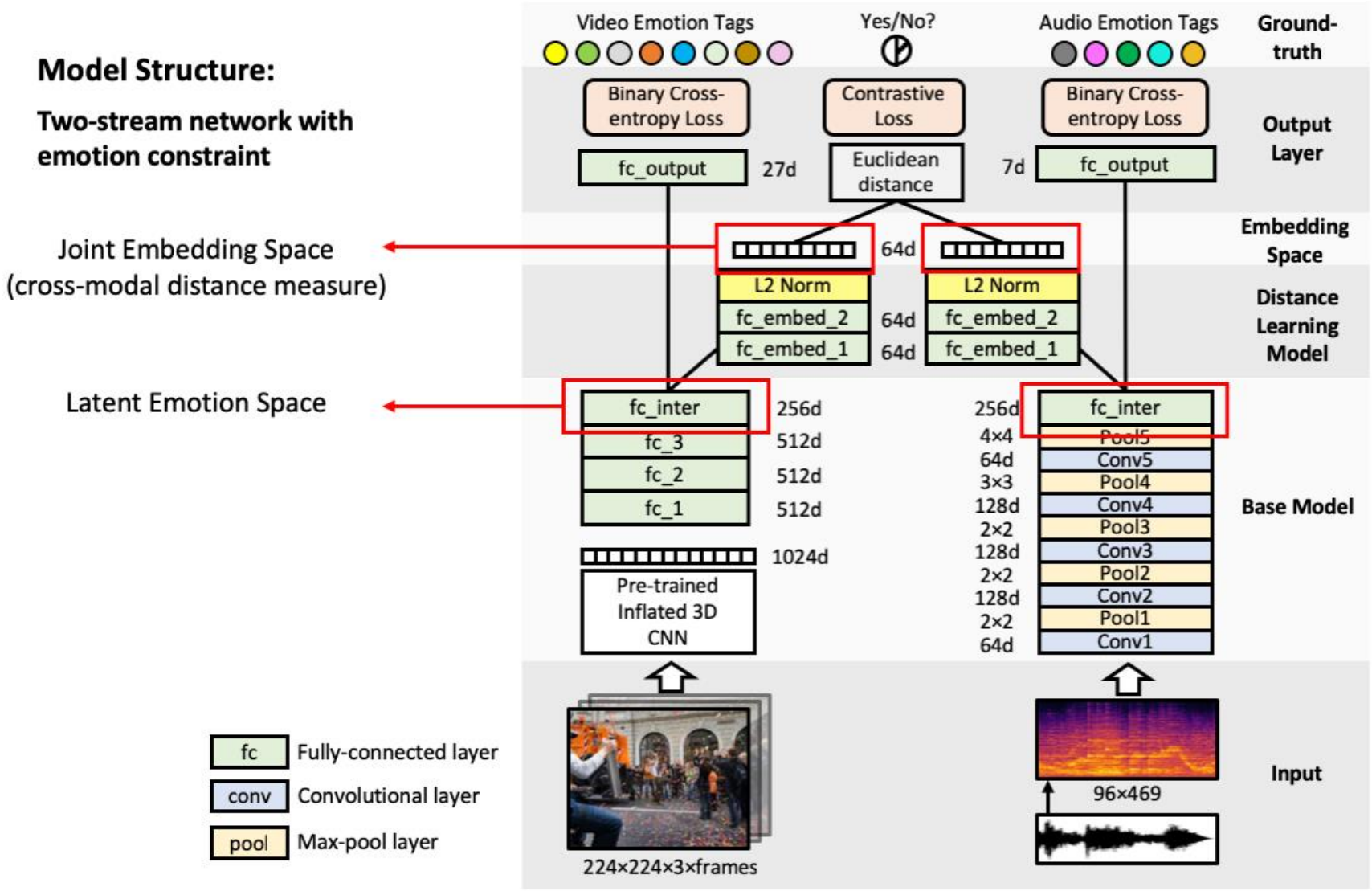
- Retrieving emotionally relevant music using video query (Li & Kumar, 2019)
  - Learning a shared latent emotion space through emotion tagging and distance constraints



# Learning Cross-Modal Latent Emotion Space

**Model Structure:**

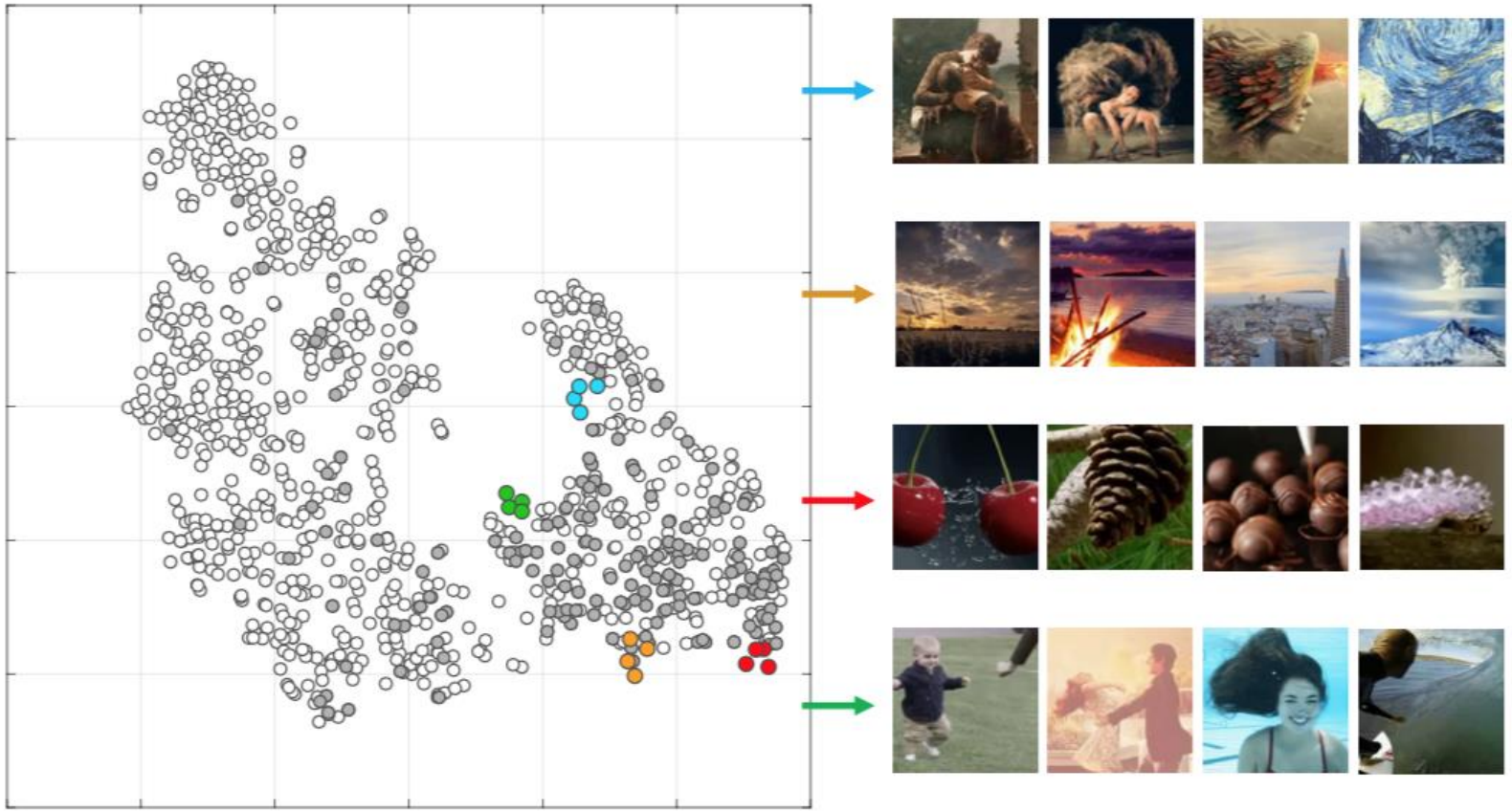
**Two-stream network with emotion constraint**




(Li and Kumar, 2019)

# Cross-Modal Emotion Representation

- Cross-modal emotion representations
  - A latent emotion space from cross-modal joint training

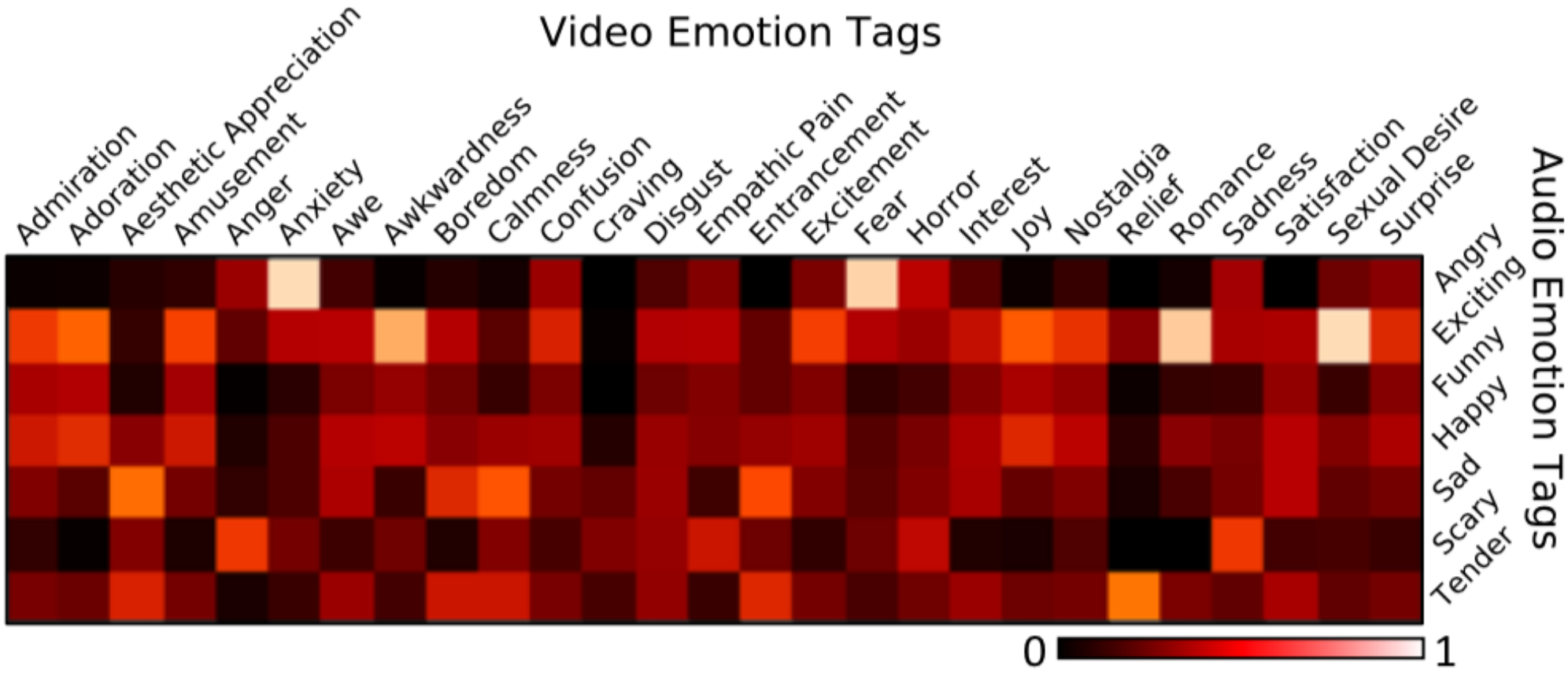


t-SNE visualization of a 64-D latent emotion space with some retrieved video thumbnails

Figure from (Li and Kuman, 2019)

# Audiovisual Emotion Tag Relations

- On crowd annotated 2000 music audio-video pairs (Li and Kuman, 2019)
  - Each audio and each video has its pre-labeled emotion tags
  - Audio-video pairs are annotated while their emotion tags are hidden



# Retrieval Examples

- User uploaded video clips from Instagram
- Retrieve music from 1195 Spotify music clips of the 30 most popular genres


Videos from http://www.ece.rochester.edu/~bli23/projects/query.html

# Audiovisual Cross-Modal Retrieval - Existing Tasks

- Music/playlist recommendation (Li and Kuman, 2019), (Sasaki et al., 2015)
  - Input query image/video, return a song or playlist
  - Recommend from a global connection (e.g., emotion, culture, genre)

