A Music Video Information Retrieval Approach to Artist Identification

Alexander Schindler[1,2] and Andreas Rauber[1]

[1] Department of Software Technology and Interactive Systems
Vienna University of Technology
{schindler,rauber}@ifs.tuwien.ac.at
[2] Intelligent Vision Systems, AIT Austrian Institute of Technology, Vienna, Austria

Abstract. We propose a cross-modal approach based on separate audio and image data-sets to identify the artist of a given music video. The identification process is based on an ensemble of two separate classifiers. Audio content classification is based on audio features derived from the Million Song Dataset (MSD). Face recognition is based on Local Binary Patterns (LBP) using a training-set of artist portrait images. The two modalities are combined using bootstrap aggregation (Bagging). Different versions of classifiers for each modality are generated from subsamples of their according training-data-sets. Predictions upon the final artist labels are based on weighted majority voting. We show that the visual information provided by music videos improves the precision of music artist identification tasks.

1 Introduction

To demonstrate the opportunities of a Music Video Information Retrieval approach, we address the problem of music artist identification - the task of identifying the performing musician of a given track. Music Video Information Retrieval (MVIR) constitutes a cross-modal approach to Music Information Retrieval (MIR) problems. 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. Musical genres can be predicted without hearing the audio, artists are recognized by their faces and even the tempo of a song can potentially be estimated by the rhythmic movements of artists or background dancers. The fact that this can be accomplished within fractions of seconds by humans implies that enough information is present to classify the content (see Figure 1). The challenging task is once again to extract this information in an appropriate way. By augmenting MIR 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).

Fig. 1. Examples of music genres that are easy to identify in images - a) Dance, b) Rock, c) Heavy Metal, d) Rap

Artist recognition is an important task for music indexing, browsing and content based retrieval originating from the music information retrieval domain. Typically it is subdivided into the tasks *artist identification*, *singer recognition*, and *composer recognition*. Recent achievements in music classification and annotation including artist recognition are summarized in [9]. A typical content based approach to this problem is to extract audio features from the corresponding tracks, train a machine learning based classifier and predict the artist name for the given track. This approach is similar to music genre classification, but whereas respectable results are reported from genre prediction, artist identification is still failing to achieve comparable levels of performance. The problem is that audio features used in these evaluations are statistical descriptions of the audio signal correlating mainly to sound properties as brightness, timbre or frequency/amplitude modulations over a period of time. All these features describe sound characteristics that are rather related to genre properties. Although an artist is mainly dedicated to a specific genre, its distinct songs are not. Tracks of a record may vary in tempo, instrumentation and rhythm. Further, stylistic orientations of the artists may change over time. The most intrinsic problem is that audio features are low level description of the audio content. Thus, two artists with similar sounding repertoire of songs will get confused, because the discriminating unique qualities of the singers voice get lost during the data reduction phase. The solution provided in [10] attempts to extract vocal segments of a song to identify the singer.

On the other side Video Information Retrieval (VIR) pursues the same goals in the video domain as MIR does in the music domain. Big effort is put into categorizing videos into different genres. A good summary of video classification is provided by [6]. Typically these approaches draw from more than one modality - the most common among them are text-based, audio-based and visual-based. Different properties of videos in conjunction with cinematic principles (e.g., light, motion, transitions from one scene to the other) are explored to estimate the genre of the video. Fast motion and short shot sequences are a good indicator for music videos. Although it is easy to distinguish music videos from other video genres, no publication is yet known to the authors, that explicitly tries to categorize the musical genres of music videos. Different acoustic styles are nevertheless used to estimate the video genre, as certain types of music are chosen to create specific emotions and tensions in the viewer [17]. VIR makes more use of time domain features, because some genres (e.g., news, sport) can already be

discriminated by their perceived loudness through computing the Root Mean Square (RMS) from the signal or by identifying voiced from unvoiced sequences through calculating the Zero Crossing Rate (ZCR) or a thresholded silence ratio.

