Piano Skills Assessment

Paritosh Parmar | Jaiden Reddy | Brendan Morris

Driving goals & questions…

1. Automating piano performance evaluation
   Can a computer assess piano player’s skill level?

2. Multimodal (visual + audio) analysis
   Combining both modes → More accurate evaluation?
   What can be learnt from audio & video information (each)?

3. Efficient performance sampling *
   How to sample clips to best reflect player performance?
   ** Long videos are difficult for CNNs to process*

Related work…

- Use of computer vision for the following:
  - Determining the accuracy of pianists
  - Automatic transcription
  - Generating audio from video/silent performance
  - Generating pianist’s body pose from audio/MIDI

- Skill assessment / action quality assessment

Novelty…

No work previously for…

- Automated evaluation of pianist’s skill level from performance
- *Multimodal* skill assessment / action quality assessment

Necessary ingredients…

1. Data to enable multimodal analysis

2. Architecture to learn evaluation such that
    a. *Data for each mode can be processed*
    b. *Features from both modes can be combined*

3. Experimental framework to test method’s efficacy

STEP 1: Training data gathering

Data acquisition…

- Expert-compiled 61 performance videos from YouTube
  - *Training set* : 31 (516 samples)
  - *Testing set* : 30 (476 samples)

- Annotations made for each performance:
  - *Player skill level*
  - *Song difficulty level*
  - *Name of the song*
  - *Bounding box around pianist's hands*

Closer look at key scored attributes…

- Player skill
  - 10 point scale (10 highest)
  - Based on a technical & repertoire syllabus by MTNA *

    ** Music Teachers National Association*
  - *Determined for dataset by trained pianist*

- Music difficulty
  - 10 point scale (10 highest)
  - Multiple syllabi referred *

    ** Con Brio, Henle & Royal Conservatory of Music*

Mitigating small dataset size…

- Dividing videos to many small equally-sized clips
  - 16 *frames per clip*
  - 992 *unique (non-overlapping) clips*

- **NOTE**: Model to be trained on clips not whole videos *

  ** Would be necessary even for a larger dataset*

  ** More specifically, only samples of clips per video are taken for evaluation*

Sampling schemes…

Addressing driving question 3:
How to sample clips to best reflect player performance?

- Sampling scheme is a detailed description of the following:
    - *What data will be obtained*
    - *How this will be done*

- Sampling schemes used:
    - Contiguous
    - Uniform

Sampling schemes (continued)…

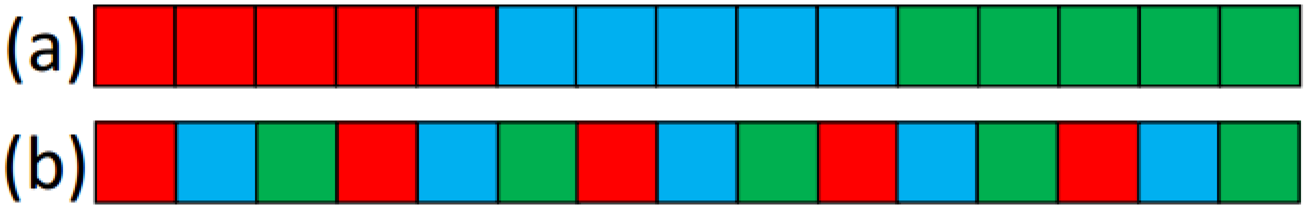


Fig. 3. **Sampling schemes**: (a) Contiguous; (b) Uniformly Distributed. Time along the x-axis. Each color represents a sample. Each square represents a clip of 16 frames.

NOTE: The entire bar is the whole video

Not addressed --- which sample to pick?

STEP 2: Defining architecture

Overview…

- Visual branch
- Aural branch
- Multimodal (visual + aural) branch
- Objective function

Visual branch – Network type used

- Network type: 3DCNN

  Why?

- Need to process short clips rather than frames

  Why?

- To detect certain high-level skills (ex. arpeggio, cadence, etc.)

