Musical Genre Recognition based on Deep Descriptors of Harmony, Instrumentation, and Segments

Igor Vatolkin[1], **Mark Gotham**[2], **Néstor Nápoles López**[3], **Fabian Ostermann**[1]

[1]Department of Computer Science, TU Dortmund University
[2]Department of Arts and Sports Sciences, TU Dortmund University
[3]Centre for Interdisciplinary Research in Music Media and Technology, McGill University

igor.vatolkin@tu-dortmund.de;mark.gotham@tu-dortmund.de;
nestor.napoleslopez@mail.mcgill.ca;fabian.ostermann@tu-dortmund.de

Overview

- Combination of deep and shallow algorithms for recognition of musical genres
- Deep convolutional neural network (CNN) models for prediction of harmonic, instrumental, and segment properties
- Shallow classifiers for prediction of 19 genres
- Significant reduction of classification errors after evolutionary feature selection compared to previous work

Classification Framework

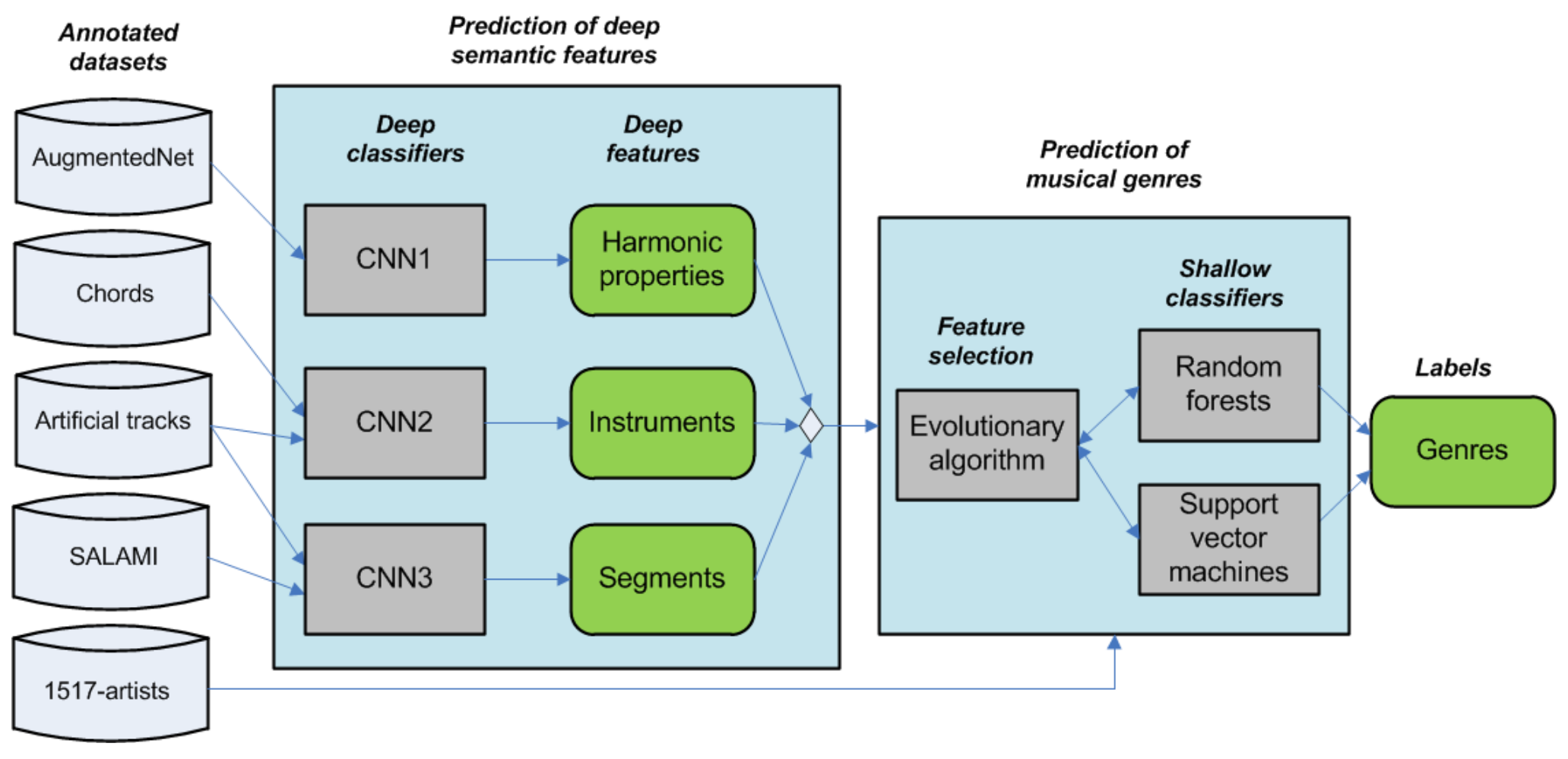

Figure 1: *Data flow in the proposed classification framework*

Deep Harmonic Properties

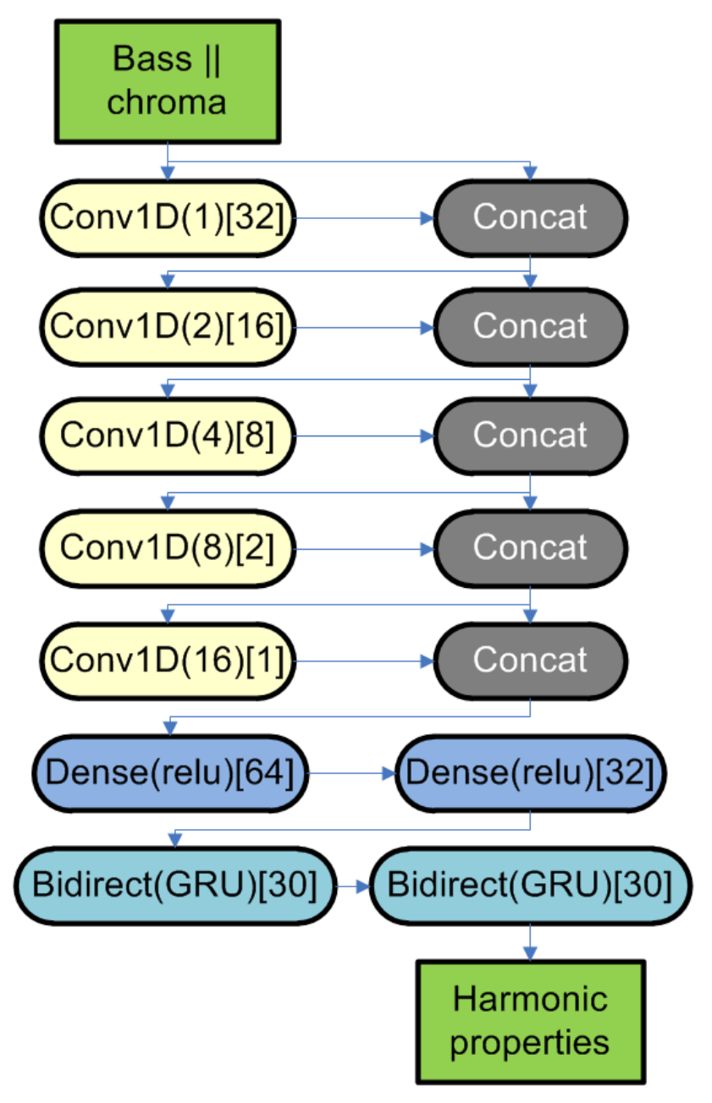

Figure 2: *Architecture of the AugmentedNet [1]*

- Multitask outputs related to harmonic rhythm, chord, and key properties
- Trained with audio chromagrams of 353 annotated music pieces instead of symbolic chromagrams from the original approach

Table 1: *Deep harmonic properties estimated for classification frames of 4s with 2s step size*

