Factor Analysis for Audio-based Video Genre Classification

Mickael Rouvier, Driss Matrouf, Georges Linarès

1LIA, University of Avignon, France

Abstract

Statistical classifiers operate on features that generally include both useful and useless information. These two types of information are difficult to separate in the feature domain. Recently, a new paradigm based on a Latent Factor Analysis (LFA) proposed a model decomposition into useful and useless components. This method was successfully applied to speaker and language recognition tasks. In this paper, we study the use of LFA for video genre classification by using only the audio channel. We propose a classification method based on short-term cepstral features and Gaussian Mixture Models (GMM) or Support Vector Machine (SVM) classifiers, that are combined with Factor Analysis (FA). Experiments are conducted on a corpus composed of 5 types of video (musics, commercials, cartoons, movies and news). The relative classification error reduction obtained by using the best factor analysis configuration with respect to the baseline system, Gaussian Mixture Model Universal Background Model (GMM-UBM), is about 56%, corresponding to a correct identification rate of about 90%.

Index Terms: video genre identification, Factor Analysis, automatic classification

1. Introduction

In statistical pattern recognition, a pattern is represented by a set of \( d \) attributes; this set is viewed as a \( d \)-dimensional feature vector. The recognition system is operated in two modes: training (estimation) and classification (testing). Preprocessing modules can be used for extracting the pattern of interest from the background, for removing noise, for performing some normalizations and other operations that are supposed to contribute to a compact and accurate modeling of the target patterns.

In spite of the use of complex feature extraction modules, the variability due to useless information introduces a bias in model parameter estimation. This bias can have a tremendous impact on the classification performances, degradation being caused mainly by the fact that the training corpora can not offer an exhaustive coverage of all the potential variabilities.

The goal of most of the usual approaches to training statistical classifiers is the estimation of the parameters that characterize the useful information, whereas the useless information is neither explicitly modelled in this process, nor implicitly captured from incomplete training corpora. However, the joint estimation of parameters concerning the useful and the useless information would allow us to better estimate the parameters related to the relevant information.

More formally, let \( O \) being an observation of a given pattern, the model can be decomposed as follows:

\[
M_O = M_{O,\text{useful}} + M_{O,\text{useless}}
\]

where \( M_O \) is the statistical model optimized on \( O \), \( M_{O,\text{useful}} \) is the statistical model for the useful information, and \( M_{O,\text{useless}} \) is the statistical model for the useless information. This decomposition allows to extract only the interesting part \( M_{O,\text{useful}} \) of the model by putting aside the useless and disturbing part \( M_{O,\text{useless}} \). The question that arises now is how to separate the useful and useless components, with respect to the desired application.

For example, if we are interested in language recognition, and if the available training data are a number of recordings coming from several speakers of a given language \( L \), then the useful information is the language, modeled by \( M_{O,L} \) and the useless information could be constituted by the speakers that are modeled in \( M_{O,T} \).

More generally, let \( O = O_1, O_2, \ldots, O_n \) be \( n \) observations for a given pattern \( P \). The common information in \( O_1, O_2, \ldots, O_n \) corresponds to the one that we wish to model. On the other hand, all the information out of \( O_1 \cap O_2 \cap \ldots \cap O_n \) corresponds to the unwanted variability. If we assume that each observation is a vector from \( R^L \) and the information not belonging in \( O_1 \cap O_2 \cap \ldots \cap O_n \) can be located in a sub-space with low dimension, then the model can be written as follows:

\[
M_{O_i} = M_{O,\text{useful}} + U x_{O_{i,\text{useless}}}
\]

where \( U \) is a matrix with \( R \) columns (\( R \) is the rank of the matrix). The sub-space generated by the vector columns of \( U \) constitutes the useless information sub-space. This paradigm is called FA; it consists in estimating the \( U \) matrix, \( M_{O,\text{useful}} \) and the vector \( x_{O_{i,\text{useless}}} \). FA has been successfully applied to speaker verification [1], [2].

In this article we transpose this approach, to apply it to video genre classification. The genre of video, in this work, can be news, movies, cartoons, musics or commercials. The only features used here are short-term cepstral coefficients corresponding to the audio part of the video recording.