Face recognition - identifying or verifying a person from still or video images - has received increased attention from academic and industrial communities over the past three decades due to potential applications in security systems, law enforcement and surveillance, and many others. A comprehensive summary of face recognition approaches is given by [29]. Despite the achievements and promising results reported of systems in relatively controlled environments, most face recognition approaches are still limited by variations in different image or face properties (e.g., pose, illumination, mimic, occlusions, age of the person, etc.) - properties that are extensively used as artistic and stylistic features of music videos. The predominating approaches to face recognition are Principal Component Analysis (PCA) (e.g., Eigenfaces [26]) and Linear Discriminant Analysis (LDA) (e.g., Fisherfaces [2]). A good summary of video-based face recognition is provided by [28].

The remainder of this paper is organized as follows. In the next section we give a brief overview of the state-of-the-art in the different domains and modalities. In Section 3 the layout of the classification approach as well its evaluation is described. Section 4 depicts the different datasets used in the evaluation. In Section 5 and 6 the separate classification approaches based on the two modalities audio and video are explained. In Section 7 the ensemble classification method that combines the previous two classifiers is outlined and the final results of the evaluation are provided which are further discussed in Section 8. Conclusions with suggestions for future work are provided in Section 9.

2 Related Work

Early approaches to artist identification are based on the Mel-Frequency Cepstral Coefficients (MFCCs) feature set in combination with Support Vector Machines (SVM) for classification [10, 15]. In [11] a quantitative analysis of the album effect - effects of post-production filters to create a consistent sound quality across a record - on artist identification was provided. A Hybrid Singer Identifier (HSI) is proposed by [22]. Multiple low-level features are extracted from vocal and non-vocal segments of an audio track and mixture models are used to statistically learn artist characteristics for classification. Further approaches report more robust singer identification through identifying and extracting the singers voice after the track has been segmented into instrumental and vocal sections [16, 25]. Good summaries of state-of-the-art approaches and challenges in face recognition are provided by [13, 19, 29]. Face detection, tracking and recognition is also used in multi-modal video retrieval [12, 23]. Faces are either used to count persons or to identify actors. Most common methods used to recognize faces are Eigenfaces [26] and Fisherfaces [2]. In [7] face tracking and text trajectories are used with Hidden Markov Models (HMM) for video classification. A face recognition approach based on real-world video data is reported in [24].

3 Classification Architecture

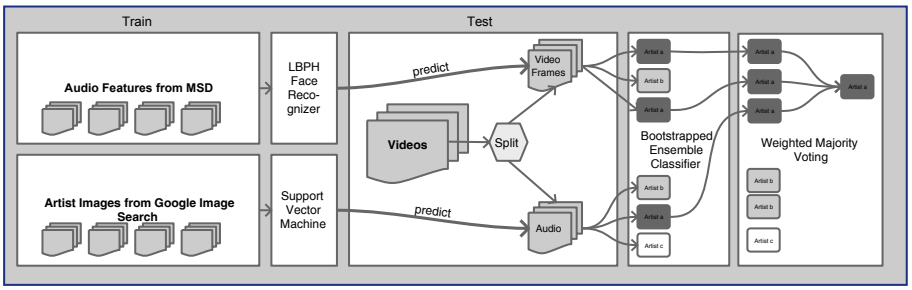


Fig. 2. Classification Architecture

Due to the expenses of producing music videos the number of productions per artist or album is marginally low compared to the number of songs recorded. Videos are typically produced for single releases of records to promote the track. As a consequence only a few videos can be collected and especially for relative young artists not enough entries to reliably train a classifier might be found. A common evaluation method in classification experiments is to use k-fold cross-validation with k usually set to 10. This requires at least 10 videos per artist, which for many artists are not available.

We present a three-folded classification approach based on two separate training data-sets to take advantage of multiple sources to predict the performing artist of a video. Figure 2 depicts the architecture of the classification system. The two modalities of the systems are trained independently on their data-sets and combined by an ensemble classifier to make a final prediction.

