Automatic Recognition of Famous Artists by Machine

Gerhard Widmer[1] and **Patrick Zanon**[2]

Abstract. The paper addresses the question whether it is possible for a machine to learn to distinguish and recognise famous musicians (concert pianists), based on their style of playing. We extract a number of low-level features related to expressive timing and dynamics from the original audio CD recordings by famous pianists, and apply various machine learning algorithms to the task of learning classifiers based on these features. Experiments show that the computer can learn to identify the performer in a new recording with a probability significantly higher than chance, despite the fact that the features only capture a very limited amount of information about a performance. An analysis of the learned classifiers reveals a number of performance features that seem particularly relevant to style differentiation, and an application of the classifiers to music of a very different style shows that the machine seems to have captured truly fundamental aspects of artistic style. One limitation of the current approach is that sequential information is totally ignored, and we briefly report on ongoing work that tries to address this problem via an interesting conversion of music performances to strings.

1 INTRODUCTION

The work presented here is part of a large investigation into the use of computational methods for studying basic principles of expressive music performance [References]. One of the questions we study is whether and to what extent aspects of *individual artistic style* can be quantified. And one of the possible approaches to this question is to investigate whether machines can learn to distinguish and recognise different performers based on their style of playing.

Recent research has shown that this seems indeed possible, to a certain extent [6]. However, those studies were limited in many respects, particularly with regard to the data (only 2 pieces) and the musicians involved (piano teachers and students of a Music University). Thus we tried to generalise this research towards the analysis of famous world-class pianists, and to work with larger collections of recordings. The present paper describes our latest results along these lines. Six different learning algorithms are applied to the task of identifying the performer in a set audio recordings, by famous pianists, of several Mozart piano sonatas. It will be shown that in a pair-wise discrimination setting, surprisingly good results can be obtained, especially given the very limited information contained in the available measurement data. We will also show that the results partly carry over to music of a very different style, and will identify a few performance features that seem particularly relevant to style differentiation.

The paper is organised as follows: after a short introduction to the notion of expressive music performance (Section 2), Section 3 describes the experimental methodology, including the data, the performance features extracted from the recordings, and the machine learning algorithms tested. Section 4 presents the main experimental results. An interesting current research direction is briefly indicated in Section 5, followed by some conclusions in Section 6.

[1] Department of Medical Cybernetics and Artificial Intelligence, Medical University of Vienna, and Austrian Research Institute for Artificial Intelligence, Vienna; E-mail: gerhard@ai.univie.ac.at

[2] Austrian Research Institute for Artificial Intelligence, Vienna

2 EXPRESSIVE MUSIC PERFORMANCE

Expressive music performance is the art of shaping a musical piece by continuously varying important parameters like tempo, dynamics, etc., particularly in classical music. Human musicians do not play a piece of music mechanically, with constant tempo or loudness, exactly as written in the printed music score. Rather, they speed up at some places, slow down at others, stress certain notes or passages by various means, and so on. The most important parameter dimensions available to a performer (a pianist, for example) are timing and continuous tempo changes, dynamics (loudness variations), and articulation (the way successive notes are connected). Most of this is not specified in the written score, but at the same time it is absolutely essential for the music to be effective and engaging. The expressive nuances added by an artist are what makes a piece of music come alive, and what distinguishes great artists from each other. Educated audiences of classical music adore certain artists because of their particular style or 'sound', though they cannot always explicitly say what it really is that makes that style. One of the goals of the research presented here is to use AI techniques to get a better understanding of what factors really contribute to personal artistic style, and to what extent they can be quantified.

3 DATA AND METHODOLOGY

For the experiments, commercial recordings by six concert pianists of piano sonatas by W.A. Mozart were collected, and a sizeable number of pieces were selected for performance measuring and analysis. The pieces, pianists, and recordings are listed in Tables 1 and 2.

From the audio recordings, rough measurements characterising the performances were obtained. More precisely, changes of *tempo* and *general loudness* were measured at the level of the beats, by determining and marking the precise onset time of each beat, e.g., each 8th note position in a piece written in 6/8 time. (Even with the help of an intelligent interactive beat tracking system [2], this was an extremely laborious process.) From the varying time intervals between successive beats, the beat-level tempo changes can be derived. Overall loudness of the performance at these time points was extracted from the audio signal and is taken as a very crude representation of the dynamics applied by the pianists. No more detailed information (e.g., about articulation, individual voices, or timing details below the level of the beat) is available.

