COMBINING CONTENT-BASED AUTO-TAGGERS WITH DECISION-FUSION

Emanuele Coviello
University of California, San Diego
ecoviell@ucsd.edu

Riccardo Miotto
University of Padova
miottori@dei.unipd.it

Gert R. G. Lanckriet
University of California, San Diego
gert@ece.ucsd.edu

ABSTRACT

To automatically annotate songs with descriptive keywords, a variety of content-based auto-tagging strategies have been proposed in recent years. Different approaches may capture different aspects of a song's musical content, such as timbre, temporal dynamics, rhythmic qualities, etc. As a result, some auto-taggers may be better suited to model the acoustic characteristics commonly associated with one set of tags, while being less predictive for other tags. This paper proposes *decision-fusion*, a principled approach to combining the predictions of a diverse collection of content-based auto-taggers that focus on various aspects of the musical signal. By modeling the correlations between tag predictions of different auto-taggers, decision-fusion leverages the benefits of each of the original auto-taggers, and achieves superior annotation and retrieval performance.

1. INTRODUCTION

The recent age of music proliferation has raised the need for automatic algorithms to efficiently search and discover music. Many successful recommendation systems rely on textual metadata provided by expert musicologists or social services in the form of semantic tags – keywords or short phrases that capture relevant characteristics of music pieces, ranging from genre and instrumentation, to mood and usage. By bridging the gap between music and human semantics, tags allow semantic retrieval based on transparent textual descriptions, or query-by-example recommendation based on semantic similarity (as opposed to acoustic similarity) to a query song.

Meta-data-based methods work well in practice, provided that enough annotations are available. However, the cold start problem and the prohibitive cost of manual labour limit their applicability to large-scale applications. Therefore, the deployment of modern music recommendation systems can benefit from the development of auto-taggers, i.e., machine-learning algorithms that automatically analyze and index music with semantic tags, which can then be used to improve the search experience and speed up the discovery of desired content.

1.1 Previous work

Most auto-taggers are based on music content analysis and are trained from a database of annotated songs (e.g., see [8, 10, 12, 20]). After extracting a set of acoustic features from each training song, a series of statistical models are estimated, each of which capturing the characteristic acoustic patterns in the songs that are associated with one of the tags from a given vocabulary. When analyzing a new song, the auto-tagger processes the time series of acoustic features of the song and outputs a vector of tag-affinities. The affinity-vector can then be transformed into a semantic multinomial (SMN), i.e., a probability distribution characterizing the relevance of each tag to a song. A song is then annotated by selecting the top-ranking tags in its SMN, or the SMN itself can be used as a high-level descriptor, e.g., for retrieving songs based on semantic similarity. A number of discriminative (e.g., see [3, 8, 9, 12, 18, 23]) and generative (e.g., see [10, 17, 20, 21]) machine learning algorithms have been proposed to model predictive acoustic patterns in audio content based on a bag-of-features (BoF) representation, which treats audio features independently and ignores their temporal order. Recently, Coviello et al. [6] proposed to leverage dynamic texture mixture (DTM) models for auto-tagging purposes. More precisely, DTM-based auto-taggers model audio fragments (i.e., time series of audio features extracted from a few seconds of musical signal) as the output of linear dynamical systems. This approach explicitly captures temporal structures in the musical signal, whereas a BoF representation discards such dynamics.

At a higher level of abstraction, contextual approaches have focused on modeling the semantic context that drives the correlation between different tags (e.g., a song tagged with "drums" is more likely to also be tagged with "electric guitar" than "violin"). While content-based models oper-

ate on low-level acoustic features to predict semantic multinomials, contextual models are designed to capture meaningful tag correlations in these SMNs, to reinforce accurate tag predictions while suppressing spurious ones. So, a contextual model naturally complements a content-based model, which usually treats tags independently. Combining them has been shown to improve performance. State-of-the-art solutions are based on discriminative approaches (e.g., support vector machines [14], boosting [1], ordinal regression [24]) as well as generative models (e.g., Dirichlet mixture models (DMM) [13]).