Visual branch – Feature processing method

- Sample-features obtained by aggregating clip-level features *

  ** Features observable across the clip's frames*

- Passed through linear layer to reduce dimensions to 128

- *Why aggregation using averaging instead of RNN?*

Prior precedent referenced:

What and How Well You Performed?
A Multitask Learning Approach to Action Quality Assessment
by Paritosh Parmar & Brendan Tran Morris

Aural branch – Network type used

- Network type: 2DCNN

  Why?

- Need to process audio as melspectograms

  Why?

- To be able to extract auditory features using visual features

Aural branch – Feature processing method

- Raw audio signal to its melspectrogram

  Raw audio signals correspond to sampled clips in visual branch

- Clip-level audio features aggregated using averaging function

  Same reasoning as with visual branch

Multimodal branch

- Visual & aural features concatenated → Multimodal features
- Passed through linear layer to reduce dimensions to 128
- No backpropagation from here to single modality backbones
  Why? To prevent cross-modality contamination!

Architecture diagram…

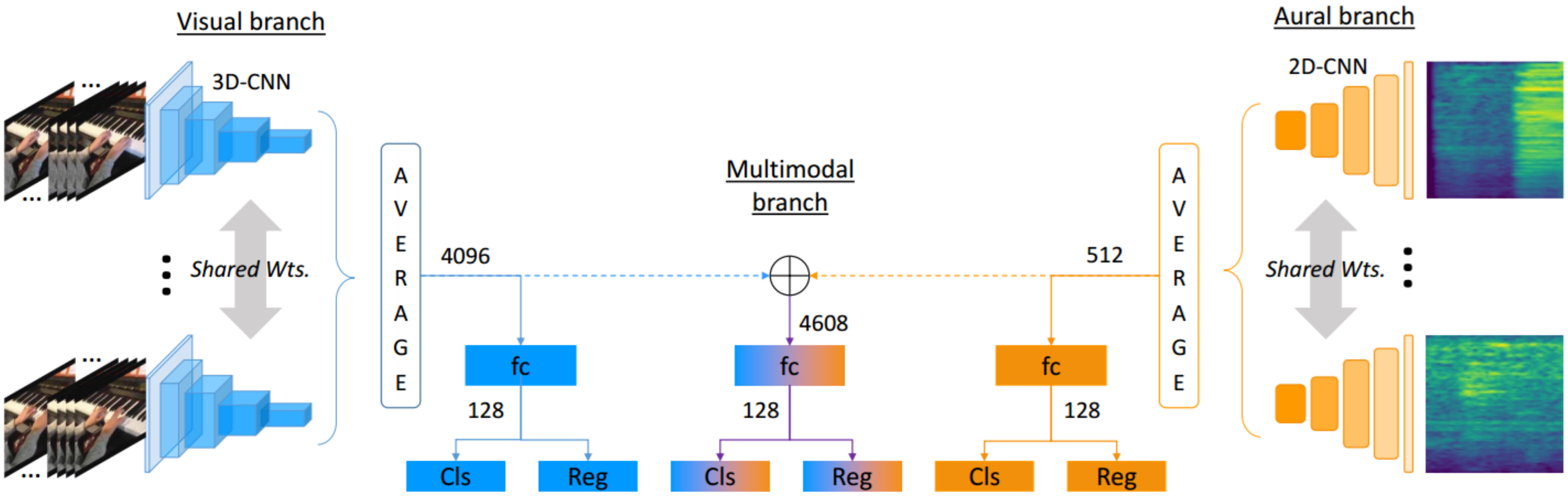


Fig. 4. Our multimodal learning architecture. ⊕ represents concatenation operation.

When considering a single modality branch, everything other than respective colored part is deactivated

fc = Fully connected layer

Objective function (considerations & components)…

1. Key considerations:
   a. We have a **multiclass classification** problem
   b. **Distance between classes** has meaning *

   ** unlike a typical classification problem*

2. Hence, objective function components:
   a. **Cross-entropy loss** (for multiclass classification error)
   b. **Distance function** (to consider distance between classes)

   Sum of L1 and L2 distances used (L1 has precedent)

Objective function (overall formula)…

$$\mathcal{L}_{total} = \alpha_1 \mathcal{L}_{CE}^{V} + \alpha_2 \mathcal{L}_{Reg}^{V} + \beta_1 \mathcal{L}_{CE}^{A} + \beta_2 \mathcal{L}_{Reg}^{A} + \gamma_1 \mathcal{L}_{CE}^{M} + \gamma_2 \mathcal{L}_{Reg}^{M}$$

