# Group Feature Selection for Audio-Based Video Genre Classification

Gerhard Sageder[1,2](✉), Maia Zaharieva[1,2](✉), and Christian Breiteneder[2]

[1] Multimedia Information Systems Group, University of Vienna, Vienna, Austria
gerhard.sageder@unvie.ac.at, maia.zaharieva@tuwien.ac.at
[2] Interactive Media Systems Group, Vienna University of Technology, Vienna, Austria

**Abstract.** The performance of video genre classification approaches strongly depends on the selected feature set. Feature selection requires for expert knowledge and is commonly driven by the underlying data, investigated video genres, and previous experience in related application scenarios. An alteration of the genres of interest results in reconsideration of the employed features by an expert. In this work, we introduce an unsupervised method for the selection of features that efficiently represent the underlying data. Performed experiments in the context of audio-based video genre classification demonstrate the outstanding performance of the proposed approach and its robustness across different video datasets and genres.

**Keywords:** Genre classification · Group feature selection · Audio features

## 1 Introduction

Video genres commonly represent a first coarse categorization of large media collections. Although partly subjective, such a categorization enables end users to efficiently access and retrieve media of potential interest. As a result, automated video genre classification is subject to active research in the context of e.g. sport events [24], TV programs [5,10,13], web videos [5,21,23]. The selection of an appropriate feature set is crucial for any approach for video genre classification. In general, feature selection strongly depends on both the underlying data and on the application scenario (genres). As a result, an alteration of the genres of interest leads to a reconsideration of the employed features. This process requires for the intervention of an expert in order to assess potential feature candidates. We facilitate and support this process by proposing a generic and unsupervised approach for the automated selection of features that best represent the underlying data. In a next step, we employ the selected features for training a classifier and performing video genre classification.

Existing approaches for video genre classification usually consider the combination of different modalities, such as visual, acoustic and textual information, in order to differentiate between video genres [5,8,13,15]. For thorough

© Springer International Publishing Switzerland 2016
Q. Tian et al. (Eds.): MMM 2016, Part I, LNCS 9516, pp. 29–41, 2016.
DOI: 10.1007/978-3-319-27671-7_3

reviews on current research, please refer to [2,18]. In this work, we focus on audio features only. Recently, audio features demonstrate competitive performance to multimodal approaches [10]. Furthermore, audio features are less computational expensive and allow for the efficient analysis of large media collections. Existing audio-based approaches for video genre classification commonly use well-established features from the temporal and frequency domains, e.g. mel filter cepstral coefficients (MFCC) [6,16], wavelet coefficients [4], short-term cepstral analysis by means of MFCC, perceptual linear prediction (PLP) and Rasta-PLP [17], acoustic topic models [10], background acoustic features [20]. Employed audio-based features often originate from the task of speech/non-speech discrimination [14].

Feature selection in the context of audio-based genre classification is usually based on previous successful experiments in similar application scenarios [14,18]. The major drawback of such a strategy is the use of prior information about the audio content, such as the existence of various content elements (e.g. speech, music). The inflexibility and resource-demanding calculations of features, that may not even contribute to a classification performance are current core limitations intended to overcome by the proposed method.

This work is organized as follows. Section 2 outlines our approach for feature selection, which we employ in order to automatically select those audio features that best describe a given dataset. Section 3 presents the evaluation setup for the performed experiments including employed features, datasets, classifiers, and performance metrics. Section 4 discusses the experimental results in detail. Eventually, Sect. 5 concludes the paper.

## 2 Feature Selection

We propose an unsupervised group feature selection approach, which makes use of canonical correlation analysis (CCA) in order to identify low-correlated and, thus, complementary and relevant features that efficiently describe the underlying data. In general, CCA is a multivariate regression method that measures the relationship between multidimensional variables in a linear manner [7]. CCA calculates correlations between features of different dimensions and provides canonical correlation coefficients for all pairs of features. We employ the term *group features* to emphasize the fact that we select multidimensional features as a whole in contrast to conventional feature selection methods, which typically ignore existing groupings of feature components.

