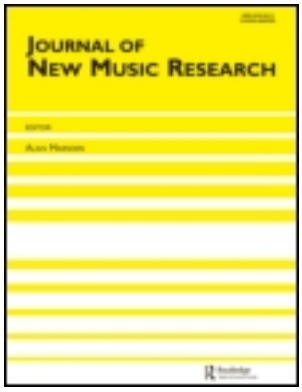

Journal of New Music Research
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/nnmr20

Multiple Viewpoint Systems for Music Classification

Darrell Conklin [a]
[a] Universidad del Pais Vasco UPV/EHU, Spain, and IKERBASQUE, Basque Foundation for Science, Spain
Version of record first published: 21 Mar 2013.

To cite this article: Darrell Conklin (2013): Multiple Viewpoint Systems for Music Classification, Journal of New Music Research, 42:1, 19-26

To link to this article: http://dx.doi.org/10.1080/09298215.2013.776611

Journal of New Music Research, 2013
Vol. 42, No. 1, 19–26, http://dx.doi.org/10.1080/09298215.2013.776611

Multiple Viewpoint Systems for Music Classification

Darrell Conklin

Universidad del País Vasco UPV/EHU, Spain, and IKERBASQUE, Basque Foundation for Science, Spain

Abstract

This paper describes a new statistical modelling method for music classification. The method is an extension of the multiple viewpoint method for music prediction and generation. A multiple viewpoint system significantly outperforms all component viewpoints on the tasks of folk tune genre and region classification. The method is successfully applied to predict the genres of unlabelled Basque folk tunes.

1. Introduction

Music classification is the task of assigning one or more class labels to pieces based on their content. Evaluation of classification methods is usually according to the accuracy of class label assignment to unseen pieces. The contrasting task of music prediction is to anticipate the events that arise during the unfolding of a piece. Music prediction methods use generative models to assign probabilities to events within pieces, and models are evaluated according to how well they can predict unseen events. Despite the fundamental differences with class label prediction, generative models for event prediction may be extended for classification tasks.

Machine learning methods for symbolic music classification can be divided into two categories. In one category are methods where pieces are converted to finite vectors of global features, each feature being a variable taking on a single value computed from the content of a piece. In this standard setting a large range of discriminative machine learning methods can be readily applied to music classification (Ponce de León & Iñesta, 2003; McKay & Fujinaga, 2004; Moreno-Seco, Iñesta, Ponce de León, & Micó, 2006; Hillewaere, Manderick, & Conklin, 2009). The other category is the language modelling approach, where statistical models generate all events in entire pieces. In this category are *n*-gram and hidden Markov models, both having great success in domains as diverse as text classification and protein sequence classification. Various forms of statistical language models have proven to be powerful for music classification and prediction (Conklin & Witten, 1995; Triviño-Rodriguez & Morales-Bueno, 2001; Pearce, Conklin, & Wiggins, 2004; Li, Ji, & Bilmes, 2006; Conklin, 2006; Gilbert & Conklin, 2007; Temperley, 2007; Pérez-Sancho, Rizo, & Iñesta, 2008).

Music presents interesting challenges for machine learning, particularly in terms of data and knowledge representation. In contrast to simple time series data, and like other types of structured sequences, events have multiple discrete attributes (in the case of notes: pitch, duration, onset time, etc.). Music is structured at multiple levels—rhythmic, melodic, and harmonic—and each of these dimensions can be described by many different facets. This observation was the basis for the development of *multiple viewpoint systems* for music prediction (Conklin & Witten, 1995), which represent each facet with a separate statistical model and combine the predictions of all models. The intuition behind multiple viewpoints is that no single music representation can be sufficient for music, and that the predictions of an ensemble of viewpoints can be combined to produce better models. The multiple viewpoint method has formed the basis of several successful music prediction and generation methods (Pachet, 2003; Assayag & Dubnov, 2004; Pearce et al., 2004; Chordia, Sastry, Mallikarjuna, & Albin, 2010) and the idea can naturally be extended to music classification, drawing on research in Bayesian text classification (Peng et al., 2004) and classifier ensembles (Kuncheva, 2004).

Ensemble methods are a general machine learning strategy that combine the outputs of several base classifiers, achieving good performance if the component classifiers tend to make uncorrelated classification errors (Dietterich, 2000). The principal decisions in designing an ensemble classifier system are the design of base classifiers and the

Correspondence: Darrell Conklin, Department of Computer Science and Artificial Intelligence, Universidad del País Vasco UPV/EHU, San Sebastián, Spain, and IKERBASQUE, Basque Foundation for Science, Bilbao, Spain. E-mail: conklin@ikerbasque.org

choice of a classifier fusion method. More detailed considerations involve the automated construction of an ensemble from base classifiers and the measurement of the diversity of an ensemble to guide this choice (Kuncheva, 2004). In music, ensemble classifiers have previously been reported for applications such as polyphonic cover song detection (Rizo, Iñesta, & Lemström, 2011) and genre classification (Moreno-Seco et al., 2006; Pérez-Sancho et al., 2008).

