

Homogeneous Segmentation and Classifier Ensemble for Audio Tag Annotation and Retrieval

Hung-Yi Lo, Ju-Chiang Wang, and Hsin-Min Wang

July 20, 2010



Spoken Language Processing Group
Natural Language and Knowledge Processing Lab.
Institute of Information Science
Academia Sinica, Taiwan
<http://sovideo.iis.sinica.edu.tw/SLG>

Social Tagging to Music

last.fm

[Music](#)
[Radio](#)
[Events](#)
[Charts](#)
[Community](#)

New! Festival recommendations based on your taste »

[English](#) | [Help](#)
Mus

Artist

Biography

Pictures

Videos

Albums

Tracks


Events

News

Charts

Similar Artists

Tags



[The Beatles](#) » [Tracks](#) » [Let It Be](#)

Tags

60s
70s
acoustic
alternative
alternative rock
amazing
awesome
ballad
ballads

beatles
beautiful
brilliant
british
british invasion
britpop
calm
chill
chillout

classic rock
classics
cool
downtempo
easy listening

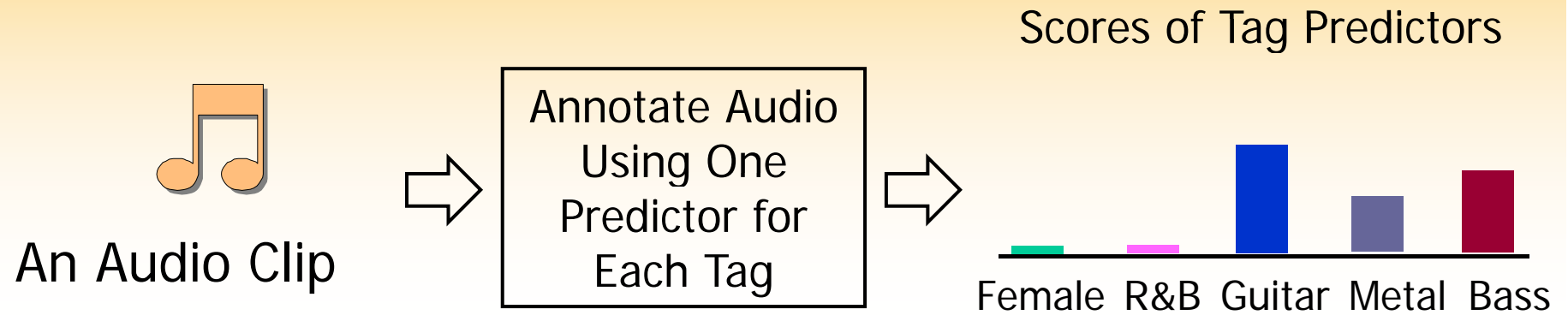
rock
rock ballad
rolling stones top 500 songs of all time
sad
singer-songwriter