1.2 Original contribution

The main contribution of this paper is to propose *decision-fusion*, which uses semantic context modeling to simultaneously leverage the benefits of different content-based auto-taggers. Using two or more content-based auto-taggers that emphasize diverse aspects of the musical signal (e.g., only timbre vs. temporal dynamics), we collect alternative opinions on each song-tag association. We expect that, besides modeling the context between tags predicted from the same auto-tagger, context modeling can capture the correlations that arise between tag predictions based on different auto-taggers, leading to a more sophisticated system.

This offers a solution to the problem of selecting or combining alternative annotation models that previous work has pointed out. Coviello et al. [6], for example, noted that even though their DTM-based auto-tagger generally outperformed a BoF approach based on Gaussian mixture models (GMM), the improvements were most significant on tags with clear temporal characteristics; for some tags, in fact, the GMM-based model was still favorable (i.e., tags where "timbre says it all").

Experimental results show that decision-fusion leads to improved annotation and retrieval performance compared to i) each individual auto-tagger, ii) each individual auto-tagger in tandem with a contextual model (the "traditional" context-based approach) and iii) various other approaches to combining multiple content-based auto-taggers, such as fixed-combination rules and the regression-based combination algorithms proposed by Tomasik et al. [19]. We note that the focus of the latter was slightly different from our work, since it investigates the combination of tags predicted from different information sources (i.e., content-based auto-tags, social tags, collaborative-filtering-based tags), rather than from different content-based auto-taggers only. In addition, as semantic context modeling is naturally complementary to any content-based auto-tagger, we corroborate the intuition that there is a benefit in combining DTM-based temporal modeling and semantic context modeling, which has not been shown before.

The remainder of this paper is organized as follows. A brief review of the automatic music tagging problem and the models used in this work are presented in Section 2. Section 3 discusses decision-fusion. Lastly, the experimental setup and results are reported in Sections 4 and 5, respectively.

2. AUTOMATIC MUSIC TAGGING

The automatic task of music tagging is widely tackled as a supervised multi-class labeling problem [2], where each class corresponds to a tag w_i of a semantic vocabulary $\mathcal{V}$ (e.g., "rock", "drum", "tender", "mellow"). The music content of a song is represented as a time series of low-level acoustic features $\mathcal{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T\}$, where each feature is extracted from a short snippet of the audio signal and T depends on the length of the song. The semantic content with respect to $\mathcal{V}$ is represented as an annotation vector $\boldsymbol{c} = (c_1, \ldots, c_{|\mathcal{V}|})$, where $c_i > 0$ only if there is a positive association between a song and the tag w_i. The goal of an auto-tagging system is to infer the relevant semantic annotations of unseen songs.

At this aim, a set of statistical models is trained to capture the patterns in the audio feature space associated with each tag in $\mathcal{V}$, from a database $\mathcal{D} = \{(\mathcal{Y}_d, \boldsymbol{c}_d)\}_{d=1}^{|\mathcal{D}|}$ of annotated songs. Based on the learned tag models, the auto-tagger can process the acoustic features extracted from a novel song $\mathcal{Y}$ and produce a vector of tag-affinities, which is mapped into a semantic multinomial $\boldsymbol{\pi} = (\pi_1, ..., \pi_{|V|})$ lying on a semantic space (i.e., $\sum_i \pi_i = 1$ with $\pi_i \geq 0$), where $\pi_i = P(w_i|\mathcal{Y})$ represents the probability that the i^{th} tag applies to song $\mathcal{Y}$.

In order to leverage high level relationships that arise in the tag predictions of content-based auto-taggers, contextual approaches additionally introduce a second modeling layer to capture meaningful tag correlations in the SMNs. In particular, a content-based auto-tagger is used to produce a SMN $\boldsymbol{\pi}_d$ for each song $\mathcal{Y}_d$ in $\mathcal{D}$, while a second layer of statistical models is trained onto $\{(\boldsymbol{\pi}_d, \boldsymbol{c}_d)\}_{d=1}^{|\mathcal{D}|}$, to capture which patterns in the SMNs are predictive for each tag. For a novel song $\mathcal{Y}$, the contextual tag models can therefore be used to refine the semantic multinomial $\boldsymbol{\pi}$ produced by the content-based auto-tagger.

Music annotation involves finding the tags that best describe a song; this is achieved by selecting the subset of tags that peak in its semantic multinomial. Retrieval given a one-tag query, requires ranking all songs in a database based on their relevance to the query, e.g., the corresponding entry in the semantic multinomials [20].