Folk tune classification is an interesting and challenging problem in music informatics, used for the inference of different properties of songs such as place name, social function, and tune family (Li et al., 2006; Conklin, 2009; Hillewaere et al., 2009; van Kranenburg, 2010). The methods developed in this paper are applied to four different folk tune datasets, illustrating the two tasks of genre and geographic region classification. For genre classification, tunes are labelled with their single annotated genre, though it is recognized that similar tunes can have different social functions (Selfridge-Field, 2006). On all datasets multiple viewpoint systems are found to significantly outperform individual viewpoints on folk tune classification. Finally, the method is successfully applied to genre prediction in a set of unlabelled Basque folk tunes.

2. Methods

This section describes the multiple viewpoint method for sequence classification. First, the viewpoint representation for music is introduced and illustrated on a monophonic texture. Second, the statistical modelling method for computing the probability of a sequence using a single viewpoint is described. Third, the method for combining the predictions of multiple viewpoints for sequence classification is presented. Finally, the four corpora used for evaluating the method are described.

2.1 Melody representation using viewpoints

Melodies are represented at the surface level as sequences of *events* that have the *basic features* of pitch, duration, and onset time (Figure 1, top). It is inadequate to model only these dimensions, as they are not robust to changes in the music surface, and practically any corpus will be too sparse to reveal distinctive frequent patterns. Particularly for the pitch dimension, any representation that is not invariant under transposition will alone be inadequate for any corpus with pieces in varying tonalities. Therefore the music surface is transformed into more abstract sequences by applying functions called *viewpoints*.

Developing this idea more precisely, an event is a conjunction of basic features. A *viewpoint* is a function mapping events to more abstract *derived features*. The function is partial, therefore it may be undefined ($\perp$) for some events. An event sequence $e_1, \ldots, e_\ell$ is transformed by iterative application of a viewpoint τ to produce the *transformed sequence* $\tau(e_1), \ldots, \tau(e_\ell)$. A viewpoint can refer to the entire contextual sequence preceding an event: for notational

Viewpoint	Transformed sequence										
pitch	67	69	71	72	69	72	64	67	72	69	b
dur	2	3	1	3	1	2	3	1	4	2	b
onset	0	2	5	6	9	10	12	15	16	20	b
int	$\perp$	2	2	1	-3	3	-8	3	5	-3	d
ioi	$\perp$	2	3	1	3	1	2	3	1	4	d
ml0	1	0	0	0	0	1	0	0	0	1	d
intref	7	9	11	0	9	0	4	7	0	9	d
pcontour	$\perp$	+	+	+	−	+	−	+	+	−	d
dcontour	$\perp$	+	−	+	−	+	+	−	+	−	d
dratio	$\perp$	3:2	1:3	3:1	1:3	2:1	3:2	1:3	4:1	1:2	d
int $\otimes$ ioi	$\perp$	$2,2$	$2,3$	$1,1$	$-3,3$	$3,1$	$-8,2$	$3,3$	$5,1$	$-3,4$	c
int $\otimes$ intref	$\perp$	$2,9$	$2,11$	$1,0$	$-3,9$	$3,0$	$-8,4$	$3,7$	$5,0$	$-3,9$	c
intref $\otimes$ ml0	$7,1$	$9,0$	$11,0$	$0,0$	$9,0$	$0,1$	$4,0$	$7,0$	$0,0$	$9,1$	c
int $\oslash$ ml0	$\perp$	$\perp$	$\perp$	$\perp$	$\perp$	5	$\perp$	$\perp$	$\perp$	-3	c
dur $\downarrow$ ml0	2	$\perp$	$\perp$	$\perp$	$\perp$	2	$\perp$	$\perp$	$\perp$	2	c
$\lhd$ pitch	0	0	0	0	-3	-2	0	-7	-3	-5	c

Fig. 1. A solution array for a small fragment from a Basque folk tune (Euskomedia code 3178). Each line shows a different viewpoint with its transformation of the event sequence. The timing resolution is one tick per sixteenth note. At the top bracket are the (b)asic viewpoints; the middle some (d)erived viewpoints; and at the bottom some (c)onstructed (linked, threaded, selected, and backreference) viewpoints.

convenience here when an event e_j is notated as an argument to a viewpoint τ, the argument $e_1, \ldots, e_j$ is implied.

The application of k viewpoints $\tau_1, \ldots, \tau_k$ to an event sequence $e_1, \ldots, e_\ell$ may be represented as a $k \times \ell$ solution array where location (i,j) holds the value $\tau_i(e_j)$. Figure 1 shows a solution array for a short fragment of a Basque folk melody. At the top are the basic features, followed by some examples of derived viewpoints, and finally some examples of viewpoints built using *constructors*, which are functions that take viewpoints as arguments and return new viewpoints. Three of the examples are *linked* viewpoints, which directly represent the interaction between other viewpoints, for example, a linked viewpoint $\mathsf{int} \otimes \mathsf{ioi}$ of melodic interval and inter-onset interval, meaning that we represent every event as a pair of its melodic interval and time difference from the start of the previous event. Next is an example of a *threaded* viewpoint $\mathsf{int} \oslash \mathsf{ml0}$, which captures relations between events not contiguous on the music surface, for example, the melodic interval between events at the strongest (metric level 0) beat in a bar. Note that this derived feature sequence contains mostly undefined ($\perp$) features. The next example is a *selected* viewpoint $\mathsf{dur} \downarrow \mathsf{ml0}$ which selects the duration feature at the strongest beat in a bar. The final example uses the *backreference* constructor $\triangleleft$, which for a particular viewpoint (in this case pitch) returns a backpointer to the closest identical value in the sequence, or 0 if the value is novel up to that point in the sequence.