- Soundtrack retrieval for videos (Shah et al., 2014), (Shin and Lee, 2017)
  - Pair silent video with proper soundtrack
  - Consider temporal correlation
    - Visual motions → music beats
    - Video shot transition → music phrase boundary

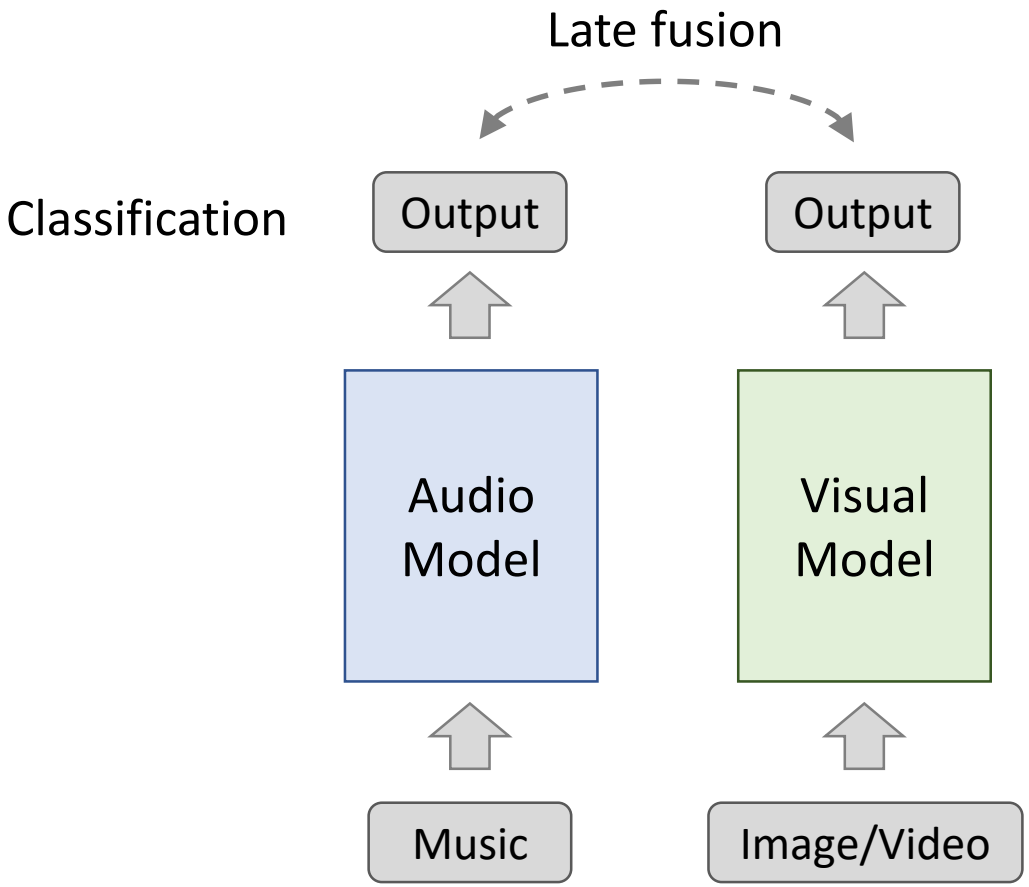
- Video retrieval for MV generation (Lin et al., 2015), (Gross et al., 2019), (Yoon et al., 2009)
  - Input query music, return video shots
  - Detect music phrase boundaries and pair a video scene for each audio segment

# Existing Cross-Modal Connections

- Low-level features
  - (Yoon et al., 2009), (Gross et al., 2019)
- Auxiliary info *(keywords, mood tags, association description, etc.)*
  - (Yu et al., 2012), (Shah and Zimmermann, 2014), (Wu et al., 2016), (Liem et al., 2012)
- Emotion modeling
  - (Chao et al., 2011)
  - Considering temporal correlation: (Wang et al., 2012), (Lin et al., 2015), (Sasaki et al., 2015), (Shin et al., 2017)
- Deep representation from neural networks
  - (Hong et al., 2018), (Li and Kumar, 2019)

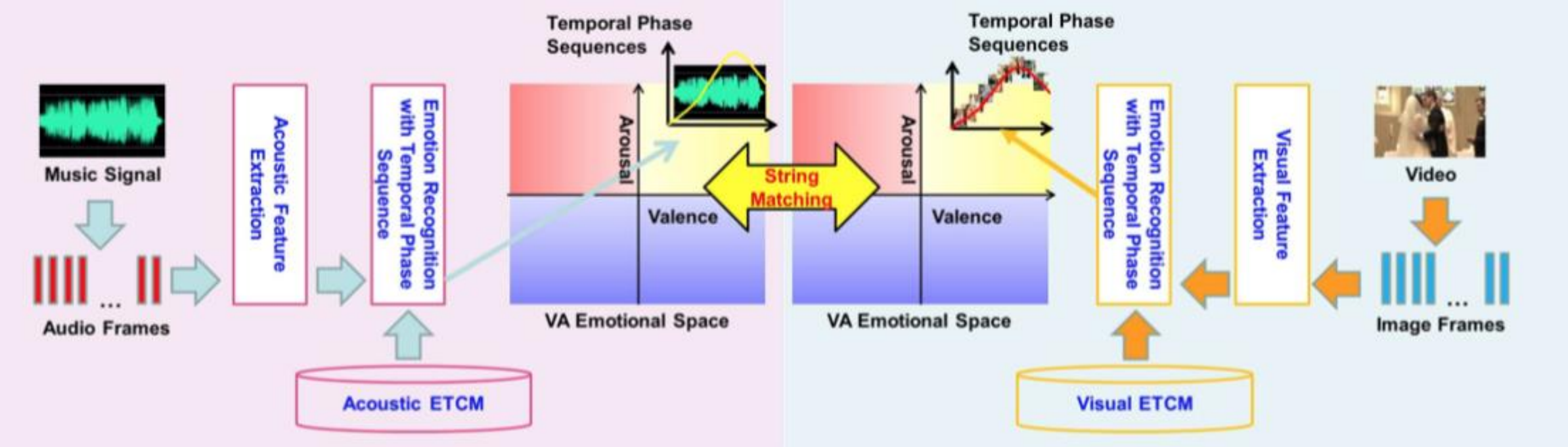
# A Typical Cross Modal Connection Approach

- Cross-Modal Connection through Late Fusion



# Example Cross-Modal Connection

(Lin et al., 2015)

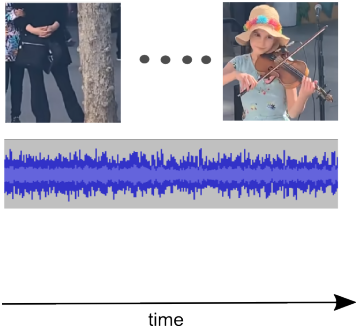



- String matching is replaced by a DNN model in their follow up work (Lin et al., 2016)

# Multimodal instrument classification and extraction

## Motivation

- Identify, locate and separate music instruments in audio and visual modalities
- Useful for various downstream tasks such as music transcription, remixing etc.



# Multimodal instrument classification and extraction

## Motivation

- Identify, locate and separate music instruments in audio and visual modalities
- Useful for various downstream tasks such as music transcription, remixing etc.

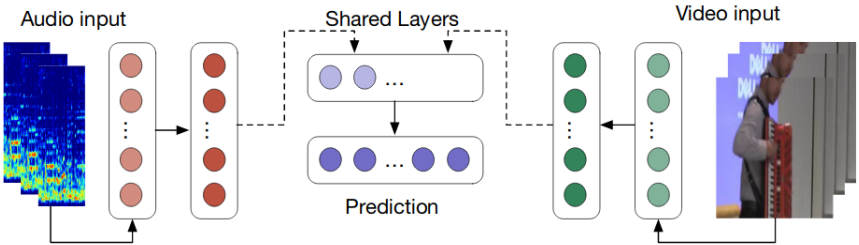
# Multimodal instrument classification and extraction

## Motivation

- Identify, locate and separate music instruments in audio and visual modalities
- Useful for various downstream tasks such as music transcription, remixing etc.

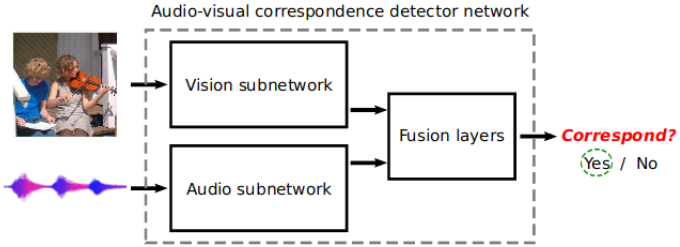

- Little to no supervision available for large-scale datasets

# Example I



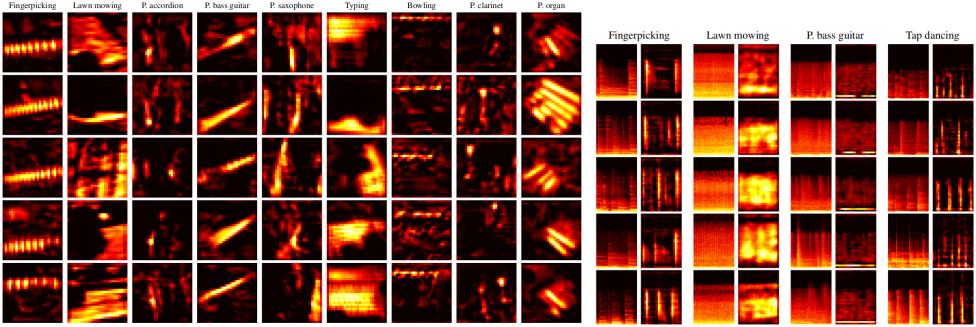

A two-stream approach to recognition
(Slizovskaia et al., 2017)

# Example II



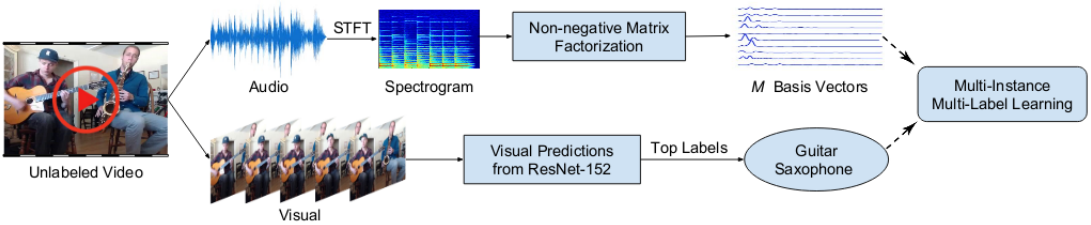

AV correspondence task: unsupervised representation learning, recognition and localization (Arandjelović and Zisserman, 2017)

# Example II




AV correspondence task: unsupervised representation learning, recognition and localization (Arandjelović and Zisserman, 2017)

# Example III

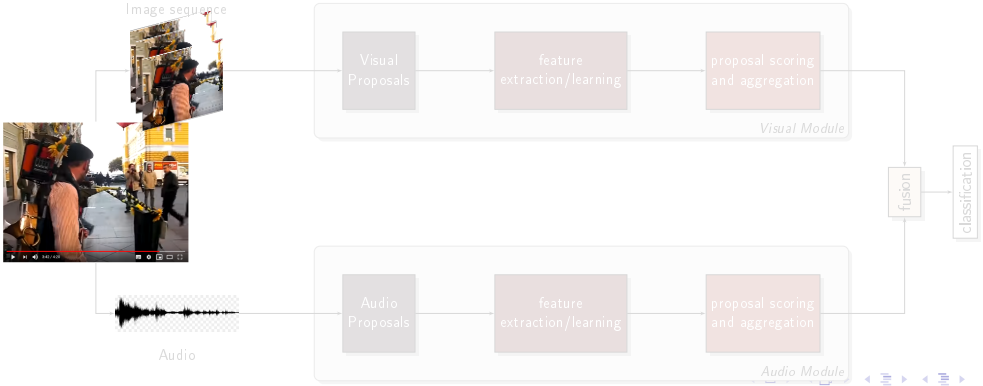



Audio source separation through unsupervised learning (Gao et al., 2018)