the beatles
sweet
uk
uplifting
1970

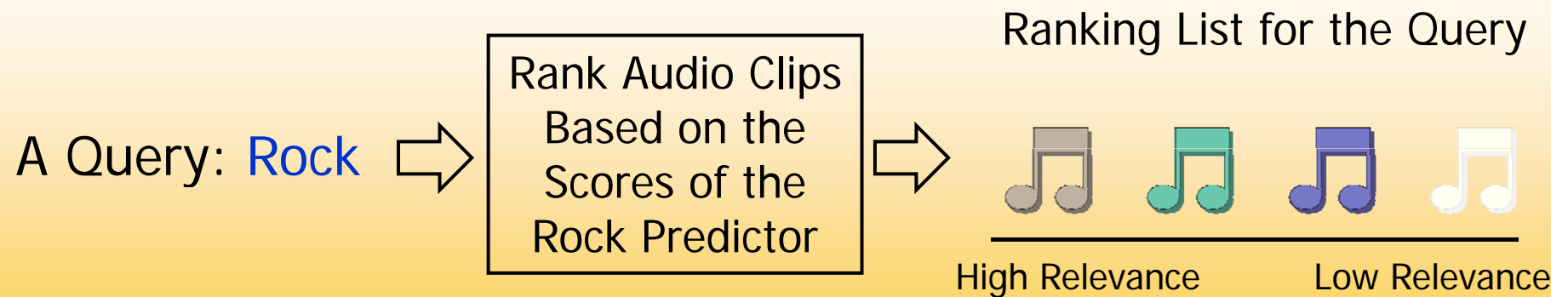
Tag

Audio Tag Annotation and Retrieval

Annotating audio clips with tags



Retrieving audio clips using a tag query

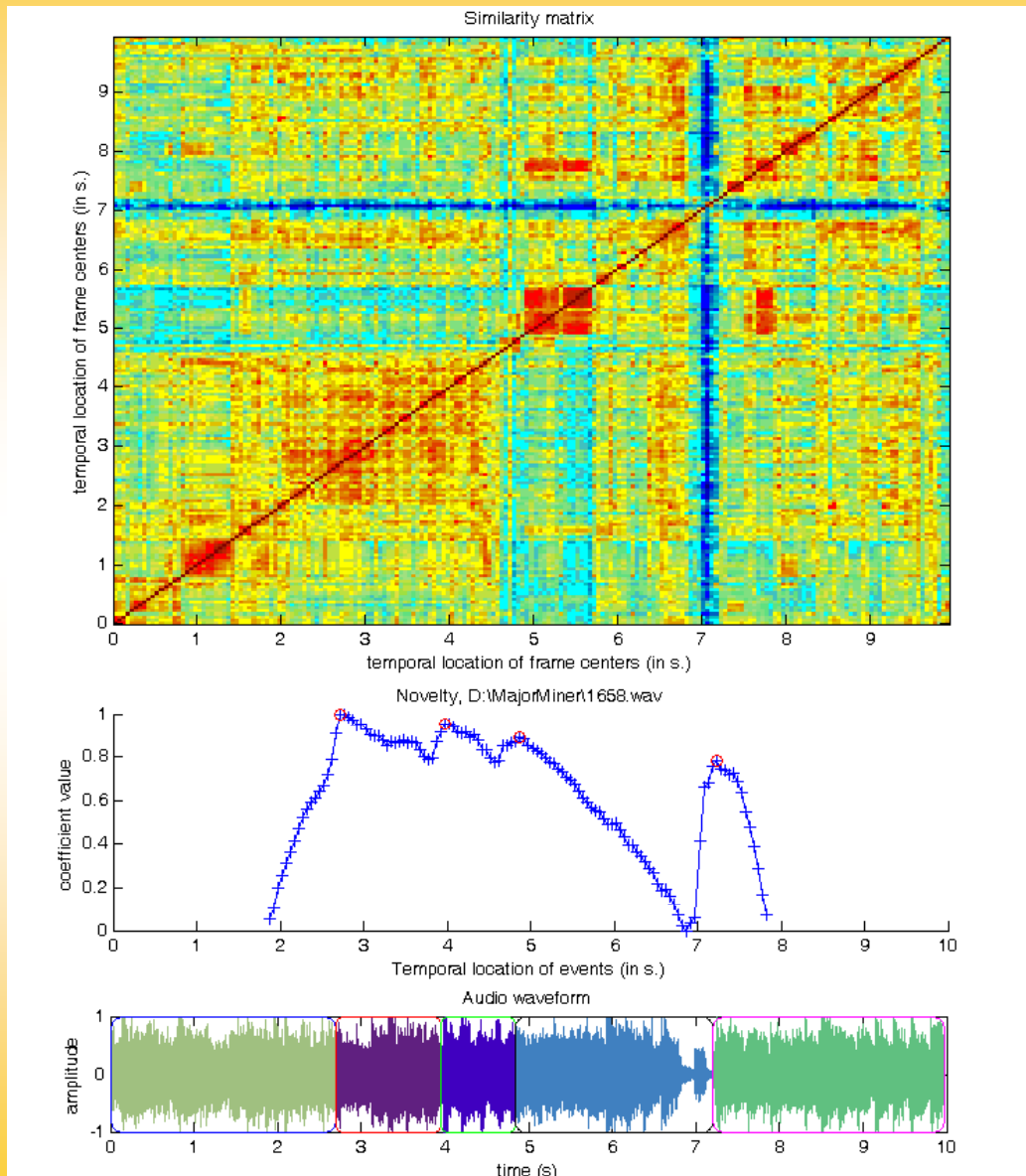


Our Contributions

1. Dividing the audio signal into homogeneous segments using an **audio novelty curve**
 2. Each tag predictor is an ensemble classifier combining two classifiers: SVM and AdaBoost
 - **Ranking Ensemble** for audio tag retrieval
 - **Probability Ensemble** for audio tag annotation
- Our **ranking ensemble** won the Audio Tagging Competition in 2009 Music Information Retrieval Evaluation eXchange (MIREX)
- **In terms of tag F-measure and the area under the ROC curve given a tag (for audio retrieval)**



