Composer Recognition based on 2D-Filtered Piano-Rolls

Velarde, Gissel; Weyde, Tillman; Cancino Chacón, Carlos; Meredith, David; Grachten, Maarten

Published in:
Proceedings of the 17th International Conference on Music Information Retrieval

Publication date:
2016

Document Version
Accepted author manuscript, peer reviewed version

COMPOSER RECOGNITION BASED ON 2D-FILTERED PIANO-ROLLS

Gissel Velarde[1] **Tillman Weyde**[2] **Carlos Cancino Chacón**[3]
David Meredith[1] **Maarten Grachten**[3]
[1] Department of Architecture Design & Media Technology, Aalborg University, Denmark
[2] Department of Computer Science, City University London, UK
[3] Austrian Research Institute for Artificial Intelligence, Austria
{gv, dave}@create.aau.dk, t.e.weyde@city.ac.uk, {carlos.cancino, maarten.grachten}@ofai.at

ABSTRACT

We propose a method for music classification based on the use of convolutional models on symbolic pitch–time representations (i.e. piano-rolls) which we apply to composer recognition. An excerpt of a piece to be classified is first sampled to a 2D pitch–time representation which is then subjected to various transformations, including convolution with predefined filters (Morlet or Gaussian) and classified by means of support vector machines. We combine classifiers based on different pitch representations (MIDI and morphetic pitch) and different filter types and configurations. The method does not require parsing of the music into separate voices, or extraction of any other predefined features prior to processing; instead it is based on the analysis of texture in a 2D pitch–time representation. We show that filtering significantly improves recognition and that the method proves robust to encoding, transposition and amount of information. On discriminating between Haydn and Mozart string quartet movements, our best classifier reaches state-of-the-art performance in leave-one-out cross validation.

1. INTRODUCTION

Music classification has occupied an important role in the music information retrieval (MIR) community, as it can immediately lead to musicologically interesting findings and methods, whilst also being immediately applicable in, for example, recommendation systems, music database indexing, music generation and as an aid in resolving issues of spurious authorship attribution.

Composer recognition, one of the classification tasks addressing musical style discrimination (among genre, period, origin identification, etc.), has aroused more attention in the audio than in the symbolic domain [13]. Particularly in the symbolic domain, the string quartets by Haydn and Mozart have been repeatedly studied [10, 12, 13, 24], since discriminating between Haydn and Mozart has been found to be a particularly challenging composer recognition task [24].

In this study, we propose a novel method and evaluate it on the classification of the string quartet movements by Haydn and Mozart. The method is based on the use of convolutional models on symbolic pitch–time representations (i.e. piano-rolls). An excerpt of a piece to be classified is first sampled to a 2D pitch–time representation which is then subjected to various transformations, including convolution with predefined filters (Morlet or Gaussian) and classified by means of Support Vector Machines (SVM).

2. RELATED WORK

Typically it is seen that computational methods use some kind of preprocessing to extract melody and harmony. Previous computational methods addressing composer discrimination of polyphonic works required defining sets of musical features or style makers, and/or relied on the encoding of separate parts or voices [10,12,13,24]. However, hard-coded musical features require musical expertise and may not perform similarly on different datasets [24], while the performance of methods relying on separate encoding of voice parts could be affected if voices are not encoded separately or even be unusable.

In order to avoid the requirements of previous methods, we aim to develop a more general approach studying the texture of pitch–time representations (i.e. piano-rolls) in the two-dimensional space. Previous studies did not address musical texture as it is proposed here.

Next, we review previous work that employs 2D music representations (2.1), and briefly sketch the background of the use of convolutional methods for machine perception and classification (2.2).

2.1 Representing music with 2D images

Visually motivated features generated from spectrograms have been successfully used for music classification (see [5, 28]). This success may be partly due to the fact that similar principles of perceptual organization operate in both vision and hearing [8]. The Gestalt principles of proximity, similarity and good continuation, originally developed to account for perceptual organization in vision, have also been used to explain the way that listeners organize

sonic events into streams and chunks [3, 7, 16]. Moreover, other studies suggest direct interaction between visual and auditory processing in common neural substrates of the human brain, which effectively integrates these modalities in order to establish robust representations of the world [9, 11, 21].