Table 1. Movements of Mozart piano sonatas selected for analysis.

ID	Sonata	Movement	Key	Time sig.
kv279_1	K.279	1st mvt.	C major	4/4
kv279_2	K.279	2nd mvt.	C major	3/4
kv279_3	K.279	3rd mvt.	C major	2/4
kv280_1	K.280	1st mvt.	F major	3/4
kv280_2	K.280	2nd mvt.	F major	6/8
kv280_3	K.280	3rd mvt.	F major	3/8
kv281_1	K.281	1st mvt.	Bb major	2/4
kv282_1	K.282	1st mvt.	Eb major	4/4
kv282_2	K.282	2nd mvt.	Eb major	3/4
kv282_3	K.282	3rd mvt.	Eb major	2/4
kv330_3	K.330	3rd mvt.	C major	2/4
kv332_2	K.332	2nd mvt.	F major	4/4

Table 2. Pianists and recordings.

ID	Name	Recording
DB	Daniel Barenboim	EMI Classics CDZ 7 67295 2, 1984
RB	Roland Batik	Gramola 98701-705, 1990
GG	Glenn Gould	Sony Classical SM4K 52627, 1967
MP	Maria João Pires	DGG 431 761-2, 1991
AS	András Schiff	ADD (Decca) 443 720-2, 1980
MU	Mitsuko Uchida	Philips Classics 464 856-2, 1987

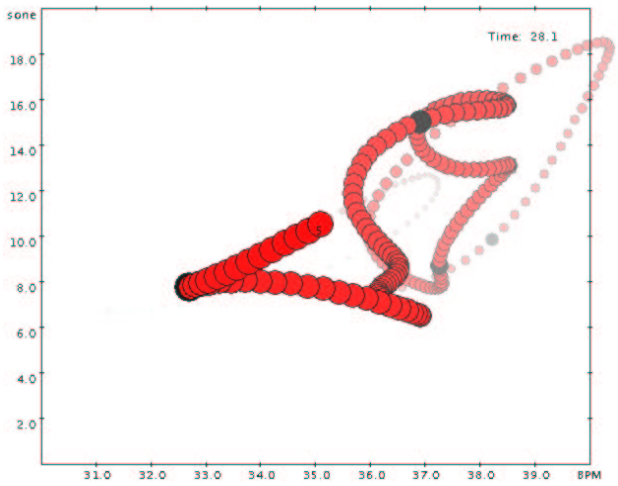

Figure 1. Snapshot of the *Performance Worm* at work: First four bars of Daniel Barenboim's performance of Mozart's F major sonata K.332, 2nd movement. Horizontal axis: tempo in beats per minute (bpm); vertical axis: loudness in *sone*. Movement to the upper right indicates a speeding up (*accelerando*) and loudness increase (*crescendo*) etc. etc. The darkest point represents the current instant, while instants further in the past appear fainter.

These sequences of measurements can be represented as two sets of performance curves — one representing variations in beat-level tempo over time, the other beat-level loudness changes — or in an integrated two-dimensional way, as trajectories over time in a 2D tempo-loudness space [4]. A graphical animation tool called the *Performance Worm* [3] displays such performance trajectories in synchrony with the music. A part of a performance as visualised by the Worm is shown in Figure 1. Note that the display is interpolated and smoothed. For the machine learning experiments reported below, only the actually measured points were used; no interpolation or smoothing was performed.

Thus, the raw data for our experiments is tempo and overall loudness values measured at specific time points in a performance (either every beat according to the time signature or, where beat tracking was performed at lower levels, at subdivisions of the beat). For each measured time point, the following is stored: t_i (absolute time in seconds), B_i (calculated local tempo in beats per minute (bpm)), and L_i (loudness level measured in the psycho-acoustic unit *sone*).

The raw data so obtained had to be refined in order to be homogeneous and usable in the learning process. In particular, most of the pianists repeated some sections, while others did not (e.g., Glenn Gould). We decided to discard all the non-common repeats, so that the learners would work with comparable data for all the performers.

