A Music Video Information Retrieval Approach to Artist Identification

Alexander Schindler[1], Andreas Rauber[2]

IMAGES

Artist 1 · Artist 2 · Artist 3 · Artist 4

LBP Face Recognizer

Build Face Recognizer

The face recognizer is based on Local Binary Patterns - a simple but efficient gray-scale invariant texture descriptor combining properties of structural and statistical texture analysis. Each pixel of an image is labelled by thresholding its 3x3-neighborhood with the center value and considering the result as an 8 bit binary number. The texture of an image can be described by a histogram representation of the frequency of the 256 different labels.

- Train face recognizer with LBP
- Extract Local Binary Patterns
- Preprocess images (normalize, crop faces)
- Preselect images (only frontal face, no smiles, occlusions, etc.)
- Collect portrait images of artists

AUDIO

Artist 1 · Artist 2 · Artist 3 · Artist 4

MILLION SONG DATASET

Support Vector Machine

Train Audio Classifier

Audio classification is based on a Support Vector Machine classifier trained with Temporal Echonest Features - statistical moments calculated from a selected set of features provided by the Million Song Dataset, describing timbral, chromatic and temporal properties of the musical track.

- Train Support Vector machine
- Calculate Temporal Echonest Features from MSD metadata
- Exclude tracks contained in test set
- Select all tracks of the artists from the MSD for training

Predict

Detect

Predict

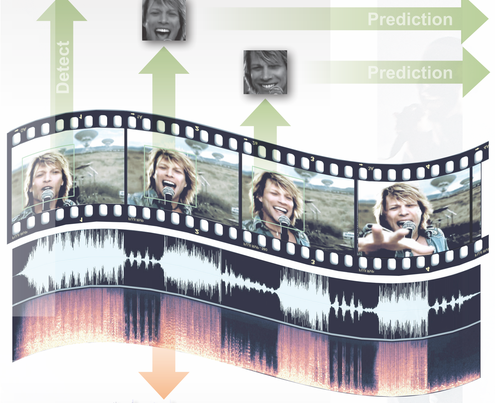
Prediction

Prediction

Prediction

Detect Faces in Video Frames

Faces are extracted using frame based face detection based on boosted cascades of Haar-like features. Local Binary Patterns are extracted and classified by the trained face recognizer.



Predict

Prediction

Extract Features from Audio

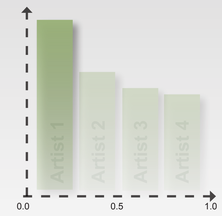
Echonest features are extracted from the separated audio stream of the music video, aggregated into Temporal Echonest Features and used to classify the underlying song with an artist label.

Preparing the Ensemble

Each of the two classifiers estimates the probability that an artist performed a certain video and finally makes a distinct decision. Further, a confidence value is calculated for each classifier, describing its ability to correctly identify the performing artist.

Summarize Face Predictions

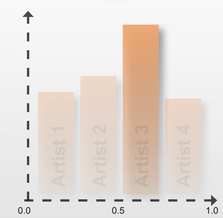
For each artist in the training set a probability is calculated based on the prediction count weighted by their confidence value. The final decision for each video is based on the highest probability.



Bootstrap Aggregation

Weighted Majority Voting

Results



Combining Audio and Video

The ensemble classifier is based on Bootstrap Aggregation which generates multiple versions of a predictor by making bootstrap replicates of the training set through random sub-sampling. Each classifier was trained on a selected sub-set and the remainder of the training set was used to estimate its confidence. The resulting predictions were aggregated through weighted majority voting using the estimated probability and the classifier's confidence as weights.

VIDEO

ENSEMBLE

Motivation

State-of-the-Art

Artist recognition is an important task for music indexing and content based retrieval. Audio features used in recent evaluations provide general descriptions of the musical track. Artists characteristics like the singing voice get lost in the aggregation process.

Demonstrate Opportunities

Artist Identification is a good example to demonstrate the opportunities of a Music Video Information Retrieval approach - a cross-modal approach to Music Information Retrieval problems

Exploit Additional Semantic Layers

Music videos like their underlying audio recordings are pieces of art and are used to accompany or augment the musical track. The visual part adds a second semantic layer to the song which may correlate with the other layers or contradict. In any case, a lot of information is provided in the visual part of music videos. By augmenting music video information retrieval technologies with solutions emerging from the video retrieval domain open research challenges could be addressed that are currently problematic to solve through audio content analysis (e.g., classifying Christmas songs).

Results

The baseline for the experiment was a typical audio content based classification approach using music features only. According to this the precision of the audio based classifier could be increased by 27% while recall values were only slightly improved by 5%. Thus, the ensemble approach did not increase the number of correctly identified tracks, but did enhance the reliability.

Video Results

Ensemble		Audio		Video		
Prec.	Recall	Prec.	Recall	Prec.	Recall	Artist
0.36	0.57	0.33	0.52	0.14	0.33	Aerosmith
0.64	0.45	0.50	0.45	0.62	0.25	Avril Lavigne
0.55	0.32	0.33	0.26	0.28	0.42	Beyonce
0.24	0.27	0.28	0.36	0.20	0.04	Bon Jovi
0.34	0.42	0.32	0.33	0.16	0.17	Britney Spears
0.33	0.50	0.48	0.71	0.18	0.43	Christina Aguilera
0.62	0.53	0.41	0.47	0.00	0.00	Foo Fighters
0.27	0.19	0.22	0.24	0.33	0.14	Jennifer Lopez
0.30	0.24	0.27	0.28	0.50	0.12	Madonna
0.35	0.70	0.20	0.10	0.12	0.80	Maroon 5
0.58	0.44	0.55	0.38	1.00	0.18	Nickelback
0.75	0.14	0.29	0.19	0.40	0.10	Rihanna
0.28	0.65	0.44	0.40	0.25	0.21	Shakira
1.00	0.16	0.60	0.32	0.50	0.06	Taylor Swift
0.47	**0.38**	**0.37**	**0.36**	**0.34**	**0.21**	**avg**

Audio Results

Problems

Most face recognition approaches are still limited by variations in different image or face properties like a) occlusions b) distortions c) video transitions d) varying illuminations e) make up and ornaments f) hair g) beards h) stylistic elements, as well as variations in pose and illumination, age of the person, etc. - properties that are extensively used as artistic and stylistic features of music videos.

Conclusions

Opportunities of Music Video Information Retrieval

By improving the baseline by 27% this evaluation is an excellent example of the opportunities provided by combining the acoustic with the visual domain for improved music classification and retrieval.

Valuable Information

Music videos provide a lot of information which is yet disregarded in the music information retrieval domain. To take advantage of the additional semantic information provided by the visual layer, approaches of the image and video retrieval domains have to evaluated for their applicability to the music research field.

TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

[1)]Department of Software Technology and Interactive Systems
Vienna University of Technology
rauber@ifs.tuwien.ac.at

[2)]Intelligent Vision Systems
AIT Austrian Institute of Technology
alexander.schindler@ait.ac.at

AUSTRIAN INSTITUTE OF TECHNOLOGY