Graphical notation systems have been used since ancient times to transmit musical information [27]. Moreover, most Western music composed before the age of recording survives today only because of transmission by graphical notation — as staff notation, tablature, neumatic notation, etc. Standard graphical musical notation methods have proved to be extremely efficient and intuitive, possibly in part due to the natural mapping of pitch and time onto two orthogonal spatial dimensions.

2.2 Convolutional models

Convolutional models have been used extensively to model the physiology and neurology of visual perception. For example, in 1980, Daugman [6] and Marčelja [17] modeled receptive field profiles in cortical simple cells with parametrized 2D Gabor filters. In 1987, Jones and Palmer [14] showed that receptive-field profiles of simple cells in the visual cortex of a cat are well described by the real parts of complex 2D Gabor filters. More recently, Kay et al. [15] used a model based on Gabor filters to identify natural images from human brain activity. In our context, the Gabor filter is equivalent to the Morlet wavelet which we have used as a filter in the experiments described below.

Filters perform tasks like contrast enhancement or edge detection. In image classification, filtering is combined with classification algorithms such as SVM or neural networks for object or texture recognition [2, 23].

In the remainder of this paper, we present our proposed method in detail (3). Then, we report the results of our experiments (4) and finally, state our conclusions (5).

3. METHOD

Figure 1 provides an overview of our proposed method. As input, the method is presented with excerpts from pieces of music in symbolic format. Then, in the *sampling* phase, a 2D image is derived from each input file in the form of a piano-roll. After the *sampling* phase, various *tranformations* are applied to the images before carrying out the final *classification* phase, which generates a class label for the input file using an SVM. Details of each phase are given below. We begin by describing the *sampling* phase, in which symbolic music files are transformed into images of piano-rolls.

3.1 Sampling piano-roll images from symbolic representations

3.1.1 MIDI note numbers encoding

Symbolic representations of music (e.g. MIDI files) encode each note's pitch, onset and duration. We encoded pitch as an integer from 1 to 128 using MIDI note numbers (MNN), where C4 or *middle C* is mapped to MNN 60. Onset and duration are temporal attributes measured in quarter notes (qn).

3.1.2 Morphetic pitch encoding

The pitch name of a note is of the form <letter name><alteration><octave number>, e.g. C♯4. By removing the <alteration> and mapping all note names with the same <letter name> and <octave number> to the same number we reduce the space to *morphetic pitch*: an integer corresponding to the vertical position of the note on a musical staff.

We use a pitch-spelling algorithm by Meredith called *PS13s1* [18], to compute the pitch names of notes. The *PS13s1* algorithm has been shown to perform well on classical music of the type considered in this study. The settings of the *PS13s1* algorithm used here are the same as in [18], [1] with the *pre-context* parameter set to 10 notes and the *post-context* set to 42 notes. These parameters define a context window around the note to be spelt, which is used to compute the most likely pitch name for the note, based on the extent to which the context implies each possible key. When transposing a pattern within a major or minor scale (or, indeed, any scale in a diatonic mode), as is common practice in tonal (and modal) music, chromatic pitch intervals within the pattern change although the transposed pattern is still recognized by listeners as an instance of the same musical motif [8]. Morphetic pitch intervals are invariant to within-scale transpositions. We hypothesize that preserving this tonal motif identity might improve the performance of our models.

3.1.3 Piano-rolls (p_{70qn})

Symbolic representations of music are sampled to 2D binary images of size $P \times T$ pixels taking values of 0 or 1, called piano-roll representations. Our piano-roll representations are sampled from the first 70 qn of each piece, using onset in qn, duration in qn and either MNN or morphetic pitch, with a sampling rate of 8 samples per qn. We denote such representations by p_{70qn}. Each note of a piece symbolically encoded is described as an ordered tuple (onset, duration, pitch). The onsets are shifted, so that the first note starts at 0 qn. The piano-roll image is initialized with zeros and filled with ones for each sampled note. Its rows correspond to pitch and columns to samples in time. For each note, its onset and duration are multiplied by the sampling rate and rounded to the nearest integer. Note that since the tempo in terms of quarter notes per minute varies across pieces in our test corpora, the resulting samples vary in physical duration.