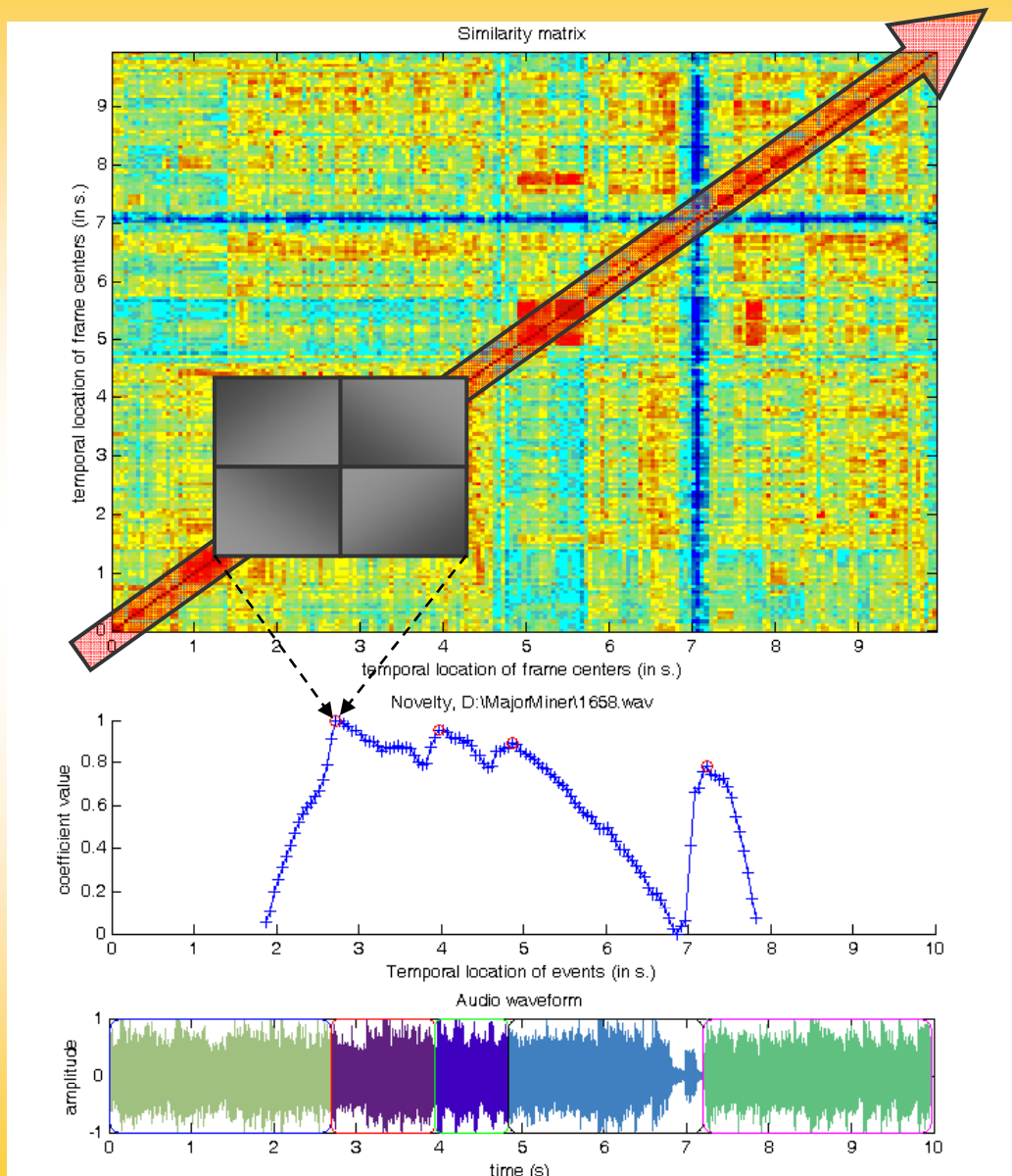
Audio Segmentation



- Feature of the Matrix:
13 Dim MFCC
- Kernel Type:
Gaussian
- Kernel Size:
128 frames

- The prediction score on the whole clip is the average of scores on each segment.

