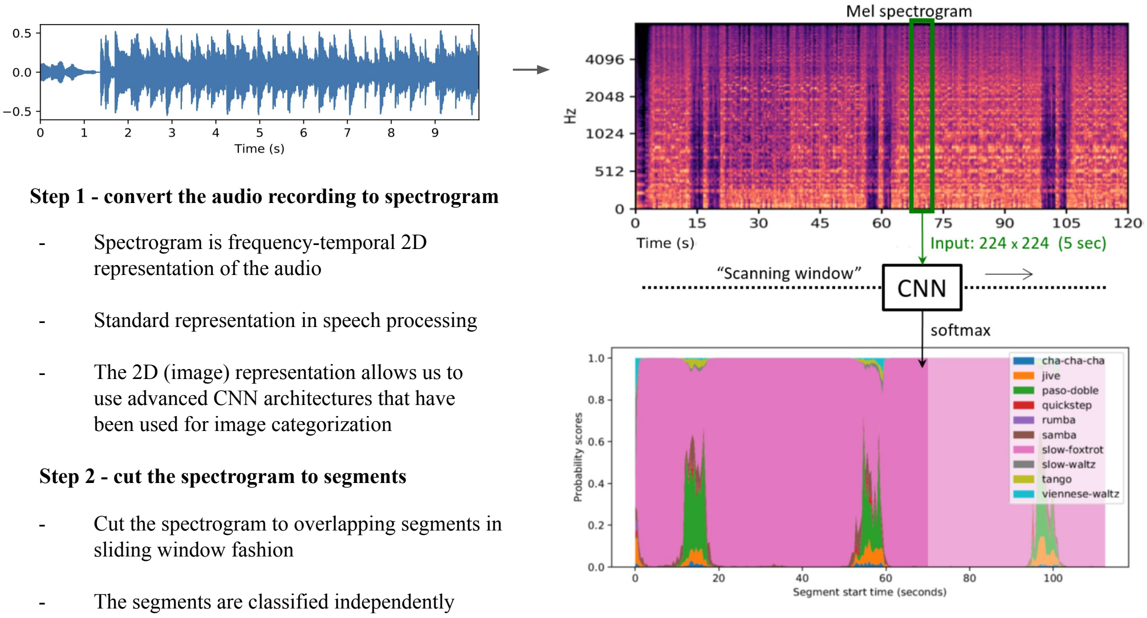
Ballroom Dance Recognition from Audio Recordings

Tomáš Pavlín, Jan Čech, Jiří Matas

Center for Machine Perception, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague



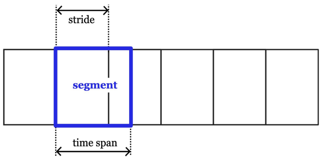
- *Cha Cha*
- *Jive*
- *Paso Doble*
- *Quickstep*
- *Rumba*
- *Samba*
- *Slow Foxtrot*
- *Slow Waltz*
- *Tango*
- *Viennese Waltz*

Step 1 - convert the audio recording to spectrogram

- Spectrogram is frequency-temporal 2D representation of the audio
- Standard representation in speech processing
- The 2D (image) representation allows us to use advanced CNN architectures that have been used for image categorization

Step 2 - cut the spectrogram to segments

- Cut the spectrogram to overlapping segments in sliding window fashion
- The segments are classified independently
- Each segment size is 224 × 224 which corresponds to **~5 seconds** (time span)
- Experiments show that ~5 seconds is long enough to predict correct dance style accurately, a dance music is "stationary"



Step 3 - convolutional neural network

- **Dense Convolutional Network (DenseNet)** [*Huang, Liu, Van Der Maaten, and Weinberger, 2017*]
- Input: spectrogram segment
- Output: probability score, vector of size 10 (number of dance classes), softmax

Step 4 - aggregation of segment results

- To predict samples that are longer than the segment duration of ~5 seconds
- The softmax outputs are averaged by arithmetic mean

Training set

Dance Genre	Count
Cha Cha Cha	711
Jive	490
Paso Doble	112
Quickstep	458
Rumba	658
Samba	721
Slow Foxtrot	421
Slow Waltz	411
Tango	395
Viennese Waltz	281
Total	**4655**

- private collection of ballroom dance music
- **~4700** audio recordings
- **10** dance classes
- the recordings are **~4 minutes** long
- studio quality

Test and validation set

- Audio extracted from public YouTube videos
- We make the dataset publicly available at http://dance.ironbrain.net/testset.zip
- Both datasets are **uniform** and consist of **10** classes of **6** recordings each (provides **60 recordings** each)
- The recordings are ~**3 minutes** long and are in studio quality
- The datasets do not overlap with each other and with training set
- Validation set is utilized for selecting epoch with highest accuracy
- Test set is used for testing resulting model only