3.1.4 Piano-rolls (p_{400n})

As an alternative to the 70 qn piano roll excerpts, p_{70qn}, defined in 3.1.3 above, we also tested the methods on piano-roll excerpts consisting of the first 400 notes of each piece.

[1] We use a Java implementation of the *PS13s1* algorithm by David Meredith that takes MIDI files as input. **kern files are first converted to MIDI. Then we use the function *writemidi_seconds* by Christine Smit: `http://www.ee.columbia.edu/~csmit/midi/matlab/html/example_script1.html#2`

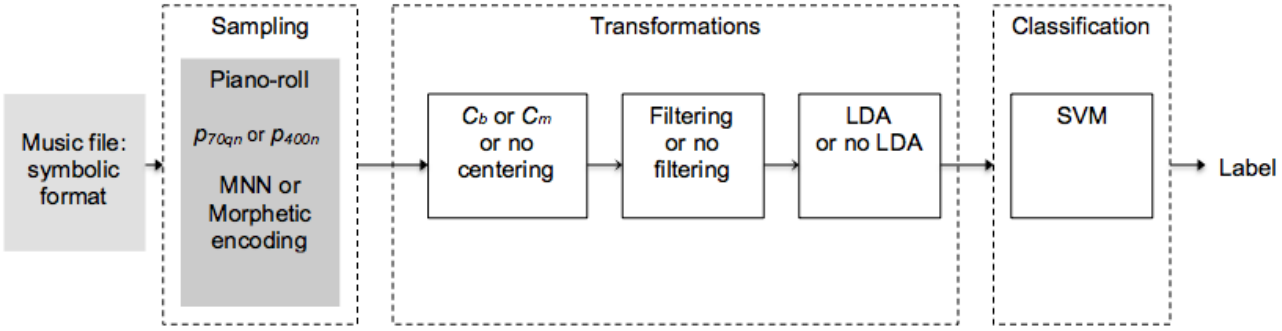

Figure 1. Overview of the method. Music, represented symbolically, is first sampled to 2D images of piano-rolls. Then, various transformations or processing steps are applied to the images, including convolution with predefined filters. The order of applying these transformations is from left to right. Finally, the images are classified with an SVM.

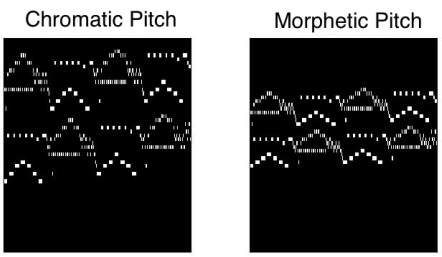

Figure 2. Piano-roll representation using MNN (left) and morphetic pitch (right) of the first 48 qn of Prelude 3 in C♯ major, BWV 848 by Bach. Note that the approximately similar inverted "V" shaped patterns in the left-hand figure are transformed into patterns of exactly the same shape in the right-hand figure.

We denote this type of representation by p_{400n}. These p_{400n} representations were produced by sampling in the way described in section 3.1.3, but using the first 400 notes instead of the first 70 qn of a piece and sampling to a size of $P \times T$ pixels. If a piece has fewer than 400 notes, all notes of the piece are represented. This representation is used to approximately normalize the amount of information per image.

In the next phase of our proposed method with a single classifier, as seen in Figure 1, various transformations or processing steps are applied which will be described as follows.

3.2 Transformations

We explore the effect of applying transformations or processing techniques to the piano-roll images. These transformations are applied in order to find a suitable normalization (i.e., alignment between the images) before classification, and to test the robustness of the method to transformations of the input data that would not be expected to reduce the performance of a human expert (cf. [22]). We now consider each of these transformations in turn.