2. Video Genre Classification

The critical need of efficient tools for structuring audio-visual databases motivated many works these last years [3]. Most of the proposed approaches rely on image analysis. Audio-based approaches were explored by automatic transcription of speech contents, or by low level audio stream analysis. However, ASR systems generally have poor performances on unexpected linguistic domains and in adverse acoustic conditions. Therefore, low-level approaches present a better robustness to the highly variable and unexpected conditions that may be encountered.
3. Factor Analysis for Genre Classification

The use of GMM in a GMM-UBM framework is a standard in the speaker verification field [10]. In this paper, the same framework will be used for video genre classification: each genre (news, movies, cartoons, musics or commercials) is modeled by using GMM. A world model (UBM-GMM) represents the whole acoustic space while genre-specific GMMs are obtained by adapting the world model. The used adaptation technique of choice is the standard Maximum A Posteriori (MAP) [11]: here (as in speaker verification) only mean vectors are adapted, the weights and the variances remaining unchanged with respect to those of the UBM.

To take into account the useless information in the modeling process, a genre-specific model can be decomposed into three different parts: a genre and session-independent component, a genre-dependent component and a session-dependent component. A GMM mean "supervector" is defined as the concatenation of the GMM component means. Let \( D \) be the dimension of the feature space, the dimension of a supervector mean is \( M D \), where \( M \) is the number of Gaussian in the GMM. A genre and session independent model is usually estimated so that it represents the inverse hypothesis: the UBM model. Let this model being parameterized by \( \theta \). In the sequel, \((h, GE)\) indicates that the genre of the recording is \( GE \) and the session is \( h \). Two different sessions corresponding to the same genre constitute different observations, differences are due to the diversity of the acoustic conditions (speaker diversity, recording materials, acoustic environments) and to the content variabilities (different kinds of music, movies, speech styles, etc.). As explained before, such differences must be located and modeled. The factor analysis model can be written as:

\[
\mathbf{m}_{(b,GE)} = \mathbf{m} + \mathbf{Dy}_{GE} + \mathbf{Ux}_{(b,GE)},
\]

where \( \mathbf{m}_{(b,GE)} \) is the session and genre-dependent supervector mean, \( \mathbf{D} \) a \( MD \times MD \) diagonal matrix, \( \mathbf{y}_{GE} \) is genre vector (a \( MD \) vector), \( \mathbf{U} \) is the session variability matrix of low rank \( R \) (a \( MD \times R \) matrix) and \( \mathbf{x}_{(b,GE)} \) are the channel factors, a \( R \) vector (theoretically \( \mathbf{x}_{(b,GE)} \) is independent of \( GE \)). Both \( \mathbf{y}_{GE} \) and \( \mathbf{x}_{(b,GE)} \) are normally distributed among \( N(0, I) \). \( \mathbf{D} \) satisfies \( \mathbf{D} = \tau \mathbf{D} \mathbf{D}^\top \mathbf{S}^{-1} \mathbf{D} \) where \( \tau \) is the relevance factor required in the standard MAP adaptation, and \( \mathbf{D} \mathbf{D}^\top \) represents the \( a \ priori \) covariance matrix of \( \mathbf{y}_{GE} \).

3.1. Factor Analysis Model Training

In the training phase, we have to estimate the useless information matrix \( \mathbf{U} \), the genre-specific components \( \mathbf{y}_{GE} \), and the useless information component \( \mathbf{x}_{(b,GE)} \). The estimation criterion is the Maximization Likelihood. The \( \mathbf{U} \) matrix is estimated on all the genres and with several sessions by genre; the \( \mathbf{y}_{GE} \) component is estimated on all the sessions that belong to genre \( GE \); the \( \mathbf{x}_{(b,GE)} \) component is estimated on session \( h \), which belong to genre \( GE \) [12].

3.2. Classification task

In this subsection we dwell on the details of the strategy used for performing the useless variability compensation. The classification task is defined as follows. A genre \( GE_{tar} \) is enrolled by the system with its training data \( \mathbf{Y}_{GE_{tar}} \). Given a sequence of speech frames \( \mathbf{Y} = \{y_1 \ldots y_T\} \) and the genre \( GE_{tar} \), the genre classification task consists in determining whether \( \mathbf{Y} \) belongs to \( GE_{tar} \) or not. Using FA decomposition in both training and testing data, one can write:

\[
\mathbf{m}_{(h_{tar},GE_{tar})} = \mathbf{m} + \mathbf{Dy}_{GE_{tar}} + \mathbf{Ux}_{(h,GE)};
\]