Figure 1 illustrates the main components and basic steps of our feature selection workflow. First, we calculate the canonical correlations for all feature combinations of the input data (feature matrix). The result of this step is a symmetrical canonical correlation matrix (CCM). We employ the pairwise correlation coefficients as a measure of redundancy. Feature pairs with high correlation coefficients are considered redundant, whereas low-correlated features provide complementary and, therefore, additional information. We prune the CCM and reduce it to the upper triangular matrix. Additionally, we remove correlations that exceed a

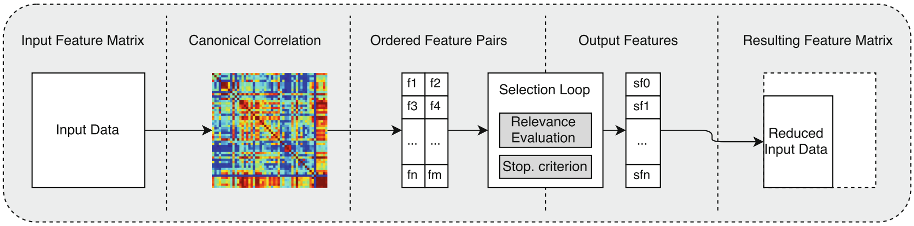



**Fig. 1.** Feature selection workflow.

certain threshold. The purpose of this threshold is to remove highly-correlated features since they are considered too redundant and, thus, non-expressive for the underlying data. The remaining pairs of features are sorted in decreasing order according to their correlation coefficients. This constitutes an initial feature ranking that is iteratively and sequentially processed. In every iteration, a candidate feature pair is evaluated whether or not to be included in the target feature set based on an internal relevance measure. In our implementation, this relevance criterion is designed in a flexible, modular manner and thus can be exchanged easily. Depending on the actual relevance evaluation, the feature selection process can terminate autonomously in every iteration. One possible relevance criterion is the entropy-based information gain (IG), which has been applied in [19]. In this work, we apply CCA directly as relevance measure. Therefore, we calculate the canonical correlation between the current candidate feature pair and the already constructed feature set. We measure the significance of a correlation following the principle of significant modes [9]. If the correlation is significant, it is considered to provide additional, descriptive, and low-redundant information to the target feature set and is added to the previously selected feature set. Eventually, an additional, optional stopping criterion can be employed to terminate the feature selection process if, for example, a feature set of a certain size is desired. In this work, we do not employ such stopping criterion but investigate all feature pairs to autonomously identify the optimal set of features for the given data.

In our experiments, we split the input data into 10 % training set and 90 % test set. The feature selection process is applied on the training data only.

## 3 Evaluation Setup

### 3.1 Audio Features

We conduct our experiments using a set of 50 high-dimensional audio features that consists of 679 feature components in total (see Table 1). This selection incorporates representative and comprehensive audio features from the temporal and frequency domains, that cover various audio aspects, such as harmonics, beat and rhythm, pitch, timbre, and loudness. For more details on the audio features please refer to [11].

**Table 1.** Overview of the employed features and the corresponding dimensions (D). The features are listed in alphabetical order.