In the following we review a variety of content-based auto-tagging strategies, where low-level acoustic content is represented either as a bag-of-features (Sections 2.1.1 and 2.1.2) or as a time series of features (Section 2.1.3). Additionally, Section 2.2 introduces a contextual approach for modeling tag correlations as well.

2.1 Content modeling

Content-based auto-taggers have been designed to model the acoustic content associated with tags and represented as a bag-of-features using both generative and discriminative models, as in Sections 2.1.1 and 2.1.2, respectively; conversely, the use of time series of audio features for music tagging has been considered in the generative approach of Section 2.1.3 only.

2.1.1 The Gaussian mixture model (GMM)

Turnbull et al. [20], proposed to capture the most prominent acoustic textures associated to each tag w_i in $\mathcal{V}$ with a probability distribution $p(\boldsymbol{y}|w_i)$ over the space of audio features $\boldsymbol{y}$, which is a Gaussian mixture model (GMM):

$$p(\boldsymbol{y}|w_i) = \sum_{r=1}^{R} a_r^{w_i} \mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}_r^{w_i}, \boldsymbol{\Sigma}_r^{w_i}) \,, \tag{1}$$

where R is the number of mixture components, $\mathcal{N}(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $a_r^{w_i}$ the mixing weights. The parameters $\{a_r^{w_i}, \boldsymbol{\mu}_r^{w_i}, \boldsymbol{\Sigma}_r^{w_i}\}_{r=1}^{R}$ of each tag model $p(\boldsymbol{y}|w_i)$ are *estimated* from the bag-of-features extracted from the songs in $\mathcal{D}$ that are positively associated with w_i, using the hierarchical expectation-maximization (EM) algorithm [22].

Given the audio content of a new song $\mathcal{Y} = \{\boldsymbol{y}_1, ..., \boldsymbol{y}_T\}$, the relevance of each tag w_i is computed using the Bayes rule:

$$\pi_i = P(w_i|\mathcal{Y}) = \frac{p(\mathcal{Y}|w_i)\, P(w_i)}{p(\mathcal{Y})}, \tag{2}$$

where $P(w_i)$ is the tag prior (assumed to be uniform) and $p(\mathcal{Y})$ the song prior, i.e., $p(\mathcal{Y}) = \sum_{j=1}^{|\mathcal{V}|} p(\mathcal{Y}|w_j)P(w_j)$. The likelihood term in (2) is computed as the geometric average of the individual sequence likelihoods, i.e., $p(\mathcal{Y}|w_i) = \prod_{t=1}^{T} p(\boldsymbol{y}_t|w_i)^{\frac{1}{T}}$.

2.1.2 Boosting (BST)

The boosting approach proposed by Eck et al. [8] is a supervised discriminative algorithm that learns a binary classifier for each tag w_i in the vocabulary $\mathcal{V}$, from both the positive and the negative training examples for that tag. More specifically, it constructs a *strong classifier* which combines a set of simpler classifiers, called *weak learners*, in an iterative way. As weak learners, according to [1], we use single stumps (i.e., binary thresholding on one low-level acoustic feature).

A novel song $\mathcal{Y}$ is classified by each of the binary classifiers and Platt scaling is applied to produce a probability estimate $\pi_i = P(w_i|\mathcal{Y})$ for each tag w_i. We will refer to this approach as BST.

2.1.3 Temporal modeling (DTM)

Coviello et al. [6] proposed a novel auto-tagger built upon the DTM model, which explicitly captures both the timbral and the temporal structures of music that are most predictive for each tag. Specifically, the dynamic texture (DT) model [7] treats an audio fragment $\boldsymbol{y}_{1:\tau}$ as output of a linear dynamical system. The model consists of a double embedded stochastic process, in which a lower dimensional Gauss-Markov process x_t encodes the dynamics (evolution) of the acoustic component $\boldsymbol{y}_t$ over time

Each tag distribution is modeled with a dynamic texture mixture (DTM) [4] probability density over sequences of audio feature vectors:

$$p(\boldsymbol{y}_{1:\tau}|w_i) = \sum_{r=1}^{R} a_r^{(w_i)} p(\boldsymbol{y}_{1:\tau}|\Theta_r^{(w_i)}) \,, \tag{3}$$

where R is the number of mixtures and $\Theta_r^{(w_i)}$ is the r^{th} DT component. The parameters $\{a_r^{(w_i)}, \Theta_r^{(w_i)}\}_{r=1}^{R}$ are estimated based on the audio fragments extracted from the songs in $\mathcal{D}$ positively associated with the tag w_i, using an efficient hierarchical EM algorithm for DTM (HEM-DTM) [5].