3.2.1 Pitch range centering (C_b)

Typically, the pitch range of a piece in a piano-roll representation does not extend over the full range of possible MIDI note number values. We hypothesized that we could improve performance by transposing each piano roll so that its pitch range is centered vertically in its image. That is, for a piano-roll image of size $P \times T$ pixels, we translated the image by $y_s = (P - (y_d + y_u))/2$ pixels vertically, where y_d and y_u are the lower and upper co-ordinates, respectively, of the bounding box of the piano roll (i.e., corresponding to the minimum and maximum pitches, respectively, occurring in the piano roll). This transformation is used to test robustness to pitch transposition.

3.2.2 Center of mass centering (C_m)

An image p of size $P \times T$ pixels is translated so that the centroid of the piano roll occurs at the center of the image. We denote the centroid by $(\bar{x}, \bar{y}) = (M_{10}/M_{00}, M_{01}/M_{00})$, where $M_{ij} = \sum_x \sum_y x^i y^j p(x, y)$. The elements of the image are shifted circularly to the central coordinates (x_c, y_c) of the image, where $(x_c = T/2)$ and $(y_c = P/2)$, an amount of $(x_c - \bar{x})$ pixels on the x-axis, and $(y_c - \bar{y})$ pixels on the y-axis. In this case, circular shift is applied to rows and columns of p. In the datasets used for the experiments, in 5% of the pieces with MNN encoding, one low-pitch note was shifted down by this transformation and wrapped around so that it became a high-pitched note (in one piece there were four low-pitch notes shifted to high pitch-notes after circular shift). However, this transformation caused most pieces to be shifted and wrapped around in the time dimension so that, on average, approximately the initial 2 quarter notes of each representation were transferred to the end.

3.2.3 Linear Discriminant Analysis

We apply Linear Discriminant Analysis (LDA) [4] solving the singularity problem by Singular Value Decomposition and Tikhonov regularization to find a linear subspace for

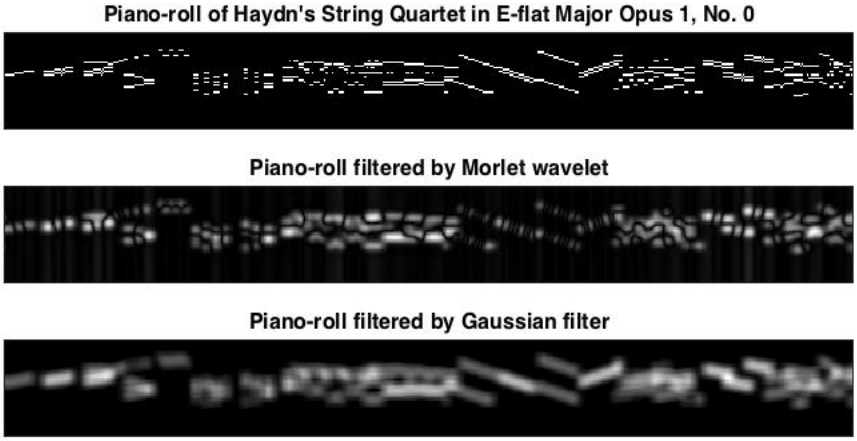

Figure 3. Piano-roll (p_{400n}) morphetic pitch representation (top) of Haydn's String Quartet in E-flat Major Opus 1, No. 0 and its transformations filtered by the Morlet wavelet at a scale of 2 pixels oriented of 90 degrees (second image), and by a Gaussian filter of size 9×9 pixels with $\sigma = 3$ (third image). p_{400n} and its filtered versions are each 56×560 pixels.

discrimination between classes. [2]

3.2.4 Filtering

Images are convolved with pre-defined filters (Morlet wavelet or a Gaussian filter). We apply the continuous wavelet transform (CWT) [1], with the Morlet wavelet ψ at fixed scale a and rotation angle θ

$$\psi_{a,\theta}(x,y) = a^{-1}\psi(a^{-1}r_{-\theta}(x,y)) \quad (1)$$

with rotation r_θ

$$r_\theta(x,y) = (x\cos\theta - y\sin\theta, x\sin\theta + y\cos\theta), 0 \leq \theta < 2\pi. \quad (2)$$

where

$$\psi(x,y) = e^{ik_0 y}e^{-\frac{1}{2}(\varepsilon^{-1}x^2+y^2)} \quad (3)$$

with frequency $k_0 = 6$ and $\varepsilon = 1$.