| | Feature | Feature Name | D | | Feature | Feature Name | D |
|---|---|---|---|---|---|---|---|
| 1 | AD | Amplitude Descriptor | 40 | 26 | M7_LAT | MPEG-7 Log Attack Time | 1 |
| 2 | BFCC | Bark-scale Frequency Cepstral Coeff. | 40 | 27 | M7_SC | MPEG-7 Spectral Centroid | 1 |
| 3 | BTHI | Beat Histogram | 7 | 28 | MFCC | Mel-scale Frequency Cepstral Coeff. | 40 |
| 4 | CRMA | Chroma CENS Features | 24 | 29 | PLP | Perceptual Linear Prediction | 38 |
| 5 | E4Hz | 4 Hz Modulation Energy | 2 | 30 | PTCH | Pitch | 2 |
| 6 | GPD | Group Delay | 40 | 31 | PTCT | Pitch Contour | 2 |
| 7 | HMDV | Harmonic Derivate | 16 | 32 | PTVB | Pitch Vibration | 1 |
| 8 | HZCR | High Zero Crossing Rate | 1 | 33 | R_ZC | Range of Zero Crossing Rate | 1 |
| 9 | LPC | Linear Predictive Coding | 40 | 34 | RMS | Root Mean Square | 2 |
| 10 | LPCC | Linear Prediction Cepstral Coefficients | 40 | 35 | ROFF | Spectral Rolloff | 2 |
| 11 | LPZC | Linear Prediction ZCR | 2 | 36 | RPLP | Raster PLP | 38 |
| 12 | LSP | Line Spectral Pairs | 40 | 37 | RYPT | Rhythm Patterns | 20 |
| 13 | M7_AFF | MPEG-7 Audio Fundamental Frequency | 4 | 38 | SBER | Subband Energy Ratio | 10 |
| 14 | M7_AH | MPEG-7 Audio Harmonicity | 4 | 39 | SF | Spectral Flux | 2 |
| 15 | M7_AP | MPEG-7 Audio Power | 2 | 40 | SONE | Loudness | 40 |
| 16 | M7_ASB | MPEG-7 Audio Spectrum Basis | 72 | 41 | SPCR | Spectral Crest | 8 |
| 17 | M7_ASC | MPEG-7 Audio Spectrum Centroid | 2 | 42 | SPCT | Spectral Center | 2 |
| 18 | M7_ASF | MPEG-7 Audio Spectrum Flatness | 34 | 43 | SPDI | Spectral Dispersion | 2 |
| 19 | M7_ASP | MPEG-7 Audio Spectrum Projection | 16 | 44 | SPEY | Spectral Entropy | 8 |
| 20 | M7_ASS | MPEG-7 Audio Spectrum Spread | 2 | 45 | SPPS | Spectral Peak Structure | 2 |
| 21 | M7_AW | MPEG-7 Audio Waveform | 4 | 46 | SPRE | Spectral Renyi Entropy | 8 |
| 22 | M7_HSC | MPEG-7 Harmonic Spectral Centroid | 1 | 47 | SPSL | Spectral Slope | 8 |
| 23 | M7_HSD | MPEG-7 Harmonic Spectral Deviation | 1 | 48 | STE | Short Time Energy | 2 |
| 24 | M7_HSS | MPEG-7 Harmonic Spectral Spread | 1 | 49 | VDR | Volume Dynamic Range | 1 |
| 25 | M7_HSV | MPEG-7 Harmonic Spectral Variation | 1 | 50 | ZCR | Zero Crossing Rate | 2 |
| | | | | | | Total dimensionality | 679 |

### 3.2 Datasets

We investigate two video datasets in our experiments: BBC documentaries and RAI TV broadcasts. Table 2 provides an overview of the employed datasets and their characteristics.

The **BBC documentaries** dataset is a self-collected set of videos from the BBC's YouTube channel[1]. It covers three sub-genres: *technical*, *nature*, and *music*. Although the semantic focus of the three sub-genres is strongly varying, all videos in this set are composed, edited, and post-processed in a very similar way, at least from a technical point of view.

The **RAI TV broadcasts** dataset contains more than 100 hours of complete broadcasted programmes of RAI television [12,13]. The data is divided into subsets of different sizes and with partly different genres, which can be investigated separately. For our experiments we employ two subsets. The first one, *RAI-6*, compromises 6 genres: *commercials*, *football*, *music*, *news*, *talk shows*, and *weather forecasts*. The second one, *RAI-3*, is a subset of RAI-6 and covers 3 genres: *commercials*, *football*, and *music*. In contrast to the BBC documentaries, the RAI broadcasts exhibit strongly varying structures and no explicit regularities among the different genres. As a result, this heterogeneous corpus corresponds to a conventional genre classification task, whereas the BBC documentaries allow for the investigation of a sub-genre classification scenario.

### 3.3 Data Preprocessing

Since the different datasets are available as different video container files, we first extract the audio tracks and convert them to PCM audio files. Next, the

[1] https://www.youtube.com/user/BBC/.

**Table 2.** Overview of the employed datasets.

| Dataset | Videos | Total Duration | Classes | Segment Size | Total Samples |
|---|---|---|---|---|---|
| BBC | 9 | 4.5 h | 3 | 2 s | 16, 140 |
| | | | | 10 s | 3, 225 |
| | | | | 30 s | 1, 070 |
| RAI-3 | 49 | 19.4 h | 3 | 2 s | 45, 935 |
| | | | | 10 s | 9, 067 |
| | | | | 30 s | 2, 998 |
| RAI-6 | 93 | 30.9 h | 6 | 2 s | 85, 517 |
| | | | | 10 s | 17, 041 |
| | | | | 30 s | 5, 636 |

audio tracks are segmented into chunks of 2, 10, and 30 s. This subdivision is carried out with an overlap of 50 % in order to maintain acoustic information near the segmentation boundaries. Especially when considering small segments of the audio signals, passages of constant silence may appear. These segments do not have any expressiveness and may cause errors in the feature extraction process. Therefore, we perform silence detection by means of a noise threshold of $-60dB$ and remove detected silent segments from the dataset. This step has a low impact on the following analysis: none of the segments from the two RAI TV datasets and only 0.31 % of the 2 s segments from the BBC documentaries are identified as silence and removed.