Given the audio fragments extracted from a new song $\mathcal{Y} = \{\boldsymbol{y}_{1:\tau}^1, \ldots, \boldsymbol{y}_{1:\tau}^F\}$, where F depends on the length of the song, the relevance of tag w_i is computed using Bayes' rule (2), with the likelihood computed as the geometric average of the individual sequence likelihoods smoothed by the sequence length τ, i.e., $p(\mathcal{Y}|w_i) = \prod_{t=1}^{F} p(\boldsymbol{y}_{1:\tau}^t|w_i)^{\frac{1}{F\tau}}$.

2.2 Context modeling (DMM)

As mentioned in Section 1.1, different approaches have been proposed to model contextual relationships in SMNs; in this work, we use the DMM [13]. The DMM is a generative model that assumes the SMNs $\boldsymbol{\pi}$ of the songs positively associated to a tag w_i are distributed accordingly to a mixture of Dirichlet distributions over the semantic space defined by $\mathcal{V}$:

$$p(\boldsymbol{\pi}|w_i; \Omega^w) = \sum_{r=1}^{R} \beta^{w_i} \mathrm{Dir}(\boldsymbol{\pi}|\alpha_r^{w_i}) \,, \tag{4}$$

where R is the number of mixtures, $\beta_k^{w_i}$ are the mixing weights, and $\mathrm{Dir}(\cdot|\boldsymbol{\alpha})$ is a Dirichelet distribution of parameters $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_{|\mathcal{V}|})$. The parameters of the DMM for each tag w_i in $\mathcal{V}$ are estimated from the semantic multinomials extracted from the songs in $\mathcal{D}$ positively associated with the tag, via the generalized EM algorithm [16].

Hence, given a new song described by the SMN $\boldsymbol{\pi} = (\pi_1, ..., \pi_{|\mathcal{V}|})$, the relevance of a tag w_i is computed using Bayes' rule to get the tag posterior probabilities in the context space:

$$\theta_i = P(w_i|\boldsymbol{\pi}) = \frac{p(\boldsymbol{\pi}|w_i)P(w_i)}{p(\boldsymbol{\pi})} \,. \tag{5}$$

All the tag posterior probabilities form the *contextual multinomial* distribution of the song, i.e., $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{|\mathcal{V}|})$, which can then be used for semantic annotation and retrieval.

3. DECISION-FUSION

Each content-based auto-tagger generally emphasizes particular aspects of the musical signal. Despite some auto-taggers could be preferred over others based on average performances (Table 1, part (a)), the spread in performances registered on specific tags (e.g., see Figure 1) makes unclear if any auto-tagger may be the best. This leaves open the problem of choosing the most appropriate method for each tag, or, indeed, the one of combining different auto-taggers.

In this paper we argue that semantic context modeling can also be used as a strategy to combine different content-based auto-taggers, which we name *decision-fusion*. Indeed, by modeling the patterns that arise from the tag predictions generated by different content-based auto-taggers, decision-fusion combines all the different opinions into a single prediction and leverages the benefits of each of the acoustic characteristics emphasized by the original auto-taggers.