The filtered images are the absolute values of the real part of the wavelet coefficients. We test a defined set of scales and angles (see section 4). The selection of scale and angle of orientation are those that yield the best classification as in [25].

We also filter images with a rotationally symmetric Gaussian low-pass filter g:

$$g(x,y) = e^{\frac{-(x^2+y^2)}{2\sigma^2}} \quad (4)$$

where x and y are the distances from the origin in the horizontal and vertical axis, respectively.

We test a defined set of filter sizes h and σ values (see section 4). The selection of the size h of the filter and the value of σ are those that yield the best classification. As an example of the effect of filtering, Figure 3 shows the piano-roll image, p_{70qn} of Haydn's String Quartet in E-flat Major Opus 1, No. 0 and the filtered images obtained by the convolution with Morlet wavelet and Gaussian filter.

[2] We use Deng Cai's LDA implementation version 2.1: `http://www.cad.zju.edu.cn/home/dengcai/Data/DimensionReduction.html`.

3.3 Classification with support vector machines

For classification, we use SVM with the Sequential Minimal Optimization (SMO) method to build an optimal hyperplane that separates the training samples of each class using a linear kernel [19]. Samples are transformed images of size $P \times T$ if they are not reduced by LDA. If LDA is applied, samples are points in 1D. Each sample is normalized around its mean, and scaled to have unit standard deviation before training. The Karush–Kuhn–Tucker conditions for SMO are set to 0.001.

4. EXPERIMENTS

We used a set of movements from string quartets by Haydn and Mozart, two composers that seemed to have influenced each other on this musical form. Walthew [26] observes that "Mozart always acknowledged that it was from Haydn that he learnt how to write String Quartets" and, in his late string quartets, Haydn was directly influenced by Mozart.

Distinguishing between string quartet movements by Haydn and Mozart is a difficult task. Sapp and Liu [20] have run an online experiment to test human performance on this task and found, based on over 20000 responses, that non-experts perform only just above chance level, while self-declared experts achieve accuracies up to around 66%.

Classification accuracy—that is, the proportion of pieces in the test corpus correctly classified—has been the established evaluation measure for audio genre and composer classification since the MIREX 2005 competition [3] and also for symbolic representations [12, 13, 24].

In our experiments we used the same dataset as in [24] consisting of 54 string quartet movements by Haydn and 53 movements by Mozart, encoded as **kern files, [4] and

[3] See `http://www.music-ir.org/mirex/wiki/2005:Main_Page`.

[4] `http://www.music-cog.ohio-state.edu/Humdrum/representations/kern.html`

	Pitch–time representation	Morlet LDA	Gauss LDA	NF LDA	Morlet	Gauss	NF
Morphetic pitch	p_{70qn}	65.4	58.9	57.9	53.3	68.2	58.9
	$C_b(p_{70qn})$	65.4	60.7	47.7	57.9	63.6	51.4
	$C_m(p_{70qn})$	53.3	60.7	52.3	64.5	59.8	56.1
	p_{400n}	67.3	**80.4**	57.0	63.6	72.9	55.1
	$C_b(p_{400n})$	62.6	72.9	54.2	61.7	66.4	53.3
	$C_m(p_{400n})$	65.4	65.4	55.1	66.4	70.1	53.3
MNN	p_{70qn}	64.5	67.3	66.4	62.6	66.4	64.5
	$C_b(p_{70qn})$	**70.1**	61.7	63.6	67.3	61.7	61.7
	$C_m(p_{70qn})$	63.6	57.9	57.0	66.4	56.1	54.2
	p_{400n}	66.4	69.2	64.5	65.4	63.6	64.5
	$C_b(p_{400n})$	54.2	64.5	52.3	58.9	58.9	49.5
	$C_m(p_{400n})$	53.3	62.6	42.1	56.1	63.6	44.9

Table 1. Haydn and Mozart String Quartet classification accuracies in leave-one-out cross validation for different configurations of classifiers (NF = no filtering).

evaluated our method's classification accuracies in leave-one-out cross-validation as it was done in [24].