### 3.4 Classification

We employ three, in the audio domain well-established classifiers: K-Nearest Neighbor (KNN) [3], Support Vector Machine (SVM) [22], and Random Forest (RF) [1]. The parameter settings for the different classifiers have been selected based on preliminary experiments with respect to classification performance. We employ KNN with $k = 2$ and the Euclidean distance as distance measure without any additional weighting. The SVM implementation uses a polynomial inhomogeneous kernel in order to support non-linear hyperplane separation: $K(x_i, x_j) = (x_i \cdot x_j + 1)^e$, with $e = 2$, a complexity factor of $c = 1$. RF as a tree-based classifier generates a forest of random trees having unlimited depth $mD = \infty$ and a maximum number of $nT = 10$ trees. Although many works employ RF with more trees (e.g. 500 trees by default in some implementations), we could not identify any significant increase in the classification performance while the runtime increased notably (about 50 times in our preliminary experiments). Therefore, we chose $nT = 10$ for all experiments in this work.

All classifications are randomly initialized, 10-fold cross-validated with respect to the underlying class distribution, and run 10 times independently.

### 3.5 Performance Metrics

We employ the *weighted F-score* to measure the accuracy of the performed classifications. The weighted F-score takes into account the varying class distribution of the datasets:

$$F_\beta^w = \frac{1}{n} \sum_{c \in C} F_\beta(c) \times n_c, \tag{1}$$

where $n_c$ denotes the number of instances per class $c$, $n$ the number of instances in total, and $F_\beta$ the standard F-score:

$$F_\beta = (1 + \beta^2) \frac{precision \times recall}{\beta^2 \times precision + recall} \tag{2}$$

In addition to the quantitative performance evaluation, the selected groups of features are investigated in terms of *semantic expressiveness* for the underlying data and *robustness* of the feature selection method. We measure the robustness of the feature selection method by considering the occurrences of the selected features averaged over the 10 independent and randomly initialized runs.

## 4 Experimental Results

### 4.1 Classification Performance

In this experiment, we focus on the classification performance in terms of F1-score for the different segment sizes for all three datasets (BBC, RAI-3, and RAI-6) and the three classifiers (KNN, SVM, and RF). Table 3 summarizes the achieved results. All three datasets and all three segment sizes achieve an outstanding performance in terms of F1-score given the notable reduction of dimensionality of the selected feature set. For the BBC data and a segment size of 30 s, for example, only 21 features covering 6 % of the full feature set are selected achieving 95 % F1-score with the KNN classifier. For all three datasets, a decrease of the segment size tends to result in a feature set of higher dimensionality. The reason for this trend is that, in general, smaller segments bear more details that need to be described and, thus, they require for more precise features. On the opposite, larger segments tend to blur details (primarily due to feature averaging) and need fewer features for their representation.

The results show a notable difference between the performance across the different segment sizes for the different datasets. The BBC dataset performs better for smaller segments, e.g. F1-score of 99 % for segments of size 2 s vs. F1-score of 94 % for segments of size 30 s using the RF classifier. On the opposite, both RAI datasets perform slightly better for increasing segment sizes, e.g. F1-score of 97 % for segments of size 2 s vs. F1-score of 99 % for segments of size 30 s using again the RF classifier on the RAI-3 dataset. This inverse tendency is primarily due to the substantial difference in the nature of the underlying data resulting in different feature selections. While the BBC dataset is very homogeneous (all

**Table 3.** Performance results for the three datasets in terms of weighted F1-scores. N: number of selected features, D: dimensionality of the corresponding feature set. Classification of the full feature set is conducted using the best performing classifier for the corresponding dataset and segment size.