2.2 Single viewpoint model

This section describes the method of viewpoints for symbolic music prediction (Conklin & Witten, 1995), a method similar to *class-based* models of natural language (Brown, Della Pietra, deSouza, Lai, & Mercer, 1992). Consider a viewpoint τ, an event sequence $\mathbf{e} = e_1, \ldots, e_\ell$, and its transformed sequence $\mathbf{v} = \tau(e_1), \ldots, \tau(e_\ell)$. The probability of the event sequence can be written in the following factored form:

$$\begin{aligned} P(\mathbf{e}) &= P(\mathbf{e},\ \mathbf{v}) \quad \text{since } \mathbf{v} \text{ is determined by } \mathbf{e} \\ &= P(\mathbf{v}) \times P(\mathbf{e} \mid \mathbf{v}) \quad \text{product rule} \\ &= \prod_{j=1}^{\ell} P(\tau(e_j) \mid \widehat{\tau}(e_j)) \times \prod_{j=1}^{\ell} P(e_j \mid \tau(e_j)), \end{aligned} \tag{1}$$

where $\widehat{\tau}$ returns the sequence of all defined features preceding the indicated event. This is done by applying τ to each event in a sequence, and then filtering out undefined ($\perp$) features. The function $\widehat{\tau}$ therefore maps concrete event sequences into equivalence classes of transformed sequences.

The first term on the right-hand side of Equation 1 is the conditional probability of the derived feature of an event given its history of derived features, thus making the assumption that each derived feature is independent of any future events in the sequence (this assumption is not strictly necessary for classification tasks, when the entire sequence is presented to a classifier at once, however none of the viewpoints used in this paper look at future events). This term is conveniently modelled by an n-gram model of transformed sequences, as will be explained further in Section 3. The second term is the probability of the concrete event given the derived feature and may be estimated by maximum likelihood over the corpus (Brown et al., 1992), or simply modelled by an uniform distribution over all events having the derived feature (Conklin & Witten, 1995).

Two simplifications (Conklin, 2006) are implicit within Equation 1: first, to handle boundary values in pitch and duration, the probability of a derived feature (first term) must take into account the event context and not only the transformed sequence. For example, if the previous event is at an upper boundary pitch, then any melodic interval feature greater than zero should have a zero probability. Second, to compute the probability of an event given some derived feature (second term) it may be necessary also to know the event context. For example, the probability of any higher pitch given a falling pitch contour feature must be zero.

The model described by Equation 1 can be given a procedural explanation in terms of generation or prediction (Conklin, 2003). To generate a sequence using a random walk, first sample a derived feature given the current history of the transformed sequence, then sample an event with respect to this transformed feature. This process is then repeated with the new event added to the generated sequence. To predict an event in a sequence, construct the probability distribution of events based on the current history of the sequence, noting the probability of the event that actually occurs. When this process of event prediction is performed for a set of pieces from a corpus (for pieces not used to train the model), the negative logarithm of the product of all probabilities gives the *cross-entropy* which measures how well a model predicts events in the corpus.

In the method for music prediction described by Conklin and Witten (1995), a model is created from an ensemble of viewpoints which are interpolated to make event predictions, a *short-term* model is constructed adaptively for the piece being predicted, and this is interpolated with a *long-term* model constructed from a large corpus. In this way repetition within the current piece can be modelled, significantly reducing cross-entropy (Conklin & Witten, 1995; Pearce et al., 2004) according to the amount of intra-opus repetition. The prediction method can be extended, as described below, to music classification (where a short-term model is unnecessary as it is assumed to be independent of the class label) using multiple viewpoints (where the class label predictions of an ensemble are interpolated).

2.3 Classification using multiple viewpoints

The model of the previous section may be extended to a multiclass classification scheme, with a Bayesian scheme similar to methods used for text classification

(Peng et al., 2004). Applying Bayes' formula to compute the posterior probability of a class label given a sequence, the following derivation can be made by the standard assumption of equal probabilities of all classes, and dropping terms that do not depend on the class label:

$$\begin{aligned} P(c \mid \mathbf{e}) &= P(c \mid \mathbf{e},\ \mathbf{v}) \quad \text{since } \mathbf{v} \text{ is determined by } \mathbf{e} \\ &= P(\mathbf{e},\ \mathbf{v} \mid c) \times P(c) \ / \ P(\mathbf{e},\ \mathbf{v}) \quad \text{Bayes' rule} \\ &\propto P(\mathbf{e},\ \mathbf{v} \mid c) \quad \text{assuming equal class priors} \\ &= P(\mathbf{v} \mid c) \times P(\mathbf{e} \mid \mathbf{v},\ c) \quad \text{product rule} \\ &= P(\mathbf{v} \mid c) \times P(\mathbf{e} \mid \mathbf{v}) \quad \text{conditional independence} \\ &= \prod_{j=1}^{\ell} P(\tau(e_j) \mid \widehat{\tau}(e_j),\ c) \times \frac{1}{Z} \\ & Z \text{ a normalizing constant} \end{aligned} \tag{2}$$

in the second last step assuming conditional independence of $\mathbf{e}$ and c given $\mathbf{v}$: conditioning an event on a derived feature does not depend on the class label.