The audio classifier is trained on all available songs of an artist, that have not been released as video. This takes advantage of the broad spectrum of the artists work and provides a richer set of information. The face recognition system is trained on artist images downloaded from Google Image Search. Like the separate audio data the image data-set constitutes the ground-truth data for our classification system. Both classifiers are cross-validated on their data-sets to assess their confidence.

The trained audio and visual classifiers are applied to the music video test data-set. In a pre-processing step the videos are split into their source components and processed separately. An ensemble classification approach based on bootstrapped aggregation is used. Instead of using the complete training-set, the classifiers for each modality are trained only on sub-samples. This classification step is repeated n times resulting in $2n$ predictions for each music video. These predictions are aggregated through a weighted majority vote, using the previously evaluated confidence values of the classifiers as weights.

4 Dataset

The evaluation data-set used for the experiments is based on 14 popular western music artists listed in Table 1. Popular artists were chosen to meet the requirement of collecting enough music videos for each musician. To demonstrate typical problems of content based artist identification the selected artists belong predominately to the two non-overlapping genres Pop and Rock.

4.1 Training Data

As described in the classification architecture in Section 3 training and test data do not originate from the same data-set. The training data is drawn from two different sources - an audio and an image data-set.

Artist Tracks For the audio modality, the artist tracks provided by the Million Song Dataset (MSD) [3] have been used. For each artist all tracks available in the MSD have been selected exuding those that are present in the music video test-set. Table 1 lists the number of tracks for each artist. The audio training set has a total size of 645 tracks.

Artist Images For each artist, portrait images have been downloaded. If the performing artist was a band, only images of the lead singer were used. Bulk download from Google Image Search was used to retrieve a huge number of images for each artist. In a second step the face detection algorithm described in Section 6.1 was applied to each image to filter out photos that do not contain detectable faces or where the resolution of the detected face was below 120x120 pixels. The resulting subset was manually analyzed to remove duplicates and images where the portrait person does not look frontal into the camera. It was also verified that the remaining images are not, in fact, screen-shots from the music videos used for the evaluation. Further images with low resolutions, occlusions, exaggerated smiles or arbitrary illuminations were removed. Such deviations from pass-photo like portrait images will degrade the performance of the recognition system by introducing too much variance. Further problems concerning face recognition in music video will be addressed in Section 6.2. The resulting set of training images contains approximately 50-150 portraits per artist (see Table 1).

4.2 Test Data

Test-data consists of music videos that have been downloaded from Youtube[3] The following requirements concerning the quality of the video and its content were used to select the test-set:

[3] `http://www.yoututbe.com`

Table 1. Artists and training data

Artist Name	MSD Tracks	Images	Music Videos
Aerosmith	83	104	23
Avril Lavigne	29	105	20
Beyonc	32	117	19
Bon Jovi	59	54	26
Britney Spears	57	160	24
Christina Aguilera	46	123	14
Foo Fighters	64	55	19
Jennifer Lopez	45	92	21
Madonna	62	47	25
Maroon 5	20	78	10
Nickelback	57	47	16
Rihanna	24	122	21
Shakira	48	123	20
Taylor Swift	19	117	19
	645	1344	277

- has to be an official music video produced by the artist
- the lead singer has to appear in the video
- a minimum resolution of 360x240 pixels
- a minimum audio bitrate of 90 kBit/s

Audio Data was retrieved directly from the video files extracting the audio stream using FFMPEG[4]. The audio was converted to mp3 format with a sample-rate of 44100 Hz and a bitrate of 128 kBit/s. The Echonest API[5] was used to extract the audio features from the files which were stored equivalent to the MSD format.

Visual Data from the videos was retrieved frame by frame using the Open Computer Vision Library (OpenCV)[6] [4] that was also used for the further video processing.

5 Audio Content Analysis

The audio content analysis task is based on audio features provided by the Million Song Dataset (MSD) [3]. The MSD provides a rich set of low level features (e.g., timbre, chroma) and mid level features (e.g., beats per minute, music key, audio segmentation). For each artist of the evaluation test-set all tracks available in the MSD that do not overlap with the test-set are used. The number of tracks used for each artist is summarized in Table 1.