| Dataset | Classifier | Segment size | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 30 s | | | | 10 s | | | | 2 s | | | |
| | | N | D | | F1 | N | D | | F1 | N | D | | F1 |
| BBC | KNN | 21 | 38 | (6 %) | 0.948 | 25 | 82 | (12 %) | 0.966 | 27 | 159 | (23 %) | 0.990 |
| | SVM | | | | 0.906 | | | | 0.984 | | | | 0.996 |
| | RF | | | | 0.940 | | | | 0.971 | | | | 0.993 |
| | Full feature set | 50 | 679 | (100 %) | 0.993 | 50 | 679 | (100 %) | 0.995 | 50 | 679 | (100 %) | 0.996 |
| RAI-3 | KNN | 19 | 89 | (13 %) | 0.986 | 22 | 117 | (17 %) | 0.985 | 28 | 221 | (33 %) | 0.953 |
| | SVM | | | | 0.995 | | | | 0.982 | | | | 0.976 |
| | RF | | | | 0.990 | | | | 0.958 | | | | 0.974 |
| | Full feature set | 50 | 679 | (100 %) | 0.998 | 50 | 679 | (100 %) | 0.997 | 50 | 679 | (100 %) | 0.989 |
| RAI-6 | KNN | 18 | 84 | (12 %) | 0.951 | 22 | 117 | (17 %) | 0.937 | 20 | 114 | (17 %) | 0.939 |
| | SVM | | | | 0.969 | | | | 0.972 | | | | 0.954 |
| | RF | | | | 0.993 | | | | 0.991 | | | | 0.975 |
| | Full feature set | 50 | 679 | (100 %) | 0.996 | 50 | 679 | (100 %) | 0.994 | 50 | 679 | (100 %) | 0.972 |
| Average over all datasets | KNN | 19 | 80 | (12 %) | 0.961 | 22 | 113 | (16 %) | 0.955 | 23 | 152 | (23 %) | 0.949 |
| | SVM | | | | 0.970 | | | | 0.976 | | | | 0.965 |
| | RF | | | | 0.986 | | | | 0.979 | | | | 0.977 |
| | Full feature set | 50 | 679 | (100 %) | 0.996 | 50 | 679 | (100 %) | 0.995 | 50 | 679 | (100 %) | 0.980 |

documentaries have similar structure and share common elements), the RAI data is distinctive to a certain degree (cp. discussion in the performed case study on full video classification). As a result, the BBC dataset requires for higher granulated segments in order to capture more descriptive information and, thus, to better distinguish across the different sub-genres of documentaries.

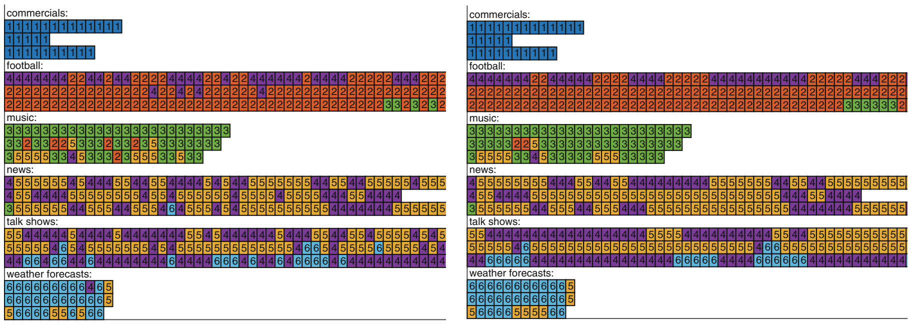
The average performance of the employed classifiers over all datasets and different segment sizes shows that the RF classifier achieves best performance over all datasets followed by SVM and lastly KNN.

Eventually, we compare our approach with related work reporting evaluation results on the RAI dataset (see Table 4). Please note, that our experiments are conducted on a subset of the dataset employed by the compared approaches and does not include the *cartoon* genre. The results indicate the outstanding performance of the selected features. The performance achieved on the employed data (F1-score of 99 %) demonstrate strong competitiveness to the top reported performance by Ekenel et al. [5]. In addition to some acoustic features, Ekenel et al. consider visual, structural, and cognitive features. The features are selected in a way to reflect the editor's process in TV production and cannot be applied for arbitrary data. In contrast, our approach autonomously selects the features that are relevant for the provided data set. The achieved performance demonstrates the quality of the selected features (see Sect. 4.2 for a detailed analysis) while at the same time the exploration of a single modality notably reduces the computational effort.

**Table 4.** Comparison with related works on the RAI dataset, used modalities (A=audio, V=video, S=structural, C=cognitive), number of genres, dataset size, and achieved accuracy in terms of F1-score.

| Authors | Modalities | # Genres | Dataset size | F1 |
|---|---|---|---|---|
| Montagnuolo et al. [12] | A, V, S, C | 7 | 6,690 min | 0.924 |
| Montagnuolo et al. [13] | A, V, S, C | 7 | 6,690 min | 0.949 |
| Ekenel et al. [5] | A, V, S, C | 7 | 6,600 min | 0.992 |
| Ekenel et al. [5] | A | 7 | 6,600 min | 0.957 |
| Kim et al. [10] | A | 7 | 4,167 min | 0.943 |
| This work | A | 6 | 1,850 min | **0.993** |


(a) Color legend for predicted genres.