# Case study: A multiple instance learning based formulation

**Propose and Learn**

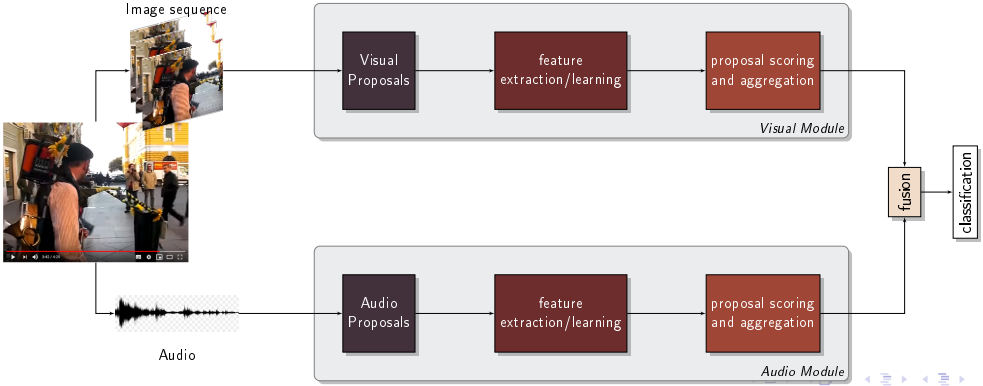
- Consider each video to be a bag of class-agnostic audio and visual proposals
- Extract features and transform them to score each according to their relevance for a particular class



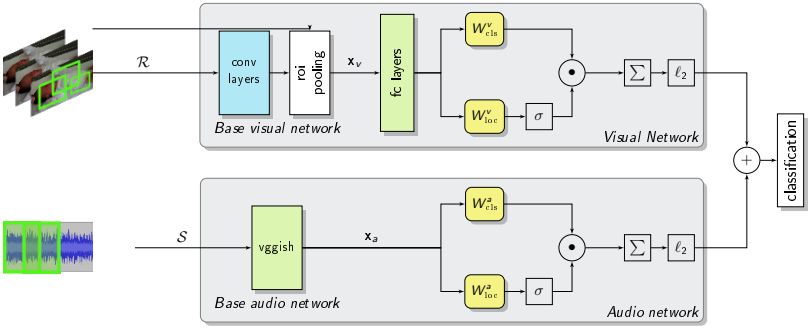
# Case study: A multiple instance learning based formulation

**Propose and Learn**

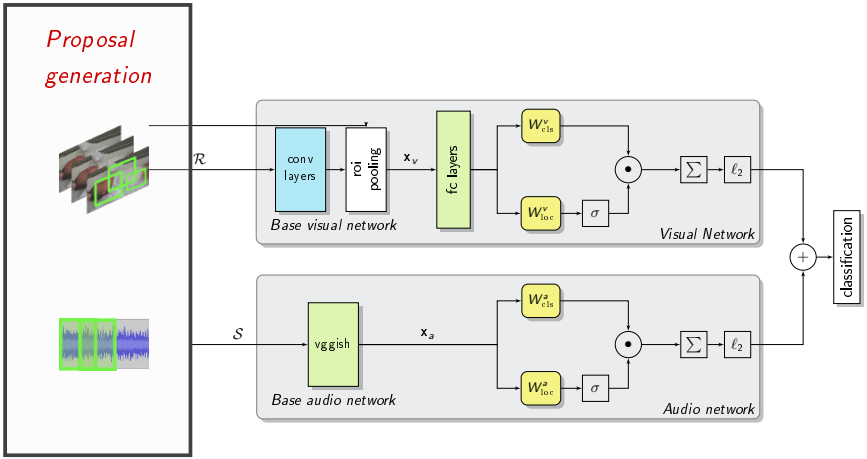
- Consider each video to be a bag of class-agnostic audio and visual proposals
- Extract features and transform them to score each according to their relevance for a particular class



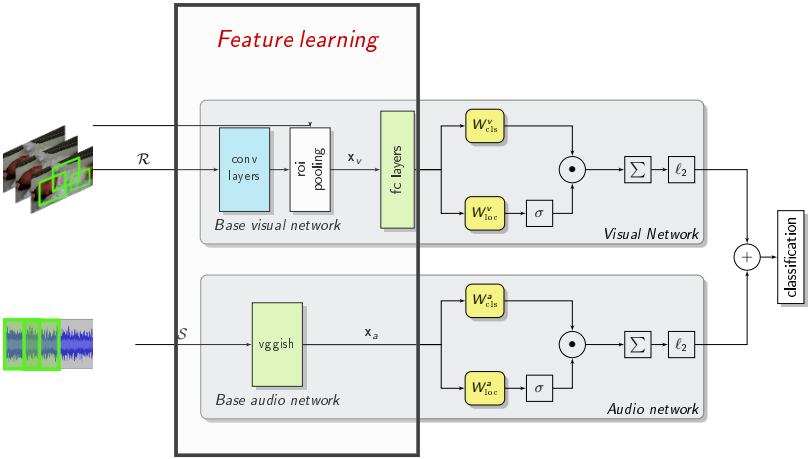
# An instantiation



# An instantiation

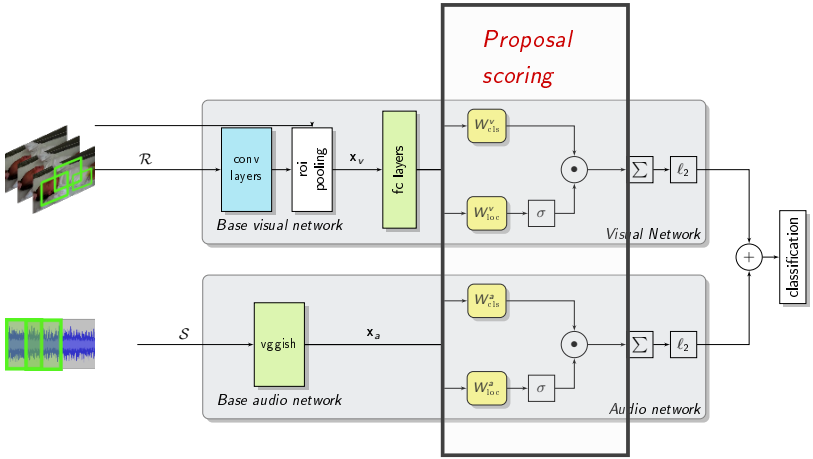


# An instantiation




- *Fine-tuning networks in green*

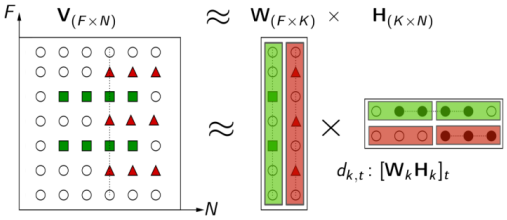
# An instantiation




- *A score for each class and proposal*

# NMF-based audio proposals

## Integrating source enhancement capability




**Source enhancement**

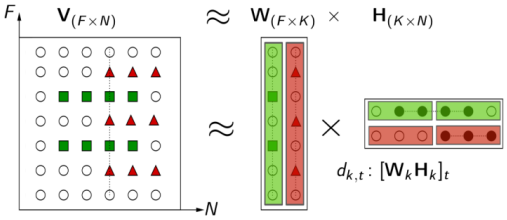
For each $d_{k,t}$ $\xrightarrow{\text{network}}$ $\beta_{k,t}$ $\xrightarrow{\text{pooling}}$ $\alpha_k$ $\xrightarrow{\text{scale}}$ $\alpha'_k$

$$\mathbf{S} = \frac{\sum_k \alpha'_k \mathbf{W}_k\mathbf{H}_k}{\mathbf{WH}}\mathbf{V}, \quad \mathbf{N} = \frac{\sum_k (1-\alpha'_k)\mathbf{W}_k\mathbf{H}_k}{\mathbf{WH}}\mathbf{V}$$

*Figure adapted from C. Fevotte*

# NMF-based audio proposals

## Integrating source enhancement capability




**Source enhancement**

For each $d_{k,t}$ $\xrightarrow{\text{network}}$ $\beta_{k,t}$ $\xrightarrow{\text{pooling}}$ $\alpha_k$ $\xrightarrow{\text{scale}}$ $\alpha'_k$

$$\mathbf{S} = \frac{\sum_k \alpha'_k \mathbf{W}_k\mathbf{H}_k}{\mathbf{WH}}\mathbf{V}, \quad \mathbf{N} = \frac{\sum_k (1-\alpha'_k)\mathbf{W}_k\mathbf{H}_k}{\mathbf{WH}}\mathbf{V}$$

*Figure adapted from C. Fevotte*

# Case study results

- Promising classification, visual localization and audio source enhancement results on Kinetics Instruments dataset
  - https://perso.telecom-paristech.fr/sparekh/ile2019_supp.html

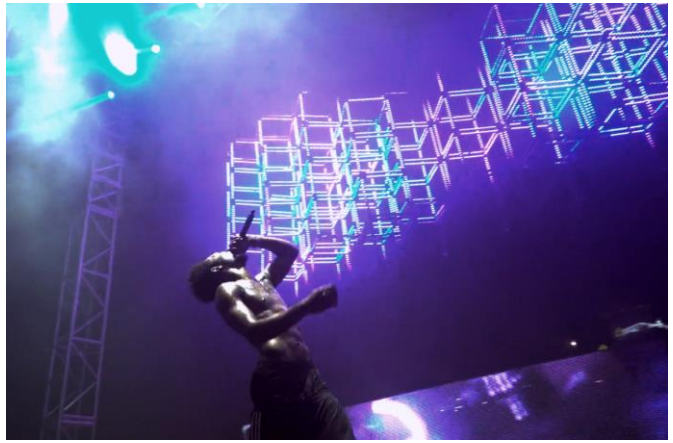
# Tutorial Outline

- Introduction
- Audiovisual Music Performance Analysis
  - Overview of Analysis Tasks
  - Audiovisual Co-Factorization for Source Separation
  - Hands-on Case Study #1: Motion Informed Audio Source Separation
- Audiovisual Content Based Classification and Retrieval
  - Genre Classification
  - Emotion Analysis
  - Cross-Modal Retrieval
  - Instrument Classification
- **Audiovisual Music Generation**
  - Hands-on Case Study #2: Skeleton Plays the Piano
- Datasets, Tools and Other Resources
- Challenges, Opportunities and Conclusions