Formally, let us assume a group $\mathcal{A}$ of different content-based auto-tagging algorithms is available. For each song d in the database $\mathcal{D}$, semantic multinomials $\boldsymbol{\pi}_d^a$ for $a = 1, \ldots, |\mathcal{A}|$ are computed (i.e., one for each auto-tagger in $\mathcal{A}$) and pooled together into the aggregated semantic multinomial:

$$\boldsymbol{\pi}_d^{\mathcal{A}} = (\boldsymbol{\pi}_d^1, \ldots, \boldsymbol{\pi}_d^{|\mathcal{A}|}) \,, \tag{6}$$

which is intended to be normalized to sum to 1. In practice, it is as we are now working with a new semantic vocabulary $\mathcal{V}^{\mathcal{A}} = \mathcal{V}^1 \times \cdots \times \mathcal{V}^{|\mathcal{A}|}$ of size $|\mathcal{A}| \cdot |\mathcal{V}|$, where each tag is replicated $|\mathcal{A}|$ times, one for each auto-tagger. Decision-fusion consists in training a set of semantic context models, i.e., $p(\boldsymbol{\pi}^{\mathcal{A}}|w_i)$ for $w_i = 1, \ldots, |\mathcal{V}|$, over the aggregated semantic multinomials $\{(\boldsymbol{\pi}_d^{\mathcal{A}}, \boldsymbol{c}_d)\}_{d=1}^{|\mathcal{D}|}$ to capture *both* intra- and inter-auto-taggers tag correlations. Note that traditional context modeling acts on the SMNs of a *single* auto-tagger, thus capturing *only* intra-auto-tagger correlations.

Decision-fusion can be implemented through a variety of context-modeling algorithms. In particular, in this work we tested the DMM presented in Section 2.2. Therefore, the aggregated SMNs $\boldsymbol{\pi}^{\mathcal{A}}$ of songs positively associated with tag w_i are assumed to be distributed accordingly to a mixture of Dirichlet distributions over the semantic space $\mathcal{V}^{\mathcal{A}}$:

$$p(\boldsymbol{\pi}^{\mathcal{A}}|w_i) = \sum_{r=1}^{R} \beta^{w_i} \mathrm{Dir}(\boldsymbol{\pi}^{\mathcal{A}}|\alpha_r^{w_i}) \,, \tag{7}$$

where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_{|\mathcal{A}|\cdot|\mathcal{V}|})$.

An unseen song $\mathcal{Y}$ is first processed by each of the content-based auto-taggers available to produce the semantic multinomials $\boldsymbol{\pi}^a$ for $a = 1, \ldots, |\mathcal{A}|$, which are then aggregated in $\boldsymbol{\pi}^{\mathcal{A}}$. Finally, Bayes' rule as in Equation 5 is applied to compute the posteriors $\theta_i^{\mathcal{A}} = p(w_i|\boldsymbol{\pi}^{\mathcal{A}})$ for each tag w_i, and to form a *decision-fusing multinomial* $\boldsymbol{\theta}^{\mathcal{A}} = (\theta_1^{\mathcal{A}}, \ldots, \theta_{|\mathcal{V}|}^{\mathcal{A}})$.

4. EXPERIMENTAL SETUP

4.1 Dataset

In our experiments, we used the CAL500 dataset [20], which consists of 502 popular Western songs by as many different artists. The CAL500 dataset provides binary annotations, which are 1 when a tag applies to the song and 0 otherwise, based on the opinions of human annotators. To accurately fit the experimental models, we restrict ourselves to the subset of 97 tags that have at least 30 songs positively associated with them (11 genre, 14 instrument, 25 acoustic quality, 6 vocal characteristics, 35 emotion and 6 usage tags).

4.2 Audio features

The acoustic content of each song in the collection is represented by computing a time series of 34-bin Mel-frequency spectral features [15], extracted over half-overlapping windows of 92 ms of audio signal. For the auto-tagger based on the DTM, Mel-frequency spectral features are grouped into fragments of approximately 6 s. (with 80% overlap), which corresponds to $\tau = 125$ consecutive feature vectors. For the auto-tagger based on the GMM, the Mel-frequency spectral features are decorrelated using the DCT, and the resulting first 13 Mel-frequency cepstral coefficients are augmented with first and second derivatives (MFCC-deltas). Lastly, for the auto-tagger based on boosting, first and second order statistics of the MFCC deltas are computed every 5 s., in order to reduce the computational burden [8] .