(b) Original segment assignments. (c) Smoothed segment assignments.

**Fig. 2.** Segment assignments of three randomly selected video sequences for each genre. Groupings in 2(b) and 2(c) correspond to the ground truth. Colors and numbers in the corresponding segments represent the predicted assignments. The timeline is cut to the right due to space limitations (Color figure online).

**Case Study: Full Video Classification.** In our first case study, we investigate the question: *Can we successfully detect the genre of a full video sequence based on the classification of its segments?* For this case study we employ the RAI-6 dataset, 30 s segments. For each genre, one of the videos is used for training and the remaining videos for testing. Figure 2 shows examples for segment assignments for videos of different genres. Since the underlying video segments have an overlap of 50 %, we smooth the predicted assignments using a sliding window of size 3 in order to remove single outliers. Figure 2 additionally indicates that while some genres such as *commercials*, *football*, and *music* can be clearly identified, segments of the remaining genres, *news*, *talk shows*, and *weather forecasts*, are often misclassified.

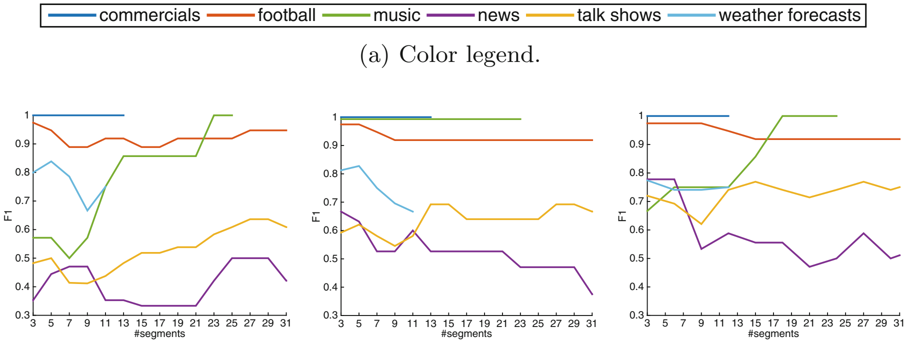
We employ majority voting as classification strategy for the assignment of a genre to a video sequence. Majority voting is a simple decision rule that selects

**Table 5.** Confusion matrix for the full video classification task. Rows correspond to the ground truth and columns to the predicted genre. Blank cells represent zero values.

| | commercials | football | music | news | talk shows | weather |
|---|---|---|---|---|---|---|
| commercials | 18 (100 %) | | | | | |
| football | | 17 (85 %) | | 2 (10 %) | 1 (5 %) | |
| music | | | 3 (100 %) | | | |
| news | | | | 3 (30 %) | 7 (70 %) | |
| talk shows | | | | 1 (9 %) | 10 (91 %) | |
| weather forecasts | | | | | 6 (35 %) | 11 (65 %) |

the genre that is in the majority in the genre assignments of the underlying video segments. The overall classification performance of the full video sequences achieves a F1-score of 80 % , which is significantly lower than the classification accuracy of single audio segments. A crucial difference between the two experiments is the amount of available data (both training and test data) which significantly influences the quality of the underlying models. Table 5 shows the confusion matrix of the classification. Due to the low number of video sequences, a single misclassification has a notable influence on the overall classification rate. For example, one video sequence from *talk shows* has been misclassified as news. As a result, the retrieval performance for *talk shows* decreases to 91 %. Furthermore, *news*, *talk shows*, and *weather forecasts* bear similar audio characteristics. Hence, multiple video sequences from these genres are incorrectly assigned within this group (predominantly as *talk shows*, cp. Fig. 2).