\[
\mathbf{m}_{(h_{test},GE_{test})} = \mathbf{m} + \mathbf{Dy}_{GE_{test}} + \mathbf{Ux}_{(h,GE)}. \tag{4}
\]

where the genre \( GE_{tar} \) in the training data and \( GE_{test} \) in the testing data have been distinguished. In this paper, an hybrid domain normalization strategy is used, aiming for withdrawing the useless component in the test and training data:

\[
\mathbf{m}_{GE_{tar}} = \mathbf{m} + \mathbf{Dy}_{GE_{tar}}
\]

\[
\mathbf{m}_{GE_{test}} = \mathbf{m} + \mathbf{Dy}_{GE_{test}}. \tag{5}
\]

3.3. Scoring

The genre function score is given by:

\[
LLK(\mathbf{Y}|\mathbf{m} + \mathbf{Dy}_{GE_{tar}}) - LLK(\mathbf{Y}|\mathbf{m}) \tag{6}
\]

where \( LLK(\cdot) \) indicate the average log-likelihood over all the frames. Here, GMMs shares the covariance matrix as well as the mixture weights (both dropped from the equation for clarity). The useless information subtraction in the testing data is performed at the frame level (in the feature domain). The following formula is used for removing the useless variability effect for each frame \( x \):

\[
\hat{x} = x - \sum_{g=1}^{M} \gamma_g(x) \cdot \{ \mathbf{U} \cdot \mathbf{x}_{(h_{test})} \}[g]. \tag{7}
\]

where \( M \) is the number of Gaussian components in the UBM, \( \gamma_g(x) \) is the a posteriori probability of Gaussian \( g \) given frame \( x \). These probabilities are estimated by using the UBM. \( \mathbf{U} \cdot \mathbf{x}_{(h_{test})} \) is a supervector with \( M \times D \) components. \( \{ \mathbf{U} \cdot \mathbf{x}_{(h_{test})} \}[g] \) is the \( g \)th \( D \) components bloc vector of \( \mathbf{U} \cdot \mathbf{x}_{(h_{test})} \).

3.4. Kernel based Scoring and SVM Modeling

By using Equation 5, FA model estimates supervectors containing only genre information, normalized with respect to the useless variability. In [13], the authors proposed a probabilistic distance kernel that computes a distance between GMMs, well suited for an SVM classifier. Let \( \mathbf{x}_{h} \) and \( \mathbf{x}_{h}' \) be two sequences of audio data corresponding to genres \( GE \) and \( GE' \); the associates the kernel formulation is given below.
4. Experimental Protocol

All the experiments were performed using the ALIZE and LIA_SpkDet toolkit\(^1\)[14]. We selected 5 categories that are commonly involved in video genre classification tasks: news, movies, cartoons, musics, and commercials. The corpus is composed of 1200 videos indexed by Dailymotion or Youtube, with durations between 2 and 5 minutes. 1000 of them are used for training, and 200 are used for testing (about 200 video per genre for training and 40 for testing). Speech contents are systematically in the French language. In our experiment, we used PLP features, extracted using a 25ms Hamming window. Each frame is composed of 39 coefficients (PLP 13, Δ PLP 13 and Δ Δ PLP 13) every 10 ms. Cepstral mean normalization is applied on each audio recording. The next subsections describe the different systems that we tested in our experiments.

4.1. GMM-UBM

The background model (GMM-UBM) is a GMM trained with the Expectation Maximization (EM) algorithm. Given a genre utterance, GMM-UBM training is performed by maximum a posteriori (MAP) adaptation of the means with a relevance factor of 14\([11]\).

4.2. GMM-UBM and FA

Given the UBM and a genre-specific utterance, Factor Analysis decomposition is performed (Equation 3). The retained model for the genre \(GE_{tar}\) is given by:

\[
m_{GE_{tar}} = m + Dy_{GE_{tar}}
\]

The classification scores are estimated as explained in Section 3.2.

4.3. SVM-UBM and FA

The LIA SpkDet toolkit now benefits from the LIBSVM library to induce SVM and to classify instances. SVM models are trained with an infinite (very large in practice) C parameter thus avoiding classification errors on the training data (hard margin behaviour). For a given genre, the negative label examples are recordings that belong to other genres. The training and testing stages are performed as explained in Section 3.4.