# Audiovisual Music Generation

# Motivation

- Music visualization / Stage light control



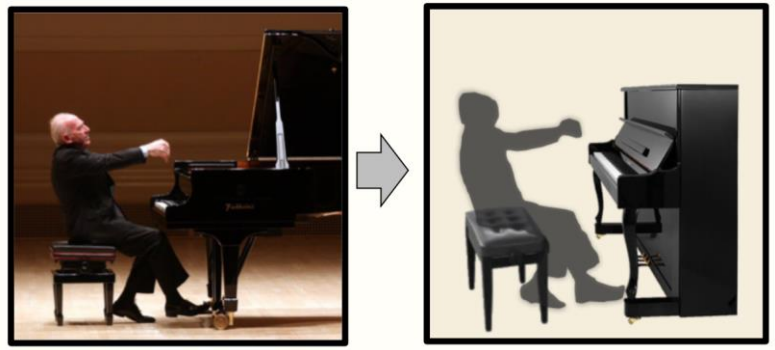
Symmetry Lab

- Education/entertainment tools

Microsoft Hololens

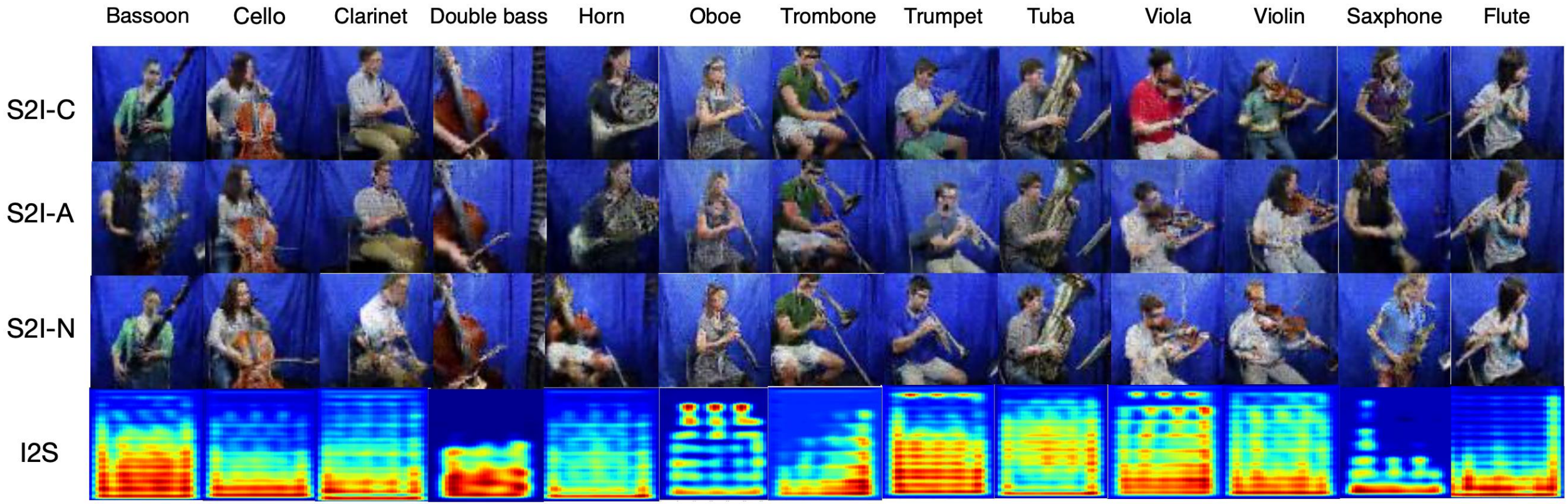
- Expressive robot ensemble

- Immersive music enjoyment



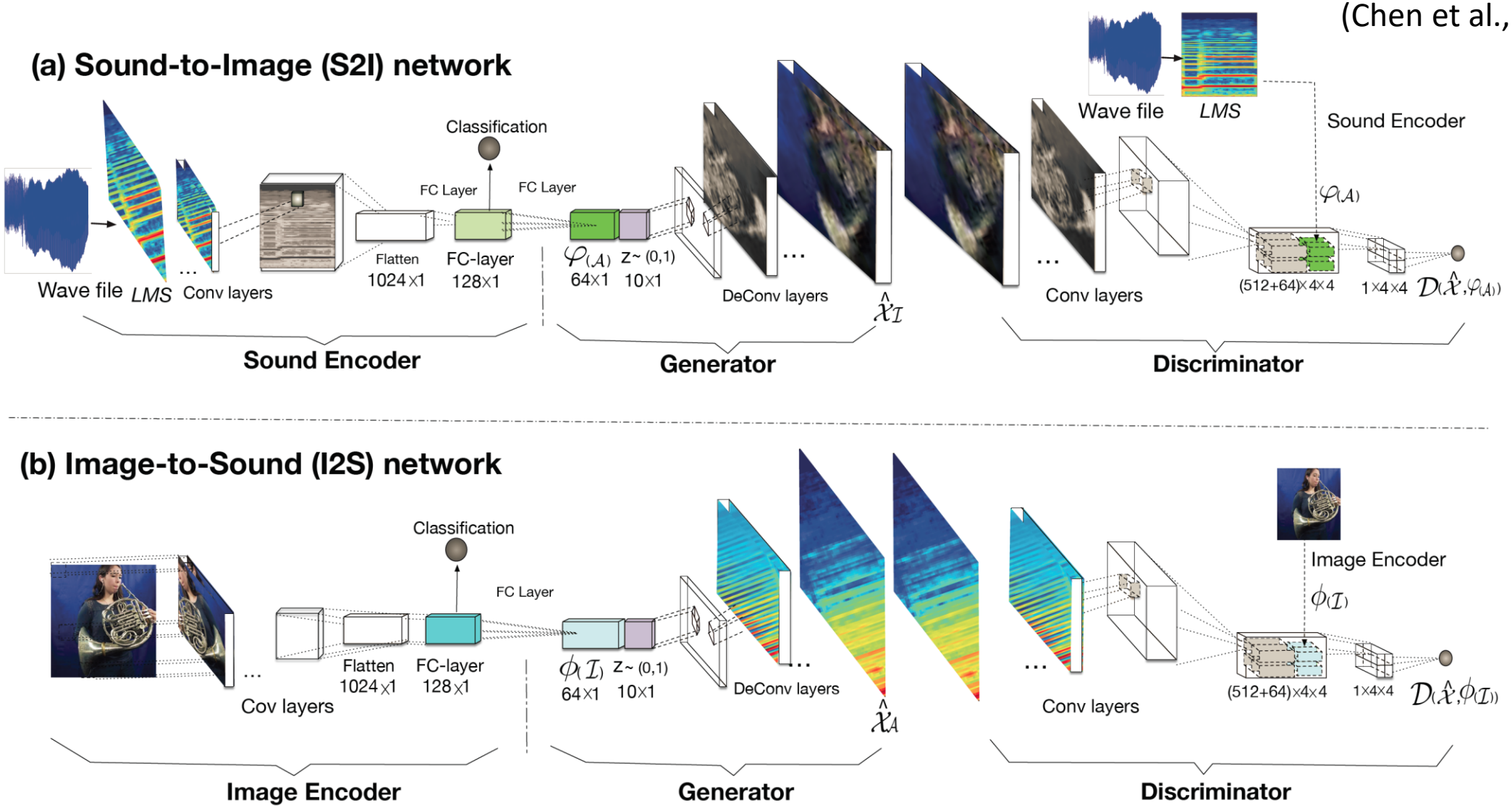
# Image vs. Sound

- Image-sound cross-modal generation for music performance (Chen et al., 2017)
  - Instrument player generation from instrument sound
  - Instrument sound generation from instrument player image
  - Focus on static image and single tone



# Model Structure

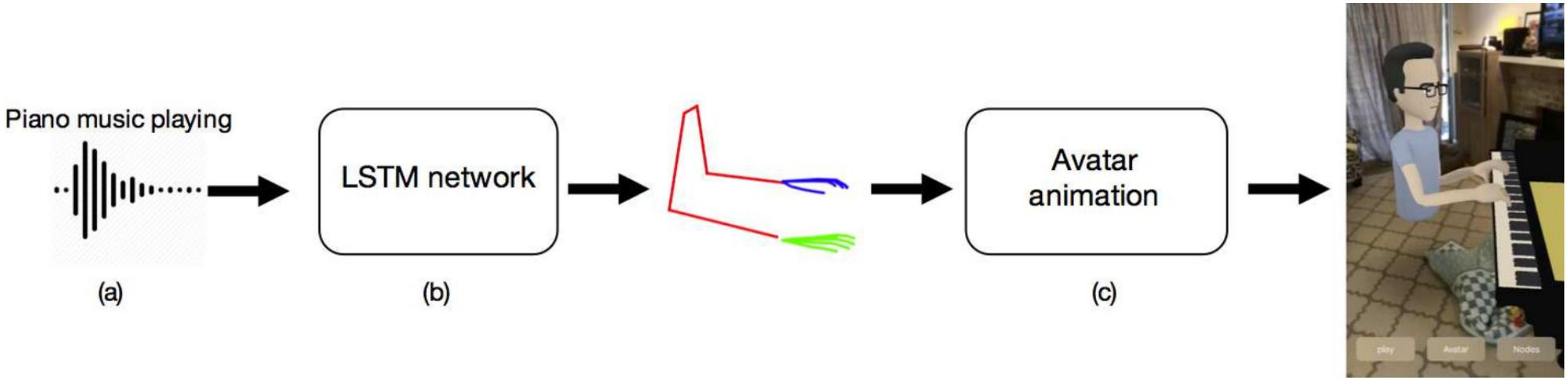
(Chen et al., 2017)



# Audio to Video

- Generate instrumentalists visual movement from music audio
  - Focus on violinist and pianist
  - Input: audio performance; Output: upper body skeleton movements
  - Generate avatar animation using a 3D body model

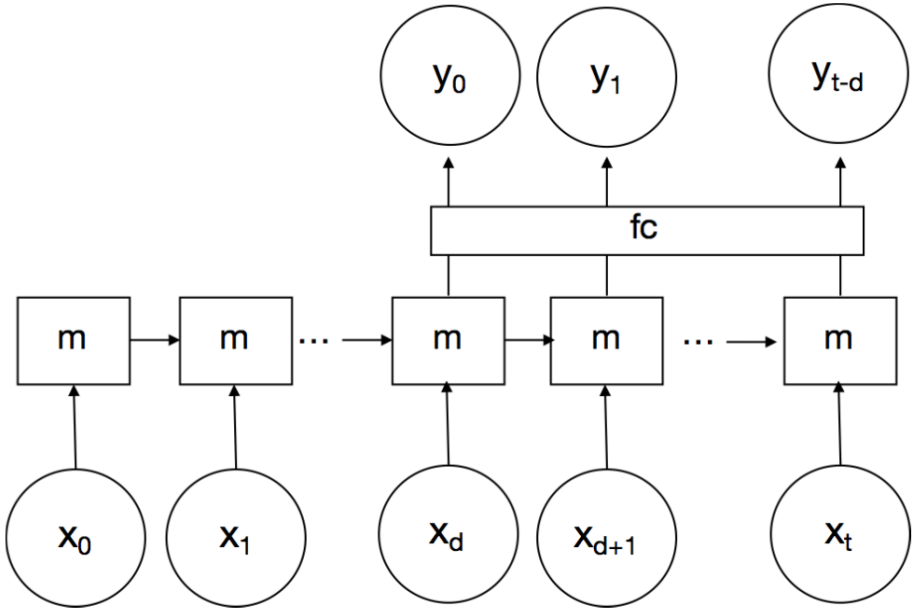
(Shlizerman et al., 2018)



# Audio to Video

- Generate instrumentalists visual movement given music audio
  - Network structure

(Shlizerman et al., 2018)

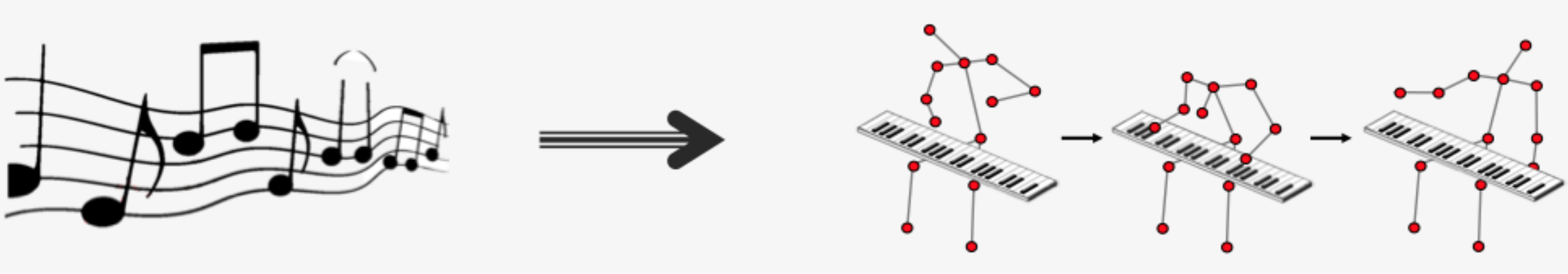



x: PCA coefficients of audio MFCC features
y: PCA coefficients of visual skeleton key points (extracted from OpenPose and MaskRCNN)

# MIDI to Video

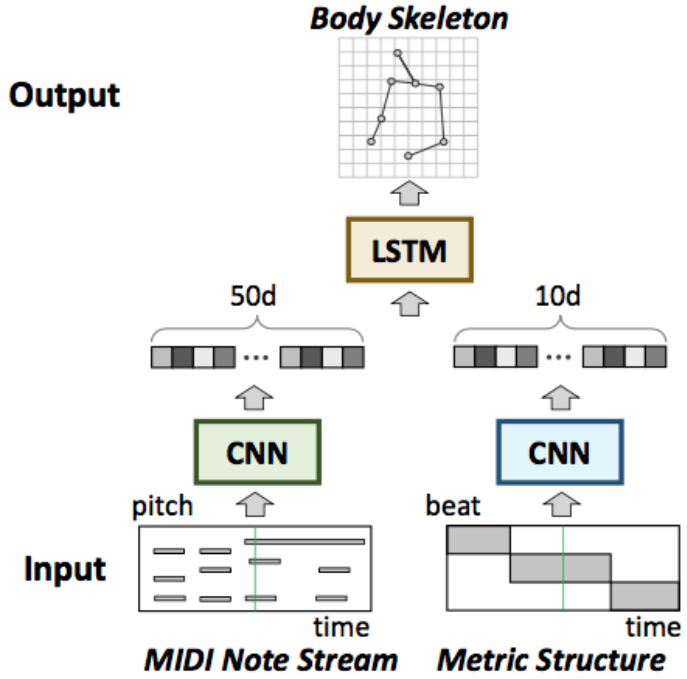
- Generate instrumentalists visual movement from MIDI performance
  - Input: MIDI (pianoroll) performance
  - Output: A sequence of coordinates of upper body skeleton points