The posterior probability described by Equation 2 can be used for multiclass classification using an ensemble of viewpoints. Suppose that we have k viewpoints $\tau_1, \ldots, \tau_k$, and therefore k corresponding transformed sequences $\mathbf{v}_1, \ldots, \mathbf{v}_k$ for an event sequence $\mathbf{e}$. Figure 2 illustrates the combination of the k viewpoints: each viewpoint τ_i computes the posterior probability $P(c \mid \mathbf{e},\ \mathbf{v}_i)$ for each class $c \in \{c_1, \ldots, c_m\}$, and their fusion gives a combined prediction $\mu_c(\mathbf{e})$. After the $\mu_c(\mathbf{e})$ have been computed for all classes, the prediction c^* of the entire ensemble is the class label c with the maximum $\mu_c(\mathbf{e})$:

$$c^* = \arg\max_c\ \mu_c(\mathbf{e}). \tag{3}$$

To compute the $\mu_c(\mathbf{e})$ as a fusion of k component viewpoints, many different classifier fusion methods are possible, such as geometric and arithmetic means, voting schemes, and various weighted variants (Kuncheva, 2004). Tests on the folk tune datasets discussed in Section 2.4 have revealed that the geometric mean of posterior probabilities (Equation 2) generally performs well:

$$\mu_c(\mathbf{e}) = \left(\prod_{i=1}^{k} P(c \mid \mathbf{e},\ \mathbf{v}_i) \right)^{1/k} \tag{4}$$

Figure 2 depicts the decision matrix of viewpoints and classes, leading to the final prediction (Equation 3) of the ensemble.

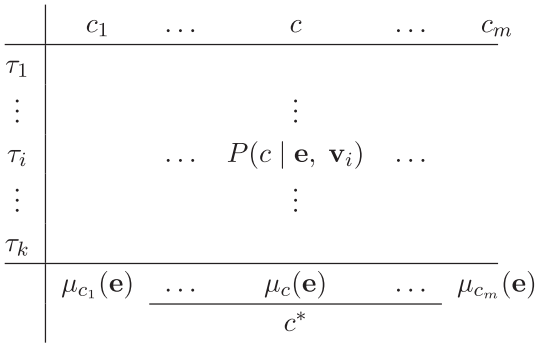

Fig. 2. Multiple viewpoint systems for classification, showing the fusion of individual classifiers for a given event sequence $\mathbf{e}$. The predictions of all viewpoints τ_i are combined, leading to a prediction strength $\mu_c(\mathbf{e})$ for each class c. The maximum of the $\mu_c(\mathbf{e})$ leads to the final ensemble prediction c^*.

2.4 Folk tune datasets

This section describes the four folk tune datasets used to evaluate the multiple viewpoint classification method. Properties of each dataset (the corpus size, the mean number of notes, and the majority class accuracy) are summarized at the bottom of Table 1. Melodic phrase boundaries and lyrics have not been captured in the encodings. For all corpora all pieces within the corpus are standardized to the same timing resolution (midi ticks per beat). Key and time signature metadata are encoded, and these features are indeed referenced by some viewpoints (`intref` and `ml0`).

The Cancionero Vasco is a collection of Basque dance and song melodies, compiled by the musicologist and priest Padre Donostia in the year 1912 as part of a Basque songbook competition run by the Basque government. Recently the entire collection has been compiled in four volumes (de Riezu, 1996) and digitized, a process overseen by the Euskomedia Foundation (Usurbil, Spain) and the Eresbil Foundation (Renteria, Spain). The collection comprises 1902 songs, each with rich metadata information containing, for each song, the classification provided by Padre Donostia into a genre or song type (e.g. dance, religious, etc.) and the geographic location where the song was collected. The Cancionero Vasco uses a total of 24 disjoint genre labels for songs, and about 3000 distinct location labels organized in a hierarchy containing the levels of territory, municipality, and town. Some pieces in the collection have homophonic texture, ranging from a few parallel or doubled notes, to longer parallel supported melodies. These textures were removed by preprocessing the entire corpus using a skyline algorithm which extracts the event with the highest pitch at each unique onset time.

To test the methods presented in this paper, the Cancionero Vasco was partitioned along two different dimensions: first in terms of geographic region, and second in terms of genre or song type. The first is a dataset called *terr-7*, with each song labelled with its territory, the highest level in the geographic region hierarchy. This dataset has 1630 songs (272 of the original 1902 songs do not have a territory annotation), with the class distribution *Alava*: 27; *Gipuzkoa*: 175; *Navarra*: 891; *Lapurdi*: 383; *Bizkaia*: 21; *Nafarroa Beherea*: 53; and *Zuberoa*: 80. An analogous corpus for European folk tunes is a corpus called *europa-6* of

3367 folk tunes from six different European countries (*Scotland*: 445 tunes; *France*; 393; *England*: 990; *Sweden*: 618; *Ireland*: 798; and *Greece*: 123). This corpus, previously studied by Hillewaere et al. (2009), is a subset of a larger corpus of 3724 folk tunes created by Li et al. (2006) which has been cleaned by removing pieces with nearly identical melodic content.