Features	Dim.
Predictions trained with AugmentedNet	
Mean and standard deviation of harmonic rhythm	1–2
Relative frequency of specific notes in the alto	3–24
Relative frequency of specific notes in the bass	25–47
Relative frequency of specific roots of local keys	48–71
Relative frequency of specific notes in the soprano	72–92
Relative frequency of specific notes in the tenor	93–112
Relative frequency of specific roots of tonicized keys	113–136
Relative frequency of specific roman numerals	137–160
Relative frequency of modes (major or minor)	161–162
Total number of different symbols	163–171

Deep Instrument Properties

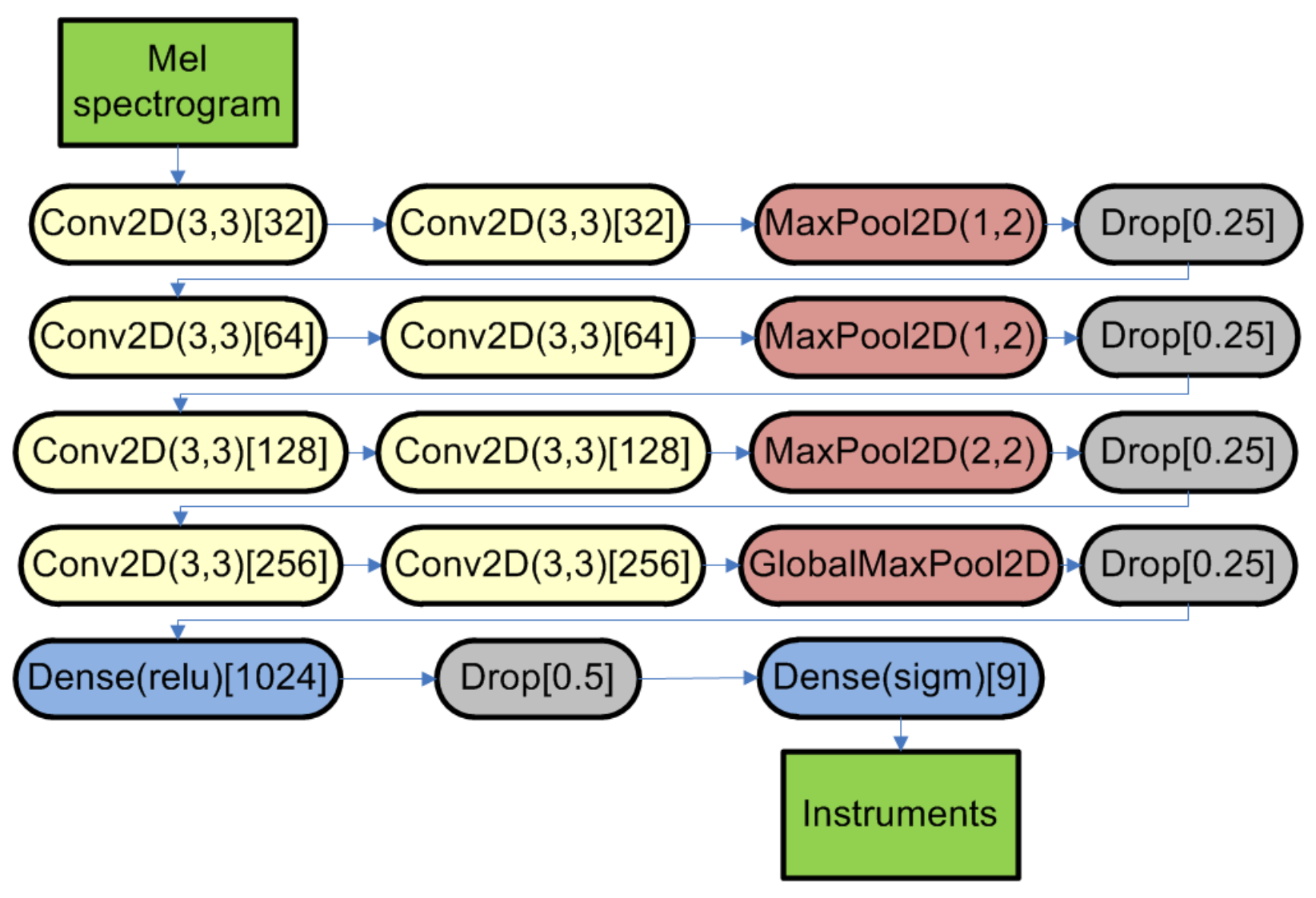

Figure 3: *Architecture of the CNN after [2]*

- Predictions of relative strengths of 51 or 31 different instruments in a 2s time frame
- Trained either with 5,000 samples and chords generated by mixing of individual samples after [3] or AAM (3,000 tracks) [4]

Table 2: *Deep instrument features estimated for classification frames of 4s with 2s step size*

Features	Dim.
Predictions trained with chords	
Mean relative strength of 51 predicted instruments (acoustic and electric guitar, organ, piano and electric piano, viola, violin, etc.)	1–51
Standard deviation of the relative strength of 51 predicted instruments	52–102
Minimum relative strength of 51 predicted instruments	103–153