In general, video durations vary strongly. For example, the RAI dataset consists of video sequences between 90 s and 53 min. The analysis of longer video sequences can easily become computationally expensive. Therefore, in a next experiment we investigate the question if it is feasible to classify a full video sequence based on the analysis of a small subsequence only. For this experiment, we select varying numbers of segments, starting with 3 segments and iteratively increasing the number of segments (step of 2) until the full video sequence is considered. Figure 3 compares the performance of segments originating from different parts of the underlying video sequences: from the beginning of a video, from the mid part, as well as a combination from the beginning, mid, and ending. Due to space limitations we only show the results for the first 30 segments, which correspond to a video subsequence of 10 min. The results demonstrate, that two genres, *commercials* and *football*, achieve an outstanding performance independently of the length of subsequences analyzed or the part of the video it is contained. On the opposite, *news*, *talk shows*, and *weather forecasts* perform poorly in general since the audio models are less discriminative in this context and the employment of visual features will definitely help to better distinguish between the three genres. Additionally, more training data can significantly improve the quality of the underlying audio models as proved by the experiments on single audio segments. Finally, *music* is well identified if either the full length of the




(a) Color legend.

(b) Beginning segments. (c) Mid segments. (d) Beginning, mid, & end.

**Fig. 3.** Performance of the full video classification task using a subset of segments taken from different video parts. 3(b): from the beginning of a video sequence; 3(c): from the mid part of a video sequence; and 3(d): from the beginning, middle, and the end part of a video sequence.

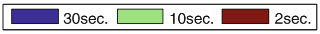
video or a subsequence from its mid part is analyzed. These results confirm the previous analyses in this case study. Genres, which are in general well recognized, require for the analysis of a small subsequence of the full video only. On the opposite, genres, which are commonly confused by the employed features, do not improve or worsen significantly with varying length of analyzed segments.

### 4.2 Feature Analysis

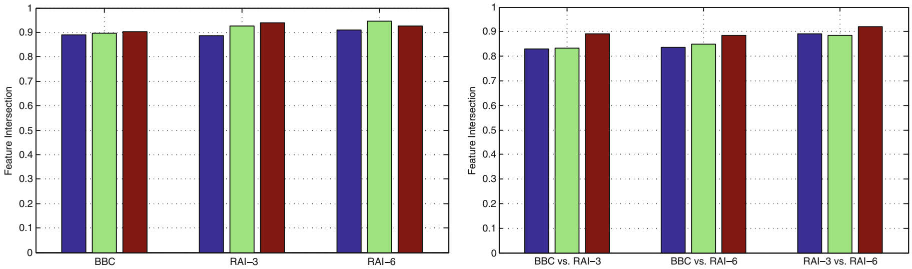
Core advantage of the proposed unsupervised feature selection approach is that it selects complete features independently of their dimensionality. Especially in the audio domain, where features usually carry a higher level of semantics as e.g. visual features, group features show an advantage in terms of interpreting the data. The purpose of this section is a brief discussion of the robustness of selected feature sets across different runs within a dataset and across the different datasets.

Figure 4(b) depicts the amount of intersection of feature sets (robustness) computed in different test runs for the different datasets and segment sizes. It can be seen that different test runs show very little differences in the computed feature sets. The overlapping of selected features sets between different datasets is illustrated in Fig. 4(c). Here, the amount of intersection is about 0.1 points lower for all combinations. One might argue that this indicates that the selected features are less depended on the datasets than expected or argued so far.

Therefore, in a small experiment we investigate the features actually selected for the RAI-3 data for a classification of the BBC dataset (30 s segments). While the original 21 BBC features achieve an F1-score of 95 % using the KNN classifier (see Table 3), the RAI-3 features obtain an F1-score of only 86 %. This notable drop in the performance stresses strengths of a high-quality feature selection




(a) Color legend.




(b) (c)

**Fig. 4.** (a) Robustness of selected feature sets for the different datasets and segment sizes across 10 test runs. (b) Overlapping of selected feature sets between different datasets.

with respect to the provided data. Even if the percentage of overlap between datasets is rather high (about 85 %) it is the remaining data-specific feature groups that contribute to excellent results.

## 5 Conclusion

This paper addressed a core question in the context of video genre classification concerning the selection process of an appropriate feature set. We proposed a generic approach for feature selection, which does not make any assumptions about the underlying data or investigated video genres, but it autonomously selects a feature set that efficiently describes the data. Performed experiments demonstrated the outstanding performance of the approach for different datasets and video genres. The analysis of the selected features showed the robustness of the approach across different runs on the same data. Additionally, the analysis demonstrated the necessity of selecting different feature sets for varying datasets, which is a core argument why a generic feature selection process is required.