Table 1 shows the classification accuracies (mean values) obtained in leave-one-out cross-validation for images of size 56×560 pixels. The standard deviation values are not presented, as they are not informative. The standard deviation can be derived from the accuracy in this case (accuracy of binary classification in leave-one-out cross-validation). The filters of the classifiers were tuned according to their classification accuracy over the different pitch–time representations. The angle of orientation of the Morlet wavelet was set to 90 degrees. This orientation was chosen out of a selection of angles (0, 45, 90 and 135 degrees). The scale was set to 2 pixels, selected varying its value from 1 to 9 pixels. The Gaussian filter was tested with pixel sizes of 1 to 10 pixels, with values of σ ranging from 1 to 4 pixels. Gaussian filters were set to 9 pixels and $\sigma = 3$. The best classifier using MNN encoding corresponds to a classifier operating on pitch–time representation $C_b(p_{70qn})$, filtered by Morlet wavelet oriented 90 degrees at a scale of 2 pixels, and LDA reduction. The best classifier of all reaches state-of-the-art performance with an accuracy of 80.4%. This classifier corresponds to a pitch–time representation p_{400n} in morphetic pitch encoding, filtered by a Gaussian filter of size 9 pixels and $\sigma = 3$, and LDA reduction. It misclassified 12 movements by Haydn and 9 by Mozart. The misclassified movements (mov.) are shown in Table 2. Due to our model section, it could be that the results present some overfitting.

From the results in Table 1 we observe that filtering significantly improves recognition at 5% significance level (Wilcoxon rank sum $= 194.5$, $p = 0.0107$, $n = 12$, with Morlet wavelet), (Wilcoxon rank sum $= 203$, $p = 0.0024$, $n = 12$, with Gaussian filter), and it is not significantly different to filter with Morlet or Gaussian filters (Wilcoxon rank sum $= 133$, $p = 0.3384$, $n = 12$). On the other side, there is not sufficient evidence to conclude that LDA improves recognition (Wilcoxon rank sum $= 154$, $p = 0.8395$, $n = 12$).

Movements by Haydn	Movements by Mozart
Op 1, N. 0, mov. 4	K. 137, mov. 3
Op 1, N. 0, mov. 5	K. 159, mov. 3
Op 9, N. 3, mov. 1	K. 168, mov. 2
Op 20, N. 6, mov. 2	K. 168, mov. 3
Op 20, N. 6, mov. 4	K. 428, mov. 3
Op 50, N. 1, mov. 3	K. 465, mov. 2
Op 64, N. 1, mov. 2	K. 465, mov. 4
Op 64, N. 4, mov. 2	K. 499, mov. 1
Op 64, N. 4, mov. 3	K. 499, mov. 4
Op 71, N. 2, mov. 2	
Op 103, mov. 1	
Op 103, mov. 2	

Table 2. Misclassified movements of our best classifier.

We study the effect of encoding (MNN vs. morphetic pitch), transposition (not centering vs. centering with C_b) and the amount of information (p_{70qn} vs. P_{400n}). The center of mass centering C_m was not evaluated, as this transformation may affect human recognition. Considering all results in Table 1 obtained with filtering and excluding the ones obtained with C_m, the difference in encoding between MNN and morphetic pitch is not significant at %5 significance level (Wilcoxon rank sum $= 269.5$, $p = 0.8502$, $n = 16$), nor are the results significantly different with or without centering C_b (Wilcoxon rank sum $= 311.5$, $p = 0.0758$, $n = 16$), neither it is significantly different to use p_{70qn} or P_{400n} (Wilcoxon rank sum $= 242$, $p = 0.4166$, $n = 16$). These findings suggest that the method based on 2D-Filtered piano-rolls is robust to transformations such as encoding, transposition, and amount of information that are considered not to affect human perception.