Turning to song type, many of the 24 genres in the Cancionero Vasco are sparsely populated (e.g. nine genres have less than 10 songs), and therefore for this study only the three largest genres of *danza* (495 tunes), *amorosa* (247 tunes), and *religiosa* (209 tunes) are considered. This smaller corpus called *vasca-3* comprises 951 songs, covering 61% of the labelled pieces. The analogous European corpus is a corpus derived from *europa-6* called *dance-9* (Hillewaere, Manderick, & Conklin, 2012) comprising 2198 folk tunes annotated with one of nine different dance types (*reel*: 453 tunes; *waltz*: 128; *marsch*: 76; *jig*: 793; *polka*: 339; *strathspey*: 123; *hornpipe*: 108; *bourree*: 59; and *schottisch*: 119).

3. Results

This section presents results with the multiple viewpoint method for music classification on the four datasets described in Section 2.4, and also some results on the prediction of song type for the unlabelled songs in the Cancionero Vasco. Relative performance of different classifiers is evaluated using either a cross-validated paired t-test (Kuncheva, 2004) between viewpoints, or alternatively using a two proportion z-test (Kuncheva, 2004) between viewpoints and other published methods.

3.1 Folk tune classification

To evaluate the methods on the four folk tune datasets, several viewpoints of Figure 1 were used to train the models for each class for each corpus. As described in Conklin and Witten (1995), for each viewpoint a suffix tree data structure is created from all transformed sequences in the training set, capturing repeated patterns for each viewpoint. Then for each piece in the test set, and for every event, a variable-length 5-gram model was used to compute the posterior probability of the class for each viewpoint (Equation 2) and thereby the final predicted class label (Equation 3), with method C smoothing (Jurafsky & Martin, 2000) used to interpolate predictions from matching contexts of various lengths, and also to avoid zero probabilities.

For each corpus 10-fold cross-validation (stratified: for each class approximately 10% of class members are present in each validation set) was used to estimate classification accuracies (the percentage of pieces assigned the correct label). In addition to the evaluation of single viewpoints as classifiers, three different multiple viewpoint systems were evaluated. The model MVS(4) comprises the first four viewpoints listed in Table 1 and is the same system which provided consistently good performance on music prediction (Conklin & Witten, 1995; Pearce et al., 2004). The model MVS(12) contains all 12 viewpoints listed in the top part of Table 1. For comparison with MVS(12), the saturated linked viewpoint SAT(12) which simply links together the same 12 viewpoints is illustrated. Finally, the model MVS(28) contains all possible dyadic linked viewpoints formed from the eight following viewpoints: pitch, int, ioi, ml0, intref, pcontour, dcontour, and dratio.

Table 1. Classification accuracy viewpoint models on the folk tune datasets, estimated by 10-fold stratified cross-validation. Top: a selection of single viewpoints; middle: three multiple viewpoint systems and the saturated viewpoint; bottom: some properties of the datasets.

	Corpus			
Viewpoint	*terr-7*	*europa-6*	*vasca-3*	*dance-9*
pitch	48.2	65.3	67.5	61.9
int ⊗ intref	48.6	64.0	67.3	64.6
intref ⊗ ml0	47.7	66.8	68.2	71.7
int ⊗ ioi	48.2	72.4	71.0	83.1
pcontour	19.8	38.7	57.6	38.9
int	48.2	64.1	69.0	66.1
ioi	30.6	56.2	66.0	76.1
ml0	8.7	33.8	43.0	40.2
dcontour	19.3	46.2	58.4	64.6
dratio	32.3	55.1	63.1	74.4
int ⊘ ml0	29.5	38.1	56.7	32.5
intref	46.6	62.7	68.1	62.6
MVS(4)	56.1	76.1	75.7	84.4
MVS(12)	56.1	76.7	77.3	87.3
SAT(12)	33.1	59.8	57.1	69.2
MVS(28)	58.8	79.2	77.6	88.7
corpus size	1630	3367	951	2198
mean events	73	98	113	102
majority class	54.7	29.4	52.1	36.1

The dashed boxed figures in Table 1 are the best single viewpoints for the particular corpus. For all datasets, a single linked viewpoint int ⊗ ioi of melodic interval and inter-onset interval attains a high accuracy relative to other single viewpoints. As expected, some single viewpoints have low accuracy: for example, the ml0 viewpoint which will mainly be modelling a phenomenon similar to the number of notes in a bar. It is interesting that the general int ⊘ ml0 viewpoint, which simply models the interval between pitches at strong beats of bars, performs above majority class accuracy on two of the datasets. The general dratio and intref viewpoints perform surprisingly well but still with lower accuracy than the int ⊗ ioi viewpoint.

The ensemble classifier MVS(4) significantly ($\alpha = 0.05$) outperforms the best of any single viewpoint on all corpora: *terr-7* ($p = 8.4\text{e}{-06}$), *europa-6* ($p = 0.00021$), *vasca-3* ($p = 0.021$), and *dance-9* ($p = 0.035$). Moving from MVS(4) to MVS(12), it is interesting that the addition of viewpoints, many with low individual accuracies on the

datasets, does not decrease the classification accuracy of the ensemble. The saturated linked viewpoint SAT(12), while performing better than many single viewpoints, performs in all cases worse than MVS(12), showing the power of unlinking component viewpoints into a distributed multiple viewpoint representation. From MVS(12) to MVS(28) the classification accuracy is further improved, in fact MVS(28) obtains the best performance of all methods on all datasets (solid boxed figures in Table 1).