$\mathcal{L}_{CE}^{V}$: **Cross-entropy loss for visual cues**

$\mathcal{L}_{Reg}^{V}$: **Distance function for visual cues**

*Likewise for superscript **A** (audio cues) and **M** (multimodal cues)*

STEP 3: Experimentation

Experimentation questions…

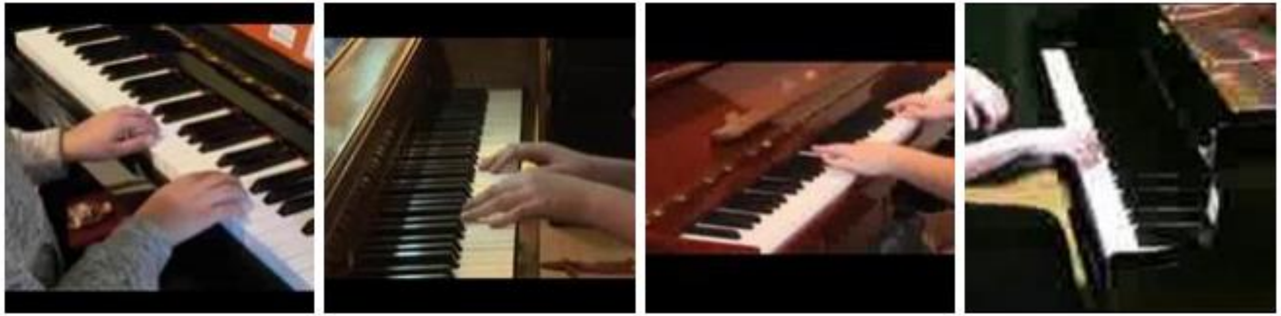
1. *Is it possible to determine pianist’s skill level with ML / CV?*

2. *What is better sampling strategy: contiguous vs. uniform?*

3. *Is multimodal assessment better than unimodal assessment?*

Preprocessing details…

- Crop to use visual information on forearms, hands & piano



- Convert the audio signal to its melspectrogram
    - *Using the* `librosa` *package of Python*
    - *Settings from another project (check reference in paper)*
    - *Amplitude expressed in decibels*

Implementation details for the whole network…

- *PyTorch* to implement networks

- Network hyperparameters:
  - Learning rate: 0.0001
  - Epochs: 100 (per network)
  - Batch size: 4

- Parameters of the objective function:
  - $\alpha_1 = \beta_1 = \gamma_1 = 1$
  - $\alpha_2 = \beta_2 = \gamma_2 = 0.1$

Implementation details for visual branch…

- Custom network for processing visual information

- Pretrain 3DCNN on UCF101 action recognition dataset
  Why? ***To avoid overfitting!***
  But why is action recognition dataset used?

- Details on input clips
  - 16 consecutive frames = 1 input clip
  - All frames resized to 112 × 112 pixels
  - Horizontal flipping applied to each frame