3.1 Instances and features

Each measured time point, along with its context, is used as a *training example* for the learning algorithms. In other words, an example or *instance* for the learners is a subsegment of a tempo-loudness trajectory (see Fig. 1), centered around a specific time point. Altogether, this procedure results in some 23,000 instances for all the six pianists.

The instances are represented by a set of *features* that are extracted from the raw trajectories. The features are calculated over a window w_i of two bars (the context size), centered on the time point of the instance. Of course, at the beginning and at the end, some of the features were calculated over a window narrower than two bars. The sliding window method produces some redundancy in the data, since the windows are overlapping; this should be an advantage in learning.

For each instance of the original raw data, the following features are computed both for tempo and loudness: the average value within the window $\mu(w_i)$, the standard deviation $\sigma(w_i)$, and the range $R(w_i) = max(w_i) - min(w_i)$. For each of these features, the corresponding normalised ones are also calculated by division by the mean. The normalised features are indicated with lower-case letters of the tempo/loudness subscripts. For example, if $\sigma_B(w_i)$ is the tempo standard deviation, then $\sigma_b(w_i) = \sigma_B(w_i)/\mu_B(w_i)$ is the corresponding normalised version. Additional features represent correlations: $\Sigma_{tB}(w_i)$ is the correlation between time and tempo, $\Sigma_{tL}(w_i)$ is the correlation between time and loudness, and $\Sigma_{BL}(w_i)$ is the correlation between tempo and loudness. *Directness of movement* is a feature that captures aspects of the curvature of a trajectory segment: it measures the ratio between the length of a direct movement between the end points of a segment and the length of the actual trajectory between the same points in a two-dimensional space. The 'directness' is computed in the time-tempo space ($\Delta_{tB}(w_i)$), in the time-loudness space ($\Delta_{tL}(w_i)$) and in the tempo-loudness space ($\Delta_{BL}(w_i)$). Finally, some derivatives are computed for tempo and loudness: the maximum of the absolute value of the derivative ($M\delta(w_i)$), the average of the absolute derivative ($\mu\delta(w_i)$), and the normalised versions too.

Caution has to be taken with some of the extracted performance information. In particular, the features derived from loudness have to be filtered in some way, because they can trivially reveal some of the performers. For example, Gould's recordings are older (1967) than the others (1980-1991), resulting in a significantly lower recording level. That would permit the learners to detect this famous performer simply by absolute loudness difference or by differences in the dynamics range. That is why normalization, especially of the loudness features, is important, and why not all of the above features were then actually used in the learning experiments. Table 3 summarises all the features and indicates which ones must not be used in learning.

Table 3. Complete set of attributes and derived features extracted from the data for each instance. $^{*)}$ feature not used by the learner in order not to trivially reveal some of the performers on the basis of the CD recording level.

Operation	Tempo	Loudness	Others
None	B_i	L_i^*	–
Average, St. Dev., Range	$\mu_B(w_i)$, $\sigma_B(w_i)$, $R_B(w_i)$	$\mu_L(w_i)^*$, $\sigma_L(w_i)^*$, $R_L(w_i)^*$	–
Normalization	$b_i(w_i)$, $\sigma_b(w_i)$, $R_b(w_i)$	$l_i(w_i)$, $\sigma_l(w_i)$, $R_l(w_i)$	–
Correlation, Directness	$\Sigma_{tB}(w_i)$, $\Delta_{tB}(w_i)$	$\Sigma_{tL}(w_i)$, $\Delta_{tL}(w_i)$	$\Sigma_{BL}(w_i)$, $\Delta_{BL}(w_i)$
Derivative	$M\delta_B$, $\mu\delta_B$, $M\delta_b$, $\mu\delta_b$	$M\delta_L^*$, $\mu\delta_L^*$, $M\delta_l$, $\mu\delta_l$	–

3.2 The learning algorithms

For the experiments, we selected a representative set of standard machine learning algorithms with different model classes and biases. All of these are available in the Waikato Environment for Knowledge Analysis (WEKA) [8].