In Table 3, we list all previous studies where machine-learning methods have been applied to this Haydn/Mozart discrimination task. A direct comparison can be made between the classification accuracy achieved by the method of van Kranenburg and Backer [24] and our proposed method, as we used the same dataset. The datasets used by the other approaches in Table 3 were not available for us to test our method and make direct comparisons. Hontanilla et al. [13] used a subset of the set used in [24]: 49 string quartets movements by Haydn and 46 string quartets movements by Mozart [13]. Hillewaere et al. [12] extended van Kranenburg and Backer's [24] dataset to almost double its size, including several movements from the period 1770–1790. Herlands et al. [10] used a dataset consisting of MIDI encodings of only the first movements of the string quartets.

Table 3 shows that our best classifier reaches state-of-the-art performance and that there is no significant dif-

Method	Accuracy
Proposed best classifier	**80.4**
Van Kranenburg and Backer (2004) [24]	79.4
Herlands et al. (2014) [10]*	80.0
Hillewaere et al. (2010) [12]*	75.4
Hontanilla et al. (2013) [13]*	74.7

Table 3. Classification accuracies achieved by previous computational approaches on the Haydn/Mozart discrimination task. * indicates that a different dataset was used from that used in the experiments reported here.

ference from the results obtained by van Kranenburg and Backer at 5% significance level (Wilcoxon rank sum $= 11449$, $p = 0.8661$, $n = 107$). Compared to previous approaches [10,12,13,24], our method is more general in that it does not need hard-coded musical style markers for each dataset as in [24], nor does it require global musical feature sets as in [12], nor does it depend on the music having been parsed into separate parts or voices as in [10, 12, 13].

5. CONCLUSION

We have shown that string quartets by Haydn and Mozart can be discriminated by representing pieces of music as 2-D images of their pitch–time structure and then using convolutional models to operate on these images for classification. Our approach based on classifying pitch–time representations of music does not require parsing of the music into separate voices, or extraction of any other predefined features prior to processing. It addresses musical texture of 2-D pitch–time representations in a more general form. We have shown that filtering significantly improves recognition and that the method proves robust to encoding, transposition and amount of information. Our best single classifier reaches state-of-the-art performance in leave-one-out cross validation on the task of discriminating between string quartet movements by Haydn and Mozart.

With the proposed method, it is possible to generate a wide variety of classifiers. In preliminary experiments, we have seen that diverse configurations of classifiers (i.e. different filter types, orientations, centering, etc.) seem to provide complementary information which could be potentially used to build ensembles of classifiers improving classification further. Besides, we have observed that the method can be applied to synthetic audio files and audio recordings. In this case, audio files are sampled to spectrograms instead of piano-rolls, and then follow the method's chain of transformations, filtering and classification. We are optimistic that our proposed method can perform similarly on symbolic and audio data, and might be used successfully for other style discrimination tasks such as genre, period, origin, or performer recognition.

6. ACKNOWLEDGMENTS

The work for this paper carried out by G. Velarde, C. Cancino Chacón, D. Meredith, and M. Grachten was done as part of the EC-funded collaborative project, "Learning to Create" (Lrn2Cre8). The project Lrn2Cre8 acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 610859. G. Velarde is also supported by a PhD fellowship from the Department of Architecture, Design and Media Technology, Aalborg University. The authors would like to thank Peter van Kranenburg for sharing with us the string quartets dataset and results that allowed as statistical tests, William Herlands and Yoel Greenberg for supporting the unsuccessful attempt to reconstruct the dataset used in their research, Jordi Gonzalez for comments and suggestions on an early draft of this paper, and the anonymous reviewers for their detailed insight on this work.

7. REFERENCES

[1] J-P Antoine, Pierre Carrette, R Murenzi, and Bernard Piette. Image analysis with two-dimensional continuous wavelet transform. *Signal Processing*, 31(3):241–272, 1993.

[2] Yoshua Bengio, Aaron Courville, and Pierre Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828, 2013.

[3] Albert S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, MA., 1990.

[4] Deng Cai, Xiaofei He, Yuxiao Hu, Jiawei Han, and T. Huang. Learning a spatially smooth subspace for face recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–7, June 2007.