Maximum relative strength of 51 predicted instruments	154–204
Predictions trained with artificial tracks	
Mean relative strength of 31 predicted instruments (subset of 51 instruments)	205–235
Standard deviation of the relative strength of 31 predicted instruments	236–266
Minimum relative strength of 31 predicted instruments	267–297
Maximum relative strength of 31 predicted instruments	298–328

Deep Segment Properties

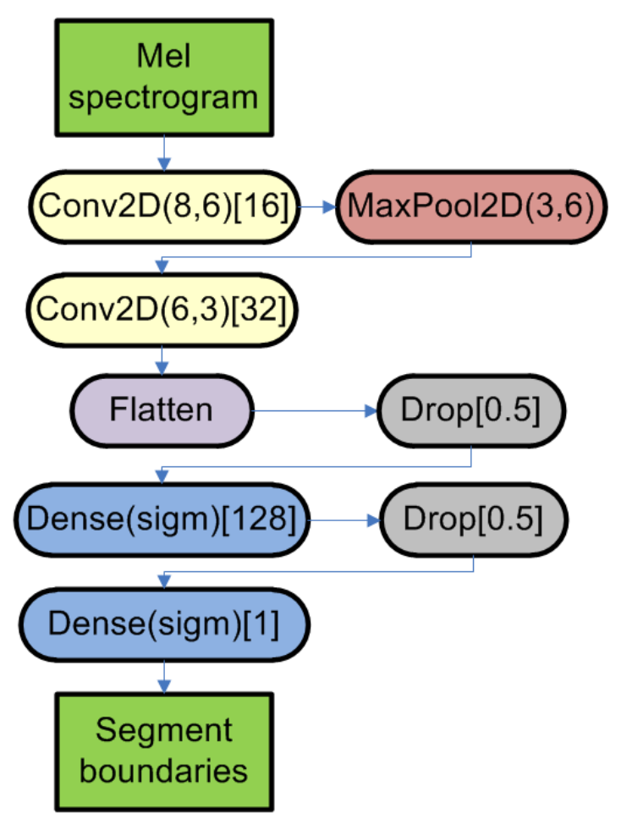

Figure 4: *Architecture of the CNN after [5]*

- Statistics of predicted segment boundaries of different types after [6]
- Trained with either SALAMI [7] (1,359 tracks) or AAM (3,000 tracks) [4]

Table 3: *Deep segment statistics estimated for complete audio tracks*

Features	Dim.
Predictions trained with SALAMI	
Number of segments	1
Mean segment length	2
Standard deviation of the segment length	3
Maximal segment length	4
Minimal segment length	5
Mean deviation of segment length	6
Predictions trained with artificial tracks	
Segment statistics as for SALAMI, trained to detect all boundaries	7–12
Segment statistics as for SALAMI, trained to detect instrument boundaries	13–18
Segment statistics as for SALAMI, trained to detect key boundaries	19–24
Segment statistics as for SALAMI, trained to detect tempo boundaries	25–30