The following learners were selected: **NaiveBayes**, a simple probabilistic classifier based on Bayes' theorem of conditional probability; **IBk**, which implements a straightforward nearest-neighbour classifier (with the number of neighbours $k = 5$, in our case); **Classification Via Regression**, a simple 'meta-learner' that induces linear discriminant functions for the individual classes and combines these into a classifier via voting; **Logistic Regression**, where the (linear) basis functions are combined and converted into a classifier through a logistic function; **J48**, a state-of-the-art decision tree learner; and **JRIP**, an efficient rule learning algorithm. Table 4 summarises the algorithms in terms of the precise call (including parameters) by which they are called in WEKA.

Table 4. Learning algorithms used for pair-wise discrimination.

ID	WEKA name
01	bayes.NaiveBayes -K
02	lazy.IBk -K 5 -W 0
03	meta.ClassificationViaRegression -W LinearRegression
04	functions.Logistic -P 1.0E-13 -R 1.0E-8 -M 200
05	trees.j48.J48 -M 2 -C 0.25
06	rules.JRip

3.3 Training and Testing Methodology

For the experiments to be reported here, the original n-class problem was converted to $n(n-1)/2$ two-class discrimination problems, one for each possible pair of pianists. That gives more insight into the discriminability of various pianists, and is easier for a classifier than n-class identification. For each pianist pair A-B, the performances by the two pianists of the selected training pieces were used for learning, and the task was then to identify the correct pianist in a new test piece, where only recordings by A and B were used for testing.

Recognition accuracy was tested via cross-validation at the level of sonata movements. Each of the algorithms was trained on all of the sonata movements except one; the learned classifiers were then tested on recordings of the remaining movement. This process was repeated in a circular fashion, so that each piece served as test piece exactly once for each classifier.

Note that, as explained above, the actual training examples are not entire pieces, but individual time points in pieces, characterised by a set of features. The learned classifiers apply to such individual instances, not to entire pieces. To make a classifier predict the pianist for a complete piece, it was applied to all instances making up the piece, and the class (pianist) predicted most often was then chosen as the final prediction for the piece.

4 RESULTS

4.1 Recognition in Mozart Performances

The results of the cross-validation experiments are summarised in Table 5, which lists the numbers and percentages of correct predictions achieved by the individual classifiers on each pair of performers. For each pianist pair, the classifiers were tested on 24 recordings (12 pieces, played by each of the two pianists). Thus, the maximum possible number of correct predictions is 24, and the *baseline* accuracy — the success rate corresponding to pure guessing — is 12, or 50%. A reasonable learner should at least obtain a recognition accuracy above the baseline.

Looking at the table in terms of learning algorithms first, we note that while all learners obtain recognition rates significantly higher than the baseline, some learners may be better suited to this task than others; the overall prediction accuracies vary between 60 and 70% (Table 5, last row). A closer look shows that not all classifiers perform well or poorly on the same learning problem (pianist pair). That indicates that it may be fruitful to join classifiers into *ensembles* [1] which combine the expertise of a diverse set of classifiers by voting on the class of a new test case.

More interesting is a look at the results in terms of individual pianist pairs. The next-to-rightmost column in Table 5 lists the average recognition accuracy, over all classifiers, for each pair. The figures range from a rather low 53.5% for the pair Pires-Uchida (MP-MU) to quite good 75% for the pair Barenboim-Batik (DB-RB). These results may not look very exciting — they are far from a perfect recognition rate of 100% — but they are significantly better than the baseline in every case (except MP-MU). That indicates that our performance features do contain information that is relevant to the identification of artists. Also, the figures are averages over all classifiers, not all of which are very effective. If we focus only on the best of the 6 classifiers — classifier 04 (last column in Table 5) —, the results are much better, with recognition rates ranging from 54.2 % (still MP-MU) to 83.3% (MP-RB), which we consider quite impressive.

The average recognition rate over all experiments where a given pianist was involved gives a rough measure of the 'distinguishability' of that pianist. Computing this over all classifiers we get the following ranking: Batik (68.2%), Barenboim (65.8%), Schiff (65.1%), Gould (64.3%), Pires (64.0%), and Uchida (60.0%). Our best classifier, 04, yields similar results, but with higher recognition rates: here, the ordering is Batik (75.0%), Gould (72.5%), Schiff (69.2%), Barenboim (68.3%), Pires (67.5%), and Uchida (65.8%). Classifier 04 ranks Glenn Gould more highly in the list of recognisable pianists, something one might have expected from the start, given his reputation as an *enfant terrible* (also in musical terms). Even more distinguishable seems to be Roland Batik (a local Viennese pianist).