4.3 Evaluation

In our experiments, we consider the models reviewed in Section 2.1, which are the content-based auto-taggers referred as GMM, BST, and DTM, and the semantic context modeling based on the DMM. We obtained the authors' code to run each algorithm. We study model combination via decision-fusion using the DMM and investigate all the possible combinations among the content-based auto-taggers considered. For instance, when combining all the three auto-taggers (i.e., when $\mathcal{A} = \{\mathrm{GMM}, \mathrm{BST}, \mathrm{DTM}\}$) Equation 7 acts on the aggregated semantic multinomials defined as:

$$\boldsymbol{\pi}_d^{\mathcal{A}} \;=\; (\boldsymbol{\pi}_d^{\mathrm{GMM}},\; \boldsymbol{\pi}_d^{\mathrm{BST}},\; \boldsymbol{\pi}_d^{\mathrm{DTM}}) \,. \tag{8}$$

To investigate the advantages of model combination via decision-fusion, we compared its performances to a variety of combination techniques, such as fixed-combination rules [11] and trained-combiners based on regression [19], all of

which are applied on the outputs of the different content-based auto-taggers (i.e., GMM, BST, DTM). We tested different fixed-combination rules (i.e., sum, product, arithmetic average, minimum and maximum rule) in preliminary experiments, with the sum rule ($\sum$ rule) being the best. So, for example, when $\sum$ rule combines GMM, BST and DTM summing the corresponding SMNs, the final semantic multinomial of each song s is:

$$\boldsymbol{\pi}_s^{\text{SUM}} = \boldsymbol{\pi}_s^{\text{GMM}} + \boldsymbol{\pi}_s^{\text{BST}} + \boldsymbol{\pi}_s^{\text{DTM}}, \quad (9)$$

which is intended to be normalized to 1.

Additionally, we implemented the trained-combiner based on linear regression (LinReg), which Tomasik et. al [19] showed to outperform alternative regression techniques. In particular, we use LinReg to learn, on a tag-by-tag bases, the optimal coefficients to combining different auto-taggers to predict a ground truth of annotated songs. We refer the reader to Section 3.3 of [19] for more details on this strategy.

Annotation and retrieval performances are measured following [20]. Test set songs are annotated with the 10 most likely tags in their SMNs, and annotation accuracy is reported by computing precision, recall and F-score for each tag. Retrieval performance are evaluated with respect to each one-tag query in our vocabulary; we report mean average precision (MAP), area under the receiver operating characteristic curve (AROC) and top-10 precision (P10). All metrics are averaged over all tags and are intended to be result of 5 fold cross validation, where each song appeared in the test set exactly once.

	retrieval			**annotation**		
Model	MAP	AROC	P10	P	R	F-score
GMM	0.417	0.686	0.425	0.374	0.205	0.213
BST	0.432	0.701	0.453	0.334	0.144	0.170
DTM	**0.446**	**0.708**	**0.460**	**0.446**	**0.217**	**0.264**
***(a)** content-based auto-taggers*						
GMM	0.447	0.711	0.465	0.436	**0.238**	0.253
BST	0.457	0.711	0.476	0.424	0.201	0.241
DTM	**0.464**	**0.723**	**0.480**	**0.461**	0.236	**0.275**
***(b)** context-modeling with DMM*						
two BoF models $\mathcal{A}$ = (GMM, BST)						
$\sum$ rule	0.440	0.709	0.463	0.369	0.153	0.185
LinReg [19]	0.444	0.708	0.459	0.371	**0.239**	0.226
context fusion	**0.460**	**0.719**	**0.475**	**0.425**	0.224	**0.255**
a BoF and a time-series model $\mathcal{A}$ = (BST, DTM)						
$\sum$ rule	0.454	0.721	0.475	0.385	0.156	0.189
LinReg [19]	0.445	0.711	0.457	0.388	**0.237**	0.228
context fusion	**0.475**	**0.729**	**0.495**	**0.434**	0.221	**0.265**
a BoF and a time-series model $\mathcal{A}$ = (GMM, DTM)						
$\sum$ rule	0.461	0.726	0.474	0.445	0.229	0.267
LinReg [19]	0.456	0.722	0.460	0.360	**0.248**	0.222
context fusion	**0.470**	**0.730**	**0.487**	**0.484**	0.230	**0.291**
two BoF and a time-series model $\mathcal{A}$ = (GMM, BST, DTM)						
$\sum$ rule	0.457	0.725	0.478	0.39	0.163	0.202
LinReg [19]	0.452	0.715	0.465	0.384	**0.242**	0.232
context fusion	**0.475**	**0.731**	**0.496**	**0.456**	0.217	**0.270**
***(c)** auto-tagger combination*						

Table 1. Annotation and retrieval for the different models on the CAL500 dataset. The best results for each scenario are indicated in bold.