(Li et al., 2018)

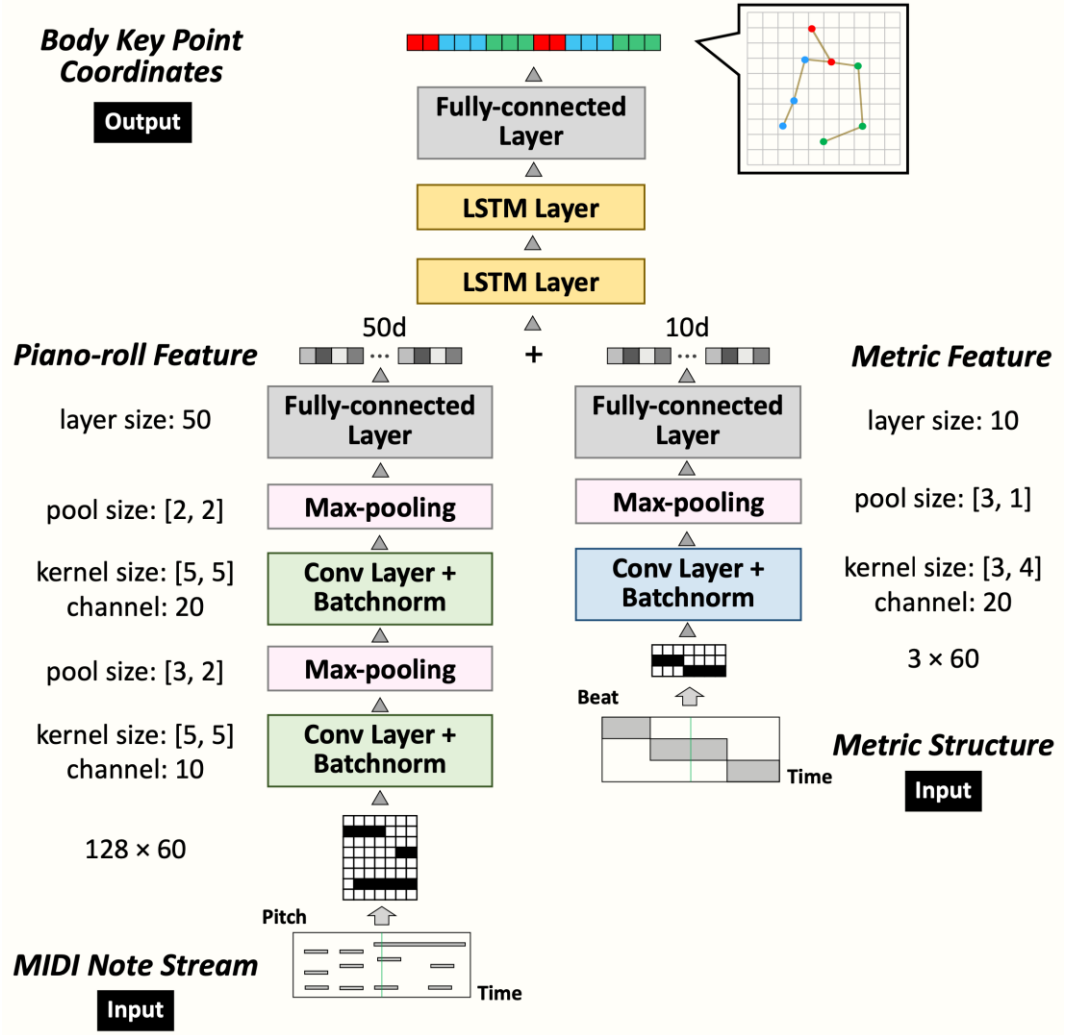


# MIDI to Video

- Generate instrumentalists visual movement given MIDI performance (Li et al., 2018)
  - Network structure



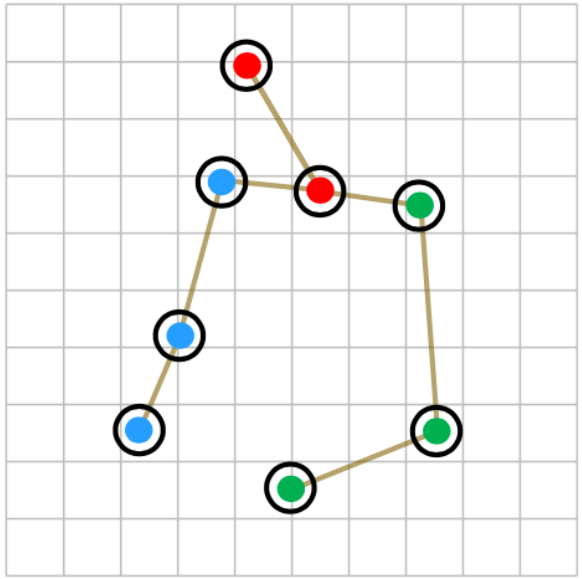

More Details



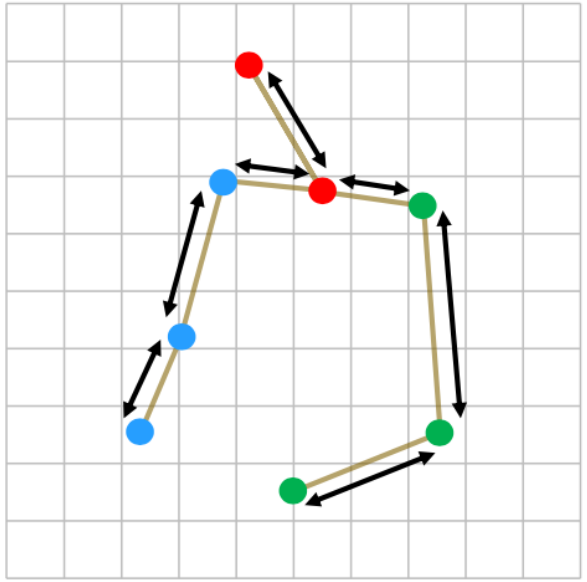
# MIDI to Video

- Generate instrumentalists visual movement given MIDI performance (Li et al., 2018)
  - Training Strategy
    - Loss function: mean absolute error
    - Joint constraint: on each individual body joint (applied in all epochs)
    - Limb constraint: ensure that relative length between adjacent joints is reasonable (applied after 30 epochs)

**Joint Constraint**



**Limb Constraint**



# MIDI to Video

- Generate instrumentalists visual movement given MIDI performance
  - Dataset collection
    - 2 players, 16 different pieces, 74 recordings

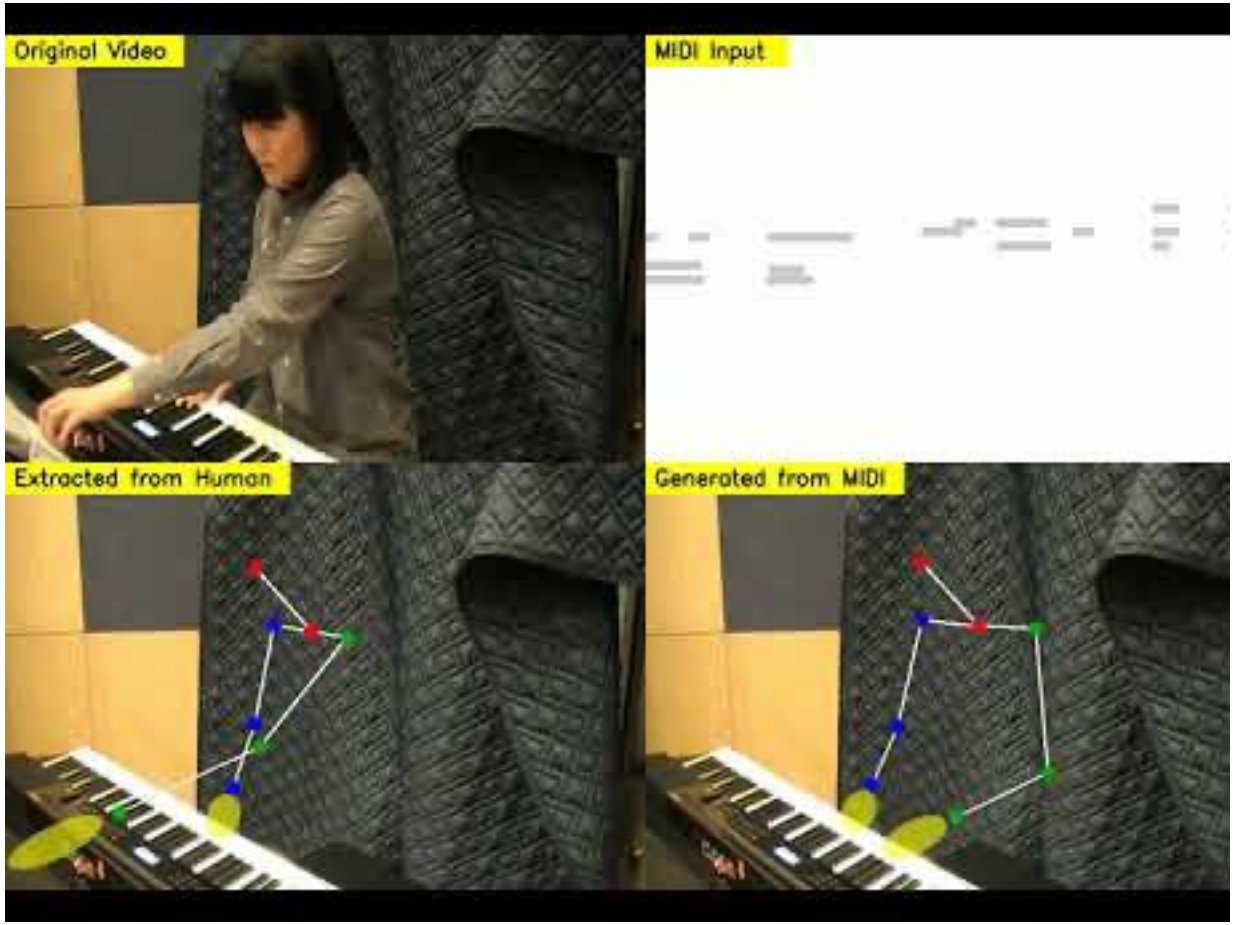
(Li et al., 2018)


Directly from MIDI keyboard

Extracted using OpenPose library

# MIDI to Video

- Generate instrumentalists visual movement given MIDI performance (Li et al., 2018)
  - Example



# Tutorial Outline

- Introduction
- Audiovisual Music Performance Analysis
  - Overview of Analysis Tasks
  - Audiovisual Co-Factorization for Source Separation
  - Hands-on Case Study #1: Motion Informed Audio Source Separation
- Audiovisual Content Based Classification and Retrieval
  - Genre Classification
  - Emotion Analysis
  - Cross-Modal Retrieval
  - Instrument Classification
- Audiovisual Music Generation
  - Hands-on Case Study #2: Skeleton Plays the Piano
- **Datasets, Tools and Other Resources**
- Challenges, Opportunities and Conclusions

# Datasets, Tools, and Resources

# Datasets about Music Performances

- Single-track

| Name | Instrument | # Pieces | Total Duration | Content |
|---|---|---|---|---|
| Multi-modal Guitar (Perez-Carrillo et al., 2015) | Guitar | 10 | 10 m | Audio, video |
| C4S (Bazzica et al., 2017) | Clarinet | 54 | 4.5 h | Audio, video, visual annotations |
| MUSIC (Zhao et al., 2018) | Multi-instrument | 685 | N/A | Audio, video (as Youtube ID) |

# Datasets about Music Performances

- Multi-track

| Name | Instrument | # Pieces | Total Duration | Content |
|---|---|---|---|---|
| ENST-Drums (Gillet and Richard, 2006) | Drum kit | N/A | 3.75 h | Audio, video (multi-camera views) |
| (Abeßer et al., 2011) | Guitar, drum, bass | N/A | 1.2 h | Audio, video (multi-camera views) |
| EEP (Marchini et al., 2014) | String quartets | 23 | N/A | Audio, Note annotation, Bow MoCap data |
| URMP (Li et al., 2018) | Multi-instrument chamber ensembles | 44 | 1.3 h | Audio, Video, Note annotation, Pitch contour |