4.4. SVM-UBM and FA with score normalization

In the case of SVM-UBM with FA, we have noticed a large score shift between different genres. This score shift yields a significant performance degradation. To tackle this issue, a score normalization is applied on the output scores. This normalization is performed as follow:

\[
score^j = \frac{P_t(S^j)}{P_t(S^j) + P_t(S^{j+1})}
\]

4.4. SVM-UBM and FA with score normalization

In the case of SVM-UBM with FA, we have noticed a large score shift between different genres. This score shift yields a significant performance degradation. To tackle this issue, a score normalization is applied on the output scores. This normalization is performed as follow:

\[
score^j = \frac{P_t(S^j)}{P_t(S^j) + P_t(S^{j+1})}
\]

5. Results

5.1. UBM size

The first experiment studies the effect of the GMM-UBM size (number of Gaussian components) on accuracy. Results are presented for model sizes of 64, 128, 256 and 512 components. The results show that a GMM-UBM with 256 components have the best performances on the video genre classification with 23\% classification errors. We use this configuration as the baseline of our next experiments.

5.2. GMM-UBM and FA

The second experiment studies the impact of the FA rank (the rank of the \(U\) matrix). The number of Gaussian components in the UBM is 256. In the best configuration (with a rank of 40) we see that the FA results in a classification error rate of 13\%. The performance is strongly improved by FA with respect to the baseline system GMM-UBM; the relative reduction of the error rate is of about 43\%.
it from all the other classes. Secondly, we can choose the one-against-others-files strategy, that consists in training, for each class, a SVM to separate a class from the others. We study the impact of the two strategies in the next subsection.

5.3.1. One-against-other-classes

We observe that the normalization is clearly required. The relative error reduction obtained by using score normalization is about 38%. Nevertheless, in this configuration (SVM-UBM-FA with score normalization), we obtain the same results as with the GMM-UBM-FA system (classification error rate of 13%).

Table 3: SVM-UBM-FA (one-against-others-class), Sys-1 is GMM-UBM-FA, Sys-2 is SVM-UBM-FA, and Sys-3 is SVM-UBM-FA with score normalization.

<table>
<thead>
<tr>
<th>System</th>
<th>Mus</th>
<th>New</th>
<th>Com</th>
<th>Car</th>
<th>Mov</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sys-1</td>
<td>3</td>
<td>3</td>
<td>37</td>
<td>13</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>Sys-2</td>
<td>3</td>
<td>16</td>
<td>66</td>
<td>10</td>
<td>8</td>
<td>21</td>
</tr>
<tr>
<td>Sys-3</td>
<td>0</td>
<td>6</td>
<td>27</td>
<td>22</td>
<td>8</td>
<td>13</td>
</tr>
</tbody>
</table>

5.3.2. One-against-other-files

Here, a supervector is trained for each target genre (positive examples). The negative examples ("black list") are composed by all the files that belong to the other genres. With this configuration (SVM-UBM-FA-NORM system), we observe a relative reduction of the classification error of about 23% (with respect to the GMM-UBM-FA system).

Table 4: SVM-UBM-FA (one-against-others-files), Sys-1 is GMM-UBM-FA, Sys-2 is SVM-UBM-FA, and Sys-3 is SVM-UBM-FA with score normalization.

<table>
<thead>
<tr>
<th>System</th>
<th>Mus</th>
<th>New</th>
<th>Com</th>
<th>Car</th>
<th>Mov</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sys-1</td>
<td>3</td>
<td>3</td>
<td>37</td>
<td>13</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>Sys-2</td>
<td>3</td>
<td>8</td>
<td>44</td>
<td>8</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>Sys-3</td>
<td>3</td>
<td>6</td>
<td>18</td>
<td>15</td>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

6. Conclusion

In this paper, we investigate the use of the FA method for video genre categorisation by acoustic space modelling. We compare various classification schemes and FA configurations. Experiments on a 5-genre identification task demonstrated the effectiveness of the proposed approach: the classification error rate is reduced by about 56% relative with respect to the standard approach based on GMM-UBM. Finally, we obtain a classification rate of about 90%, similar to the one classically obtained by genre-identification methods that combine audio and vido information, and outperforming classical audio-only based techniques.

These results suggest that, even if it was mainly applied to speaker identification, FA is a general approach to variability reduction, which could be successfully applied to various pattern recognition tasks.

Starting from this idea, we now plan to generalize the proposed approach of genre identification by integrating FA with combined audio and video features.

7. References