[4] `http://www.ffmpeg.org/`
[5] `http://developer.echonest.com/`
[6] `http://opencv.org`

5.1 Audio Features

Content based artist classification is based on the audio features provided by the Million Song Dataset (MSD). The features provided by the MSD are extracted by the Echonest[7]. According to the documentation of the Echonest Analyzer[8] their proprietary algorithms uses onset detection to localize beats and extract a feature vector for each beat. This approach returns a list of feature vectors of varying length. To make this information applicable for standard machine learning algorithms, it has to be aggregated into a fixed length single vector representation.

In this evaluation we use Temporal Echonest Features (TEN) as proposed by [21]. These audio descriptors summarize an empirically selected set of features provided by the MSD by calculating all statistical moments of the MSD features *Segments Pitches*, *Segments Timbre*, *Segments Loudness Max*, *Segments Loudness Max Time* and lengths of segments calculated from *Segments Start*. The resulting feature vector has 224 dimensions. In [21] we have shown that these features outperform state-of-the-art music feature sets in genre classification tasks on conventional data-sets.

5.2 Audio Classification Results

Audio classification was conducted using the Python machine learning library Scikit Learn[9]. Training and test data was separately normalized to have zero mean and unit variance. A Support Vector Machine (SVM) with a linear kernel and a penalty parameter of $C = 0.1$ was trained on the data from the MSD and used to classify the audio test-data set of the music videos. The results showed, that the audio data from the videos can be predicted with a precision of 37% and a recall of 36%. Such a value was to be expected do to the high variance in musical style of some of the artists. This can be seen in the high variance of the distinct values for all artists in Table 2 and is also illustrated by the corresponding confusion-matrix of the classification result in Figure 5.

6 Visual Content Analysis

The visual content analysis part of this evaluation is focused on face recognition. Generally face recognition systems can be classified into the two groups of recognizing faces from still images or from video. In this approach we use frame-by-frame analysis of still images - thus, ignoring spatio-temporal relationships. First faces from the training-set, i.e. the images collected from the Web, were detected and extracted to train a face recognizer. In a second step faces in video frames were detected and the recognizer was used to compute predictions concerning the artist.

[7] `http://echonest.com/`
[8] `http://developer.echonest.com/docs/v4/_static/AnalyzeDocumentation.pdf`
[9] `http://scikit-learn.org`

6.1 Face Detection

Detecting faces in the video data is the first step of the recognition task. Frame based face detection using boosted cascades of Haar-like features as proposed by Viola and Jones [27] and Lienhart [14] was used. Their method uses a set of simple features based on pixel value differences between neighboring adjacent rectangles. These features are rapidly calculated from an intermediate representation - the *integral image* - that already pre-computes neighborhood statistics for each pixel of the original image. The classifier for the face detection task is constructed by selecting a small subset of important features using AdaBoost [8]. Finally more complex classifiers are combined in a cascade structure. This approach for object detection minimizes computation time while achieving high prediction accuracy.

In order to eliminate false positives, detected faces were further analyzed in a post-processing step. Additional cascaded detectors were used to locate eye-pairs, noses and mouths within the region of the detected face. If all sub-components were recognized, the face was verified and added to the test-set. Face detection and extraction are the main pre-processing steps for face recognition and are limited by the same problems that are summarized in the following section and listed in Figure 3.

6.2 Obstacles in Face Detection / Recognition

Although remarkable progress in face recognition in the last decades [13, 19, 28, 29] most of the reported work has been evaluated in laboratory environments. The most influencing factors for the accuracy of face recognition systems are illumination, occlusions and distortions - properties that are common in music videos. See Figure 3.