Table 5. Pair-wise recognition results: classification accuracy in terms of correctly classified pieces. The maximum possible number of correct predictions in each pianist pair is 24 (12 pieces $\times$ 2 pianists); the *baseline accuracy* is 50%. Rightmost column: accuracy (percentage) achieved by classifier **04**.

	Classifiers								
Pair	**01**	**02**	**03**	**04**	**05**	**06**	**Ave.**	**Ave.** [%]	**Classifier 04** [%]
AS - DB	16	14	15	16	16	15	15.3	63.9	66.7
AS - GG	14	16	17	17	14	13	15.2	63.2	70.8
AS - MP	15	17	18	16	17	16	16.5	68.8	66.7
AS - MU	14	14	15	16	11	13	13.8	57.6	66.7
AS - RB	15	17	17	17	19	19	17.3	72.2	70.8
DB - GG	16	15	17	18	14	13	15.5	64.6	75.0
DB - MP	16	16	16	15	15	15	15.5	64.6	62.5
DB - MU	15	16	15	15	14	13	14.7	61.1	62.5
DB - RB	17	19	19	19	17	17	18.0	75.0	79.2
GG - MP	15	17	16	17	14	15	15.7	65.3	70.8
GG - MU	16	17	16	18	13	14	15.7	65.3	75.0
GG - RB	12	15	17	17	14	16	15.2	63.2	70.8
MP - MU	12	15	13	13	13	11	12.8	53.5	54.2
MP - RB	14	16	20	20	13	15	16.3	68.1	83.3
MU - RB	14	17	16	17	14	12	15.0	62.5	70.8
Average	14.7	16.1	16.5	16.7	14.5	14.5	–	–	–
Average [%]	61.4	66.9	68.6	69.7	60.6	60.3	–	–	–

Listening to his performances with well-trained ears, we find that his Mozart indeed sounds quite different compared to the more famous artists (though he is a highly skilled pianist). The least easily distinguishable pianists seem to be Pires and (particularly) Uchida. Indeed, our learners achieve the lowest recognition rates when trying to distinguish the two directly (see row MP-MU in Table 5).

4.2 'Closed-world' Classification

The evaluation presented above is suboptimal, from the classifiers' point of view, because they were not able to use all the information they have available. Remember that classification of a piece is done by classifying all the instances (time points) that make up the piece, and finally predicting the class that is predicted more often for the instances. The *ratio* of votes for class A vs. votes for class B, over all instances, would actually give the classifier a notion of relative *confidence* in its prediction. One situation where they could exploit this information is what me might call *'closed-world classification'*.

Assume the classifier is always given a *pair* of recordings and is told that one is by pianist A and the other by pianist B. In such a situation the classifier could do the following: collect the class predictions over the instances of both pieces, check for which piece the ratio of class A predictions vs. class B predictions is more unbalanced, i.e., for which piece it feels more confident in predicting, make the coresponding prediction for that piece, and automatically assign the opposite class to the other piece. As a consequence, the learner will either get both predictions right, or both wrong.

Table 6 shows the result of this prediction procedure for our best classifier from above (04). The results are dramatically better, with the classifier achieving perfect identification rates for more than half of the pianist pairs. The average number of correct predictions per pair is 22.67 or **94.4%**. Whether this kind of pairwise classification is a realistic application scenario is a matter of debate, but at least the result clearly shows that the classifier manages to capture some important structure in the data.

Table 6. Results of 'closed-world' classification (classifier 04). Perfect results are printed in bold.

Pair	**hits**	**Pair**	**hits**	**Pair**	**hits**
AS-DB	22	DB-GG	**24**	GG-MU	**24**
AS-GG	20	DB-MP	**24**	GG-RB	22
AS-MP	**24**	DB-MU	22	MP-MU	22
AS-MU	18	DB-RB	**24**	MP-RB	**24**
AS-RB	22	GG-MP	**24**	MU-RB	**24**