Setup of Experiments

- Dataset
  - 1517-artists [8]: 19 genres
  - 16 “positive” + 18 “negative” training tracks per genre
  - 228 test tracks
  - 228 optimization sets for feature selection (see below)
- Features
  - Instrument- and timbre-related features from [3]
  - Mel frequency cepstral coefficients (MFCCs)
  - All deep harmonic features
  - All deep instrument features
  - All deep segment features
  - All deep features
  - Best sets after evolutionary feature selection following [9]
- Classifiers
  - Random forests
  - Support vector machines
- Evaluation
  - Let tp be true positives, tn true negatives, fp false positives, and fn false negatives
  - Balanced relative error:

$$e_b = \frac{1}{2}\left(\frac{fn}{tp+fn} + \frac{fp}{tn+fp}\right) \qquad (1)$$

Results: Tables

Table 4: *Test e_b for 19 musical genre recognition tasks. [3]: the best results reported in that work; MFCCs: Mel frequency cepstral coefficients; Harm: deep harmonic features listed in Table 1; Inst: deep instrument features listed in Table 2; Segm: deep segment features listed in Table 3; All: all deep features; All-FS: the best feature set after evolutionary feature selection. Bolded values are the best (smallest) for each genre in the current study. A bolded value using italic font marks a sole case where an error of [3] was lower than the lowest error in our study.*

	Random forests						
Genre	**[3]**	**MFCCs**	**Harm**	**Inst**	**Segm**	**All**	**All-FS**
Alternative	0.1928	0.2847	0.4861	0.3148	0.4375	0.3218	0.2431
Blues	0.3170	0.4028	0.3727	0.3495	0.4954	0.4259	**0.1921**
Childrens	0.3880	0.5069	0.5116	0.4329	0.3148	0.3102	0.2685
Classical	0.0929	0.1250	0.5231	0.0995	0.2106	0.2083	0.0833
Comedy	0.2214	0.3333	0.3634	0.3125	0.2894	0.3125	0.2407
Country	0.2350	0.3472	0.4190	0.2199	0.3843	0.3403	**0.1273**
Easy List.	0.2904	0.2894	0.4537	0.3542	0.3542	0.3866	**0.2245**
Electronic	0.1487	0.3843	0.2731	0.0926	0.3472	0.2454	**0.0370**
Folk	0.2682	0.3935	0.4236	0.3449	0.5440	0.3264	0.1852
Hip-Hop	0.1240	0.3495	0.4954	0.1065	0.2477	0.2824	0.0880
Jazz	0.3123	0.3889	0.3681	0.3519	0.5231	0.4514	0.2523
Latin	0.3049	0.5069	0.5694	0.4028	0.5231	0.3704	0.2940
New Age	0.2349	0.3056	0.5139	0.2731	0.3773	0.3750	**0.1505**
R’n’B	0.2534	0.2731	0.4144	0.2500	0.4213	0.2616	**0.1898**
Reggae	0.1941	0.3194	0.5069	0.2454	0.4375	0.3912	0.1875
Religious	0.3759	0.4352	0.3634	0.3912	0.5093	0.3611	0.2523
Rock/Pop	0.2346	0.2870	0.5579	0.2894	0.6273	0.2963	**0.1343**
Soundtr.	0.2652	0.2708	0.5926	0.3079	0.4190	0.3750	0.2616
World	0.4059	0.3403	0.4144	0.5069	0.5046	0.4745	**0.2662**
	Support vector machines						
Alternative	***0.1656***	0.2593	0.4282	0.2546	0.5000	0.2639	**0.2060**
Blues	0.3030	0.4074	0.3449	0.2546	0.5000	0.2847	0.2153
Childrens	0.3366	0.5185	0.5162	0.4769	0.5000	0.5000	**0.1944**
Classical	0.0885	0.0903	0.4190	**0.0810**	0.5000	0.1574	0.0833
Comedy	0.2360	0.3542	0.3519	0.3426	0.5000	0.2431	**0.1782**
Country	0.2247	0.3565	0.4352	0.2407	0.5000	0.2940	0.1319
Easy List.	0.2980	0.2315	0.4514	0.4259	0.5000	0.5000	0.2477
Electronic	0.1448	0.2245	0.3380	0.1412	0.5000	0.1806	0.0532
Folk	0.2621	0.3449	0.4190	0.3495	0.5000	0.4167	**0.1736**
Hip-Hop	0.1201	0.2431	0.5185	0.0671	0.5000	0.5000	**0.0810**
Jazz	0.2680	0.4190	0.2708	0.3588	0.5000	0.3356	**0.2338**
Latin	0.3168	0.4514	0.5509	0.4838	0.5000	0.5602	**0.2593**
New Age	0.2122	0.2685	0.4745	0.2894	0.5000	0.5000	0.1921
R’n’B	0.2594	0.3380	0.4514	0.2593	0.5000	0.4236	0.2014
Reggae	0.1872	0.2546	0.5301	0.2523	0.5000	0.3449	**0.1690**
Religious	0.3751	0.3981	0.4005	0.3935	0.5000	0.3912	**0.2269**
Rock/Pop	0.2390	0.2014	0.5648	0.2917	0.5000	0.4120	0.1389
Soundtr.	0.3108	0.2593	0.5231	0.3773	0.5000	0.3403	**0.2431**
World	0.3604	0.3472	0.4954	0.4306	0.5000	0.3495	0.2731