5. RESULTS

Annotation and retrieval results are presented in Table 1. Results for (a) individual auto-taggers are in the first block of the table, results for (b) standard contextual approaches are in the second block, and results for (c) content-based auto-tagger combination are in the last four blocks.

First, we notice that for each combination of the content-based auto-taggers considered, decision-fusion outperforms all the other combination techniques, except in recall, where LinReg is generally the best one. Second, differently from $\sum$ rule and LinReg, decision-fusion always improves with respect to the original content-based auto-taggers combined.

Decision-fusion performs better by capturing the correlations that arise between tag predictions based on different auto-taggers and, consequently, by indirectly leveraging various aspects of the musical signal emphasized by each of those auto-taggers. Indeed, decision-fusion of BoF auto-taggers with the DTM has major benefits, as it takes advantage of predictions that are based on different fundamentals, i.e., timbre and temporal dynamics vs. only timbre. On the other hand, decision-fusion of GMM and BST, which both model only the timbre, does not achieve comparable improvements over the corresponding standard context-models. In addition, the combination of all three auto-taggers with decision-fusion leads to the best retrieval performance; yet the modest improvements over the combination of BST and DTM in retrieval are compensated by improvements in precision and F-score over the same method.

Figure 1 depicts the MAP score achieved by a subset of tags, for the content-based auto-taggers (i.e., GMM, BST, DTM) and for decision-fusion using GMM, BST and DTM. Even if DTM could be preferred over both GMM and BST based on the *average* performances reported in Table 1, the *fluctuation* in performance on specific tags shown in Figure 1 suggests that each content-based auto-tagger may be better suited for a subset of the tags than the others. However, leveraging a rich contextual information that benefits from various acoustic characteristics of the musical signal, decision-fusion using GMM, BST and DTM performs best on the majority of all the tags reported.

Finally, part (b) of Table 1 also reports that standard context modeling always improves over the individual performance of the original content-based auto-taggers. While

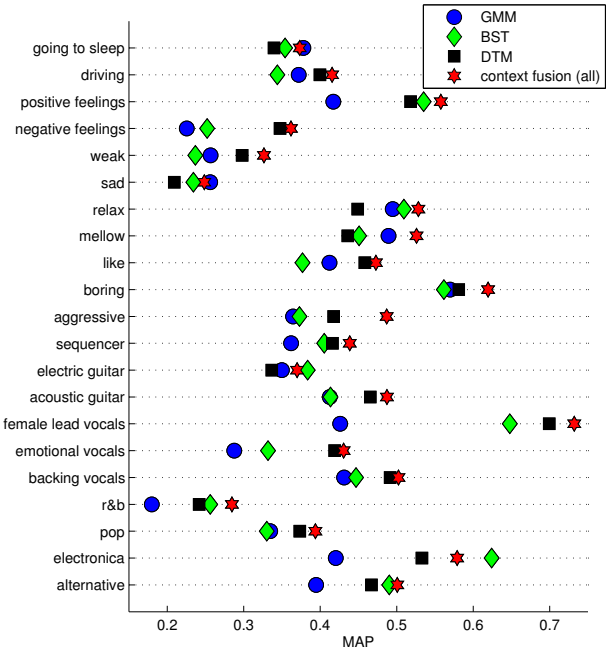

Figure 1. Retrieval performance (MAP) for a subset of the CAL500 vocabulary for GMM, BST, DTM, and decision-fusion of GMM, BST and DTM. Among the content-based auto-tagger, each one appears to be best on a subset of tags. However, decision-fusion is superior on the majority of tags.

Miotto et al. [13] already showed this for the BoF models (i.e., GMM and BST), we have demonstrated that it holds true for the DTM as well.

6. CONCLUSION

In this paper we have proposed *decision-fusion* as a strategy for combining different content-based auto-taggers. It uses semantic context modeling to simultaneously leverage the benefits of different content-based auto-taggers. Experimental results demonstrate especially that it achieves better annotation and retrieval performance than individual auto-taggers and various other techniques to combining multiple content-based auto-taggers.