4.3 Relevant Features

An inspection of the learned models as well as more extensive experiments with automatic feature selection methods reveals those features that contribute most strongly to correct prediction and may thus point to some important facets of personal performance style. Among the most informative features seem to be the following: the *instantaneous tempo* at a particular time point (B_i), the *maximum value of the tempo derivative* within a window ($M\delta_B$), and the *correlation between tempo and loudness* ($\Sigma_{BL}(w_i)$). In the learned models, B_i usually serves as a kind of 'secondary' classification criterion that acts in conjunction with others, helping distinguish the artists based on what they do at different tempi. The apparent importance of $M\delta_B$ as a disciminator makes intuitive sense; the feature obviously captures one aspect of a pianist's "expressivity" in the timing domain, namely, quick and extreme local tempo changes. $\Sigma_{BL}(w_i)$, finally, measures to what extent a pianist synchronises tempo and loudness, that is, combines loudness increases and decreases (*crescendi* and *decrescendi*) with corresponding changes in tempo (*accelerandi* and *ritardandi*, respectively). This is particularly interesting, because it corroborates findings from a different study on characteristic differences between pianists [7].

4.4 Extrapolation: From Mozart to Chopin

An interesting question is how general and robust the induced classifiers are relative to *different styles* of music. We happen to also have measured a few recordings, by two of the pianists considered above, of pieces by the Romantic composer *Frédéric Chopin*: the *Nocturnes* op.9 No.2 in E♭ major, op.15 No.1 in F major, op.27 No.1 in C♯ minor, and op.27 No.2 in D♭ major. The two pianists are Daniel Barenboim (DB) and Maria João Pires (MP). The four pieces were segmented into sections of different musical character (2 sections for op.9/2 and three each for op.15/1, op.27/1 and op.27/2). Thus we have 22 test cases: 11 Chopin pieces played by 2 pianists.

The following experiment was performed: the learners were trained on all 12 Mozart pieces as performed by DB and MP. The

induced classifiers were then tested on the Chopin recordings. The best classifier in this case turned out to be the nearest-neighbour classifier (02), with a recognition rate of $15/22 = 68.2\%$, followed by logistic regression (04, $14/22 = 63.6\%$) and Naive Bayes and CVR (01, 03; $13/22 = 59.1\%$). The other two learners failed to learn anything that seems transferable to Chopin and produced prediction accuracies around the baseline.

And again, the results improve considerably when we consider a 'closed-world' classification scenario (see above): here the recognition rates rise to an astounding $18/22 =$ **81.1%** for classifer 02 (IBk).

Recognition rates of 68.2% or even 81.1% are quite remarkable, given the differences between the training corpus (Mozart) and the test pieces (Chopin); the Chopin pieces are much more romantic, and all very slow. Indeed, it would be a very difficult task for a human listener to do the same: listen to a few Mozart recordings by two different artists, and then recognise the artists in Chopin recordings (although the human listener would hear much more detail in the recording than what is captured in our features).

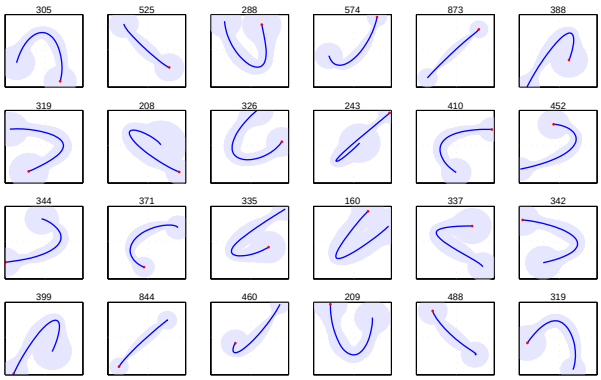

Figure 2. A 'Mozart performance alphabet' (cluster prototypes) computed by segmentation, mean and variance normalization, and clustering, from Mozart performances by Daniel Barenboim, Roland Batik, Vladimir Horowitz, Maria João Pires, András Schiff, and Mitsuko Uchida. To indicate directionality, dots mark the end points of segments. Shaded regions indicate the variance within a cluster.