- **Occlusions** of the face are one of the biggest problems in face recognition and unfortunately very common in music videos (e.g., microphone, hands touching the face, sunglasses, caps, hats, etc.) (see Figure 3a). Makeup and jewelry (see Figure 3e) pose a similar problem.
- **Distortions** of the face due to singing, screaming, expressive mimic or fast movements (see Figure 3b).
- **Pose** deviations. Face recognition systems work optimal when subjects look frontal into the camera, but in video or photography frontal face shots are not considered to flatter the photographed person. Further, poses are used for acting purposes to express emotions such as grief, sorrow or thinking.
- **Illumination** changes are a stylistic tool in many music videos. Typically stage lighting is used to create the impression of live performance. This results in fast illumination changes even within a short sequence of video frames (see Figure 3d).
- **Facial Hair** in the form of locks of hair hanging into the face is a similar problem to occlusions (see Figure 3f). Another, more severe problem are beards of male artists. Those may change over time or disappear completely.

Fig. 3. Examples of problematic faces - a) occlusions b) distortions c) video transitions d) varying illuminations e) make up and ornaments f) hair g) beards h) stylistic elements

Because they are not excluded during the face extraction process, beards influence the training-set or prediction. Figure 3g shows the same artist with different beard styles and no beard.

Special video related problems:

– **Blending Effects** between scenes or cuts. Smooth transitions with image-cross fading effects can overlay the content of consecutive frames onto the face (see Figure 3c). In such cases the face detector recognizes valid properties

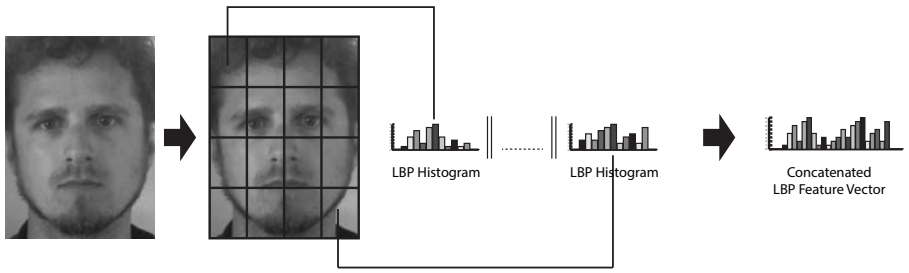


Fig. 4. Face recognition with Local Binary Patterns

of a face, but the overlaid content distorts the face similar to make-up or illumination changes.
- **Overlays** of visual blocks (e.g., text, images, logos) have similar effects as occlusions.

Further problems arise through aging of the artist. Music videos are typically produced in combination with new records which are released in a time-span of one to three years on average [18]. Artists that have begun to produce videos in the early stages of the music video trend and are still actively doing so have aged more than thirty years now. The effects of aging are reflected in the face and even when surgery is used to overcome them, the effects on the face recognizer are the same - the person might get misclassified.

6.3 Face Recognition

Face recognition used in this evaluation is based on Local Binary Patterns as proposed by Ahonen et al [1] due to their robustness against different facial expressions, illumination changes and aging of the subjects. LBP is a simple but very efficient gray-scale invariant texture descriptor that combines properties of structural and statistical texture analysis. It labels each pixel of an image by thresholding the 3x3-neighborhood of each pixel 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. For efficient face recognition the image is divided into regions to retain also spatial information. As depicted in Figure 4 the resulting histograms of the different image regions are normalized and concatenated to form the final face descriptor. Recognition based on these descriptors is performed using a nearest neighbor classifier in the corresponding feature space with Chi square as a dissimilarity measure.