The results of multiple viewpoint systems on *europa-6* can be compared with an earlier study on the same corpus. Another general technique for classifying music is to first convert pieces to vectors of global features amenable to standard machine learning methods. In this context Hillewaere et al. (2009) studied five different machine learning methods and four different global feature sets on the *europa-6* corpus. The best result obtained in that study was a classification accuracy of 69.7, using a support vector machine classifier on a pooled set of 150 global features. By contrast, the model MVS(28) obtains an accuracy of 79.2, significantly exceeding global feature vector methods on the *europa-6* dataset ($p = 4.0\text{e}{-19}$). The model MVS(28) significantly ($p = 5.1\text{e}{-11}$) outperforms the statistical language modelling method of Li et al. (2006) which achieved an accuracy of 72.5 on a superset of the *europa-6* corpus even though their larger corpus of 3724 pieces contains many similar pieces which should be relatively easy to classify.

In general *dance-9* is the easiest of the four datasets, as even simple rhythmic viewpoints (e.g. ioi, dratio) perform adequately relative to the majority class accuracy. In fact, the relative performance of a basic melodic (int) and rhythmic (ioi) feature is reversed on this corpus. Also particularly interesting is that only the MVS(28) classifier significantly improves upon the majority class accuracy on the *terr-7* corpus, in contrast to the analogous *europa-6* corpus, on which all single and multiple viewpoints achieve good results. This surprising observation suggests that the melodic and rhythmic content carry little information predictive of the geographic region of collection of Basque folk tunes. This could be explained by the sharing of tunes throughout the territories of the Basque country and by the lack of a distinctive melodic or rhythmic style within territories.

3.2 Genre prediction for unlabelled songs

In addition to the annotated songs, in the Cancionero Vasco there are 341 songs that were not annotated with a genre at the time of song collection. It is hypothesized that predictive data mining methods can be used to predict a possible class for each of these, and also for new songs collected by musicologists in the future.

Table 2 shows in detail the results of MVS(12) on the *vasca-3* corpus, along with the classifier precision (probability of class correct given class predicted) and recall (probability of class predicted given actual class) for each class. It can be seen that the numbers vary considerably, with higher precision on the class *danza* than on either *amorosa* or *religiosa*. The classifier precision values can be used to indicate the confidence of class predictions for unlabelled songs.

Given the high precision of MVS(12) on the *danza* class, the decision was made to focus on which of the 341 unlabelled songs might be dances. Therefore from the entire corpus of 1902 pieces, the labelled 1561 pieces were used to create the new corpus *vasca-2* containing just two classes: *danza* (495 tunes) and *other* (1066). Estimated by 10-fold cross-validation, the classification accuracy of MVS(12) on *vasca-2* is 88.5. The precision of MVS(12) on the *danza* class is 0.851, acceptably high to make predictions for unlabelled songs. The recall on the *danza* class in *vasca-2* is 0.772, somewhat lower than on the *vasca-3* corpus (Table 2) due to the presence of all 23 other genres in the *other* class.

During the evaluation of MVS(12) on *vasca-2*, the unlabelled 341 songs were not used for training but were retained in stratified validation sets. A total of 44 unlabelled pieces were thereby predicted to be in the class

Table 2. Confusion matrix and MVS(12) precision and recall on the *vasca-3* corpus, estimated by 10-fold stratified cross-validation.

	Predicted			
	religiosa	*amorosa*	*danza*	Recall
religiosa	101	75	33	0.483
amorosa	36	176	35	0.713
danza	11	26	458	0.925
Precision	0.682	0.635	0.871	

Table 3. Five unlabelled songs predicted to be dances.

Code	Title	Comments
7590	Makil-dantza	the song title clearly indicates dance type: probably missing annotation.
3409	Para publicar bandos	this song and the next song (3410) are members of a group of four 'Duos de trompeta', which includes 3407 ('Marcha de procesion') and 3408 ('Para salida y entrada del exmo. ayuntamiento'), both annotated as dances.
3410	Para salida del toro	see song 3409
2658	La trin kuli kuli kun	melodic similarity comparisons show that this melody is identical to 3008, which is annotated as a dance: probably missing annotation.
3178	Danza	the song title clearly indicates dance type: probably missing annotation.

danza, and given the recall on the *danza* class it is predicted that $44/0.772 \approx 57$ dances are present in the set of unlabelled songs. Interestingly, this is less than the expected number ($495/1561 \times 341 \approx 108$), suggesting that the unlabelled part of the Cancionero Vasco does not follow the same class distribution as the labelled portion. Given the precision of 0.851 on the *danza* class in *vasca-2*, $44 \times 0.851 \approx 37$ of the predicted dances are expected to actually be dances.