5 CURRENTLY ONGOING WORK

The representation and learning approach described above is extremely simplistic in that it completely ignores the *sequential nature* of music. Each time point in a performance is described and classified in isolation. The features attached to a time point do capture some of the local context, but clearly, this is very limited information, and one would expect a lot of artist-specific information to reside in the succession of expressive tempo-loudness variations over time. Taking sequential information into account requires different learning algorithms. We are currently exploring ways of doing this with learning methods based on *string kernels* [5]. The basic approach is a follows: we create so-called *performance alphabets* (see Fig. 2) by cutting the performance trajectories into short segments of a fixed length (e.g., four beats), applying various normalisation operations on the segments to make them invariant with respect to absolute tempo and loudness, clustering the segments into a fixed number of categories, and computing a prototype (the centroid) for each cluster. These prototypes then represent prototypical elementary tempo-loudness shapes that the original trajectories are basically composed of. In other words, full performance trajectories can be approximated as sequences of such tempo-loudness patterns; if we associate each prototype with a letter, a full performance trajectory then is a *string*.

Clearly, we lose additional information in this process of clustering and prototype computation, but we gain the ability to consider sequential information. Very preliminary results obtained on our performance strings with string kernel methods (this is work being carried out by J. Shawe-Taylor and C. Saunders at the University of Southampton) are quite promising and indicate that sequential information may substantially improve the recognition rates achievable.

6 CONCLUSION

This paper has presented experimental evidence that machines may be capable of recognising famous artists on the basis of their style, at least to some extent. The machine learning algorithms achieved recognition rates significantly above chance level. At first sight, some of the absolute accuracy figures may look rather poor. However, our current training data (performance measurements) are extremely crude and incomplete, comprising only beat-level tempo and beat-level overall loudness. A lot of information that listeners hear and that influences their judgement is missing, for example, information about the relative loudness and timing of the individual voices, about articulation (legato vs. staccato), etc. — in short, a lot of things that relate to the specific 'sound' of a pianist. This information is very hard to obtain from audio recordings; in fact, it is almost impossible to measure in an automated way.

Considering this limitation, the results are actually quite surprising. We believe it would be hard, if not impossible, for even highly educated listeners to achieve comparable recognition rates under the same conditions — though it is difficult to specify exactly how to produce such 'comparable conditions'; listeners bring a lot of knowledge and experience to bear (about musical patterns and styles, etc.) that would have to be eliminated somehow from the experiment.

An analysis of the learned models gives first hints as to which aspects of an artist's performances may contribute to his or her personal and recognisable style. We are currently conducting a parallel investigation, using new visualisation techniques combined with statistical analysis, in order to get more direct insight into what exactly distinguishes great artists. We are confident that by continuing this line of research, we may make progress towards better understanding some of the 'mysteries' that surround the style of great artists.

ACKNOWLEDGEMENTS

This research was and is supported by a generous START research prize by the Austrian Federal Government (FWF project no. Y99-INF), an ERASMUS scholarship to the second author, and by the EU Project SIMAC (6th FP, 507142). The Austrian Research Institute for Artificial Intelligence acknowledges basic financial support by the Austrian Federal Ministry for Education, Science, and Culture, and the Federal Ministry of Transport, Innovation, and Technology.

REFERENCES

[1] T. G. Dietterich, *Ensemble Methods in Machine Learning*, First International Workshop on Multiple Classifier Systems, Springer Verlag, New York, 2000.

[2] S. Dixon, 'Automatic extraction of tempo and beat from expressive performances', *Journal of New Music Research*, **30**(1), 39–58, (2001).

[3] S. Dixon, W. Goebl, and G. Widmer, *The Performance Worm: Real Time Visualization of Expression Based on Langner's Tempo-Loudness Animation*, Proceedings of the International Computer Music Conference (ICMC 2002), Goeteborg, Sweden, 2002.

[4] J. Langner and W. Goebl, 'Visualizing expressive performance in tempo-loudness space', *Computer Music Journal*, **27**(4), 69–83, (2003).
[5] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, 'Text classification using string kernels', *Journal of Machine Learning Research*, **2**, 419–444, (2002).
[6] E. Stamatatos and G. Widmer, *Music Performer Recognition Using an Ensemble of Simple Classifiers*, Proceedings of the 15th European Conference on Artificial Intelligence (ECAI'2002), Lyon, France, 2002.
[7] G. Widmer, S. Dixon, W. Goebl, E. Pampalk, and A. Tobudic, 'In Search of the Horowitz Factor', *AI Magazine*, **24**(3), 111–130, (2003).
[8] I. H. Witten and E. Frank, *Data Mining*, Morgan Kaufmann, San Francisco, CA, 1999.