**Acknowledgments.** This work has been partly funded by the Vienna Science and Technology Fund (WWTF) through project ICT12-010. The authors are thankful to Marcus Hudec for pointing our interest towards CCA. The authors would also like to thank Maurizio Montagnuolo from RAI Centre for Research and Technological Innovation for providing the RAI TV dataset.

## References

1. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
2. Brezeale, D., Cook, D.: Automatic video classification: a survey of the literature. IEEE Trans. Syst. Man Cybern. **38**(3), 416–430 (2008)

3. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Trans. Inf. Theor. **13**(1), 21–27 (1967)
4. Dinh, P.Q., Dorai, C., Venkatesh, S.: Video genre categorization using audio wavelet coefficients. In: Asian Conference on Computer Vision (2002)
5. Ekenel, H.K., Semela, T.: Multimodal genre classification of TV programs and YouTube videos. Multimedia Tools Appl. **63**(2), 547–567 (2013)
6. Guo, J., Gurrin, C.: Short user-generated videos classification using accompanied audio categories. In: ACM International Workshop on Audio and Multimedia Methods for Large-scale Video Analysis, pp. 15–20 (2012)
7. Hotelling, H.: Relations between two sets of variates. Biometrika **28**, 321–377 (1936)
8. Huang, Y.-F., Wang, S.-H.: Movie genre classification using SVM with audio and video features. In: Huang, R., Ghorbani, A.A., Pasi, G., Yamaguchi, T., Yen, N.Y., Jin, B. (eds.) AMT 2012. LNCS, vol. 7669, pp. 1–10. Springer, Heidelberg (2012)
9. Jolliffe, I.: Principal Component Analysis. Springer Series in Statistics. Springer, Heidelberg (2002)
10. Kim, S., Georgiou, P., Narayanan, S.: On-line genre classification of TV programs using audio content. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 798–802 (2013)
11. Mitrovic, D., Zeppelzauer, M., Breiteneder, C.: Features for content-based audio retrieval. Adv. Comput.: Improving Web **78**, 71–150 (2010)
12. Montagnuolo, M., Messina, A.: TV genre classification using multimodal information and multilayer perceptrons. In: Basili, R., Pazienza, M.T. (eds.) AI*IA 2007. LNCS (LNAI), vol. 4733, pp. 730–741. Springer, Heidelberg (2007)
13. Montagnuolo, M., Messina, A.: Parallel neural networks for multimodal video genre classification. Multimedia Tools Appl. **41**(1), 125–159 (2009)
14. Natarajan, R., Chandrakala, S.: Audio-based event detection in videos - a comprehensive survey. Int. J. Eng. Technol. **6**(4), 1663–1674 (2014)
15. Roach, M., Mason, J., Xu, L.Q.: Video genre verification using both acoustic and visual modes. In: IEEE Workshop on Multimedia Signal Processing, pp. 157–160 (2002)
16. Roach, M., Mason, J.: Classification of video genre using audio. Eurospeech **4**, 2693–2696 (2001)
17. Rouvier, M., Linares, G., Matrouf, D.: On-the-fly video genre classification by combination of audio features. In: IEEE International Conference on Acoustics Speech and Signal Processing, pp. 45–48 (2010)
18. Rouvier, M., Oger, S., Linares, G., Matrouf, D., Merialdo, B., Li, Y.: Audio-based video genre identification. IEEE/ACM Audio, Speech, Lang. Process. **23**(6), 1031–1041 (2015)
19. Sageder, G., Zaharieva, M., Zeppelzauer, M.: Unsupervised selection of robust audio feature subsets. In: SIAM International Conference on Data Mining, pp. 686–694 (2014)
20. Saz, O., Doulaty, M., Hain, T.: Background-tracking acoustic features for genre identification of broadcast shows. In: IEEE Spoken Language Technology Workshop, pp. 118–123 (2014)
21. Song, Y., Zhao, M., Yagnik, J., Wu, X.: Taxonomic classification for web-based videos. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 871–878 (2010)
22. Vapnik, V.: Nature Statistical Learning Theory. Springer, Heidelberg (1995)

23. Wu, X., Zhao, W.L., Ngo, C.W.: Towards google challenge: Combining contextual and social information for web video categorization. In: ACM International Conference on Multimedia, pp. 1109–1110 (2009)
24. Zhang, N., Duan, L.Y., Li, L., Huang, Q., Du, J., Gao, W., Guan, L.: A generic approach for systematic analysis of sports videos. ACM Trans. Intell. Syst. Technol. **3**(3), 46:1–46:29 (2012)