Table 3 presents a few examples of songs that have strong supporting evidence for being dances. This evidence ranges from an analogous annotation (code 2658) supported by melodic similarity (detected by alignment of melodic interval sequences), through to information contained within song annotations or title. Particularly interesting is the prediction of 3409 and 3410 as *danza*: these two pieces have no detectable melodic similarity to either 3407 or 3408, the other two members of the group of four pieces.

4. Conclusions

The method of viewpoints for music (Conklin & Witten, 1995) is a knowledge representation method for derived event features and transformed sequences, and a statistical modelling method for event prediction and generation. Multiple viewpoint systems provide a distributed representation for prediction of music by combining the predictions of a set of component viewpoints. This paper has further extended the method of multiple viewpoints to an ensemble classifier method for music classification. Evaluated on four folk tune datasets, on the tasks of genre and region classification, it was found that multiple viewpoint systems using a simple classifier fusion scheme perform significantly better than any individual viewpoint. This indicates that the abstraction of the music surface provided by viewpoints enables the detection of distinctive patterns within classes, and that linking and combining the predictions of viewpoints is an effective modelling method for symbolic music classification. This paper also shows that folk song melodies can be effectively classified based on melodic and rhythmic content alone, without reference to other presumably relevant information such as lyrics, tempo, dynamics, timbre, social context of the performance, and performer or instrumentation, though the interpolation with such features may further improve the results.

For the Basque folk tune corpus, the results suggest that geographic region classification based on melodic content alone appears to be a difficult classification problem. For genre classification the multiple viewpoint model performs well, suggesting that genres contain distinctive melodic and rhythmic patterns that can be represented by viewpoints. The method was used with success to predict potential dance tunes in the unlabelled portion of Basque folk tunes.

The viewpoint model has some similarities to the factored language model (Li et al., 2006) in that linked viewpoints model the joint distribution of two or more attributes. With factored language models, multiple attributes of events in a sequence can be coupled in ways more flexible than the simple mechanism provided by linked viewpoints. However, the factored language model does not accommodate abstract derived features of events, nor the fusion of multiple models. Another method which has similarities to the multiple viewpoint model is the multi-stream statistical model of Kirchhoff and Parandekar (2001), which was effectively applied for language identification from a speech signal. In that approach an abstract feature for an event can be conditioned on its context in the same stream though in another stream only to the feature at the same event position.

The viewpoint model can also be compared and contrasted with the hidden Markov model, which models the joint probability of an event sequence and a parallel, hidden state sequence. A major difference here is that though a transformed sequence is also hidden, an event has only one feature for a viewpoint (which is a function, not a general relation). Therefore in a viewpoint model there is no expensive hidden state sequence decoding (referring to Equation 1, there is no need for a summation over the possible transformed sequences for a given viewpoint) and the n-gram models for viewpoints need not be restricted to bigram models. Furthermore, in a multiple viewpoint system an event may have multiple derived sequences. The factorial hidden Markov model (Ghahramani & Jordan, 1997) is an extension of the hidden Markov model to handle multi-attribute event sequences. Similar to standard hidden Markov models, the state sequences are not determined uniquely by the event sequences, and learning and inference is a complex process.

The methods presented here on folk tune classification may be adapted for other music classification tasks where events have multiple basic features, and where derived features can be designed and computed for events within sequences. The method is not restricted to melodic textures, as viewpoints can be used to represent homophonic (Conklin, 2002) and polyphonic (Conklin & Bergeron, 2010) textures, segmented melodies (Conklin, 2006), and chord sequences (Conklin, 2010). An intriguing idea for future research is to combine predictions for different textures, using different multiple viewpoint systems for each texture type. Future work could also explore integrating sequence classification with methods for event prediction.

Acknowledgments

The Fundación Euskomedia and the Fundación Eresbil are graciously thanked for making the Cancionero Vasco available for study. Thanks to Ruben Hillewaere for compiling the *europa-6* and *dance-9* datasets, and to Kerstin Neubarth for comments on the manuscript. This

research was partially supported by a grant *Análisis Computacional de la Música Folclórica Vasca* (2011–2012) from the Diputación Foral de Gipuzkoa, Spain.

References

Assayag, G., & Dubnov, S. (2004). Using factor oracles for machine improvisation. *Soft Computing, 8*, 604–610.

Brown, P., Della Pietra, V., deSouza, P., Lai, J., & Mercer, R. (1992). Class-based n-gram models of natural language. *Computational Linguistics, 18*(4), 467–479.

Chordia, P., Sastry, A., Mallikarjuna, T., & Albin, A. (2010). Multiple viewpoints modeling of tabla sequences. In *ISMIR 2010: 11th International Society for Music Information Retrieval Conference*, Utrecht, The Netherlands, pp. 381–386.

Conklin, D. (2002). Representation and discovery of vertical patterns in music. In C. Anagnostopoulou, M. Ferrand, & A. Smaill (Eds.), *Music and Artificial Intelligence* (Lecture Notes in Artificial Intelligence Vol. 2445, pp. 32–42). Berlin: Springer-Verlag.

Conklin, D. (2003). Music generation from statistical models. In *Proceedings of the AISB Symposium on Artificial Intelligence and Creativity in the Arts and Sciences*, Aberystwyth, Wales, pp. 30–35.

Conklin, D. (2006). Melodic analysis with segment classes. *Machine Learning, 65*(2–3), 349–360.