Conclusions

- Performance of complete individual deep feature groups or also all of them often rather poor, because of too many irrelevant dimensions
- After feature selection identifying the most relevant features, the errors are lowest for 17 of 19 genres
- Future work: integration of other deep predictors and more robust classification models

References

[1] N. Nápoles López, M. Gotham, and I. Fujinaga: A Roman Numeral Analysis Network with Synthetic Training Examples and Additional Tonal Tasks. Proc. ISMIR, 404–411 (2021)

[2] Y. Han, J.-H. Kim, and K. Lee: Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(1):208–221 (2017)

[3] I. Vatolkin, B. Adrian, and J. Kuzmic: A Fusion of Deep and Shallow Learning to Predict Genres Based on Instrument and Timbre Features. Proc. EvoMUSART, 313–326 (2021)

[4] F: Ostermann, I. Vatolkin, M. Ebeling: AAM: a Dataset of Artificial Audio Multitracks for Diverse Music Information Retrieval Tasks. EURASIP Journal on Audio, Speech, and Music Processing volume, 2023, 13 (2023)

[5] T. Grill and J. Schlüter: Music Boundary Detection Using Neural Networks on Combined Features and Two-Level Annotations. Proc. ISMIR, 531–537 (2015)

[6] I. Vatolkin, F. Ostermann, and M. Müller: An Evolutionary Multi-objective Feature Selection Approach for Detecting Music Segment Boundaries of Specific Types. Proc. GECCO, 1061–1069 (2021)

[7] J. B. L. Smith, J. A. Burgoyne, I. Fujinaga, D. De Roure, J. S. Downie: Design and Creation of a Large-scale Database of Structural Annotations. Proc. ISMIR, 555-560 (2011)

[8] K. Seyerlehner, G. Widmer, and P. Knees: Frame Level Audio Similarity - a Codebook Approach. Proc. DAFx (2008)

[9] I. Vatolkin, G. Rudolph, C. Weihs: Evaluation of Album Effect for Feature Selection in Music Genre Recognition. Proc. ISMIR, 169–175 (2015)

EvoMUSART 2023 12th International Conference on Artificial Intelligence in Music, Sound, Art and Design (EvoMUSART), Brno, Czech Republic, April 12-14, 2023