Results

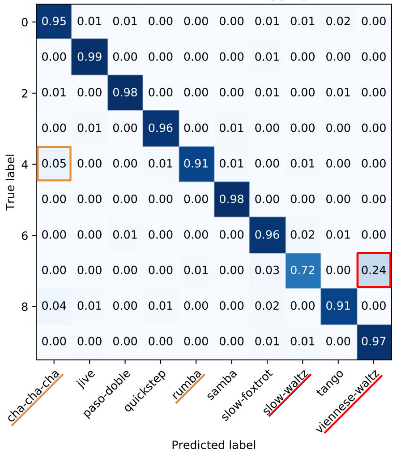
Results on Youtube test set

Method	Top-1 accuracy	Top-2 accuracy
Our method with aggregation	96.7%	100.0%
Our method without aggregation	92.2%	-

Confusion of similar dances:

- **Waltz x Viennese Waltz**
- **Rumba x Cha-cha-cha**

Confusion matrix (without aggregation)



	cha-cha-cha	jive	paso-doble	quickstep	rumba	samba	slow-foxtrot	slow-waltz	tango	viennese-waltz
0	0.95	0.01	0.01	0.00	0.00	0.00	0.01	0.01	0.02	0.00
	0.00	0.99	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
2	0.01	0.00	0.98	0.00	0.00	0.00	0.01	0.00	0.01	0.00
	0.00	0.01	0.00	0.96	0.00	0.01	0.00	0.00	0.00	0.00
4	0.05	0.00	0.00	0.01	0.91	0.01	0.00	0.01	0.00	0.00
	0.00	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.00
6	0.00	0.00	0.01	0.00	0.00	0.00	0.96	0.02	0.01	0.00
	0.00	0.00	0.00	0.00	0.01	0.00	0.03	0.72	0.00	0.24
8	0.04	0.01	0.00	0.01	0.00	0.00	0.02	0.00	0.91	0.00
	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.97

Experiments

Architecture	Top-1 accuracy	Top-2 accuracy	Top-1 without aggregation
VGG 16	25.0%	41.7%	24.8%
ResNet-18	**96.7%**	**100.0%**	89.9%
ResNeXt-50 32x4d	95.0%	**100.0%**	89.6%
DenseNet 161	**96.7%**	**100.0%**	**92.8%**

Configuration	Top-1 accuracy	Top-2 accuracy	Top-1 without aggregation
DenseNet-TL-C	63.3%	76.7%	42.9%
DenseNet-TL-DB4 (half)	80.0%	85.0%	62.7%
DenseNet-TL-DB4 (full)	83.3%	91.7%	64.5%
DenseNet-TL-DB4-N (n=24)	70.0%	85.0%	62.0%
DenseNet-TL-DB4-N (n=48)	76.7%	80.0%	63.3%
DenseNet-TL-DB4-N (n=72)	75.0%	86.7%	64.5%
DenseNet-FT	**96.7%**	**100.0%**	**92.8%**
DenseNet-RW	95.0%	**100.0%**	91.3%
DenseNet-RW-1C7x7	93.3%	**100.0%**	89.2%
DenseNet-RW-1C16x3	93.3%	**100.0%**	88.2%
DenseNet-RW-1C40x3	91.7%	98.3%	88.2%

Baseline:

- hand-crafted features classifier
- relies on hand-crafted audio features instead of a spectrogram
- classification using simple *SVM*
- accuracy **40%**

Cross-dataset tests

Dataset	Top-1 accuracy	Top-2 accuracy	Top-1 without aggregation
Extended ballroom	93.9%	97.5%	86.6%
YouTube test set	96.7%	100.0%	92.2%
Dance competitions	87.9%	98.6%	70.6%
StarDance	68.0%	78.0%	45.2%
Low Quality Recordings	72.7%	86.7%	58.0%

Extended ballroom

- publicly available dataset
- 4180 recordings
- each recording is 30 seconds long

YouTube test dataset

- 6 x 10 = 60 recordings

Dance competitions

- 360 recordings
- extracted from YouTube videos of dance competitions of *World DanceSport Federation (WDSF)*

StarDance

- extracted from 10th season of Czech TV show similar to Dancing with the Stars
- popular music
- 50 recordings

Low Quality Recordings

- recorded using mobile phone camera in dance competitions
- low audio quality (echo, people applauding, dancers steps)
- 128 recordings

Demonstration: dance.ironbrain.net