Conklin, D. (2009). Melody classification using patterns. In *MML 2009: International Workshop on Machine Learning and Music*, Bled, Slovenia, pp. 37–41.

Conklin, D. (2010). Discovery of distinctive patterns in music. *Intelligent Data Analysis, 14*(5), 547–554.

Conklin, D., & Bergeron, M. (2010). Discovery of contrapuntal patterns. In *ISMIR 2010: 11th International Society for Music Information Retrieval Conference*, Utrecht, The Netherlands, pp. 201–206.

Conklin, D., & Witten, I. (1995). Multiple viewpoint systems for music prediction. *Journal of New Music Research, 24* (1), 51–73.

de Riezu, P.J. (1996). Cancionero Vasco P. Donostia. *Revista Internacional de los Estudios Vascos, 41*, 189–190.

Dietterich, T.G. (2000). Ensemble methods in machine learning. In J. Kittler & F. Roli (Eds.), *First International Workshop on Multiple Classifier Systems* (Lecture Notes in Computer Science Vol. 1857, pp. 1–15). Berlin: Springer Verlag.

Ghahramani, Z., & Jordan, M. (1997). Factorial hidden Markov models. *Machine Learning, 29*, 245–275.

Gilbert, E., & Conklin, D. (2007). A probabilistic context-free grammar for melodic reduction. In *IJCAI 2007: International Workshop on Artificial Intelligence and Music*, Hyderabad, India, pp. 83–94.

Hillewaere, R., Manderick, B., & Conklin, D. (2009). Global feature versus event models for folk song classification. In *ISMIR 2009: 10th International Society for Music Information Retrieval Conference*, Kobe, Japan, pp. 729–733.

Hillewaere, R., Manderick, B., & Conklin, D. (2012). String methods for folk tune genre classification. In *ISMIR 2012: 13th International Society for Music Information Retrieval Conference*, Porto, Portugal, pp. 217–222.

Jurafsky, D., & Martin, J. (2000). *Speech and Language Processing*. Englewood Cliffs, NJ: Prentice-Hall.

Kirchhoff, K., & Parandekar, S. (2001). Multi-stream statistical n-gram modeling with application to automatic language identification. In *Proceedings of Eurospeech*, Aalborg, Denmark, pp. 803–806.

Kuncheva, L.I. (2004). *Combining Pattern Classifiers: Methods and Algorithms*. New Jersey: Wiley.

Li, X., Ji, G., & Bilmes, J. (2006). A factored language model for quantized pitch and duration. In *International Conference on Computer Music*, New Orleans, LA, pp. 556–563.

McKay, C., & Fujinaga, I. (2004). Automatic genre classification using large high-level musical feature sets. In *ISMIR 2004: Proceedings of the International Conference on Music Information Retrieval*, Barcelona, Spain, pp. 525–530.

Moreno-Seco, F., Iñesta, J.-M., Ponce de León, P., & Micó, L. (2006). Comparison of classifier fusion methods for classification in pattern recognition tasks. In *Structural, Syntactic, and Statistical Pattern Recognition, Joint IAPR International Workshops, SSPR 2006 and SPR 2006*, Hong Kong, pp. 705–713.

Pachet, F. (2003). The Continuator: musical interaction with style. *Journal of New Music Research, 32*(3), 333–341.

Pearce, M., Conklin, D., & Wiggins, G. (2004). Methods for combining statistical models of music. In *Computer Music Modeling and Retrieval: Second International Symposium*, Esbjerg, Denmark, pp. 295–312.

Peng, F., Schuurmans, D., & Wang, S. (2004). Augmenting naive Bayes classifiers with statistical language models. *Information Retrieval, 7*(3), 317–345.

Pérez-Sancho, C., Rizo, D., & Iñesta, J.-M. (2008). Stochastic text models for music categorization. In N.D. Vitoria Lobo, T. Kasparis, M. Georgiopoulos, F. Roli, J. Kwok, G.C. Anagnostopoulos, & M. Loog (Eds.), *Proceedings of the 12th International Workshop on Structural and Syntactic Pattern Recognition* (Lecture Notes in Computer Science (Vol. 5342, pp. 55–64). Berlin: Springer-Verlag.

Ponce de León, P., & Iñesta, J.-M. (2003). Feature-driven recognition of music styles. In *Proceedings of the 1st Iberian Conference on Pattern Recognition and Image Analysis: Lecture Notes in Computer Science*, (vol. 2652, pp. 773–781). Springer-Verlag.

Rizo, D., Iñesta, J.-M., & Lemström, K. (2011). Polyphonic music retrieval with classifier ensembles. *Journal of New Music Research, 40*(4), 313–324.

Selfridge-Field, E. (2006). Social cognition and melodic persistence: Where metadata and content diverge. In *ISMIR 2006: 7th International Conference on Music Information Retrieval*, Victoria, Canada, pp. 272–275.

Temperley, D. (2007). *Music and Probability.* Cambridge, MA: The MIT Press.

Triviño-Rodriguez, J., & Morales-Bueno, R. (2001). Using multiattribute prediction suffix graphs to predict and generate music. *Computer Music Journal, 25*(3), 62–79.

van Kranenburg, P. (2010). *A computational approach to content-based retrieval of folk song melodies* (PhD thesis). Utrecht University.