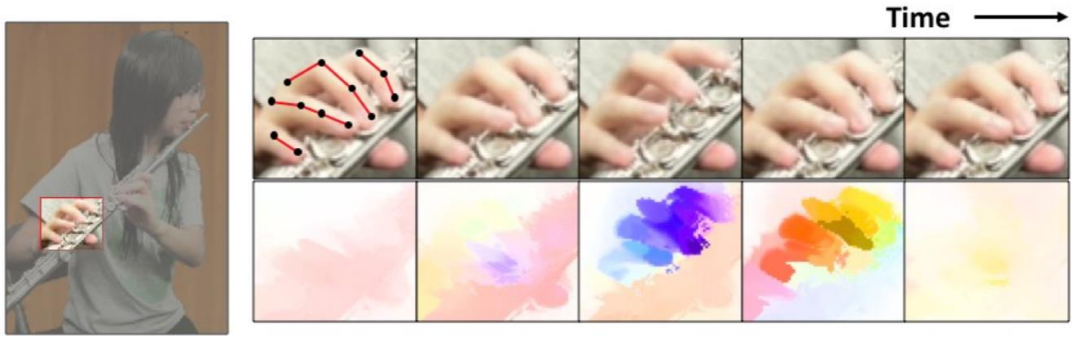
# Datasets about General Music Videos

| Name | Content | Tasks |
|---|---|---|
| (Hong et al., 2018) | Music videos (as Youtube ID) | Cross-modal retrieval |
| (Choi, 2017) | Music videos (as Youtube ID) | Cross-modal retrieval |
| MuMu (Oramas et al., 2017) | Album reviews for Million Song Dataset | Genre classification |
| RAVDESS (Livingstone and Russo, 2018) | Audiovisual recording of speech and singing, emotion labels | Emotion recognition |
| DEAP dataset (Koelstra et al., 2011) | Music videos, emotion annotations, annotators' EEG signal and face recordings | Cross-modal retrieval, multi-modal emotion analysis |

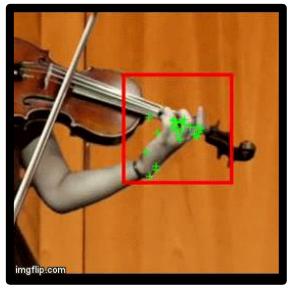
- Other datasets that contain audiovisual music subsets:
  - Audioset, Youtube-8M, Moments in Time, etc.

# Low-Level Video Feature Extractors

- Optical flow
  - Pixel-wise velocity across frames




- KLT tracker
  - Initialize bounding box for the 1st frame
  - Detect feature points
  - Track feature points



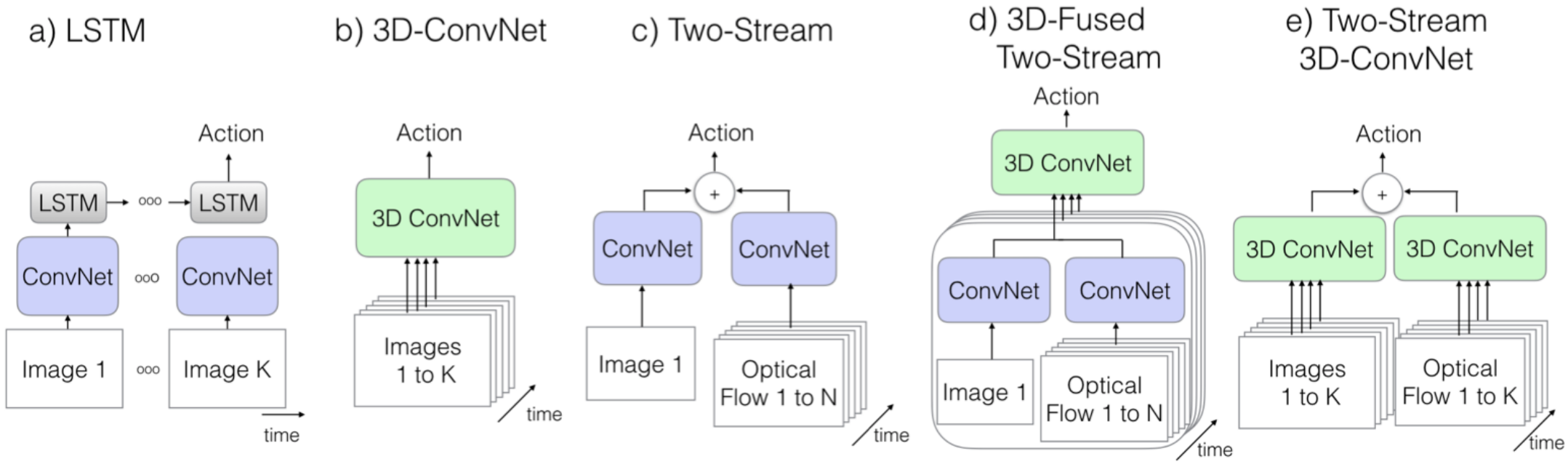
- Supervoxel (Xu et al., 2016)
  - Video temporal-spatial segmentation

# High-Level Video Feature Extractors

- Pretrained image classification models
  - ResNet (He et al., 2016)
  - VGGNet (Carreira and Zisserman, 2017)
  - DenseNet (Huang et al., 2017)
- Network architectures (Image to Video)

Pretrained video processing models

- C3D (Tran et al., 2015)
  - Pretrained on sport activity classification
- I3D (Carreira and Zisserman, 2017)
  - Pretrained on human action recognition
  - SoTA in Kinetics dataset




(Carreira and Zisserman, 2017)

# Open-Source Projects/Code

- Facebook Detectron
  - Object detection
  - Object segmentation
  - Human-object interaction recognition

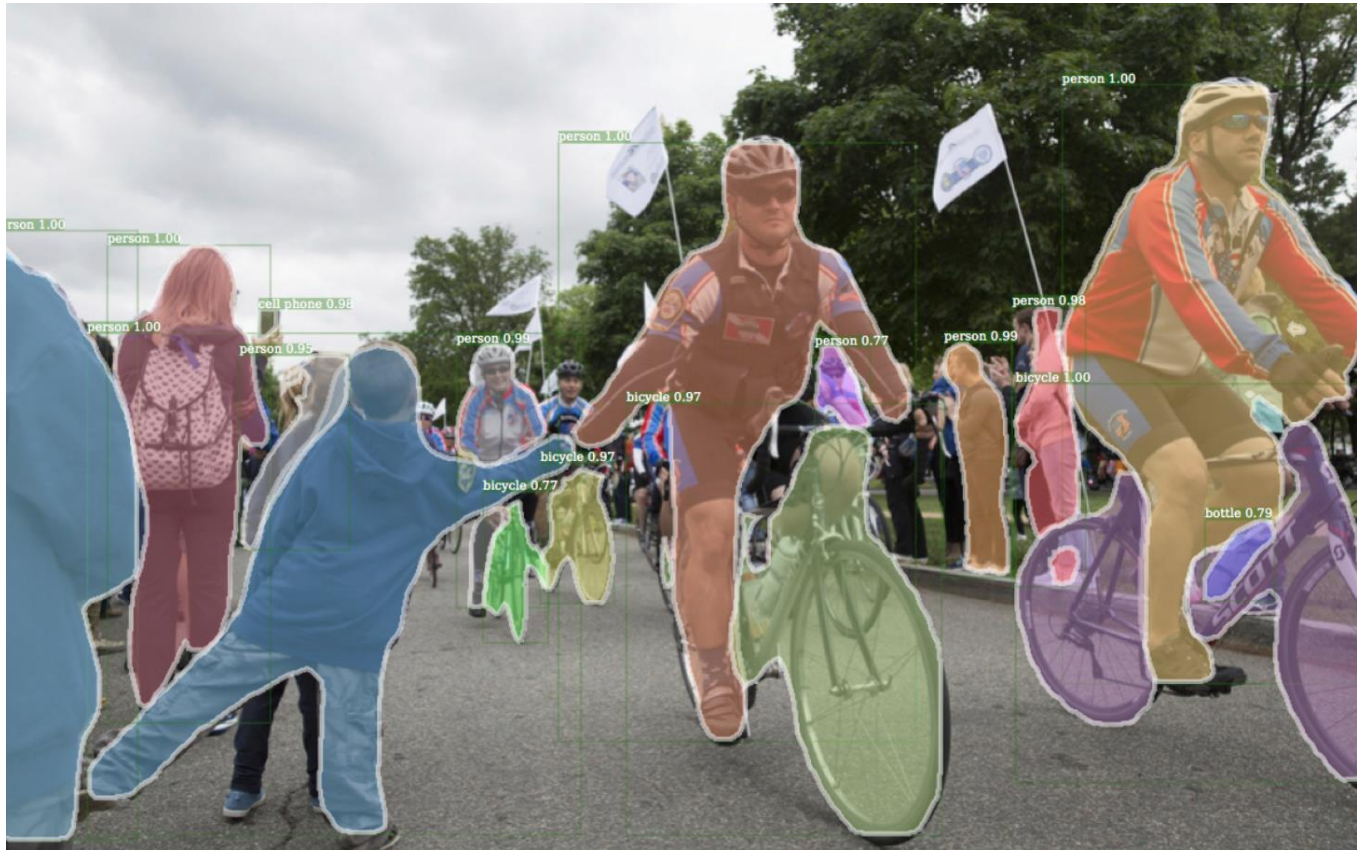



Figure from: https://research.fb.com/downloads/detectron

# Open-Source Projects/Code

- OpenPose
  - Body joint landmark detection
  - Facial landmark detection
  - Finger joint landmark detection
  - Single-person tracking

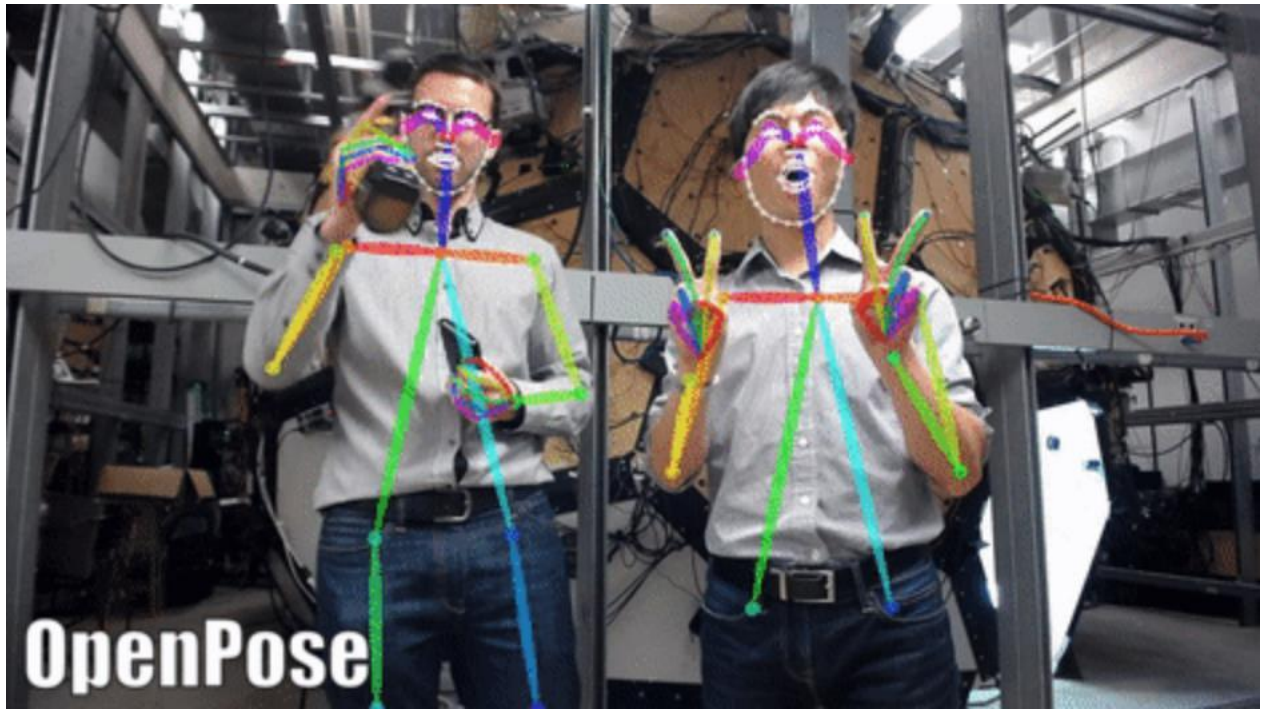



Figure from: https://github.com/CMU-Perceptual-Computing-Lab/openpose

- Introduction
- Audiovisual Music Performance Analysis
- Audiovisual Content Based Classification and Retrieval
- Audiovisual Music Generation
- Datasets, Tools and Other Resources
- Challenges, Opportunities and Conclusions