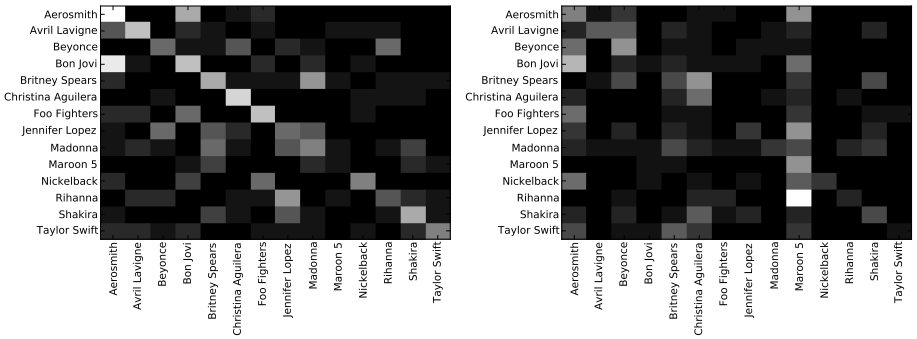
6.4 Video Classification Results

The face recognition based visual classifier was implemented using the Python programming language bindings of the OpenCV library [4]. This library provides implementations for the cascaded classifier based on Haar-like features which is used for face detection (see Section 6.1). In a preceding step the images were converted to gray-scale and their color histograms were normalized to retrieve better results from the face detector. The detected and verified faces were extracted. Contrast Limited Adaptive histogram equalization (CLAHE) [20] was applied to the face images to further enhance contrasts and normalize the images for the identification step. The LBP face recognition approach described in Section 6.3 which is also provided by OpenCV was used to predict the corresponding artist name. The recognizer was initiated with the radius of 1 and 8 neighbors used for building the Circular Local Binary Pattern. A grid of 8x8 cells was applied to the image resulting in a LBP descriptor consisting of 64 concatenated histograms. Each extracted and post-processed face gets an artist name label assigned by the

Table 2. Classification results of the separate modalities

Artist Name	**Audio**			**Video**		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Aerosmith	0.33	0.52	0.39	0.14	0.33	0.20
Avril Lavigne	0.50	0.45	0.47	0.62	0.25	0.36
Beyonce	0.33	0.26	0.29	0.28	0.42	0.33
Bon Jovi	0.28	0.36	0.32	0.20	0.04	0.07
Britney Spears	0.32	0.33	0.33	0.16	0.17	0.16
Christina Aguilera	0.48	0.71	0.57	0.18	0.43	0.26
Foo Fighters	0.41	0.47	0.44	0.00	0.00	0.00
Jennifer Lopez	0.22	0.24	0.22	0.33	0.14	0.20
Madonna	0.27	0.28	0.24	0.50	0.12	0.19
Maroon 5	0.20	0.10	0.13	0.12	0.80	0.20
Nickelback	0.55	0.38	0.44	1.00	0.18	0.30
Rihanna	0.29	0.19	0.23	0.40	0.10	0.15
Shakira	0.44	0.40	0.41	0.25	0.21	0.23
Taylor Swift	0.60	0.32	0.41	0.50	0.06	0.10
avg	0.37	0.36	0.35	0.34	0.21	0.20

face recognizer. For each label the average prediction confidence is calculated. To punish supposed isolated mis-classifications and to favor frequent assignments the average confidence is divided by the natural logarithm of the number of how often this label has been assigned. The results showed, that the visual data from the videos can be predicted with a precision of 34% and a recall of 21% where *Precision* describes the confidence a video classified as artist a to be truly from a whereas *Recall* describes how reliably all videos of artist a are recognized to be from a. The distinct values of the artists are listed in Table 2. The corresponding confusion-matrix of the classification result is depicted in Figure 5.



(a) Audio Classifier (b) Video Classifier

Fig. 5. Confusion matrices of the classification results.

7 Cross-Modal Results

The previous chapters 5 and 6 explicated the functionality, implementation and performance of the classifiers for the separate audio and video modalities. In this chapter a combined classification approach is explained that is used to combine the distinct modalities and provide enhanced classification accuracy.

7.1 Ensemble Classification Method

The ensemble classifier is based on the Bootstrap Aggregation (Bagging) as introduced by Breiman [5]. Bagging generates multiple versions of a predictor by making bootstrap replicates of the learning set through random sub-sampling. In our approach subset selection on $Train_{Audio}$ and $Train_{Video}$ was applied to generate $i = 10$ classifiers for each modality. Each classifier C_{Audio_i} and C_{Video_i} was trained on a selected sub-set $Train_{Audio_i}$ and the remainder of the training set $Train_{Audio} - Train_{Audio_i}$ was used to estimate its confidence $Conf_{Audio_i}$.