7. ACKNOWLEDGEMENTS

The authors thank L. Barrington and T. Bertin-Mahieux for providing the code of [20] and [8] respectively, and acknowledge support from Qualcomm, Inc., Yahoo! Inc., the Hellman Fellowship Program, and NSF Grants CCF-0830535 and IIS-1054960. This research was supported in part by the UCSD FWGrid Project, NSF Research Infrastructure Grant Number EIA-0303622. R.M. thanks Nicola Orio for helpful discussion.

8. REFERENCES

[1] T. Bertin-Mahieux, D. Eck, F. Maillet, and P. Lamere. Autotagger: a model for predicting social tags from acoustic features on large music databases. *Journal of New Music Research*, 37(2):115–135, June 2008.

[2] G. Carneiro, A.B. Chan, P.J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):394–410, 2007.

[3] M. Casey, C. Rhodes, and M. Slaney. Analysis of minimum distances in high-dimensional musical spaces. *IEEE Transactions on Audio, Speech and Language Processing*, 16(5):1015–1028, 2008.

[4] A. B. Chan and N. Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5):909–926, 2008.

[5] A.B. Chan, E. Coviello, and G. Lanckriet. Clustering dynamic textures with the hierarchical EM algorithm. In *Proc. IEEE CVPR*, 2010.

[6] E. Coviello, A. Chan, and G. Lanckriet. Time Series Models for Semantic Music Annotation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(5):1343–1359, July 2011.

[7] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto. Dynamic textures. *Intl. J. Computer Vision*, 51(2):91–109, 2003.

[8] D. Eck, P. Lamere, T. Bertin-Mahieux, and S. Green. Automatic generation of social tags for music recommendation. In *Advances in Neural Information Processing Systems*, 2007.

[9] A. Flexer, F. Gouyon, S. Dixon, and G. Widmer. Probabilistic combination of features for music classification. In *Proc. ISMIR*, pages 111–114, 2006.

[10] M. Hoffman, D. Blei, and P. Cook. Easy as CBA: A simple probabilistic model for tagging music. In *Proc. ISMIR*, pages 369–374, 2009.

[11] J. Kittler. Combining classifiers: A theoretical framework. *Pattern Analysis and Applications*, 1(1):18–27, 1998.

[12] M.I. Mandel and D.P.W. Ellis. Multiple-instance learning for music information retrieval. In *Proc. ISMIR*, pages 577–582, 2008.

[13] R. Miotto, L. Barrington, and G. Lanckriet. Improving auto-tagging by modeling semantic co-occurrences. In *Proc. ISMIR*, pages 297–302, 2010.

[14] S.R. Ness, A. Theocharis, G. Tzanetakis, and L.G. Martins. Improving automatic music tag annotation using stacked generalization of probabilistic svm outputs. In *Proc. ACM MULTIMEDIA*, pages 705–708, 2009.

[15] L. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Upper Saddle River (NJ, USA), 1993.

[16] N. Rasiwasia and N. Vasconcelos. Holistic context modeling using semantic co-occurences. In *Proc. IEEE CVPR*, pages 1889–1895, 2009.

[17] J. Reed and C.H. Lee. A study on music genre classification based on universal acoustic models. In *Proc. ISMIR*, pages 89–94, 2006.

[18] M. Slaney, K. Weinberger, and W. White. Learning a metric for music similarity. In *Proc. ISMIR*, pages 313–318, 2008.

[19] B. Tomasik, J.H. Kim, M. Ladlow, M. Augat, D. Tingle, R. Wicentowski, and D. Turnbull. Using regression to combine data sources for semantic music discovery. In *Proc. ISMIR*, pages 405–410, 2009.

[20] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2):467–476, February 2008.

[21] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002.

[22] N. Vasconcelos and A. Lippman. Learning mixture hierarchies. In *Advances in Neural Information Processing Systems*, pages 606–612, 1998.

[23] B. Whitman and D. Ellis. Automatic record reviews. In *Proc. ISMIR*, pages 470–477, 2004.

[24] Y.H. Yang, Y.C. Lin, A. Lee, and H. Chen. Improving musical concept detection by ordinal regression and context fusion. In *Proc. ISMIR*, pages 147–152, 2009.