Audio Segmentation



- Feature of the Matrix:
13 Dim MFCC
- Kernel Type:
Gaussian
- Kernel Size:
128 frames

- The prediction score on the whole clip is the average of scores on each segment.

Audio Feature Extraction Using MIRToolbox

Classes	Features
Dynamics	<ul style="list-style-type: none"> ▪ Rms
Rhythm	<ul style="list-style-type: none"> ▪ Peak and centroids of the fluctuation summary ▪ Tempo ▪ Attack slop and attack time of the onset
Timbre	<ul style="list-style-type: none"> ▪ Zero-crossing rate ▪ Spectral centroid, spread, skewness and kurtosis ▪ Brightness ▪ Rolloff with 95% threshold ▪ Rolloff with 85% threshold ▪ Spectral entropy and flatness ▪ Roughness ▪ Irregularity ▪ Inharmonicity ▪ MFCCs, delta-MFCCs, and delta-delta-MFCCs ▪ Low energy rate ▪ Spectral flux
Pitch	<ul style="list-style-type: none"> ▪ Pitch ▪ Chromagram and its centroids and highest peak
Tonality	<ul style="list-style-type: none"> ▪ Key clarity ▪ Key mode ▪ Harmonic change



Classification Methods and The Difficulties

- The tag predictor is an ensemble that combines the outputs of two classifiers
 - SVM: Linear SVM implemented by the LIBLINEAR package
 - AdaBoost: decision stump as the base learner

Two methods to merge the two prediction scores

1. Ranking Ensemble for the retrieval task

- The scales of the two classifiers' prediction scores are rather different

2. Probability Ensemble for the annotation task

- The scores of different tag predictors are not comparable



Ranking Ensemble

AdaBoost SVM

1.9	7.1
-0.5	6.5
1.1	3.9
-2.3	-0.3
0.2	12

Prediction
Scores

AdaBoost SVM

1	2
4	3
2	4
5	5
3	1

Respective
Rankings

Merged
Prediction

1.5
3.5
3
5
2

Average
Ranking

Probability Ensemble

- In the audio annotation task, we need to compare the scores of all tag predictors
 - The raw scores of different tag classifiers are not comparable
- We transform the output scores of SVM and AdaBoost into **probability scores** with a **sigmoid function**:

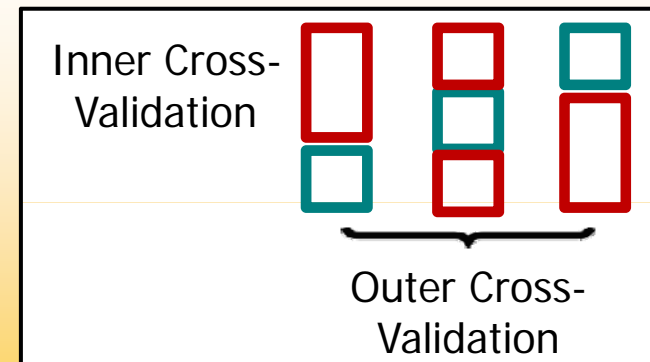
$$\Pr(y = 1 | \mathbf{x}) \approx \frac{1}{1 + \exp(Af + B)}$$

- f : the output score of a classifier
 - A, B : can be learned by solving a regularized maximum likelihood problem
- Then average the two probability score.



Model Selection

- MIREX evaluates submitted algorithms by **3-fold cross-validation**
- **Inner cross-validation** on the training set to determine the classifier parameters
 - The cost parameter C in the linear SVM
 - The number of base learners in AdaBoost
- **Re-train** the classifiers with the complete training set and the selected parameters
- Model selection criterion: **AUC-ROC**
 - Since the class distributions for some tags are imbalanced



MIREX 2009 Results on The MajorMiner Dataset

	Tag F-measure	Tag Accuracy	Tag AUC-ROC	Clip AUC-ROC
No Seg	0.289	0.900	0.782	0.751
Seg	0.311	0.903	0.807	0.774
BP1				
BP2				
CC1				
CC2	0.241	0.905	0.791	0.882
CC3	0.170	0.913	0.721	0.854
CC4	0.263	0.890	0.749	0.854
GP	0.012	0.891		
GT1	0.290	0.850	0.784	0.872
GT2	0.293	0.850	0.786	0.876
HBC	0.044	0.914	0.736	0.851