The resulting 20 predictions were aggregating through weighted majority voting. Each classifier C_{Audio_i} and C_{Video_i} predicts a music video mv of the test-set. Each prediction is now assigned a weight that is defined through the confidence of the used classifier $Conf_{Audio_i}$ or $Conf_{Video_i}$.

$$weight_{mv_i} = Conf_{Audio_i}$$

For each music video mv the sum of the weights of all labels is calculated. The label with the highest sum wins the vote and is the result of the ensemble classifier for the music video mv.

7.2 Results

The bootstrap aggregation ensemble classification approach as described in the previous section has been implemented using the Python Scientific Machine Learning Kit (SciKit Learn). For each modality bootstrapped sub-sampling with 10 iterations and 10% test-set size was applied to the according training-set. The results are summarized in Table 3 show an improve in precision using the multi-modal ensemble classification approach.

8 Discussion

The presented approach should demonstrate how to improve common approaches to artist identification through information extracted from music videos. The baseline for this experiment was a typical audio content based approach using the audio feature set presented in [21]. 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.

Table 3. Results of the Ensemble Classification

Artist Name	Precision	Recall	f1-score
Aerosmith	0.36	0.57	0.44
Avril Lavigne	0.64	0.45	0.53
Beyonce	0.55	0.32	0.40
Bon Jovi	0.24	0.27	0.25
Britney Spears	0.34	0.42	0.38
Christina Aguilera	0.33	0.50	0.40
Foo Fighters	0.62	0.53	0.57
Jennifer Lopez	0.27	0.19	0.22
Madonna	0.30	0.24	0.27
Maroon 5	0.35	0.70	0.47
Nickelback	0.58	0.44	0.50
Rihanna	0.75	0.14	0.24
Shakira	0.28	0.65	0.39
Taylor Swift	1.00	0.16	0.27
avg	0.47	0.38	0.37

As described in Section 3 this evaluation was intentionally based on two simple approaches. The audio classifier uses song-level features describing temporal statistics of timbral and chromatic music properties. Using audio segmentation to separate voiced from un-voiced sections [16, 25] may enhance the performance of the audio classifier. The visual classification approach was based on frame-by-frame face recognition and prediction was made by a majority vote. This approach might be improved through considering spatio-temporal relationships. By applying a shot-detection music videos can be segmented and faces tracked within one shot could be verified and summarized more reliably. A further limiting factor of this evaluation was the low resolution of the music videos which has been chosen as a compromise to collect enough videos. Face recognition systems highly depend on the information provided in the images. The minimum resolution of 120x120 pixel is sub-optimal and a verification-test-set using high-definition videos might provide better results.

The presented results showed that the performance of audio based artist identification can be improved through information extracted from music videos.

9 Conclusion and Future Work

We presented a cross-modal approach to music artist identification. Audio content and visual based classifiers were combined using an ensemble classifier. The audio classifier used Temporal Echonest Features [21] to predict artist labels. Its precision of 37% and recall of 36% was used as benchmark for the further experiments. The visual content classifier used face recognition based on a Local Binary Patterns (LBP) predictor. The two modalities are combined through bootstrap aggregation. For each modality 10 classifiers are created and trained on sub-samples of their according training-sets. The final prediction for a music

video is calculated on the basis of weighted majority voting of the resulting 20 predictions. The proposed cross-modal approach showed that the initial audio content based baseline could be increased by 27% through information extracted from the visual part of music videos.

The presented approach relies on a predefined dataset of artist images - thus, still requiring manual interaction. Future work will include automatic identification of lead singers to train the face recognition algorithm directly on faces extracted from the music videos. Such an approach would provide the possibility to use k-fold cross-validation on a single music video dataset.