Implementation details for aural branch…

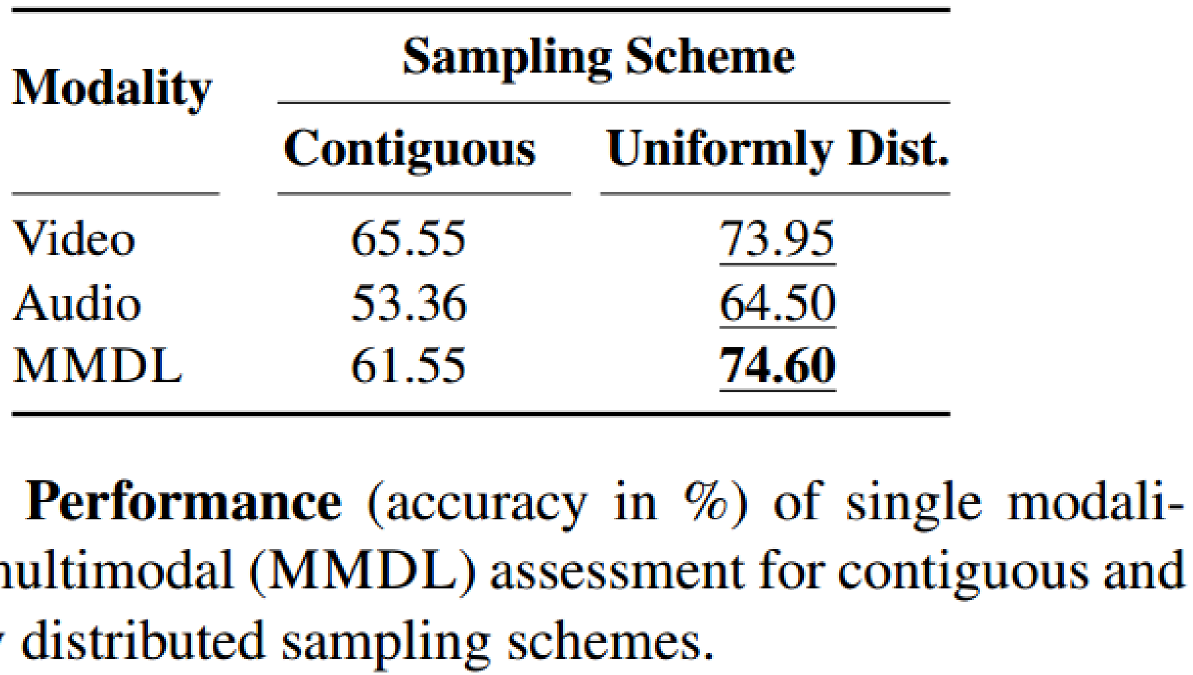
- Res-Net 18 (R18) network for processing melspectograms
  - R18 is a deep residual networks for image recognition
  - Has version pretrained on 1000000+ images from *ImageNet* *Initialising weights using ImageNet helped significantly!*
  - 18 ⇒ 18 layers

- Input details:
  - Changed no. of input channels of 1st conv. layer from 3 to 1
  - Converted melspectrogram to single channel images
  - Images <u>resized </u>to 224 × 224 pixels (not cropped!)

Implementation details for aural branch (continued)…

“We found that applying random cropping hurt the performance. This may be because the useful information is present in the lowest and highest frequencies and removing those in the process of cropping adversely affects the performance.”

Results…



Modality	Sampling Scheme	
	Contiguous	Uniformly Dist.
Video	65.55	73.95
Audio	53.36	64.50
MMDL	61.55	**74.60**

Table 1. Performance (accuracy in %) of single modalities vs a multimodal (MMDL) assessment for contiguous and uniformly distributed sampling schemes.

Understanding the results…

- Visual analysis significantly better than aural analysis
  Maybe more significant indicators in visual information?

- Multimodal analysis slightly better than unimodal analysis
  - *Skill-level indicators present in visual & audio information*
  - *Significance of visual mode ⇒ small boost in accuracy?*

- Uniform sampling significantly better than contiguous
  - ⇒ *Networks not biased to local or static cues in streams*

Final remarks

Merits…

- Small dataset size & overtraining mitigations
  - *Small dataset size mitigated by dividing into clips*
  - *Overtraining avoided by pretraining networks*

- Multimodal no worse than the best unimodal, *maybe better*
  - *Promising indicator for further research in this path*

- No bias toward local or static cues in the performance
  - *Indicates robustness of the network & lack of overtraining*

Limitations…

- Small dataset size (*despite mitigations*, *may pose challenges)*
- Small significance of improvement due to multimodal analysis?
  - *Even if small, is the improvement consistent?*
  - *Is the small improvement worth the complexity?*
- Even best accuracy maybe too low for reliable application
  - *No mention if accuracy considers distance in classification*
- How is clip size chosen? Could different sizes do better?
- Target is ordinal; is classification the best approach?

Possible avenues for improvement…

- Extracting more useful information from auditory inputs
  - *Maybe custom network design needed (as with visual)?*

- Organising larger & diverse dataset for better training

- Testing with different clip sizes

- Exploring other sampling schemes for better accuracy
  - *Uniform random sampling*
  - *Probability distribution-based sampling*

End