Better Than

Audio Annotation:
Given a tag, correct clips should have higher scores

MIREX 2009 Results on The Mood Dataset

	Tag F-measure	Tag Accuracy	Tag AUC-ROC	Clip AUC-ROC
No Seg	0.204	0.882	0.667	0.678
Seg	0.219	0.887	0.701	0.704
BP1	0.195	0.837	0.648	0.854
BP2	0.193	0.829	0.632	0.859
CC1	0.172	0.878	0.652	0.849
CC2	0.180	0.882	0.681	0.848
CC3	0.147	0.882	0.629	0.812
CC4	0.183	0.862	0.646	0.812
GP	0.084	0.863		
GT1	0.211	0.823	0.649	0.860
GT2	0.209	0.824	0.655	0.861
HBC	0.063	0.909	0.664	0.861



Extended Experiments

- We extensively evaluate the classifiers and the ensemble methods on the downloaded MajorMiner dataset
 - MajorMiner is a web-based music labeling game: <http://majorminer.org/>
- Our extended experiments basically follow the MIREX 2009 setup
 - Use the same 45 tags and download all the audio clips that are associated with these tags
 - The dataset might be slightly different from that used in MIREX 2009
 - The resulting audio database contains 2,472 clips
- Repeat cross-validation twenty times to reduce variance

metal	instrumental	horns	piano	guitar
ambient	saxophone	house	loud	bass
fast	keyboard	vocal	noise	british
solo	electronica	beat	80s	dance
jazz	drum machine	strings	pop	r&b
female	distortion	voice	rap	male
slow	electronic	quiet	techno	drum
funk	acoustic	rock	organ	soft
country	hip hop	synth	trumpet	punk



Results of The Audio Retrieval Task

Mean \pm Standard Deviation	Tag AUC-ROC		Tag F-measure	
	Without Seg.	With Seg.	Without Seg.	With Seg.
AdaBoost	0.7520 ± 0.0026	0.7943 ± 0.0024	0.2856 ± 0.0036	0.3034 ± 0.0051
Linear SVM	0.7848 ± 0.0029	0.7990 ± 0.0030	0.3092 ± 0.0028	0.3169 ± 0.0038
Probability Ensemble	0.7894 ± 0.0030	0.8108 ± 0.0020	0.3163 ± 0.0037	0.3296 ± 0.0039
Ranking Ensemble	0.7997 ± 0.0022	0.8189 ± 0.0017	0.3211 ± 0.0032	0.3332 ± 0.0038

Performance differences (Ranking Ensemble vs. others):
 - vs. AdaBoost: 4.23% (AUC-ROC)
 - vs. Linear SVM: 1.42% (AUC-ROC)
 - vs. Probability Ensemble: 2.14% (AUC-ROC)
 - vs. Linear SVM: 6.69% (F-measure)

Better Than (Ranking Ensemble vs. Linear SVM)

Results of The Audio Annotation Task

Mean \pm Standard Deviation	Clip AUC-ROC		Tag Accuracy	
	Without Seg.	With Seg.	Without Seg.	With Seg.
AdaBoost	0.8627 ± 0.0009	0.8774 ± 0.0009	0.9162 ± 0.0004	0.9184 ± 0.0004
Linear SVM	0.8788 ± 0.0009	0.8828 ± 0.0012	0.9191 ± 0.0004	0.9200 ± 0.0003
Probability Ensemble	0.8788 ± 0.0007	0.8848 ± 0.0007	0.9191 ± 0.0002	0.9201 ± 0.0003
Ranking Ensemble	10.34% 0.7626 ± 0.0012	0.7814 ± 0.0010	0.9016 ± 0.0004	0.9057 ± 0.0003



Conclusion

- This paper has presented our methods for audio tag annotation and retrieval
- Major contributions:
 - Use a novelty curve to divide audio clips into homogeneous segments
 - Exploit two classifier ensembles: ranking ensemble and probability ensemble
- The **ranking ensemble** performs very well in the MIREX 2009 audio tag classification task in terms of **audio retrieval metrics**
 - But not very good in terms of **audio annotation metrics**
- The **probability ensemble** method performs very well in terms of **audio annotation metrics**



Thank You