References

1. Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. Face recognition with local binary patterns. In *Computer Vision-ECCV*, pages 469–481. Springer, 2004.
2. Peter N. Belhumeur, Joao P. Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.
3. Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, pages 591–596, Miami, USA, 2011.
4. G. Bradski. The opencv library. *Dr. Dobb's Journal of Software Tools*, 2000.
5. Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.
6. Darin Brezeale and Diane J Cook. Automatic video classification: A survey of the literature. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 38(3):416–430, 2008.
7. Nevenka Dimitrova, Lalitha Agnihotri, and Gang Wei. Video classification based on hmm using text and faces. In *European Conf. on Sig. Proc.* Citeseer, 2000.
8. Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.
9. Zhouyu Fu, Guojun Lu, Kai Ming Ting, and Dengsheng Zhang. A survey of audio-based music classification and annotation. *Multimedia, IEEE Transactions on*, 13(2):303–319, 2011.
10. Youngmoo E Kim and Brian Whitman. Singer identification in popular music recordings using voice coding features. In *3rd International Conference on Music Information Retrieval (ISMIR 2002)*, volume 13, page 17, 2002.
11. Youngmoo E Kim, Donald S Williamson, and Sridhar Pilli. Towards quantifying the album effect in artist identification. In *7th International Conference on Music Information Retrieval (ISMIR 2006)*, volume 18, page 145, 2006.
12. Michael S Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 2(1):1–19, 2006.
13. Stan Z Li. *Handbook of face recognition*. Springerverlag London Limited, 2011.
14. Rainer Lienhart and Jochen Maydt. An extended set of haar-like features for rapid object detection. In *Int. Conf. Image Processing*, pages I–900. IEEE, 2002.
15. Michael I Mandel and Daniel PW Ellis. Song-level features and support vector machines for music classification. In *6th International Conference on Music Information Retrieval (ISMIR 2005)*, pages 594–599, London, UK, 2005.

16. Annamaria Mesaros, Tuomas Virtanen, and Anssi Klapuri. Singer identification in polyphonic music using vocal separation and pattern recognition methods. In *8th Int. Conf. on Music Information Retrieval (ISMIR 2007)*, pages 375–378, 2007.
17. Simon Moncrieff, Svetha Venkatesh, and Chitra Dorai. Horror film genre typing and scene labeling via audio analysis. In *International Conference on Multimedia and Expo (ICME'03)*, volume 2, pages II–193. IEEE, 2003.
18. Julie Holland Mortimer, Chris Nosko, and Alan Sorensen. Supply responses to digital distribution: Recorded music and live performances. *Information Economics and Policy*, 24(1):3–14, 2012.
19. P Jonathon Phillips, Patrick J Flynn, Todd Scruggs, Kevin W Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the face recognition grand challenge. In *Computer vision and pattern recognition (CVPR 2005)*, pages 947–954. IEEE, 2005.
20. Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368, 1987.
21. Alexander Schindler and Andreas Rauber. Capturing the temporal domain in echonest features for improved classification effectiveness. In *Adaptive Multimedia Retrieval*, Copenhagen, Denmark, October 24-25 2012.
22. Jialie Shen, John Shepherd, Bin Cui, and Kian-Lee Tan. A novel framework for efficient automated singer identification in large music databases. *ACM Transactions on Information Systems (TOIS)*, 27(3):18, 2009.
23. Cees GM Snoek and Marcel Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia tools and applications*, 25(1):5–35, 2005.
24. Johannes Stallkamp, Hazım K Ekenel, and Rainer Stiefelhagen. Video-based face recognition on real-world data. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.
25. Wei-Ho Tsai and Hsin-Min Wang. Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):330–341, 2006.
26. Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.
27. Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, volume 1, pages I–511. IEEE, 2001.
28. Huafeng Wang, Yunhong Wang, and Yuan Cao. Video-based face recognition: A survey. *World Academy of Science, Engineering and Technology*, 60:293–302, 2009.
29. Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *Acm Computing Surveys*, 35(4):399–458, 2003.