# Challenges and Opportunities

## Audio-visual correspondence

- Determining the what, where and how of fusion
  - May differ at different spatio-temporal locations!
  - Integration of static and dynamic correpondence models (Owens and Efros, 2018)
- Better modeling of AV associations
  - Often non-linear: exploring deep architectures
  - Distinguishing between cause and correlation
- Complexity of visual analysis
  - Difficult to capture and extract subtle *musically* relevant movements
    *for e.g.* playing hand movement on string instruments
    - Combining other sources of information (e.g. motion capture)

# Challenges and Opportunities

## Audio-visual correspondence

- Determining the what, where and how of fusion
  - May differ at different spatio-temporal locations!
  - Integration of static and dynamic correpondence models (Owens and Efros, 2018)
- Better modeling of AV associations
  - Often non-linear: exploring deep architectures
  - Distinguishing between cause and correlation
- Complexity of visual analysis
  - Difficult to capture and extract subtle *musically* relevant movements *for e.g.* playing hand movement on string instruments
    - Combining other sources of information (e.g. motion capture)

Cause

Correlation

# Challenges and Opportunities

## Audio-visual correspondence

- Determining the what, where and how of fusion
  - May differ at different spatio-temporal locations!
  - Integration of static and dynamic correpondence models (Owens and Efros, 2018)
- Better modeling of AV associations
  - Often non-linear: exploring deep architectures
  - Distinguishing between cause and correlation
- Complexity of visual analysis
  - Difficult to capture and extract subtle *musically* relevant movements *for e.g.* playing hand movement on string instruments
    - Combining other sources of information (e.g. motion capture)

# Challenges and Opportunities

## Data

- Scarcity of large-scale clean, annotated datasets
  - Amateurish quality and unconstrained nature of user–generated videos
    - Adapt to changing camera view-points, illumination
  - Artificially edited content (video frames unrelated to audio etc.)
    - Flexibly tackle intermittent or long range correspondences
  - Lack of annotations
    - Towards unsupervised, weakly supervised and semi-supervised methods
    - An opportunity to develop robust annotation tools
  - Small data size
    - Transfer learning based approaches (Aytar et al., 2016)

# Challenges and Opportunities

## Data

- Scarcity of large-scale clean, annotated datasets
  - Amateurish quality and unconstrained nature of user–generated videos
    - Adapt to changing camera view-points, illumination
  - Artificially edited content (video frames unrelated to audio etc.)
    - Flexibly tackle intermittent or long range correspondences
  - Lack of annotations
    - Towards unsupervised, weakly supervised and semi-supervised methods
    - An opportunity to develop robust annotation tools
  - Small data size
    - Transfer learning based approaches (Aytar et al., 2016)

# Challenges and Opportunities

## Data

- Scarcity of large-scale clean, annotated datasets
  - Amateurish quality and unconstrained nature of user–generated videos
    - Adapt to changing camera view-points, illumination
  - Artificially edited content (video frames unrelated to audio etc.)
    - Flexibly tackle intermittent or long range correspondences
  - Lack of annotations
    - Towards unsupervised, weakly supervised and semi-supervised methods
    - An opportunity to develop robust annotation tools
  - Small data size
    - Transfer learning based approaches (Aytar et al., 2016)

# Challenges and Opportunities

## Data

- Scarcity of large-scale clean, annotated datasets
  - Amateurish quality and unconstrained nature of user–generated videos
    - Adapt to changing camera view-points, illumination
  - Artificially edited content (video frames unrelated to audio etc.)
    - Flexibly tackle intermittent or long range correspondences
  - Lack of annotations
    - Towards unsupervised, weakly supervised and semi-supervised methods
    - An opportunity to develop robust annotation tools
  - Small data size
    - Transfer learning based approaches (Aytar et al., 2016)

# Challenges and Opportunities

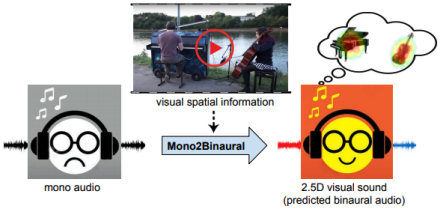
## Tasks and applications

- Audio-visual music synthesis
  - An end goal for artists and users (Owens et al., 2016; Shlizerman et al., 2018; Li et al., 2018b; Zhou et al., 2018)
    - Interactive machine learning (Fiebrink et al., 2016)
    - Assisting those with vision or hearing impairment
  - A tool for analysis
    - Requires learning meaningful scene representations (Owens et al., 2016)
- Enhancing music education and production
  - Movement, posture and playing style analysis
  - Assiting with remixing, transcription etc.

# Challenges and Opportunities

## Tasks and applications

- Audio-visual music synthesis
  - An end goal for artists and users (Owens et al., 2016; Shlizerman et al., 2018; Li et al., 2018b; Zhou et al., 2018)
    - Interactive machine learning (Fiebrink et al., 2016)
    - Assisting those with vision or hearing impairment
  - A tool for analysis
    - Requires learning meaningful scene representations (Owens et al., 2016)
- Enhancing music education and production
  - Movement, posture and playing style analysis
  - Assiting with remixing, transcription etc.




(Gao and Grauman, 2019a)

# Credits

- L. Chen
- K. Dinesh
- A. Kumar
- C. Liem
- X. Liu
- A. Maezawa
- G. Sharma
- C. Xu
- A. Ozerov
- G. Richard
- N. Q. K. Duong
- P. Perez
- C. Févotte
- N. Seichepine
- O. Cappé

# Bibliography I


J. Abeßer, O. Lartillot, C. Dittmar, T. Eerola, and G. Schuller. Modeling musical attributes to characterize ensemble recordings using rhythmic audio features. In *Proc. IEEE Intl. Conf. Acoustics, Speech and Sig. Process. (ICASSP)*, pages 189–192, 2011.

G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *Proc. of International Conference on Machine Learning*, pages 1247–1255, 2013.

R. Arandjelović and A. Zisserman. Look, listen and learn. In *IEEE International Conference on Computer Vision*, 2017.

R. Arandjelovic and A. Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–451, 2018.

Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, pages 892–900, 2016.

Z. Barzelay and Y. Y. Schechner. Harmony in motion. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55, 2015.

A. Bazzica, C. C. S. Liem, and A. Hanjalic. On detecting the playing/non-playing activity of musicians in symphonic music videos. *Computer Vision and Image Understanding*, 144:188–204, 2016.

A. Bazzica, J. van Gemert, C. Liem, and A. Hanjalic. Vision-based detection of acoustic timed events: a case study on clarinet note onsets. *arXiv preprint arXiv:1706.09556*, 2017.

M. G. Boltz, B. Ebendorf, and B. Field. Audiovisual interactions: The impact of visual information on music perception and memory. *Music Perception: An Interdisciplinary Journal*, 27(1):43–59, 2009.

A.-M. Burns and M. M. Wanderley. Visual methods for the retrieval of guitarist fingering. In *Proc. International Conference on New Interfaces for Musical Expression (NIME)*, 2006.

# Bibliography II


J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

J. Chao, H. Wang, W. Zhou, W. Zhang, and Y. Yu. Tunesensor: A semantic-driven music recommendation service for digital photo albums. In *Proceedings of the International Semantic Web Conference (ISWC)*, 2011.

L. Chen, S. Srivastava, Z. Duan, and C. Xu. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 349–357. ACM, 2017.

E. Coutinho and K. R. Scherer. The effect of context and audio-visual modality on emotions elicited by a musical performance. *Psychology of music*, 45(4):550–569, 2017.

A. S. Cowen and D. Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *National Academy of Sciences*, 114(38):E7900–E7909, 2017.

K. Dinesh, B. Li, X. Liu, Z. Duan, and G. Sharma. Visually informed multi-pitch analysis of string ensembles. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3021–3025, 2017.

A. Dremeau and S. Essid. Probabilistic dance performance alignment by fusion of multimodal features. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013.

S. Essid, X. Lin, M. Gowing, G. Kordelas, A. Aksay, P. Kelly, T. Fillon, Q. Zhang, A. Dielmann, V. Kitanovski, R. Tournemenne, A. Masurelle, E. Izquierdo, N. E. O'Connor, P. Daras, and R. G. A multi-modal dance corpus for research into interaction between humans in virtual environments. *Journal on Multimodal User Interfaces: Special issue on multimodal corpora*, 2012.

R. Fiebrink, B. Caramiaux, R. Dean, and A. McLean. *The machine learning algorithm as creative musical tool*. Oxford University Press, 2016.

J. Fisher, T. Darrell, W. T. Freeman, P. Viola, and J. W. Fisher III. Learning Joint Statistical Models for Audio-Visual Fusion and Segregation. In *Advances in Neural Information Processing Systems*, pages 772–778, 2001.

R. Gao and K. Grauman. 2.5 d visual sound. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2019a.

# Bibliography III


R. Gao and K. Grauman. Co-separating sounds of visual objects. *arXiv preprint arXiv:1904.07750*, 2019b.

R. Gao, R. Feris, and K. Grauman. Learning to separate object sounds by watching unlabeled video. In *ECCV*, September 2018.

J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017.

O. Gillet and G. Richard. Automatic transcription of drum sequences using audiovisual features. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

O. Gillet and G. Richard. ENST-Drums: an extensive audio-visual database for drum signals processing. In *Proc. Intl. Soc. for Music Info. Retrieval (ISMIR)*, pages 156–159, 2006.

O. Gillet, S. Essid, and G. Richard. On the correlation of automatic audio and visual segmentations of music videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(3):347–355, 2007.

D. Gorodnichy and A. Yogeswaran. Detection and tracking of pianist hands and fingers. In *Proc. Canadian Conference on Computer and Robot Vision*, 2006.

S. Gross, X. Wei, and J. Zhu. Automatic realistic music video generation from segments of youtube videos. *arXiv preprint arXiv:1905.12245*, 2019.

A. Hanjalic and L.-Q. Xu. Affective video content representation and modeling. *IEEE Transactions Multimedia*, 7(1):143–154, 2005.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

S. Hong, W. Im, and H. S. Yang. CBVMR: Content-based video-music retrieval using soft intra-modal structure constraint. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, pages 353–361, 2018.

X. Hu and J. S. Downie. Exploring mood metadata: Relationships with genre, artist and usage metadata. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 67–72, 2007.

# Bibliography IV


G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

P. N. Juslin and P. Laukka. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, 33(3):217–238, 2004.

S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski, et al. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 10(2):99–111, 2016.

H. Kaya, F. Gürpınar, and A. A. Salah. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing*, 65:66–75, 2017.

C. Kerdvibulvech and H. Saito. Vision-based guitarist fingering tracking using a Bayesian classifier and particle filters. In *Advances in Image and Video Tech.*, pages 625–638. Springer, 2007.

M. Keuper, B. Andres, and T. Brox. Motion trajectory segmentation via minimum cost multicuts. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3271–3279, 2015.

E. Kidron, Y. Schechner, and M. Elad. Pixels that sound. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 88–95 vol. 1, June 2005. doi: 10.1109/CVPR.2005.274.

S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

B. Li and A. Kumar. Query by video: Cross-modal music retrieval. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

B. Li, K. Dinesh, Z. Duan, and G. Sharma. See and listen: Score-informed association of sound tracks to players in chamber music performance videos. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2906–2910, 2017a.

B. Li, K. Dinesh, G. Sharma, and Z. Duan. Video-based vibrato detection and analysis for polyphonic string music. In *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, pages 123–130, 2017b.

# Bibliography V


B. Li, C. Xu, and Z. Duan. Audiovisual source association for string ensembles through multi-modal vibrato analysis. In *Proc. Sound and Music Computing (SMC)*, 2017c.

B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma. Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Transactions on Multimedia*, 21(2):522–535, 2018a.

B. Li, A. Maezawa, and Z. Duan. Skeleton plays piano: Online generation of pianist body movements from midi performance. In *ISMIR*, pages 218–224, 2018b.

B. Li, K. Dinesh, C. Xu, G. Sharma, and Z. Duan. Online audio-visual source association for chamber music performances. *Transactions of the International Society for Music Information Retrieval*, 2(1), 2019.