[5] Yandre MG Costa, LS Oliveira, Alessandro L Koerich, Fabien Gouyon, and JG Martins. Music genre classification using lbp textural features. *Signal Processing*, 92(11):2723–2737, 2012.

[6] John G Daugman. Two-dimensional spectral analysis of cortical receptive field profiles. *Vision Research*, 20(10):847–856, 1980.

[7] Diana Deutsch. Grouping mechanisms in music. In Diana Deutsch, editor, *The Psychology of Music*, pages 299–348. Academic Press, San Diego, 2nd edition, 1999.

[8] Diana Deutsch. *Psychology of Music*. Academic Press, San Diego, 3rd edition, 2013.

[9] Marc O Ernst and Heinrich H Bülthoff. Merging the senses into a robust percept. *Trends in Cognitive Sciences*, 8(4):162–169, 2004.

[10] William Herlands, Ricky Der, Yoel Greenberg, and Simon Levin. A machine learning approach to musically meaningful homogeneous style classification. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI)*, pages 276–282, 2014.

[11] Souta Hidaka, Wataru Teramoto, Yoichi Sugita, Yuko Manaka, Shuichi Sakamoto, Yôiti Suzuki, and Melissa Coleman. Auditory motion information drives visual motion perception. *PLoS One*, 6(3):e17499, 2011.

[12] Ruben Hillewaere, Bernard Manderick, and Darrell Conklin. String quartet classification with monophonic models. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, pages 537–542, Utrecht, The Netherlands, 2010.

[13] María Hontanilla, Carlos Pérez-Sancho, and Jose M Iñesta. Modeling musical style with language models for composer recognition. In *Pattern Recognition and Image Analysis*, pages 740–748. Springer, 2013.

[14] Judson P Jones and Larry A Palmer. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233–1258, 1987.

[15] Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.

[16] Fred Lerdahl and Ray S. Jackendoff. *A Generative Theory of Tonal Music*. MIT Press, Cambridge, MA., 1983.

[17] S Marčelja. Mathematical description of the responses of simple cortical cells. *Journal of Neuophysiology*, 70(11):1297–1300, 1980.

[18] David Meredith. The *ps13* pitch spelling algorithm. *Journal of New Music Research*, 35(2):121–159, 2006.

[19] John C. Platt. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods - Support Vector Learning*. MIT Press, January 1998.

[20] Craig Sapp and Yi-Wen Liu. The Haydn/Mozart String Quartet Quiz, 2015. `http://qq.themefinder.org` (Accessed 26 December 2015).

[21] Daniele Schön and Mireille Besson. Visually induced auditory expectancy in music reading: a behavioral and electrophysiological study. *Journal of Cognitive Neuroscience*, 17(4):694–705, 2005.

[22] Bob L Sturm. A simple method to determine if a music information retrieval system is a horse. *IEEE Transactions on Multimedia*, 16(6):1636–1644, 2014.

[23] Devis Tuia, Michele Volpi, Mauro Dalla Mura, Alain Rakotomamonjy, and Remi Flamary. Automatic feature learning for spatio-spectral image classification with sparse svm. *Geoscience and Remote Sensing, IEEE Transactions on*, 52(10):6062–6074, 2014.

[24] Peter Van Kranenburg and Eric Backer. Musical style recognition—a quantitative approach. In *Proceedings of the Conference on Interdisciplinary Musicology (CIM)*, pages 106–107, 2004.

[25] Gissel Velarde, Tillman Weyde, and David Meredith. An approach to melodic segmentation and classification based on filtering with the haar-wavelet. *Journal of New Music Research*, 42(4):325–345, 2013.

[26] Richard H Walthew. String quartets. *Proceedings of the Musical Association*, pages 145–162, 1915.

[27] Martin Litchfield West. The babylonian musical notation and the hurrian melodic texts. *Music & Letters*, pages 161–179, 1994.

[28] Ming-Ju Wu, Zhi-Sheng Chen, Jyh-Shing Roger Jang, Jia-Min Ren, Yi-Hsung Li, and Chun-Hung Lu. Combining visual and acoustic features for music genre classification. In *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*, volume 2, pages 124–129. IEEE, 2011.