J. Libeks and D. Turnbull. You can judge an artist by an album cover: Using images for music annotation. *IEEE MultiMedia*, 18(4):30–37, 2011.

C. Liem, A. Bazzica, and A. Hanjalic. Musesync: standing on the shoulders of hollywood. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1383–1384. ACM, 2012.

J.-C. Lin, W.-L. Wei, and H.-M. Wang. EMV-matchmaker: emotional temporal course modeling and matching for automatic music video generation. In *Proc. ACM International Conference on Multimedia*, pages 899–902, 2015.

S. R. Livingstone and F. A. Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

L. Lu, D. Liu, and H.-J. Zhang. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on audio, speech, and language processing*, 14(1):5–18, 2006.

M. Marchini, R. Ramirez, P. Papiotis, and E. Maestre. The sense of ensemble: a machine learning approach to expressive performance modelling in string quartets. *J. of New Music Res.*, 43(3):303–317, 2014.

K. McGuinness, O. Gillet, N. E. O'Connor, and G. Richard. Visual analysis for drum sequence transcription. In *Proc. IEEE European Signal Processing Conference*, pages 312–316, 2007.

# Bibliography VI


J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proc. of International Conference on Machine Learning*, pages 689–696, 2011.

A. Oka and M. Hashimoto. Marker-less piano fingering recognition using sequential depth images. In *Proc. Korea-Japan Joint Workshop on Frontiers of Comp. Vision (FCV)*, 2013.

S. Oramas, O. Nieto, F. Barbieri, and X. Serra. Multi-label music genre classification from audio, text, and images using deep features. In *Proc. International Society for Music Information Retrieval (ISMIR)*, 2017.

S. Oramas, F. Barbieri, O. Nieto, and X. Serra. Multimodal deep learning for music genre classification. *Transactions of the International Society for Music Information Retrieval.*, 1(1):4–21, 2018.

A. Owens and A. A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.

A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman. Visually indicated sounds. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2405–2413, 2016.

M. Paleari, B. Huet, A. Schutz, and D. Slock. A multimodal approach to music transcription. In *Proc. International Conference on Image Processing (ICIP)*, 2008.

S. Parekh, S. Essid, A. Ozerov, N. Q. K. Duong, P. Perez, and G. Richard. Guiding audio source separation by video object information. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017a.

S. Parekh, S. Essid, A. Ozerov, N. Q. K. Duong, P. Perez, and G. Richard. Motion informed audio source separation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017b.

S. Parekh, A. Ozerov, S. Essid, N. Duong, P. Pérez, and G. Richard. Identify, locate and separate: Audio-visual object extraction in large video collections using weak supervision. In *WASPAA*, October 2019.

R. Parke, E. Chew, and C. Kyriakakis. Quantitative and visual analysis of the impact of music on perceived emotion of film. *Computers in Entertainment (CIE)*, 5(3):5, 2007.

# Bibliography VII


A. Perez-Carrillo, J.-L. Arcos, and M. Wanderley. Estimation of guitar fingering and plucking controls based on multimodal analysis of motion, audio and musical score. In *Proc. Intl. Symp. Comput. Music Multidisciplinary Res. (CMMR)*, pages 71–87, 2015.

F. Platz and R. Kopiez. When the eye listens: A meta-analysis of how audio-visual presentation enhances the appreciation of music performance. *Music Perception: An Interdisciplinary Journal*, 30(1):71–83, 2012.

Á. Sarasúa and E. Guaus. Beat tracking from conducting gestural data: a multi-subject study. In *Proc. ACM International Workshop on Movement and Computing*, page 118, 2014.

S. Sasaki, T. Hirai, H. Ohya, and S. Morishima. Affective music recommendation system based on the mood of input video. In *International Conference on Multimedia Modeling*, pages 299–302. Springer, 2015.

J. Scarr and R. Green. Retrieval of guitarist fingering information using computer vision. In *Proc. International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 2010.

M. Schedl, N. Orio, C. Liem, and G. Peeters. A professionally annotated and enriched multimodal data set on popular music. In *Proceedings of the 4th ACM Multimedia Systems Conference*, pages 78–83. ACM, 2013.

A. Schindler and A. Rauber. An audio-visual approach to music genre classification through affective color features. In *European Conference on Information Retrieval*, pages 61–67. Springer, 2015.

E. M. Schmidt and Y. E. Kim. Modeling and predicting emotion in music. *emotion*, 5:6, 2012.

F. Sedighin, M. Babaie-Zadeh, B. Rivet, and C. Jutten. Multimodal Soft Nonnegative Matrix Co-Factorization for Convolutive Source Separation. *IEEE Transactions on Signal Processing*, 65(12):3179–3190, jun 2017. ISSN 1053-587X. doi: 10.1109/TSP.2017.2679692. URL http://ieeexplore.ieee.org/document/7874217/.

N. Seichepine, S. Essid, C. Fevotte, and O. Cappe. Soft nonnegative matrix co-factorization. *IEEE Transactions on Signal Processing*, PP(99), 2014. doi: 10.1109/TSP.2014.2360141.

R. R. Shah, Y. Yu, and R. Zimmermann. ADVISOR: Personalized video soundtrack recommendation by late fusion with heuristic rankings. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 607–616. ACM, 2014.

# Bibliography VIII


K.-H. Shin and I.-K. Lee. Music synchronization with video using emotion similarity. In *Proc. International Conference on Big Data and Smart Computing (BigComp)*, pages 47–50, 2017.

T. Shiratori, A. Nakazawa, and K. Ikeuchi. Detecting dance motion structure through music analysis. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 857–862. IEEE, 2004.

E. Shlizerman, L. Dery, H. Schoen, and I. Kemelmacher-Shlizerman. Audio to body dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7574–7583, 2018.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

J. Skowronek, M. F. McKinney, and S. Van De Par. A demonstrator for automatic music mood estimation. In *ISMIR*, pages 345–346, 2007.

O. Slizovskaia, E. Gómez, and G. Haro. Musical instrument recognition in user-generated videos using a multimodal convolutional neural network architecture. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 226–232. ACM, 2017.

P. Smaragdis and M. Casey. Audio/visual independent components. In *Proc. of ICA*, pages 709–714, 2003.

M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang. 1000 songs for emotional analysis of music. In *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*, pages 1–6. ACM, 2013.

D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

C.-J. Tsay. Sight over sound in the judgment of music performance. *National Academy of Sciences*, 110(36):14580–14585, 2013.

J.-C. Wang, Y.-H. Yang, I.-H. Jhuo, Y.-Y. Lin, H.-M. Wang, et al. The acousticvisual emotion guassians model for automatic generation of music video. In *Proc. ACM international conference on Multimedia*, pages 1379–1380, 2012.

J.-C. Wang, H.-M. Wang, and G. Lanckriet. A histogram density modeling approach to music emotion recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 698–702. IEEE, 2015.

# Bibliography IX


T.-L. Wu and S.-K. Jeng. Automatic emotion classification of musical segments. In *Proceedings of the 9th International Conference on Music Perception & Cognition, Bologna*, 2006.

X. Wu, Y. Qiao, X. Wang, and X. Tang. Bridging music and image via cross-modal ranking analysis. *IEEE Transactions on Multimedia*, 18(7): 1305–1318, 2016.

B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal. Video emotion recognition with transferred deep feature encodings. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 15–22. ACM, 2016.

Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen. A regression approach to music emotion recognition. *IEEE Transactions on audio, speech, and language processing*, 16(2):448–457, 2008.

A. Yazdani, K. Kappeler, and T. Ebrahimi. Affective content analysis of music video clips. In *Proceedings of the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pages 7–12. ACM, 2011.

J.-C. Yoon, I.-K. Lee, and S. Byun. Automated music video generation using multi-level feature-based segmentation. *Multimedia Tools and Applications*, 41(2):197, 2009.

Y. Yu, Z. Shen, and R. Zimmermann. Automatic music soundtrack generation for outdoor videos from contextual sensor information. In *Proc. ACM international conference on Multimedia*, pages 1377–1378, 2012.

B. Zhang and Y. Wang. Automatic music transcription using audio-visual fusion for violin practice in home environment. Technical Report TRA7/09, The National University of Singapore, 2009.

S. Zhang, Q. Huang, S. Jiang, W. Gao, and Q. Tian. Affective visualization and retrieval for music video. *IEEE Transactions on Multimedia*, 12(6): 510–522, 2010.

H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba. The sound of pixels. In *The European Conference on Computer Vision (ECCV)*, September 2018.

H. Zhao, C. Gan, W.-C. Ma, and A. Torralba. The sound of motions. In *Proc. International Conference on Computer Vision (ICCV)*, October 2019.

# Bibliography X

S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun. Exploring principles-of-art features for image emotion recognition. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 47–56. ACM, 2014.

C. Zhen and J. Xu. Multi-modal music genre classification approach. In *Proceedings of the International Conference on Computer Science and Information Technology*, volume 8, pages 398–402, 2010.

Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg. Visual to sound: Generating natural sound for videos in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

# Soft matrix co-factorisation

(Seichepine et al., 2014)

Solve the problem:

$$\min_{\mathbf{W}_1,\mathbf{H}_1,\mathbf{W}_2,\mathbf{H}_2} D_1(\mathbf{V}_1 \,|\, \mathbf{W}_1\mathbf{H}_1) + \beta_2 D_2(\mathbf{V}_2 \,|\, \mathbf{W}_2\mathbf{H}_2) + \beta_c \|\mathbf{H}_1 - \mathbf{H}_2\|_p$$

- **Remarks:**
  - the initial formulation cannot lead to sensible solutions:
    - there is a scale ambiguity on $\mathbf{W}_1\mathbf{H}_1$ and $\mathbf{W}_2\mathbf{H}_2$;

    → the problem needs to be rescaled
    - $\mathbf{H}_1$ and $\mathbf{H}_2$ are not necessarily directly comparable

    → introduce scaling matrix $\mathbf{S}$

# Soft matrix co-factorisation

(Seichepine et al., 2014)

Solve the problem:

$$\min_{\mathbf{W}_m,\mathbf{H}_m} D_1(\mathbf{V}_1 \,|\, \mathbf{W}_1\mathbf{H}_1) + \beta_2 D_2(\mathbf{V}_2 \,|\, \mathbf{W}_2\mathbf{H}_2) + \beta_c \left\| \mathbf{\Lambda}_1\mathbf{H}_1 - \mathbf{\Lambda}_2\mathbf{H}_2 \right\|_p$$

$$\mathbf{\Lambda}_1 = \mathrm{diag}(\lambda_{m,1}, \cdots, \lambda_{m,K}) \;;\; \lambda_{m,k} = \textstyle\sum_f w_{m,fk} \;;\; m \in \{1,2\}$$

- **Remarks:**
  - the initial formulation cannot lead to sensible solutions:
    - there is a scale ambiguity on $\mathbf{W}_1\mathbf{H}_1$ and $\mathbf{W}_2\mathbf{H}_2$;

    $\rightarrow$ the problem needs to be rescaled
    - $\mathbf{H}_1$ and $\mathbf{H}_2$ are not necessarily directly comparable

    $\rightarrow$ introduce scaling matrix $\mathbf{S}$

# Soft matrix co-factorisation

(Seichepine et al., 2014)

Solve the problem:

$$\min_{\mathbf{W}_m,\mathbf{H}_m,\mathbf{S}} D_1(\mathbf{V}_1 \,|\, \mathbf{W}_1\mathbf{H}_1) + \beta_2 D_2(\mathbf{V}_2 \,|\, \mathbf{W}_2\mathbf{H}_2) + \beta_c \left\|\mathbf{\Lambda}_1\mathbf{H}_1 - \mathbf{S}\mathbf{\Lambda}_2\mathbf{H}_2\right\|_p$$

$$\mathbf{\Lambda}_1 = \mathrm{diag}(\lambda_{m,1}, \cdots, \lambda_{m,K}) \;;\; \lambda_{m,k} = \textstyle\sum_f w_{m,fk} \;;\; m \in \{1,2\}$$

- **Remarks:**
  - the initial formulation cannot lead to sensible solutions:
    - there is a scale ambiguity on $\mathbf{W}_1\mathbf{H}_1$ and $\mathbf{W}_2\mathbf{H}_2$;

    $\rightarrow$ the problem needs to be rescaled
    - $\mathbf{H}_1$ and $\mathbf{H}_2$ are not necessarily directly comparable

    $\rightarrow$ introduce scaling matrix $\mathbf{S}$

# Behavior of the soft co-factorisation

## Using synthetic data